Przegląd TELEKOMUNIKACYJNY

WIADOMOŚCI TELEKOMUNIKACYJNE

4/2025

Dwumiesięcznik Stowarzyszenia Elektryków Polskich

ISSN 1230-3496, e-ISSN 2449-7487

Cena: 60 zł (w tym 8% VAT)

WYDAWNICTWO *SIGMA-NOT*

KRiT 2025

KONFERENCJA RADIOKOMUNIKACJI I TELEINFORMATYKI

GDAŃSK, 10 - 12 WRZEŚNIA 2025

WWW.KRIT.COM.PL

POLITECHNIKA GDAŃSKA

WYDZIAŁ ELEKTRONIKI, TELEKOMUNIKACJI I INFORMATYKI

Sekretariat

tel.: +48 58 347 2562

email: kontakt@krit.com.pl

Szanowni Państwo!

W tym roku mamy wielką przyjemność być po raz pierwszy organizatorami Konferencji Radiokomunikacji i Teleinformatyki. Co prawda na powrót konferencji tego formatu do Gdańska musieliśmy czekać kilka lat, ponieważ poprzednio Politechnika Gdańska była organizatorem Krajowej Konferencji Radiokomunikacji, Radiofonii i Telewizji w 2018 roku oraz Krajowego Sympozjum Telekomunikacji i Teleinformatyki w 2013 roku, ale wierzymy, że opłacało się czekać.

Kilka nadchodzących dni będzie obfitowało w treści obejmujące zagadnienia z dwóch dyscyplin naukowych, tj. informatyki technicznej i telekomunikacji oraz automatyki, elektroniki, elektrotechniki i technologii kosmicznych.

Łącznie zostaną wygłoszone 122 referaty, 6 spośród nich w ramach sesji plenarnych, 111 w ramach 26 tematyk i obszarów naukowych w tym 5 w ramach ogólnopolskiego konkursu o nagrodę Fundacji Wspierania Rozwoju Radiokomunikacji i Technik Multimedialnych (FWRRiTM) za najlepszą pracę doktorską z dziedziny radiokomunikacji i technik multimedialnych. W tej edycji konferencji zdecydowany prym wiodą zagadnienia związane z technikami multimedialnymi, aspektami radiowymi sieci 5G i 6G, kryptografią oraz mechanizmami cyberbezpieczeństwa.

Jak w poprzednich edycjach konferencji, w tej również, zostaną zrealizowane konkursy na najlepszy referat młodych naukowców organizowany przez FWRRiTM oraz na najlepszy referat organizowany przez polski oddział IEEE Communications Society. Dodatkowo, zrealizowany zostanie nowy konkurs na najlepszy referat związany z realizacją doktoratu bądź wynikami obronionej rozprawy doktorskiej organizowany przez FWRRiTM i firmę Orange Polska S.A.

Dziękujemy wszystkim sponsorom, partnerom honorowym, partnerom medialnym i Wam, drodzy uczestnicy, że zdecydowaliście się wziąć udział w tegorocznej edycji Konferencji Radiokomunikacji i Teleinformatyki.

Na szczególne podziękowania zasługuje również Komitet Programowy oraz Komitet Organizacyjny. To dzięki Państwa pracy ta konferencja ma szansę się odbyć.

Mamy nadzieję, że nadchodzący czas dostarczy Państwu wielu interesujących tematów do rozmów i że podczas całego Państwa pobytu będzie towarzyszyła sprzyjająca aura pogodowa, dopełniająca to kilkudniowe spotkanie w Gdańsku.

Z wyrazami szacunku

prof. dr hab. inż. Jacek Stefański
Dziekan Wydziału Elektroniki,
Telekomunikacji i Informatyki

dr hab. inż. Jacek Rak, prof. PG
Kierownik Katedry Teleinformatyki

4/2025

DWUMIESIĘCZNIK STOWARZYSZENIA
ELEKTRYKÓW POLSKICH
WYDAWANY PRZY WSPÓŁPRACY
KOMITETU
ELEKTRONIKI I TELEKOMUNIKACJI
POLSKIEJ AKADEMII NAUK

Przegląd TELEKOMUNIKACYJNY

TELE–RADIO–ELEKTRONIKA–INFORMATYKA

ROK ZAŁOŻENIA 1928 • ROCZNIK XCVIII • ISSN 1230-3496

ZESPÓŁ REDAKCYJNY

dr hab. inż. Janusz Dudczyk, prof. WAT
– redaktor naczelny

Redaktorzy tematyczni:
dr inż. Piotr Grzejszczak,
dr hab. inż. Marek Suproniuk, prof. WAT,
dr hab. inż. Marcin Wesołowski

Bożena Lachowicz – sekretarz redakcji
tel. (22) 827 38 79; 604 446 028
e-mail: **elektronika@red.pl.pl**

RADA PROGRAMOWO-NAUKOWA

Prof. dr hab. inż. Józef Woźniak – przewodniczący
dr hab. inż. Andrzej Bęben, prof. PW; mgr inż. Jacek Cyrek, mgr inż. Paweł Biskupski; dr hab. inż. Jerzy Domżał, prof. AGH; dr inż. Andrzej Dulka; dr hab. inż. Sławomir Hausman, prof. PŁ; prof. dr hab. inż. Andrzej Jajszczyk; mgr inż. Stefan Kamiński; dr hab. inż. Jan Kelner, prof. WAT; dr hab. inż. Adrian Kliks, prof. PP; prof. dr hab. inż. Józef Modelski; dr hab. inż. Marek Natkaniec, prof. AGH; prof. dr hab. Stanisław Piątek; dr hab. inż. Jacek Rak, prof. PG; dr inż. Łukasz Rybak; dr hab. inż. Kajetana Snopek, prof. PW; prof. dr hab. inż. Kamil Staniec; prof. dr hab. inż. Jacek Stefański; dr hab. inż. Krzysztof Szczypiorski, prof. PW; prof. dr hab. inż. Krzysztof Wesołowski; prof. dr hab. inż. Tadeusz Więckowski; dr inż. Andrzej Wilk, mgr inż. Dariusz Zmysłowski

REDAKCJA
ul. Czackiego 3/5, 00-043 Warszawa
www.przegladtelekomunikacyjny.pl
przeg.tel@sigma-not.pl
facebook.com/przegladtelekom/
linkedin.com/company/przegladtelekom/

WYDAWCA
WYDAWNICTWO SIGMA-NOT
ul. Ratuszowa 11, 03-450 Warszawa
tel. 22 818 09 18
e-mail: sekretariat@sigma-not.pl
www.sigma-not.pl

Zamówienia na reklamę przyjmuje Redakcja
lub Dział Reklamy i Marketingu,
tel. 22 827 43 65, e-mail: **reklama@sigma-not.pl**

SKŁAD I DRUK
Zakład Poligrafii i Kolportażu
Wydawnictwa SIGMA-NOT Sp. z o. o
ul. ks. J. Popiełuszki 19/21, 01-595 Warszawa,
tel. 22 833 40 69
e-mail: **drukarnia@sigma-not.pl**

Publikowane artykuły naukowe są recenzowane przez samodzielnych pracowników nauki

Redakcja nie ponosi odpowiedzialności za treść ogłoszeń. Zastrzega sobie prawo do skracania i adiustacji nadesłanych materiałów.

Indeks 35722 Nakład do 500 egz (w tym wersja elektroniczna)

Dwumiesięcznik **Przegląd Telekomunikacyjny – Wiadomości Telekomunikacyjne** znajduje się w wykazie czasopism naukowych Ministerstwa Nauki i Szkolnictwa Wyższego.

Wersja papierowa Przeglądu Telekomunikacyjnego – WiadomościTelekomunikacyjnych jest wersją pierwotną.

SPIS TREŚCI • CONTENTS

TECHNOLOGIA 5G W ZASTOSOWANIACH MILITARNYCH
5G TECHNOLOGY FOR MILITARY USAGE
MAREK SUCHAŃSKI, ANDRZEJ BĘBEN 9

PRZEGLĄD METOD WIELODOSTĘPU NIEORTOGONALNEGO I MOŻLIWOŚCI ICH ZASTOSOWANIA W SYSTEMACH 5G I 6G
OVERVIEW OF NON-ORTHOGONAL MULTIPLE ACCESS METHODS AND THEIR POSSIBLE APPLICATIONS IN 5G AND 6G SYSTEMS
KRZYSZTOF WESOŁOWSKI 17

MODELE NUMERYCZNE CIAŁA CZŁOWIEKA UMOŻLIWIAJĄCE ANALIZĘ I PROJEKTOWANIE NOWOCZESNYCH SYSTEMÓW ŁĄCZNOŚCI BEZPRZEWODOWEJ
NUMERICAL HUMAN BODY MODELS FOR THE ANALYSIS AND DESIGN OF MODERN WIRELESS COMMUNICATION SYSTEMS
ŁUKASZ JANUSZKIEWICZ 25

KOMPRESJA WIZJI DLA MASZYN
VIDEO COMPRESSION FOR MACHINES
MAREK DOMAŃSKI 32

GENERATYWNA SZTUCZNA INTELIGENCJA A STUDIOWANIE TELEINFORMATYKI
GENERATIVE ARTIFICIAL INTELLIGENCE AND STUDYING INFORMATION AND COMMUNICATION TECHNOLOGY
ARTUR JANICKI 37

„SECURITY-BY-DESIGN" - BEZPIECZEŃSTWO SYSTEMOWE ICT - W PRZEPISACH, NORMACH I PRAKTYCE
"SECURITY-BY-DESIGN" – SYSTEMIC ICT SECURITY – IN REGULATIONS, STANDARDS, AND PRACTICE
PAWEŁ K. KOSTKIEWICZ 42

pełne wydanie zeszytu dostępne jest na sigma-not.pl

RADA PROGRAMOWO-NAUKOWA

Prof. dr hab. inż. Józef Woźniak – przewodniczący

Dr hab. inż. Andrzej Bęben, prof. PW – Politechnika Warszawska
Mgr inż. Jacek Cyrek – WB Electronics
Mgr inż. Paweł Biskupski – Systemics PAB
Dr hab. inż. Jerzy Domżał, prof. AGH – Akademia Górniczo-Hutnicza w Krakowie
Dr inż. Andrzej Dulka – PIIT
Dr hab. inż. Sławomir Hausman, prof. PŁ – Politechnika Łódzka
Prof. dr hab. inż. Andrzej Jajszczyk – Akademia Górniczo-Hutnicza w Krakowie
Mgr inż. Stefan Kamiński – KIGEiT
Dr hab. inż. Jan Kelner, prof. WAT – Wojskowa Akademia Techniczna
Dr hab. inż. Adrian Kliks, prof. PP – Politechnika Poznańska
Prof. dr hab. inż. Józef Modelski – Politechnika Warszawska
Dr hab. inż. Marek Natkaniec, prof. AGH – Akademia Górniczo-Hutnicza w Krakowie
Prof. dr hab. Stanisław Piątek – Uniwersytet Warszawski
Dr hab. inż. Jacek Rak, prof. PG – Politechnika Gdańska
Dr inż. Łukasz Rybak – Wojskowa Akademia Techniczna
Dr hab. inż. Kajetana Snopek, prof. PW – Politechnika Warszawska
Prof. dr hab. inż. Kamil Staniec – Politechnika Wrocławska
Prof. dr hab. inż. Jacek Stefański – Politechnika Gdańska
Dr hab. inż. Krzysztof Szczypiorski, prof. PW – Politechnika Warszawska
Prof. dr hab. inż. Krzysztof Wesołowski – Politechnika Poznańska
Prof. dr hab. inż. Tadeusz Więckowski – Politechnika Wrocławska
Dr inż. Andrzej Wilk
mgr inż. Dariusz Zmysłowski – Wojskowa Akademia Techniczna

WYDAWNICTWO *SIGMA-NOT*

WARIANTY PRENUMERATY 2025

Prenumerata papierowa – czasopismo tylko w wersji papierowej, za jego dostarczenie doliczamy opłatę roczną;
Prenumerata cyfrowa – czasopismo wyłącznie w wersji cyfrowej, dostępne na Portalu Informacji Technicznej www.sigma-not.pl, prenumerator otrzyma indywidualny link dostępu do logowania na Portalu;
PRENUMERATA PAKIET PLUS – czasopismo w wersji papierowej (bez opłaty za dostarczanie prasy) oraz cyfrowej, a także dostęp do archiwum zaprenumerowanego tytułu na Portalu Informacji Technicznej www.sigma-not.pl wraz z indywidualnym linkiem do logowania.

Cena jednego egzemplarza – 60,00 zł

CENY PRENUMERATY ROCZNEJ:

- prenumerata papierowa: 348 zł brutto + (koszty wysyłki 21 zł);
- prenumerata cyfrowa: 300 zł brutto;
- prenumerata w wersji Pakiet PLUS (papierowa + cyfrowa + archiwum) 450,- zł.

PORTAL INFORMACJI TECHNICZNEJ www.sigma-not.pl – to największa polska internetowa baza czasopism i artykułów technicznych, wyposażona w szybką i skuteczną wyszukiwarkę tematyczną, umożliwiającą dostęp on-line do ponad 129 000 publikacji z różnych dziedzin.
Dla Państwa wygody logowanie do Portalu Informacji Technicznej stało się łatwiejsze. Portal zyskał nową, przejrzystą szatę graficzną, pojawiły się nowe funkcje ułatwiające zakup pojedynczych artykułów i całych zeszytów. Każdy prenumerator wersji cyfrowej oraz prenumeraty Pakietu PLUS otrzyma indywidulany link z kodem dostępu do zamówionego czasopisma.

Cena brutto zawiera 8% VAT na czasopisma w wersji papierowej oraz cyfrowej. W przypadku zmiany przez ustawodawcę stawki VAT na czasopisma i w konsekwencji zmiany ceny brutto, prenumerator zobowiązany jest do dopłaty różnicy.
Dla prenumeratorów zagranicznych obowiązuje cena według kursu waluty NBP (z dnia bezpośrednio poprzedzającego datę wystawienia faktury) plus koszty wysyłki.

PRENUMERATĘ ZAMAWIAMY

22 840 35 89
prenumerata@sigma-not.pl
www.sigma-not.pl

LISTA RECENZENTÓW

Prof. dr hab. Adam Abramowicz
Prof. dr hab. Marek Amanowicz
Prof. dr hab. Andrzej Dobrowolski
Prof. dr hab. Piotr Gajewski
Prof. dr hab. Marcin Iwanowski
Dr hab. Jacek Izydorczyk
Prof. dr hab. Tomasz Kacprzak
Prof. dr hab. Andrzej Karwowski
Prof. dr hab. Ryszard Katulski
Prof. dr hab. Bogdan Kwolek
Prof. dr hab. Roman Kubacki
Prof. dr hab. Lidia Łukasiak
Prof. dr hab. Stanisław Osowski
Dr hab. inż. Andrzej Paszkiewicz
Prof. dr hab. Krzysztof Perlicki
Prof. dr hab. Grzegorz Różański
Prof. dr hab. Jacek Stefański
Dr hab. Zenon Szczepaniak
Prof. dr hab. Maciej Walkowiak
Prof. dr hab. Krzysztof Wesołowski
Prof. dr hab. Ryszard Zieliński

Niniejszy zeszyt (4/2025) naszego dwumiesięcznika jest związany z KONFERENCJĄ RADIOKOMUNIKACJI I TEIELEINFORMATYKI 2025. Konferencja powstała w wyniku połączenia Krajowej Konferencji Radiokomunikacji, Radiofonii i Telewizji (KKRRiT) oraz Krajowego Sympozjum Telekomunikacji i Teleinformatyki (KSTiT). Organizatorem konferencji są Katedra Systemów i Sieci Radiokomunikacyjnych oraz Katedra Teleinformatyki Politechniki Gdańskiej.

Zeszyt zawiera całość materiałów konferencyjnych. Referaty plenarne są drukowane, a pełne teksty wszystkich referatów są zamieszczone na Portalu Informacji Technicznej SIGMA-NOT z nieodpłatnym dostępem https://sigma-not.pl/zeszyt-7863-przeglad-telekomunikacyjny-2025-4.html.

Stanowią one integralną część zeszytu. Wszystkie artykuły i referaty są recenzowane. Podstawowe informacje na temat konferencji znajdą Państwo w artykule wstępnym (II str. okładki), przygotowanym przez prof. dr. hab. inż. Jacka Stefańskiego oraz dr. hab. inż. Jacka Raka, prof. Politechniki Gdańskiej, współprzewodniczących Komitetu Programowego KRiT 2025.

LISTA RECENZENTÓW

Marek Amanowicz, Sławomir Ambroziak, Andrzej Bęben, Edyta Biernacka, Piotr Boryło, Piotr Chołda, Grzegorz Danilewicz, Marek Domański, Janusz Dudczyk, Sławomir Hausman, Łukasz Januszkiewicz, Wojciech Kabaciński, Sylwester Kaczmarek, Ryszard Katulski, Jan Kelner, Adrian Kliks, Mirosław Klinkowski, Katarzyna Kosek-Szott, Rafał Krenz, Paweł Kryszkiewicz, Sławomir Jerzy Kula, Mikołaj Leszczuk, Jerzy Łopatka, Józef Modelski, Jordi Mongay Batalla, Marek Natkaniec, Marcin Niemiec, Zdzisław Papir, Zbigniew Piotrowski, Jarosław Sadowski, Kajetana Snopek, Jacek Stefański, Marek Suchański, Wojciech Sułek, Szymon Szott, Krzysztof Walkowiak, Wojciech Wojtasiak, Robert Wójcik,Yevhen Yashchyshyn, Zbigniew Zakrzewski, Ryszard Zieliński.

Referaty sesji tematycznych zamieszczone na Portalu Informacji Technicznej SIGMA-NOT

ALGORYTMY SZTUCZNEJ INTELIGENCJI W SYSTEMACH RADIOWYCH

ALICJA OLEJNICZAK, JAROSŁAW SADOWSKI
ANALIZA WPŁYWU STRUKTURY SYGNAŁÓW RADIOWYCH NA SKUTECZNOŚĆ GŁĘBOKIEJ SIECI NEURONOWEJ 54

ŁUKASZ RYBAK, JANUSZ DUDCZYK, JAKUB OLSZEWSKI, PIOTR MAZUR
HIERARCHICZNE, NIEHIERARCHICZNE I GĘSTOŚCIOWE ALGORYTMY ANALIZY SKUPIEŃ W GRUPOWANIU SYGNAŁÓW RADAROWYCH 58

ADRIAN KUBOWICZ, JERZY ŁOPATKA
KONCEPCJA WIELOKANAŁOWEGO SYSTEMU WYBORU KANAŁU RADIOWEGO W OPARCIU O PREDYKCJĘ ZAJĘTOŚCI KANAŁÓW 62

KRZYSZTOF K. CWALINA, PIOTR RAJCHOWSKI, JAROSŁAW SADOWSKI
KONTRADYKTORYJNE INTERFERENCJE NEUROINTERFEJSU RADIOWEGO 5G-NR 66

PIOTR MAZUR, ŁUKASZ RYBAK, JANUSZ DUDCZYK, JAKUB OLSZEWSKI
WPŁYW WYMIAROWOŚCI DANYCH ROZPOZNANIA RADIOELEKTRONICZNEGO NA DOKŁADNOŚĆ ALGORYTMÓW KLASYFIKACJI 70

PIOTR OLESIŃSKI, MICHAŁ PANEK, RÓŻA GOŚCIEŃ
WYKORZYSTANIE UCZENIA MASZYNOWEGO DO EFEKTYWNEGO WYKRYWANIA ANOMALII ORAZ PREDYKCJI PRZEPUSTOWOŚCI W SIECIACH 5G 74

APLIKACJE I USŁUGI W INTERNECIE

ANDRZEJ MATIOLAŃSKI
ARCHITEKTURA PRZETWARZANIA BRZEGOWEGO DLA EFEKTYWNEJ ANALIZY WIDEO W SYSTEMACH MONITORINGU WIZYJNEGO 78

ARCHITEKTURY I PROTOKOŁY TELEKOMUNIKACYJNE

MARCIN CIOŁEK
IDENTYFIKACJA NIESTACJONARNYCH KANAŁÓW KOMUNIKACYJNYCH 82

MARCIN KUCHARCZYK, GRZEGORZ DZIWOKI, WOJCIECH SUŁEK
OPROGRAMOWANE Z OTWARTYMI ŹRÓDŁAMI W NAUCZANIU TELEKOMUNIKACJI 86

DAWID BUDNAROWSKI, AGNIESZKA CZAPIEWSKA, ANDRZEJ ŁUKSZA, RYSZARD STUDAŃSKI, ŁUKASZ WOJEWÓDKA, ANDRZEJ ŻAK
PROTOKÓŁ KOMUNIKACYJNY W MOBILNYCH SYSTEMACH ŁĄCZNOŚCI PODWODNEJ 90

ZUZANNA KONOPEK, JERZY DOMŻAŁ, ROBERT WÓJCIK
TUNELOWANIE ICMP JAKO MEDIUM SZYFROWANEJ I NIEZAWODNEJ KOMUNIKACJI 94

ASPEKTY PRAWNE, EKONOMICZNE I SPOŁECZNE W TELEKOMUNIKACJI

MAGDALENA ŚLUSARCZYK, DAWID JUSZKA
COMPARISON BETWEEN EUROPEAN UNION'S AI ACT AND ISO/IEC 42001 REQIREMENTS 99

WALDEMAR KARCZ
METROLOGIA PRAWNA W ASPEKCIE USTAW I ROZPORZĄDZEŃ KRAJOWYCH DOTYCZĄCYCH PÓL ELEKTROMAGNETYCZNYCH 103

ASPEKTY RADIOWE SIECI 5G I 6G

PIOTR RAJCHOWSKI, KRZYSZTOF CWALINA, OLGA BŁASZKIEWICZ, ALICJA OLEJNICZAK, JAROSŁAW SADOWSKI
ANALIZA WPŁYWU SYGNAŁÓW INTERFERUJĄCYCH NA PRACĘ INTERFEJSU RADIOWEGO 5G-NR 108

KRZYSZTOF MALON, PAWEŁ SKOKOWSKI, OLGA BŁASZKIEWICZ, KRZYSZTOF CWALINA, ALICJA OLEJNICZAK, PIOTR RAJCHOWSKI
ANALIZA WPŁYWU ŚRODOWISKA TESTOWEGO NA SKUTECZNOŚĆ DETEKCJI INTERFERENCJI W INTERFEJSIE RADIOWYM 5G-NR 112

MARCIN PAKUŁA, MARCIN HOFFMANN, PAWEŁ KRYSZKIEWICZ
BUDOWA BLIŹNIAKA CYFROWEGO ŚRODOWISKA RADIOWEGO Z UŻYCIEM METODY ŚLEDZENIA PROMIENI 117

FILIP FERENC, MICHAŁ SZCZĄCHOR
DETEKCJA PREAMBUŁ PRACH JAKO PROBLEM KLASYFIKACJI WIELOKLASOWEJ: PODEJŚCIE UCZENIA MASZYNOWEGO Z WYKORZYSTANIEM SVM 121

JAKUB BORKOWSKI, JACEK WSZOŁEK, MICHAŁ WĄGROWSKI, JUKKA LEMPIAINEN
FWA PERFORMANCE IMPROVEMENT IN 5G NETWORKS THROUGH MIMO PRECODING OPTIMIZATION 125

DAWID ROSOŁOWSKI, DANIEL GRYGLEWSKI, WOJCIECH WOJTASIAK, WOJCIECH KAZUBSKI, MICHAŁ KAJCZUK
IMPLEMENTACJA MECHANIZMÓW OSZCZĘDZANIA ENERGII W NADAJNIKACH 5G – STUDIUM 129

KONRAD BOŻEK, PAWEŁ SKOKOWSKI, JAN M. KELNER
IMPLEMENTACJA SYNCHRONICZNEJ GENERACJI ZAKŁÓCEŃ CELOWYCH SYSTEMÓW LTE I 5G W ŚRODOWISKU GNU RADIO Z WYKORZYSTANIEM PLA TFORMY USRP X310 133

BARTŁOMIEJ BRYŚ, PIOTR PYDA
MECHANIZM PRZECIWDZIAŁANIA ZAKŁÓCENIOM W SIECI 5G 137

ADAM SAMORZEWSKI, ADRIAN KLIKS
METODY REDUKCJI STRAT PROPAGACYJNYCH W SIECIACH WSPOMAGANYCH TECHNOLOGIĄ IPR 141

PAWEŁ HATKA, KAROLINA LENARSKA, ADRIAN KLIKS
OCENA DOKŁADNOŚCI ESTYMACJI CHARAKTERYSTYKI ODBICIOWEJ MATRYCY RMA NA PODSTAWIE MODELU TANGA 145

MARTA SIERADZKA, ŁUKASZ KUŁACZ, ADRIAN KLIKS
SYSTEM STEROWANIA REKONFIGUROWALNYMI MACIERZAMI ANTENOWYMI: DETEKCJA ZAJĘTOŚCI WIDMA. 149

SŁAWOMIR HAUSMAN, PIOTR KORBEL
WPŁYW ZJAWISK ATMOSFERYCZNYCH NA BILANS ŁĄCZA W SYSTEMACH 6G WYKORZYSTUJĄCYCH PASMA SUBTERAHERCOWE. 153

ŁUKASZ JOPEK, SŁAWOMIR HAUSMAN
WPŁYW WZAJEMNYCH SPRZĘŻEŃ POMIĘDZY META-ATOMAMI NA PARAMETRY META-POWIERZCHNI NA PRZYKŁADZIE INTELIGENTCH POWIERCHNI ODBIJAJĄCYCH. 157

ASPEKTY SIECI RDZENIOWYCH 5G I 6G

MIROSŁAW KLINKOWSKI
ANALIZA DYNAMICZNEJ REALOKACJI JEDNOSTEK DU W SIECIACH XHAUL Z KOMUTACJĄ PAKIETÓW PRZY ZMIENNYM RUCHU 161

BEZPIECZEŃSTWO SIECI I SYSTEMÓW TELEINFORMATYCZNYCH ORAZ BEZPIECZEŃSTWO NORMATYWNE

NORBERT ROCZNIAK, KRZYSZTOF RUSEK
ANALIZA SKUTECZNOŚCI REGRESJI LOGISTYCZNEJ W KLASYFIKACJI ATAKÓW DDOS 165

PIOTR KRAWIEC, JORDI MONGAY BATALLA
LEKKA METODYKA EWALUACJI CYBERBEZPIECZEŃSTWA KOMPONENTÓW OPEN RAN. 169

ANDRES VEJAR, PIOTR CHOŁDA
MACHINE LEARNING BASED DETECTION OF ATTACKS AIMED AT THE PACKET FORWARDING CONTROL PROTOCOL WITHIN THE 5G CONTROL PLANE 173

BEZPRZEWODOWE SIECI SENSOROWE I AD-HOC

TOMASZ JÓŹWIK
WYZNACZANIE ENERGOOSZCZĘDNYCH ŚCIEŻEK ROUTINGU W SIECIACH FANET Z WYKORZYSTANIEM MODELI LLM SONAR 177

KOMPATYBILNOŚĆ ELEKTROMAGNETYCZNA SYSTEMÓW I URZĄDZEŃ RADIOWYCH

MICHAŁ REWIEŃSKI
ANALIZA KOMPATYBILNOŚCI ELEKTROMAGNETYCZNEJ OPARTA NA MAKROMODELACH KAWAŁKAMI LINIOWYCH PO TRAJEKTORIACH 181

EMILIA SAKOWSKA, LESZEK NOWOSIELSKI
SYMULACJA PROCESU REKONSTRUKCJI OBRAZU NA PODSTAWIE EMISJI UJAWNIAJĄCEJ SYGNAŁU WIDEO. 185

OLGA BŁASZKIEWICZ, ALICJA OLEJNICZAK, JAROSŁAW SADOWSKI, JACEK STEFAŃSKI
WYTWARZANIE ZNIEKSZTAŁCEŃ MULTIPLIKATYWNYCH ANTENĄ OBCIĄŻONĄ ZMIENNĄ IMPEDANCJĄ 189

KRYPTOGRAFIA I MECHANIZMY CYBERBEZPIECZEŃSTWA

MIKITA VOSTRY, PIOTR CHOŁDA
A COMPARISON OF QAOA VS. VQE ALGORITHMS . . 193

BENIAMIN JANKOWSKI, KAMIL SZYDŁOWSKI, PIOTR CHOŁDA
ALGORYTM KRYPTOGRAFII POSTKWANTOWEJ OPARTY NA IZOGENIACH KRZYWYCH ELIPTYCZNYCH: IMPLEMENTACJA PRAKTYCZNA 197

ZUZANNA PIETRZAK, FILIP OPIŁKA, MARCIN NIEMIEC
ANALIZA WYBRANYCH IMPLEMENTACJI ALGORYTMÓW KRYPTOGRAFII POSTKWANTOWEJ. 201

KACPER GÓRA, JAKUB ZIĘBA, ANIELA CABAK, PAWEŁ RYDZ, ARINA PASHKEVICH, JAN DERKACZ
EFEKTYWNOŚĆ ALGORYTMÓW SZYFROWANIA W SYMULOWANYM ŚRODOWISKU CHMUROWYM: OPTYMALIZACJA I DYNAMICZNE ZARZĄDZANIE ZASOBAMI 205

JAKUB ŁOMZIK, PIOTR CHOŁDA
IMPLEMENTACJA I PORÓWNANIE ATAKÓW KLASYCZNYCH NA KRYPTOGRAFIĘ POSTKWANTOWĄ OPARTĄ NA KODACH. 209

MACIEJ KACZYŃSKI, ZBIGNIEW PIOTROWSKI
NEURONOWY MECHANIZM SZYFROWANIA RUCHU SIECIOWEGO UŻYTKOWNIKA W SIECIACH TELEKOMUNIKACYJNYCH 213

MARIUSZ BOROWSKI
NIEŁATWA DROGA DO 256 BITOWEGO BEZPIECZEŃSTWA W INTERFEJSIE RADIOWYM SIECI 5G NA PRZYKŁADZIE RODZINY ALGORYTMÓW SNOW 217

PRZEMYSŁAW PAWLITKO, NATALIA MOĆKO, MARCIN NIEMIEC, PIOTR CHOŁDA
PORÓWNANIE ALGORYTMÓW SHORA I REGEVA . . . 221

ROBERT WICIK, MAREK LEŚNIEWICZ, PIOTR KOMOROWSKI, MARCIN GRZONKOWSKI
PRAKTYCZNE METODY GENEROWANIA I TESTOWANIA BINARNYCH CIĄGÓW LOSOWYCH NA POTRZEBY SYSTEMÓW KRYPTOGRAFICZNYCH 225

LOKALNE SIECI BEZPRZEWODOWE WLAN

MONIKA DROZDOWSKA, NARCIS CARDONA
O INTELIGENTNYM PROJEKTOWANIU SIECI KONFERENCYJNYCH W PAŚMIE FAL MILIMETROWYCH W ŚRODOWISKACH ZABUDOWANYCH 229

OPTOTELEKOMUNIKACJA

KRZYSZTOF PYRCHLA, ROBERT BOGDANOWICZ
BADANIA AB INITIO WŁAŚCIWOŚCI ELEKTROOPTYCZNYCH NANOWARSTW B-ASXP1−X JAKO MATERIAŁÓW DO NOWEJ GENERACJI FOTODETEKTORÓW DLA PASM TELEKOMUNIKACYJNYCH 233

JULIUSZ BOJARCZUK, JAKUB KASJANOWICZ, MICHAŁ MARZĘCKI, GRZEGORZ STĘPNIAK
KOMPENSACJA NIELINIOWOŚCI DLA KOMUNIKACJI W ŚWIETLE WIDZIALNYM. 237

ZBIGNIEW ZAKRZEWSKI
KONCEPCJA SIECI RADIOWO-FOTONICZNEJ OPARTEJ NA OTWARTEJ ARCHITEKTURZE URZĄDZEŃ WĘZŁOWYCH 241

MARCIN GNYBA
SPEKTROSKOPIA RAMANOWSKA Z SONDAMI ŚWIATŁOWODOWYMI – WYBRANE ZASTOSOWANIA 245

OPTYMALIZACJA I ALGORYTMY SZTUCZNEJ INTELIGENCJI W SIECIACH TELEKOMUNIKACYJNYCH

SEBASTIAN ZARĘBSKI, KRZYSZTOF RUSEK, PIOTR CHOŁDA
AN APPROACH FOR STATISTICAL VALIDATION OF MACHINE LEARNING MODELS IN TELECOMMUNICATION. 249

SZYMON BETKA, SYLWESTER KACZMAREK, MAREK M. LANDOWSKI
KLASYFIKACJA RUCHU INTERNETOWEGO ZA POMOCĄ PRZETWARZANIA NEUROMORFICZNEGO . 253

MATEUSZ BURCZANIUK, TOMASZ KURYŁOWICZ, MICHAŁ OSTAPOWICZ, HOANG NGHIA LE
OPTYMALIZACJA SIECI DOSTĘPOWYCH PRZY UŻYCIU ALGORYTMÓW EWOLUCYJNYCH. 257

MAKSYMILIAN WOJNAR
UCZENIE REPREZENTACJI SIECI IEEE 802.11 Z WYKORZYSTANIEM GRAFOWYCH SIECI NEURONOWYCH 263

MICHAŁ BŁASZCZAK, IRENEUSZ JABŁOŃSKI, ALEKSANDRA KNAPIŃSKA, KRZYSZTOF WALKOWIAK
WYKORZYSTANIE MULTIMODALNEGO UCZENIA MASZYNOWEGO ORAZ WYJAŚNIALNEGO AI DO AUTOMATYZACJI MONITOROWANIA I ZARZĄDZANIA STANEM SIECI KOMÓRKOWEJ 5G+. 267

KAMIL SZCZĘCH
SAMOLUBNE ZACHOWANIA AGENTÓW UCZENIA ZE WZMOCNIENIEM W SIECIACH IEEE 802.11 271

POMIARY W TELEKOMUNIKACJI

MACIEJ BARZOWSKI, SYLWESTER KACZMAREK, BARTOSZ MATWIEJCZYK, MAGDALENA MŁYNARCZUK
KONFIGUROWALNY GENERATOR RUCHU SIECIOWEGO W TECHNOLOGII FPGA DO TESTOWANIA SIECI DO 100 GBIT/S. 275

PIOTR GRALL
POPRAWA DOKŁADNOŚCI OKREŚLENIA GŁĘBOKOŚCI AKWENU ZA POMOCĄ ECHOSONDY INTERFEROMETRYCZNEJ 280

PROPAGACJA FAL RADIOWYCH

SEBASTIAN DZIEDZIEWICZ
ALGORYTM ZNAJDOWANIA ZER I BIEGUNÓW FUNKCJI ZESPOLONEJ W ANALIZIE PROPAGACJI FAL ELEKTROMAGNETYCZNYCH. 284

TOMASZ A. MIŚ
DETEKCJA LOGICZNO-KORELACYJNA EFEKTU LUKSEMBURSKIEGO 288

TOMASZ JAWORSKI, CYRYL PRENTKI, PAWEŁ KRYSZKIEWICZ, ADRIAN KLIKS
GENEROWANIE I ZASTOSOWANIE OGRANICZONEJ KSIĄŻKI KODOWEJ DLA RIS W SCENARIUSZU WEWNĄTRZBUDYKOWYM 292

TOMASZ GRASZA, PIOTR ZDANKOWSKI, JAROSŁAW WOJTUŃ, RAFAŁ PRZESMYCKI, JAN KELNER, CEZARY ZIÓŁKOWSKI
PARAMETRY TRANSMISYJNE STANOWISKA DO SONDOWANIA KANAŁÓW RADIOWYCH W ZAKRESIE FAL MILIMETROWYCH 296

SŁAWOMIR J. AMBROZIAK, AGNIESZKA CZAPIEWSKA, KRZYSZTOF K. CWALINA, PIOTR RAJCHOWSKI, FILIPE D. CARDOSO, MANUEL M. FERREIRA, MARIELLA SÄRESTÖNIEMI, ANNA PIETRENKO-DĄBROWSKA, SŁAWOMIR KOZIEŁ, LUIS M. CORREIA
POMIARY PARAMETRÓW KANAŁU RADIOWEGO W ZAKRESIE FAL MILIMETROWYCH DLA SIECI BAN WEWNĄTRZ POJAZDU 300

PAWEŁ HATKA, MARCEL GARCZYK, PAWEŁ PŁACZKIEWICZ, DAWID BRZĄKAŁA, MIKOŁAJ WYPYCH, PIOTR SKRZYPCZAK, MARCIN PAKUŁA, MACIEJ NIEDŹWIECKI, PAWEŁ SROKA, MICHAŁ SYBIS
PROJEKTOWANIE KSIĄŻKI KODOWEJ RIS Z WYKORZYSTANIEM DANYCH POMIAROWYCH 304

BOGDAN ULJASZ
ZESTAW BADAWCZY DO POMIARU PARAMETRÓW CHARAKTERYSTYCZNYCH ŁĄCZA JONOSFERYCZNEGO. 308

RADIOKOMUNIKACJA MORSKA, LOTNICZA I SATELITARNA

RYSZARD ZIELIŃSKI
ANALIZA MOŻLIWOŚCI WDROŻENIA USŁUG SATELITARNYCH W SIECI KOMÓRKOWEJ 5G I 6G . . . 314

JAROSŁAW MAGIERA
ANALIZA TRANSMISJI RADIOWEJ KOMERCYJNYCH BSP. 318

RYSZARD KATULSKI
RADIOKOMUNIKACJA MORSKA W ŚWIETLE PRZEPISÓW IMO 322

ROZWIĄZANIA DLA SYSTEMÓW I SIECI CHMUROWYCH

TOMASZ ŚMIAŁKOWSKI, WOJCIECH SŁOWIŃSKI, PIOTR SZCZUKO
INFOLIGHT: WIELOSENSOROWA INTELIGENTNA PLATFORMA STEROWANIA OŚWIETLENIEM 327

SYSTEMY I SIECI KOMÓRKOWE

OLGA BŁASZKIEWICZ, PIOTR RAJCHOWSKI, JAROSŁAW SADOWSKI
DOBÓR POSTACI SYGNAŁU SYNCHRONIZACYJNEGO NPSS W NB-IOT PRZY UŻYCIU ALGORYTMU GENETYCZNEGO 331

SYSTEMY I SIECI RADIOKOMUNIKACYJNE DO ZASTOSOWAŃ SPECJALNYCH

JAROMIR GAJEWSKI, JERZY ŁOPATKA
KONCEPCJA STANOWISKA POMIAROWEGO DO BADANIA MOŻLIWOŚCI LOKALIZACJI EMITERÓW EM W ŚRODOWISKU ZURBANIZOWANYM 335

KRZYSZTOF K. CWALINA, PIOTR RAJCHOWSKI, JAROSŁAW SADOWSKI, PAWEŁ SKOKOWSKI, KRZYSZTOF MALON, MICHAŁ KRYK, KRZYSZTOF MAŚLANKA, JAN M. KELNER
ZAKRES STOSOWALNOŚCI OPRACOWANEGO DETEKTORA SYGNATUR RADIOWYCH BSP 339

SYSTEMY I SIECI TELEINFORMATYCZNE

TOMASZ RACHWALIK, ROBERT BIAŁAS, PRZEMYSŁAW DĄBROWSKI
IMPLEMENTACJA AUTOENKODERA DO ETYKIETOWANIA DANYCH I DO DETEKCJI NOWYCH ATAKÓW SIECIOWYCH 344

MARCIN KOZŁOWSKI, MIROSŁAW PŁAZA, MAŁGORZATA LUCIŃSKA, MARCIN KOWALCZYK
METODA PREDYKCJI RUCHU WYCHODZĄCEGO W SYSTEMACH KLASY CALL/CONTACT CENTER WYKORZYSTUJĄCA ALGORYTM UCZENIA ZE WZMOCNIENIEM 348

GRZEGORZ DANILEWICZ, REMIGIUSZ RAJEWSKI
NOWY ALGORYTM PRZESTROJEŃ DLA PÓL KOMUTACYJNYCH TYPU SWS1 352

WOJCIECH GUMIŃSKI, ROBERT LATOSZEWSKI
TRANSMISJA DANYCH W SIECIACH O WYSOKIEJ STOPIE BŁĘDÓW Z WYKORZYSTANIEM PROTOKOŁÓW QUIC I HTTP/3. 357

MICHAŁ MAJ, PIOTR BORYŁO
WERYFIKACJA MOŻLIWOŚCI OPRACOWANIA SKANERA PASMA PMR Z WYKORZYSTANIEM RADIA DEFINIOWANEGO PROGRAMOWO 360

SYSTEMY RADIONAWIGACYJNE I RADIOLOKALIZACYJNE

KRZYSZTOF GROMADA
AUTONOMICZNY LOT BEZ GNSS I ŁĄCZNOŚCI: ZASTOSOWANIE RADARU Z SYNTETYCZNĄ APERTURĄ ORAZ AI W NAWIGACJI BEZZAŁOGOWCÓW 366

JACEK STEFAŃSKI, JAROSŁAW SADOWSKI
ANALIZA PORÓWNAWCZA ASYNCHRONICZNYCH METOD RADIOLOKALIZACYJNYCH 370

GABRIEL JEKATERYŃCZUK, ZBIGNIEW PIOTROWSKI
ANALIZA WPŁYWU ROZMIESZCZENIA MIKROFONÓW I CZĘSTOTLIWOŚCI PRÓBKOWANIA SYGNAŁU AUDIO NA DOKŁADNOŚĆ LOKALIZACJI ŹRÓDŁA DŹWIĘKU 376

MARCIN KOŁAKOWSKI, VITOMIR DJAJA-JOŚKO
DYNAMICZNA ESTYMACJA BŁĘDU LOKALIZACJI W SYSTEMIE OPARTYM NA WI-FI. 380

MACIEJ KAWKA, JAKUB KOŁODZIEJ, PAWEŁ SKOKOWSKI
NAWIGACJA INERCYJNA PRZY POMOCY SEKWENCYJNEJ SIECI NEURONOWEJ Z WYKORZYSTANIEM NIEWYSPECJALIZOWANYCH CZUJNIKÓW IMU 384

PIOTR KOCIK, GRZEGORZ CICHOCKI, STANISŁAW NOWAK, ŁUKASZ JEZIERSKI, KRZYSZTOF CICHOŃ
PORÓWNANIE ALGORYTMÓW POZYCJONOWANIA DLA SYSTEMU LOKALIZACJI WEWNĄTRZBUDYNKOWEJ CZASU RZECZYWISTEGO Z WYKORZYSTANIEM SIECI WI-FI 389

BARTOSZ CZAJA, KRZYSZTOF MAŚLANKA
SAL - LOKALIZACJA Z SYNTETYCZNĄ KOTWICĄ 393

MARCIN ŻYWEK
WYKORZYSTANIE PRAKTYCZNIE ZWERYFIKOWANYCH MODELI PREDYKCJI ZASIĘGOWEJ RADARU PASYWNEGO DO OPTYMALIZACJI WYBORU NADAJNIKÓW OKAZJONALNYCH. 398

TECHNIKA ANTENOWA

ŁUKASZ JANUSZKIEWICZ, REMIGIUSZ DANYCH, MACIEJ ŁASKI
ANTENA INTELIGENTNEGO ZEGARKA PRACUJĄCA W PAŚMIE 2,4 GHZ. 402

KORNELIA BENDZEL, ŁUKASZ JANUSZKIEWICZ
ANTENA NA PASMO 2,4 GHZ ZINTEGROWANA Z HEŁMEM 406

RAFAŁ PRZESMYCKI, MAREK BUGAJ, KLAUDIA JANUSZCZAK
MIKROPASKOWE UKŁADY ANTENOWE NA PASMO HIGH-BAND SYSTEMU 5G 410

RAFAŁ PRZESMYCKI, MAREK BUGAJ, KLAUDIA JANUSZCZAK
SZEROKOPASMOWA ANTENA MIKROPASKOWA DO ZASTOSOWAŃ W SYSTEMACH 5G MMWAVE I 6G 414

STANISŁAW SŁOWIŃSKI, MICHAŁ BANDYCH, GRZEGORZ BOGDAN, YEVHEN YASHCHYSHYN
SZESNASTOELEMENTOWY SZYK ANTENOWY Z MODULACJĄ CZASOWĄ I SKANOWANIEM WIĄZKI W PAŚMIE ISM 418

MARIUSZ ZAMŁYŃSKI, PIOTR SŁOBODZIAN
ZASADA WZAJEMNOŚCI W KONTEKŚCIE MODELOWANIA SPRZĘŻEŃ WZAJEMNYCH W UKŁADACH ANTENOWYCH. 422

TECHNIKI MULTIMEDIALNE

DOĞUKAN ÖZTÜRK, MIKOŁAJ LESZCZUK
COMPARISON OF IMPROVEMENT ON IQA METRICS WITH DEEP LEARNING-BASED VISUAL QUALITY ENHANCEMENT MODELS 426

SYED UDDIN, MICHAL GREGA, MIKOŁAJ LESZCZUK, WAQAS UR RAHMAN
DATASET EVALUATING QOE IN SEGMENT BASED ADAPTIVE BITRATE ALGORITHMS. 430

SERATOWICZ ŁUKASZ, DERKACZ JAN
ANALIZA I OPTYMALIZACJA MODELU SZTUCZNEJ INTELIGENCJI DO ODWZOROWANIA SYLWETKI CZŁOWIEKA NA PODSTAWIE OBRAZU Z KAMERY INTERNETOWEJ. 434

JAKUB STANKOWSKI, ADRIAN ŚLIWIŃSKI, ADAM GRZELKA, TOMASZ GRAJEK
ANALIZA MOŻLIWOŚCI ZRÓWNOLEGLENIA OBLICZEŃ W KODERZE I DEKODERZE JPEG 438

PAWEŁ DUSZEJKO, ZBIGNIEW PIOTROWSKI
DETEKCJA MANIPULACJI W OBRAZACH WYSOKIEJ ROZDZIELCZOŚCI Z UŻYCIEM TRANSPARENTNEGO ZNAKU WODNEGO. 442

TOMASZ KONASZYŃSKI, AVRAJYOTI DUTTA, DAWID JUSZKA, MIKOŁAJ LESZCZUK
DZIAŁANIA MOTYWUJĄCE TESTERA JAKO ELEMENT PRZYGOTOWANIA EKSPERYMENTU QOE 447

FILIP KORUS, DAWID JUSZKA, MICHAŁ GREGA, MIKOŁAJ LESZCZUK, MICHAŁ SUCHOŃ, AVRAJYOTI DUTTA, DOMINIKA WANAT
HD-DSEQUA: HIGH DEFINITION DIVERSE VIDEO SEQUENCE DATASET FOR QUALITY ASSESSMENT WITH INDICATORS 451

ADAM GRZELKA
SYSTEMY WIZYJNE SWOBODNEGO PUNKTU WIDZENIA O WYSOKIEJ POSTRZEGANEJ JAKOŚCI USŁUG 455

JÓZEF KOTUS, MACIEJ SZCZODRAK, PIOTR SZCZUKO, GRZEGORZ SZWOCH
KLASYFIKACJA ZDARZEŃ AKUSTYCZNYCH W SYSTEMIE MONITORINGU BEZPIECZEŃSTWA OPARTA NA UCZENIU MASZYNOWYM 459

JAKUB STANKOWSKI, ARTUR FOJUT, DAMIAN KARWOWSKI, TOMASZ GRAJEK
ŁĄCZNE WYKORZYSTANIE TECHNIKI RDOQ I METODY OPTYMALIZACJI KODÓW HUFFMANA W KODERZE JPEG 463

MATEUSZ OLSZEWSKI, LUCJAN JANOWSKI, RAFAŁ FIGLUS, MIKOŁAJ LESZCZUK
OCENA JAKOŚCI DOŚWIADCZENIA W WARUNKACH ZWIĘKSZONEGO REALIZMU ŻYCIOWEGO 467

MARTA BISTROŃ, ZBIGNIEW PIOTROWSKI
OD GENERACJI DO OCHRONY: MODELE DYFUZYJNE W ZNAKOWANIU WODNYM OBRAZÓW 471

WERONIKA SOCZEK, JAN DERKACZ
PORÓWNANIE ALGORYTMÓW ZNAKOWANIA WODNEGO I ICH ODPORNOŚĆ NA ZNIEKSZTAŁCENIA 475

AVRAJYOTI DUTTA, BURAK GIZLICE, DAWID JUSZKA, MIKOŁAJ LESZCZUK
REJECTION OUTLIERS DETECTION IN VIDEO QUALITY ASSESSMENT: A CROWDSOURCING STUDY479

NATALIA JAKUBIEC, LUCJAN JANOWSKI, JAKUB NAWAŁA, KAMIL KONIUCH
REKOMENDACJA POLSKIEJ SKALI ACR NA PODSTAWIE BADANIA JAKOŚCI DOŚWIADCZEŃ Z WYKORZYSTANIEM PROTOKOŁU GŁOŚNEGO MYŚLENIA 483

TOMASZ WALCZYNA, ZBIGNIEW PIOTROWSKI
REKONSTRUKCJA OBRAZU NA PODSTAWIE ZNAKOWANIA WODNEGO ZERO-BIT Z WYKORZYSTANIEM SIECI NEURONOWEJ 487

AVRAJYOTI DUTTA, TOMASZ KONASZYNSKI, DAWID JUSZKA, MIKOŁAJ LESZCZUK
STATISTICAL EVALUATION OF SUBJECTIVE VIDEO QUALITY ON A CROWDSOURCING PLATFORM 491

DOMINIKA WANAT, DAWID JUSZKA
SUBIEKTYWNA OCENA QOE WIDEO NA PODSTAWIE REAKCJI NA DEGRADACJĘ JAKOŚCI W RÓŻNYCH GRUPACH WIEKOWYCH 495

ADRIANNA SOBCZAK, ZBIGNIEW PIOTROWSKI
SYNTION – TERMINOLOGIA I KLASYFIKACJA 499

MEHRUNNISA, LALLOUANI BOUCHAKOUR, DAWID JUSZKA, YI ZHANG, MIKOŁAJ LESZCZUK
UNDERWATER IMAGE ENHANCEMENT WITH DEEP LEARNING: A MOS-GUIDED, MULTI-CRITERIA LOSS FUNCTION APPROACH 503

JAKUB SUDER
VISION INSPECTION USING ARTIFICIAL INTELLIGENCE TO IMPROVE SAFETY OF AERODROME 507

JAKUB KWAŚNIAK, MATEUSZ MAJTKA, MATEUSZ LORKIEWICZ, TOMASZ GRAJEK
ZASTOSOWANIE SIECI RESNET DO WYBORU PODZIAŁU JEDNOSTKI CTU W HEVC ALL INTRA 511

TECHNIKI WARSTWY FIZYCZNEJ W TRANSMISJI RADIOWEJ

OLGA BŁASZKIEWICZ, KRZYSZTOF K. CWALINA, ALICJA OLEJNICZAK, PIOTR RAJCHOWSKI, JAROSŁAW SADOWSKI
DETEKCJA POCIĄGÓW PRZY UŻYCIU SYGNAŁÓW OKAZJONALNYCH SIECI LTE 515

WOJCIECH SUŁEK, JAKUB HYLA, MARCIN KUCHARCZYK
NIEBINARNE KODOWANIE LDPC TYPU RAPTOR O KRÓTKIEJ DŁUGOŚCI BLOKU 519

TRANSMISJA RADIOWA W SYSTEMACH INTERNETU RZECZY

BROWARSKI JAKUB, NIEPLA DANIEL, NIEPLA ROBERT, DEMKOWICZ JERZY
PROJEKT I REALIZACJA SYSTEMU ROZPROSZONEGO Z WYKORZYSTANIEM TECHNOLOGII LORA 523

ZAAWANSOWANE TECHNOLOGIE ZARZĄDZANIA SIECIĄ (NFV, SDN, NETWORK SLICES)

BARTŁOMIEJ MASTEJ
INTELIGENCJA ROZPROSZONA W MIGRACJI FUNKCJI VNF 527

BARTOSZ KOPEĆ-PERSIŃSKI, ANDRZEJ BĘBEN
METODA ORKIESTRACJI WIRTUALNYCH ZAPÓR SIECIOWYCH W SIECIACH OPERATORSKICH 531

DOI: 10.15199/59.2025.4.1

TECHNOLOGIA 5G W ZASTOSOWANIACH MILITARNYCH

5G TECHNOLOGY FOR MILITARY USAGE

dr hab. inż. Marek Suchański, prof. WIŁ, Wojskowy Instytut Łączności – Państwowy Instytut Badawczy
dr hab. inż. Andrzej Bęben, prof. PW, Politechnika Warszawska

STRESZCZENIE: Technologia 5G wprowadza nowe możliwości dla mobilnych systemów łączności. Wykorzystuje ona nowe pasma częstotliwości, zapewnia znacząco większe prędkości transmisji i mniejsze opóźnienia przekazu danych oraz umożliwia skalowanie wydajności i obsługę dużej liczby urządzeń. Te właściwości stały się źródłem analiz przydatności tej technologii dla zastosowań militarnych. Dlatego w prezentowanym opracowaniu przedstawiono stan prac analitycznych i eksperymentalnych mających na celu dokonanie oceny możliwości wykorzystania technologii 5G w specyficznych scenariuszach właściwych dla domeny militarnej.

SŁOWA KLUCZOWE: wojskowa sieć 5G, luki technologiczne 5G, sieci testowe 5G i eksperymenty, kierunki rozwoju 5G/6G

ABSTRACT: The 5G technology brings many new features to mobile communications. It uses new radio frequencies, provides significantly higher transmission rates and lower latencies, and allows for scaling performance for many devices. These properties motivate analyses of the usability of the 5G technology for military applications. In this paper, we present the state of analytical and experimental studies focused on assessing the possibilities of applying 5G technology in specific military scenarios.

KEYWORDS: military 5G networks, 5G technology gaps, 5G testbeds and experiments, research directions 5G/6G

Prace nad 5 generacją (5G) systemów bezprzewodowych rozpoczęto w 2012 r. i w efekcie w roku 2015 Międzynarodowa Unia Telekomunikacyjna ITU (ang. *International Telecommunication Union*) opublikowała wizję kierunków rozwoju systemów radiokomunikacji mobilnej „*IMT for 2020 and beyond*" (ang. *International Mobile Telecommunication*) [1]. Wizja ta zawiera podstawowe scenariusze działania systemów 5G oraz zestaw 13. zdolności, takich jak np. szczytowa prędkość transmisji, efektywność widmowa, opóźnienie itp., które umożliwiają realizację tych scenariuszy. Opublikowane przez ITU-R (R oznacza sektor radiokomunikacyjny ITU) scenariusze stanowią syntezę szerokiego spektrum przypadków użycia sugerowanych przez producentów, organizacje regionalne i operatorów. Scenariusze te dotyczą trzech podstawowych sposobów użytkowania systemu 5G obejmujących:

- rozszerzony mobilny szerokopasmowy dostęp do Internetu eMBB (ang. *Enhanced Mobile Broadband*);
- ultraniezawodną komunikację o niskich opóźnieniach URLLC (ang. *Ultra Reliable Low Latency* Communications);
- masową komunikację pomiędzy maszynami mMTC (ang. *Massive Machine Type Communications*).

Zaawansowanie prac nad rozwojem systemów 5G spowodowało, że w ostatnich latach drugiej dekady XXI wieku podjęto analizy ich przydatności do zastosowań militarnych.

W 2018 r. amerykańskie centrum badań strategicznych (ang. *US Center for Strategic and International Studies*) opublikowało raport zatytułowany „*5G technology will reshape the innovation and security environment*" [2]. W raporcie tym podkreślono, że technologia 5G posiada wyjątkowe walory dla sfery militarnej i może być wykorzystana w robotach, systemach wykorzystujących sztuczną inteligencję, a także rojach zaawansowanych sensorów i dronów stając się podstawową technologią tworzącą nowe zdolności systemów militarnych.

W 2020 r. Agencja NATO prowadząca badania w obszarze systemów łączności i informatyki NCIA (ang. *NATO Communications and Information Agency*) opublikowała raport zatytułowany „*Potential of 5G technologies for NATO applications*" [3]. W raporcie tym wyspecyfikowano właściwości i parametry systemów 5G stanowiące o dużym potencjale tej technologii dla zastosowań militarnych.

W 2021 r. Europejska Agencja Obrony EDA (ang. *European Defence Agency*) opublikowała raport „*5G technologies for defence*" [4] o podobnych konkluzjach rekomendujących zastosowanie 5G dla sfery militarnej.

Przeprowadzona w NCIA analiza charakterystyk systemów 5G doprowadziła do wytypowania pięciu atrybutów czyniących te systemy perspektywicznymi dla zastosowań militarnych. Są to:

- szeroki zakres wykorzystywanych częstotliwości,
- nowy interfejs radiowy (5G NR),
- sieć rdzeniowa (5G CN),
- usługi bezpośrednie/bliskie ProSe (ang. *Proximity Services*),
- *sieci wykorzystujące wdrożenia nienaziemne NTN (ang. Non-Terrestrial Networks)*, czyli infrastrukturę znajdującą się poza powierzchnią Ziemi – głównie satelity, platformy latające HPAS (ang. *High Altitude Platform Stations*).

W związku z rozproszeniem badań dotyczących militarnych aplikacji technologii 5G prowadzonych przez kraje członkowskie NATO, agencja NCIA podjęła inicjatywę utworzenia wspólnego przedsięwzięcia o nazwie **MULTINATIONAL COLLABORATION ON 5G (MN5G) PROJECT [5].** Zakłada się bowiem, że wielonarodowa kooperacja zapewni uzyskanie synergii wysiłków pozwalając osiągnąć masę krytyczną w tym złożonym obszarze i zwiększy możliwości stymulowania rozwoju przemysłowego w znacznie większym stopniu niż miałoby to miejsce w przypadku autonomicznych działań poszczególnych krajów.

Analiza literatury wykazuje, że podejmowane w różnych krajach działania mające na celu przygotowanie technologii 5G do eksploatacji są inicjowane poprzez budowę instalacji testowych (testbedów) przeznaczonych do weryfikacji urządzeń, aplikacji i rozwiązań dedykowanych dla systemów 5G przewidywanych do wdrożenia w siłach zbrojnych danego kraju. Testbedy te służą również do prowadzenia eksperymentów mających na celu wprowadzenie nowych funkcjonalności „uzdatniających" rozwiązania komercyjne do zastosowań militarnych. Wynika to z faktu, że militarne sieci łączności charakteryzują się istotnymi różnicami scenariuszy operacyjnych w porównaniu z odpowiednimi dla sieci cywilnych. Różnice te wyrażają się np. przez brak rozległej infrastruktury stacjonarnej, dynamicznymi zmianami topologii sieci, ich nomadycznym charakterem i funkcjonowaniem w skomplikowanym środowisku elektromagnetycznym. Z tych względów rozwiązania komercyjne nie są w stanie sprostać rygorom eksploatacji w takim środowisku. Niezbędne jest dla takich zastosowań wprowadzenie modyfikacji lub istotnych zmian eliminujących te niedostatki, zwane w literaturze „lukami technologicznymi" (ang. *5G technology gaps*). Szczegółową charakterystykę tych luk przedstawiono w [4].

ATRYBUTY TECHNOLOGII 5G ISTOTNE DLA ZASTOSOWAŃ MILITARNYCH

Zakres wykorzystywanych częstotliwości

W 5G wykorzystuje się widmo częstotliwości radiowych w szerokim zakresie, w szczególności pasmo powyżej 7,125 GHz. Pasma częstotliwości dostępne dla IMT-2020 klasyfikuje się w oparciu o kryterium uwzględniające zasięgi łączności oraz pojemności komórek:

- Pasma FR1 od 1,5 do 7,125 GHz (FR1) – zapewniają wystarczająco duże przydziały ciągłego widma (do 100 MHz), co umożliwia realizację scenariusza eMBB. Oferują odporność na tłumienie deszczu i dobry kompromis między zasięgiem a przepustowością;
- Pasma FR2 powyżej 7,125 GHz (FR2) – umożliwiają duże przydziały ciągłego widma (nawet do 400 MHz), jednak propagacja w zakresach fal milimetrowych podlega tłumieniu przez deszcz, dużym stratom propagacji i wymaga bezpośredniej widzialności anten.

Wykorzystanie fal milimetrowych otwiera jednak możliwość tworzenia zaawansowanych, wysokowydajnych (w sensie pojemności i szybkości transmisji danych) i skrytych elektromagnetycznie systemów wojskowych. Ponadto w kontekście wojskowym wyzwania związane z propagacją fal milimetrowych (silne tłumienie) można uznać za czynnik umożliwiający stosunkowo łatwe ukrycie transmisji radiowej. Potencjał ten jest zwiększany przez technologie wieloantenowe MIMO wykorzystujące formowanie wiązki (*beamforming*), które może również przyczynić się do zmniejszenia zakłóceń zewnętrznych i zmniejszenia prawdopodobieństwa przechwycenia sygnałów radiowych.

Interfejs radiowy (5G NR)

5G NR (ang. *New Radio*) jest nowym interfejsem radiowym między stacją bazową a sprzętem użytkownika. Z punktu widzenia zastosowań militarnych najważniejszymi jego cechami są:

- **Obsługa wielu zakresów częstotliwości**. 5G NR obsługuje zakresy FR1 i FR2 i pozwala na użycie odpowiednio kanałów o szerokości 100 i 400 MHz. Różne podejścia do wykorzystania widma są możliwe dzięki technikom agregacji nośnych lub technologii multiRAT (ang. *Multi-Radio Access Technology*);
- **Warstwa fizyczna**. 5G NR wykorzystuje modulacje OFDM z protokołami dostępu OFDMA i adaptacyjną alokacją zasobów radiowych. 5G NR wykorzystuje zarówno dupleks z podziałem częstotliwości (FDD), jak i dupleks z podziałem czasu (TDD), przy czym w zakresie FR2 przewidziano jedynie tryb TDD. 5G NR dodatkowo rozszerza dotychczasowe zastosowania MIMO na *massive* MIMO oraz wprowadza aktywne anteny umożliwiające formowanie i sterowanie wiązką. Formowanie wiązki przyczynia się do znacznego zmniejszenia zakłóceń własnych oraz do poprawy budżetów łączy radiowych;
- ***Zintegrowany dostęp i dosył radiowy IAB** (ang. Integrated Access Backhaul)*. Rozwiązanie IAB polega na wykorzystaniu części zasobów radiowych stacji bazowej

gNB dla sieci dosyłowej (ang. *backhaul)*, umożliwiając komunikację pomiędzy stacjami gNB przez interfejs radiowy 5G NR. Ułatwia to wdrażanie gęstych sieci komórkowych bez konieczności podłączania wszystkich stacji gNB bezpośrednio do sieci rdzeniowej za pomocą łączy kablowych, co jest szczególnie ważne w sieciach polowych o dużej zmienności topologii.

Sieć rdzeniowa (5G CN)

Technika wirtualizacji funkcji sieci oraz możliwość tworzenia wydzielonych sieci (ang. *network slicing*), stanowiące immanentne cechy sieci 5G CN, oferują nowe możliwości, ale także związane z nimi wyzwania dla zastosowań militarnych. W szczególności pozwalają na adaptację struktury oraz wydajności sieci 5G do aktualnego scenariusza działań w ramach dostępnych w danej chwili zasobów. Ponadto, właściwości te pozwalają na wykorzystanie fragmentu publicznej infrastruktury do stworzenia na wybranym obszarze sieci prywatnej dla potrzeb wojskowych poprzez jej udostępnienie przez operatora publicznego. Mogą to być rozwiązania typu *system-in-a-box*, zapewniające realizację scenariuszy eMBB i URLLC w tym samym prywatnym taktycznym systemie 5G.

Komunikacja typu Sidelink oraz IAB

Usługi komunikacji bezpośredniej typu *5G Sidelink* i *D2D* (ang. *Device-to-Device)* oraz zintegrowany dostęp i dosył radiowy IAB są atrakcyjne z punktu widzenia hierarchicznych (wielopoziomowych) i/lub kratowych (*mesh*) struktur sieci taktycznych opartych na 5G. Usługi 5G Sidelink oraz D2D w przypadku łączy radiowych UE-UE oraz 5G IAB dla łączy radiowych gNB-gNB ułatwiają realizację szeregu scenariuszy taktycznych realizowanych „bez infrastruktury". W szczególności IAB oferuje możliwości szybkiego rozwinięcia autonomicznych klastrów 5G na polu bitwy (bez połączeń kablowych).

Sieci NTN 5G

Sieci nienaziemne NTN (ang. *Non-Terrestrial Networks*) 5G mogą stworzyć nowe możliwości aplikacyjne w dziedzinie taktycznej. Specyfikacja architektury 5G NTN jest obecnie przedmiotem prac w procesie standaryzacji. Może jednak stwarzać możliwości rozszerzenia łączności taktycznej i usług aplikacyjnych na obszar działań wspólnych za pomocą środków lub systemów 5G umieszczonych w środkach latających, np. typu UFO (ang. *Unmanned Flying Objects*) lub na satelitach. Powietrzne systemy 5G NTN mogą oświetlać obszar wspólnych operacji taktycznych, zapewniając łączność szerokopasmową i inne zastosowania (np. śledzenie sił przeciwnika, wywiad itp.) mobilnym oddziałom na Ziemi poprzez nawiązanie łączności z odległymi macierzystymi stanowiskami dowodzenia.

LUKI TECHNOLOGICZNE

Militarne sieci łączności charakteryzują się istotnymi różnicami scenariuszy operacyjnych w porównaniu z odpowiednimi dla sieci cywilnych. Różnice te wyrażają się np. przez brak rozległej infrastruktury stacjonarnej, dynamicznymi zmianami topologii sieci, ich nomadycznym charakterem i funkcjonowaniem w skomplikowanym środowisku elektromagnetycznym. Sytuacja komplikuje się ponadto w scenariuszach dotyczących eksploatacji w środowisku morskim, przestrzeni powietrznej i w obszarze zajmowanym przez przeciwnika, gdzie brak jest infrastruktury stacjonarnej, a w szczególności stacji bazowych. Z tych względów rozwiązania komercyjne nie są w stanie zapewnić niezawodnej, utajnionej i bezprzerwowej łączności w zastosowaniach militarnych. Niezbędne jest dla takich zastosowań wprowadzenie modyfikacji lub istotnych zmian eliminujących te niedostatki zwane w literaturze „lukami technologicznymi" (ang. *5G technology gaps*). W opublikowanym w 2021 r. przez Agencję EDA (ang. *European Defence Agency*) raporcie technicznym [4] przedstawiono katalog tych luk – niezbędnych do „uzdatnienia" rozwiązań komercyjnych dla zastosowań militarnych. Poniżej przedstawiono listę tych luk i scharakteryzowano niezbędne prace prowadzące do wyeliminowania tych niedostatków.

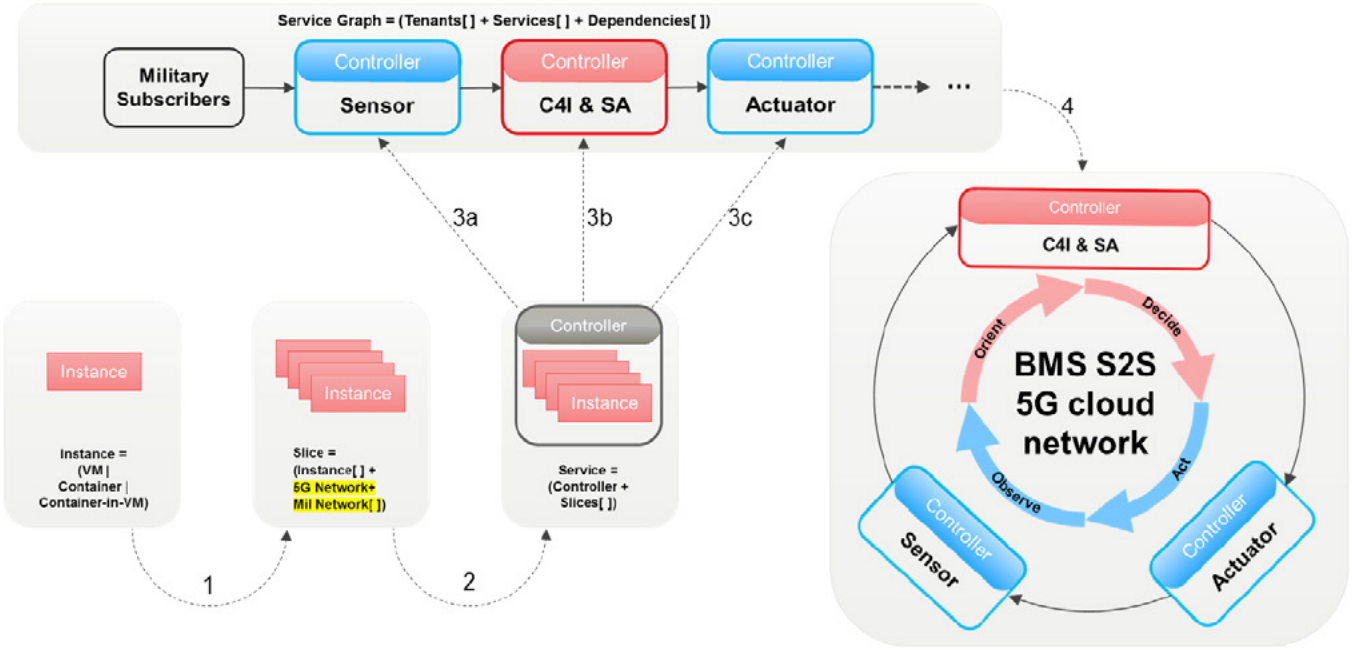
Wykorzystanie technik przetwarzania w chmurze

Technika 5G, a także przyszłe rozwiązania 6G, ewoluują w kierunku rozwiązań programowych, wdrażanych i uruchamianych wykorzystując techniki chmur obliczeniowych (ang. *cloud native 5G)*. Rozwiązania te bazują na architekturze mikrousług SOA (ang. *Service Oriented Architecture*), która cechuje się znacznie większą elastycznością we wdrażaniu i rekonfiguracji oferowanych usług w porównaniu do rozwiązań bazujących na dedykowanym sprzęcie. Wynika to z faktu, że usługi i aplikacje wojskowe wspierające między innymi funkcje dowodzenia i kierowania, zapewnienia łączności, cyberbezpieczeństwa, wywiadu, budowania świadomości sytuacyjnej, a także funkcje wykonawcze, są implementowane w postaci łańcuchów mikrousług, których instancje są wdrażane w ramach dostępnych w danym scenariuszu operacyjnym zasobów obliczeniowych. Zasoby te, które ze względu na dynamikę działań i nomadyczny charakter sieci, istotnie zmieniają się w czasie realizacji operacji.

Główną luką technologiczną i jednocześnie wyzwaniem jest opracowanie efektywnych metod orkiestracji przeznaczonych dla sieci wojskowych zapewniających efektywne działanie aplikacji i usług wojskowych w rozproszonym, dynamicznie zmiennym środowisku zintegrowanej infrastruktury sieciowo-obliczeniowej uwzględniając nomadyczność i mobilność zarówno użytkowników jak i elementów infrastruktury. W szczególności należy zwrócić uwagę, że zintegrowana infrastruktura sieciowo-obliczeniowa będzie tworzona dynamicznie wykorzystując zasoby obliczeniowe wielu systemów chmurowych, w tym systemów mobilnych, zasobów dostępnych na brzegu sieci, a także w miarę możliwości, zasobów infrastruktury stacjonarnej. Przykładowe wdrożenie systemu pola walki w architekturze wielochmurowej przedstawia *rys. 1.*

Bezpieczeństwo

Technika 5G zapewnia znacznie wyższy poziom bezpieczeństwa przekazu informacji w porównaniu do poprzednich generacji systemów mobilnych, ponieważ zagadnie-



» *Rys.1. Przykład systemu zarządzania pola walki (BMS) wykorzystującego cloud-native 5G z architekturą mikrousługową [4]*
» *Fig. 1. Example of Battle Management System (BMS) using cloud-native 5G with microservice architecture [4]*

nia bezpieczeństwa zostały uwzględnione już na etapie jej projektowania. Należy, jednakże zwrócić uwagę, że zastosowanie nowych technik radiowych i informatycznych jest potencjalnym źródłem nowych wektorów ataków i podatności. Bezpieczeństwo systemów 5G w zastosowaniach wojskowych można zwiększyć poprzez rozszerzenie i dostosowanie dostępnego sprzętu 5G do wymagań wojskowych. Innym podejściem może być ograniczenie użycia standardowego cywilnego sprzętu 5G wyłącznie do scenariuszy wojskowych wymagających niższego poziomu bezpieczeństwa. Ważnym elementem jest również odpowiednia konfiguracja sieci 5G, która pozwoli zwiększyć poziom bezpieczeństwa i odporności. W szczególności, rozproszenie instalacji, częste zmiany konfiguracji, zróżnicowanie wykorzystywanych pasm częstotliwości, heterogeniczne techniki komunikacji, a także techniki formowania wiązek radiowych stanowią przykłady cech techniki 5G, które można potencjalnie wykorzystać w celu zmniejszenia podatności na zagrożenia bezpieczeństwa.

Luka technologiczna: Nie ulega wątpliwości, że zastosowanie techniki 5G w systemach wojskowych wymaga przeprowadzenia analizy bezpieczeństwa systemów, ich podatności na nowe wektory ataków, analizy zapewnienia izolacji (czasowej i przestrzennej) oferowanych usług i wykorzystywanych zasobów, a także możliwości wykorzystania środków pasywnego i aktywnego zapewnienia bezpieczeństwa. Istotnym zagadnieniem jest również zapewnienie wielopoziomowego zabezpieczenia przesyłanych zastrzeżonych danych, we współpracy z obecnie wykorzystywanymi systemami łączności, a także w scenariuszach współpracy w misjach federacyjnych FMN (ang. *Federated Mission Networks*).

Scentralizowana architektura sieci 5G

Technika 5G zakłada architekturę sieci, w której rozproszone radiowe sieci dostępowe RAN są dołączone do logicznie scentralizowanego rdzenia sieci 5G (5GC). Rdzeń 5GC jest odpowiedzialny za realizację większości funkcji sieci 5G, w tym m.in. za obsługę sygnalizacji, uwierzytelnienie użytkownika, sterowanie połączeniami, obsługę mobilności użytkowników, a także przekaz danych i zapewnienie współpracy z innymi systemami oraz utrzymanie i zarządzanie całą infrastrukturą. Należy podkreślić, że rdzeń sieci 5G wraz z siecią szkieletową 5G zostały zaprojektowane jako systemy stacjonarne, w których zmiany zachodzą jedynie sporadyczne, np. w przypadku wystąpienia awarii lub w wyniku modernizacji sieci.

Luka technologiczna: Zastosowanie techniki 5G w systemach wojskowych wymaga jej dostosowania do warunków użycia sieci wojskowych, w szczególności systemów pola walki. Istotne cechy tych systemów to zdecentralizowany, rozproszony charakter działania systemu, mobilność użytkowników i elementów infrastruktury, konieczność adaptacji do ciągle zmieniającej się topologii oraz brak hierarchicznej struktury. Taktyczne sieci 5G będą wykorzystywać własne, lokalnie zarządzane stacje bazowe, które mogą działać autonomicznie lub współdziałać z innymi elementami sieci wykorzystując heterogeniczne środki łączności do ich dołączenia do systemu, np. łącza radioliniowe, kanały satelitarne, łącza mikrofalowe, światłowody czy też inne łącza nienaziemne. Oznacza to, że wojskowe 5G można postrzegać jako połączenie lub integrację potencjalnie niezależnych sieci 5G lub wdrażanych podsystemów. Z tego względu koniecznym jest opracowanie efektywnych metod wdrożenia funkcjonal-

ności rdzenia sieci 5G i zapewnienia współpracy dla sieci wojskowych 5G.

Niezawodność

Jak wspomniano powyżej, funkcjonalności sieci 5G zostały zaprojektowane i są głównie wdrożone w stacjonarnej infrastrukturze. Działanie sieci 5G wymaga zapewnienia niezawodnej komunikacji pomiędzy sieciami RAN a rdzeniem sieci 5G, ze względu na fakt, iż decyzje dotyczące sterowania i zarządzania siecią są podejmowane przez funkcje rdzenia sieci 5G. Obejmują one między innymi uwierzytelnienie użytkowników, sterowanie dostępem, czy też obsługę mobilności.

Luka technologiczna: Ponieważ sieci wojskowe charakteryzują się dynamicznymi zmianami topologiii wynikającej z tego płaskiej struktury systemu, jest niezbędnym zapewnienie wysokiej niezawodności działania sieci, w szczególności działania rdzenia sieci. Metody te muszą automatycznie podążać i reorganizować sieć na potrzeby danej operacji. W szczególności metody te powinny zapewnić odpowiedni poziom redundancji zasobów, połączeń, usług w celu wyeliminowania występowania pojedynczych punktów awarii (ang. single point of failure). Efekt ten uzyskuje się dzięki rozproszeniu i zapewnieniu redundancji zasobów oraz automatycznych metod skalowania, wdrażania i orkiestracji mikrousług wykorzystywanych przez wojskowe aplikacje i usługi. Ponadto, niezbędnym jest również zapewnienie rozproszenia i redundancji wykorzystywanych danych.

Integracja różnych technik sieciowych i zapewnienie współpracy

Technika 5G umożliwia łatwą integrację z innymi technikami sieciowymi i systemami. Współpraca ta jest realizowana przez implementację w rdzeniu sieci 5G funkcji pośredniczących IWF (ang. *Inter-Working Function*), które zapewniają współpracę zarówno w płaszczyźnie przekazu danych jak i w płaszczyźnie starowania. Ponadto, sieć 5G zapewnia łatwą integrację z różnymi technikami sieci dostępowych (w szczególności 5G NR, LTE, WiFi, Ethernet) oraz technologii sieci dosyłowych (*backhaul*), w tym łączy satelitarnych. Rdzeń 5G zapewnia również możliwość współpracy z zewnętrznymi systemami przez funkcje NEF (ang. *Network Exposure Function*).

Luka technologiczna: Wykorzystanie techniki 5G w systemach wojskowych wymaga zapewnienia współpracy rozwiązań 5G z istniejącymi systemami przez implementację dedykowanych funkcji IWF w rdzeniu sieci 5G. Możliwość elastycznego dostosowania sieci 5G do aktualnego scenariusza działania, automatycznego dopasowania jej funkcjonalności i wydajności wykorzystując mechanizmy skalowania i migracji mikrousług w ramach dostępnych zasobów są kluczowe dla wykorzystania techniki 5G w wojsku. Ponadto, istotna jest również możliwość dynamicznego tworzenia wydzielonych sieci (ang. *network slices*) dla dedykowanych grup użytkowników, obszarów zastosowań czy też działań w misjach koalicyjnych. Należy zwrócić uwagę, że wojskowa sieć 5G musi również umożliwić integrację starszych technik łączności, które niekoniecznie opierają się na protokole IP, zapewnić współpracę metod zabezpieczenia informacji, połączenie z wykorzystywanymi rozwiązaniami Internetu Rzeczy czy sieci czujników oraz współpracę z publicznymi sieciami operatorów.

Zarządzanie tożsamością

Technika 5G zakłada logicznie scentralizowane zarządzanie tożsamością. Informacje o użytkownikach są przechowywane w bazie danych przechowywanej w rdzeniu sieci. Są one wykorzystywane do uwierzytelnienia użytkowników podczas dołączenia do sieci, w tym dołączenia do sieci wydzielonych (ang. *network slices*) i zestawiania połączeń. Usługa zarządzania tożsamością nie umożliwia delegowania funkcjonalności do lokalnych podsystemów.

Luka technologiczna: Wojskowe wykorzystanie techniki 5G wymaga zaprojektowania zdecentralizowanego systemu kontroli dostępu i zarządzania tożsamością. Informacje o tożsamości użytkowników powinny być przechowywane lokalnie w ramach autonomicznych podsystemów i udostępniane w ramach współpracy w działaniu koalicyjnym. Ponadto, wykorzystywanie globalnie znanych tożsamości jest nieuzasadnione w przypadku operacji nomadycznych lub tajnych. Niezbędne jest zatem zaprojektowanie rozproszonego systemu zarządzania tożsamością dostosowanego do dynamiki działania systemów wojskowych, które w tym zakresie istotnie różnią się od systemów cywilnych bazujących na stacjonarnej infrastrukturze.

Massive MIMO

Massive MIMO (mMIMO) odnosi się do sytuacji, w której stosowane są duże szyki antenowe i techniki kształtowania wiązki promieniowania anten. Jest to kluczowa technologia zapewniająca wyższe przepływności w systemach 5G i zwiększająca odporność na zakłócenia celowe.

Luka technologiczna polega na tym, że systemy mMIMO wymagają dokładnej estymacji kanału propagacji realizowanej w czasie rzeczywistym co staje się wyzwaniem przy bardzo dużych szybkościach przemieszczania się terminali użytkownika (np. powyżej 500 km/h).

Techniki kształtowania wiązki promieniowania anten pozwalają kreować wąskie wiązki w kierunku własnego korespondenta i tłumienia promieniowania w innych kierunkach („zerowanie wiązki"). W ten sposób dzięki zerowaniu wiązki zmniejsza się interferencje dla innych użytkowników sieci i redukuje się możliwość niepożądanego przejęcia informacji przez przeciwnika. Dla realizacji takich właściwości jest niezbędne przesyłanie informacji zwrotnej przez swoich korespondentów o ich lokalizacji w celu określenia kierunków zerowania wiązki. Ta funkcjonalność wprowadza dodatkowe wyzwania dotyczące algorytmów kształtowania wiązki promieniowania anten.

Fale milimetrowe

Fale milimetrowe przewidywane dla systemów 5G mieszczą się w zakresie od 24 GHz do 300 GHz. Fale o takich częstotliwościach podlegają znacznie większemu tłumieniu i stąd maleją zasięgi łączności. Ta właściwość stwarza możliwość ponownego użycia kanałów częstotliwości w relatywnie mniejszej odległości (około 1 km), co prowadzi do zwiększenia pojemności sieci. Ponadto ze względu na duże

tłumienie fale milimetrowe są trudniejsze do niepożądanego przechwycenia i zakłócenia. Dlatego fale z tego zakresu częstotliwości są odpowiednie do realizacji połączeń w obszarze zajmowanym przez małe zgrupowania wojsk oraz w konwojach tworząc swego rodzaju funkcjonalne „bańki".

Luka technologiczna zawiera się w wyzwaniu polegającym na konieczności zapewnienia ciągłości połączeń pomiędzy ww. „bańkami", które systematycznie zmieniają swoją lokalizację. Ponadto pewne scenariusze militarne wymagają zapewnienia łączności w półsferze wokół stacji bazowych (gNB), co umożliwi realizację połączeń we wszystkich kierunkach – w tym w górę w kierunku UAVs. Takiej możliwości nie przewidziano bowiem w cywilnych systemach 5G.

Efekt Dopplera

Zjawisko Dopplera jest jednym z immanentnych efektów towarzyszących propagacji fal elektromagnetycznych podczas ruchu terminali. Zjawisko to staje się źródłem rozproszenia widma i w efekcie pogorszenia jakości odbieranego sygnału. Zjawisko to intensyfikuje się ze wzrostem częstotliwości transmitowanego sygnału oraz ze wzrostem prędkości przemieszczania się odbiornika względem nadajnika. Oddziaływanie tego zjawiska jest zauważalne w szczególności dla częstotliwości powyżej 300 MHz i w przypadku systemów 5G dla prędkości przemieszczania się powyżej 500 km/h.

Luka technologiczna wynika ze specyfiki militarnych scenariuszy, w których trzeba przewidywać instalowanie technologii 5G na helikopterach, samolotach czy dronach i w nieodległej przyszłości na obiektach hipersonicznych. Kompensacja efektu Dopplera w takich warunkach jest dla projektantów specyficznym wyzwaniem.

Integrated Access and backhaul (IAB)

Komunikacja IAB oznacza zintegrowany dostęp i dosył radiowy. Funkcjonalność tę wprowadzono w sieciach 5G dla podwyższenia elastyczności i adaptowalności systemu. Zasada jego działania polega na wykorzystaniu tej samej techniki dostępu radiowego dla łączenia terminali użytkownika ze stacjami bazowymi (gNB) i realizacji łączy dosyłowych, czyli połączeń stacji gNB z sieciami szkieletowymi bądź między stacjami gNB. W ten sposób eliminuje się z sieci kłopotliwe w eksploatacji i kosztowne połączenia kablowe.

Luka technologiczna wynika z konieczności rozwiązania problemu interferencji własnych wynikających z korzystania tego samego zakresu częstotliwości dla radiowego dostępu i dosyłu. Dodatkowe badania są niezbędne dla oszacowania efektywności systemu w dynamicznym scenariuszu, w którym topologia sieci ulega zmianom wynikającym z właściwej dla sieci militarnych, mobilności stacji bazowych, pojawianiu się przeszkód (przesłonięć) między węzłami i lokalnych przeciążeń stanowiących wynik zmian intensywności ruchu radiowego.

Połączenia bezpośrednie terminali (Sidelink)

Ta funkcjonalność umożliwia połączenie urządzeń końcowych D2D (ang. *Device-to-Device*) bez pośrednictwa infrastruktury sieci (stacji bazowych). W standardach 3GPP przewiduje się dwa modele realizacji funkcjonalności D2D, a mianowicie Model 1. z nadzorem stacji bazowej, gdzie stacja bazowa inicjuje taki tryb pracy lub Model 2. bez takiego nadzoru i wówczas współpraca dwóch urządzeń końcowych inicjowana jest autonomicznie przez te urządzenia. Funkcjonalność ta zapewnia odciążenie stacji bazowych oraz jest sposobem na zwiększenie zasięgu łączności w przypadku, gdy jeden z terminali znajduje się poza zasięgiem stacji bazowej oferując możliwość pośredniczenia terminali w przekazie danych.

Luka technologiczna polega na ograniczonej dostępności chipsetów zapewniających tę funkcjonalność i braku zainteresowania tą funkcjonalnością wykazywanego przez operatorów komercyjnych sieci komórkowych. Wysiłki mające na celu przekonanie dostawców i operatorów by wprowadzili tę funkcjonalność są trudnym wyzwaniem ze względu na mało interesującą dla nich skalę ekonomiczną takiego przedsięwzięcia.

CHARAKTERYSTYKA MILITARNYCH INSTALACJI TESTOWYCH (TESTBEDÓW)

Wprowadzanie nowych rozwiązań technologicznych w sferze militarnej zazwyczaj jest poprzedzone pracami koncepcyjnymi a następnie budową instalacji pilotażowych służących do weryfikacji opracowanych koncepcji i walidacji elementów przewidzianych do budowy docelowych sieci 5G. Potwierdzają to działania podejmowane w NATO oraz w poszczególnych państwach członkowskich Sojuszu, mających na celu wdrożenie technologii 5G do sił zbrojnych [6]. Agencja NCIA opublikowała dokument "Provision of Development of a NATO 5G Testbed for experimentation and proof-of-concept activities", w którym zdefiniowano następujące cele:

- opracowanie poligonu testowego NATO 5G w środowisku laboratoryjnym, w tym pozyskanie, integracja i konfiguracja odpowiednich komponentów sprzętowych i programowych. Poligon ten ma zwiększyć zdolności testowe NATO korzystającej do tej pory z poligonu na Łotwie;
- przedstawienie funkcjonalności poligonu testowego NATO 5G;
- planowanie i opracowanie harmonogramu wdrożenia kluczowych funkcji 5G w poligonie testowym NATO 5G w celu przeprowadzenia działań *proof-of-concept* (PoC).

Również Europejska Agencja Obrony (EDA) przygotowuje program, którego cele są zbliżone do powyższego, a mianowicie:

- stworzenie europejskiej platformy dla bezpiecznych urządzeń 5G i infrastruktury, łącznie z sieciami hybrydowymi (wykorzystującymi zarówno komercyjne jak i militarne technologie radiowe), zapewniającej narzędzia do konfiguracji i zarządzania oraz cyberbezpieczeństwo w zastosowaniach militarnych;
- demonstracja i adaptacja odpowiednich standardów przemysłowych do wymagań wojskowych;
- orkiestracja usług w środowisku federacyjnym.

Instalacje przeznaczone do testowania i weryfikacji urządzeń, aplikacji i rozwiązań dedykowanych dla zastosowań militarnych 5G umożliwiają lub mają umożliwić:

- testowanie funkcjonalności militarnych sieci 5G, w tym m.in.:
  - nowych aplikacji wynikających z ewolucyjnego rozwoju systemów 5G do 5G Advanced (Standard Release 18 – AI, ML);

- nowych, militarnych aplikacji 5G i metod zapewnienia jakości przekazu ruchu generowanego przez te aplikacje;
- rozwiązań kryptograficznych w środowisku sieci 5G;
- obsługi mobilności w militarnych systemach 5G, w tym mobilności urządzeń, użytkowników, usług;

- testowanie niezawodności działania systemu, w tym:
  - odporności systemu na częściowe zniszczenie jego elementów;
  - odporności na zakłócenia radioelektroniczne (w tym możliwości wykrycia aktywności sytemu);
  - odporności na awarie i błędy konfiguracji sprzętu i oprogramowania;
- testowanie metod orkiestracji oferowanych usług, obejmujące m.in. badania:
  - efektywności alokacji usług i ocenę efektywności wykorzystania dostępnych zasobów;
  - metod skalowania wydajności oferowanych usług w warunkach zmieniającego się obciążenia;
  - adaptacji i rekonfiguracji oferowanych usług do zmieniającej się topologii systemu i dostępności zasobów sieciowych i obliczeniowych;
  - automatycznej aktualizacji oprogramowania oferowanych usług (*live upgrade/downgrade*);
- testowanie procedur automatyzacji zarządzania i utrzymania systemu, obejmujące również testy czasu „rozwinięcia" (uruchomienia) i „zwinięcia" (przeniesienia) systemu 5G;
- testowanie podatności systemu, efektywności metod rozpoznawania i przeciwdziałania cyberatakom oraz bezpieczeństwa systemu;
- testowanie interoperacyjności sieci 5G z aktualnie eksploatowanymi sieciami polowymi;
- testowanie współpracy z komercyjnymi sieciami 5G oraz współpracy z innymi technikami łączności;
- testowanie funkcjonalności ragidyzowanych elementów sieci 5G, w tym dostosowania do częstotliwości wykorzystywanych w domenie militarnej.

Bardzo intensywnie jest rozbudowywana infrastruktura służąca do testowania technologii 5G przeznaczonych dla sfery militarnej w USA. Departament Obrony utworzył Zespół Funkcjonalny 5G oraz FutureG (ang. *5G and Future Generation Wireless Cross-Functional Team*), który ma za zadanie przyspieszyć wdrażanie przełomowych technologii bezprzewodowych 5G i przyszłych generacji. Wyasygnowano w tym celu 600 milionów dolarów na uruchomienie instalacji testowych na pięciu amerykańskich poligonach wojskowych [7]. W każdej z tych lokalizacji zaplanowano uruchomienie sieci testowej służącej do rozwijania innych zdolności operacyjnych sił zbrojnych USA. Jednym z najbardziej rozbudowanych wśród nich jest testbed Sił Powietrznych USA w którym prowadzone są eksperymenty wykorzystujące sieci 5G rozwinięte na obszarze około 7000 km kwadratowych w Nevadzie (baza lotnicza *Nellis Air Force Base*), gdzie znajduje się jedno z pięciu centrów dowodzenia Siłami Powietrznymi USA.

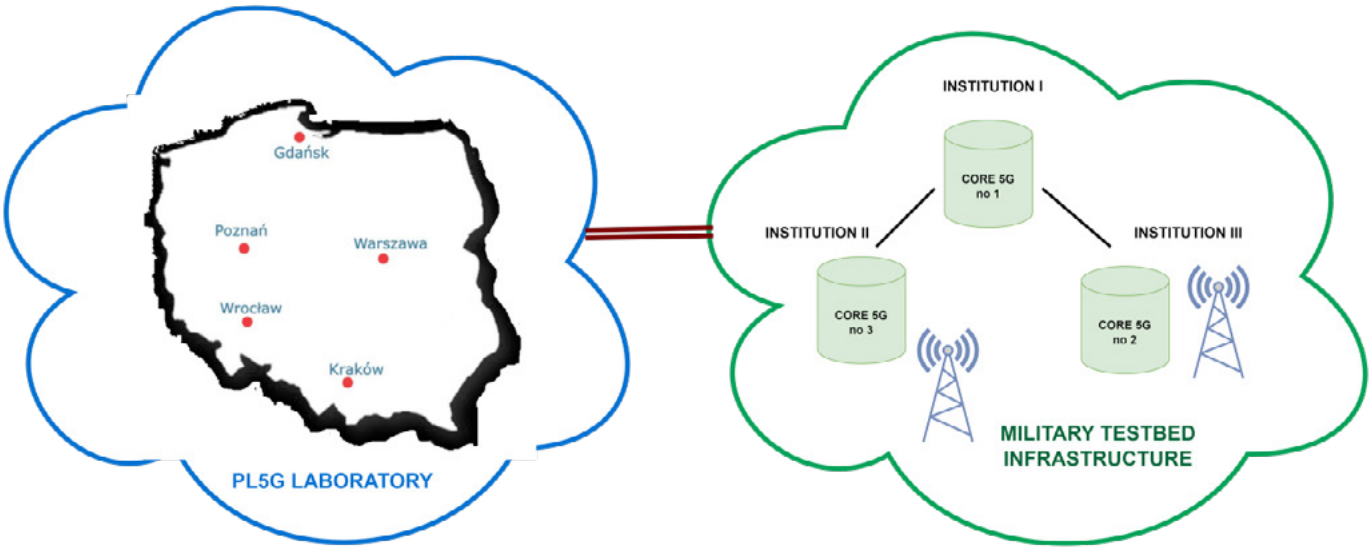
Podobne przedsięwzięcia dotyczące testów technologii 5G przeznaczonej dla wykorzystania w domenie militarnej zainicjowano w laboratoriach krajowych, m.in. Hiszpanii, Norwegii i Finlandii.

W naszym kraju w bieżącym 2025 roku Ministerstwo ON zainicjowało prace nad budową militarnego testbedu 5G. Projekt badawczy „Opracowanie koncepcji architektury militarnego testbedu 5G" jest obecnie realizowany w Wojskowym Instytucie Łączności - Państwowym Instytucie Badawczym we współpracy z Politechniką Warszawską i Wojskową Akademią Techniczną.

Zasadniczym celem budowy testbedu jest utworzenie infrastruktury badawczo-eksperymentalnej, która umożliwi realizację badań dotyczących rozwoju techniki 5G w obszarze zastosowań militarnych, zapewnienia współpracy rozwiązań 5G z obecnie wykorzystywanym systemami wojskowymi i cywilnymi, jak również umożliwi przeprowadzenie testów proponowanych dla wojska rozwiązań 5G w zakresie ich zgodności z wymaganiami wojskowymi dot. m.in. funkcjonalności, wydajności, bezpieczeństwa oraz możliwości współpracy. W szczególności testbed ten umożliwi: 1) praktyczną weryfikację rozważanych koncepcji wdrożenia techniki 5G w wojsku polskim, uwzględniając testowanie różnych scenariuszy operacyjnych, 2) przeprowadzenie eksperymentów umożliwiających zweryfikowanie korzyści oraz zidentyfikowanie potencjalnych zagrożeń i podatności techniki 5G w zastosowaniach wojskowych, 3) rozwój rozwiązań 5G o dodatkowe funkcjonalności niezbędne w systemach wojskowych, 4) budowanie kompetencji, doświadczenia i praktycznych umiejętności kadry odpowiedzialnej za wdrożenie techniki 5G w wojsku, 5) udział zespołów badawczych w ćwiczeniach ze sprzętem 5G, w tym ćwiczeniach prowadzonych w ramach działań koalicyjnych, 6) udział zespołów w krajowych i międzynarodowych projektach badawczo-rozwojowych dot. rozwoju i standaryzacji techniki 5G i przyszłej 6G, a także 7) szeroko rozumianą edukację zorientowaną na zdobywanie praktycznych umiejętności i doświadczenia kadry przyszłych specjalistów w obszarze wojskowych sieci 5G przez uczelnie techniczne.

Projektowany testbed obejmie 4 węzły laboratoryjne połączone w jedną, spójną, rozległą infrastrukturę badawczo-eksperymentalną, którą przedstawiono na *rys. 2*. Infrastruktura ta umożliwi przeprowadzenie eksperymentów w środowisku zbliżonym do rzeczywistych warunków pracy wojskowej sieci 5G, dotyczących zróżnicowanych scenariuszy operacyjnych, m.in. praca w warunkach polowych w trybie autonomicznym i we współpracy z wojskową infrastrukturą stacjonarną, współpraca ad hoc wielu jednostek operacyjnych, praca w trybie stacjonarnym, współpraca z sieciami cywilnymi, w tym z wydzieloną siecią wojskową utworzoną w sieciach cywilnych, tzw. military slice, współpraca w działaniach koalicyjnych.

Węzły zostaną wyposażone w urządzenia i oprogramowanie sieci 5G, obejmujące stacje bazowe, głowice radiowe, terminale i modemy, rdzeń sieci 5G, urządzenia transmisyjne i utajniające oferujące pełną funkcjonalność sieci 5G. Zakłada się, że testbed zostanie zbudowany w większości wykorzystując programowalne urządzenia 5G oraz oprogramowanie oferujące dostęp do kodu źródłowego co umożliwi rozwój oferowanej funkcjonalności sieci 5G, dostosowanie jej do wymagań sieci wojskowych, a także szczegółową analizę implementacji. Zbiór tych urządzeń zostanie uzupełniony o komercyjne wyposażenie, np. głowice radiowe, zapewniające wydajność na poziomie rozwiązań komercyjnych.



» *Rys. 2 Topologia projektowanego testbedu*
» *Fig. 2. Topology of designed testbed*

PODSUMOWANIE

Technologia 5G wprowadza wiele nowych właściwości do radiokomunikacji mobilnej bowiem wykorzystuje nowe, znacznie szersze pasma radiowe, oferuje większe prędkości transmisji i mniejsze opóźnienia przekazu danych oraz zapewnia masową obsługę łączących się urządzeń. Z tego względu technologia ta stała się obszarem intensywnych analiz dotyczących jej przydatności dla zastosowań militarnych. W analizach tych rozważa się różnorodne sposoby wykorzystania technologii 5G w aplikacjach militarnych – od zastosowania rozwiązań komercyjnych – po przystosowanie do eksploatacji w domenie militarnej rozwiązań „z półki" (ang. *off the shelf*) czy docelowo – zaprojektowanie i wykonanie militarnych rozwiązań elementów systemu 5G.

W artykule przedstawiono zasadnicze rezultaty tych prac przedstawiające walory technologii 5G dla takich zastosowań. Jednocześnie w analizach tych podkreśla się, że specyfika wymagań stawianych systemom radiokomunikacji przewidzianym do zastosowań militarnych (wrażliwość przesyłanych informacji, duża dynamika topologii sieci, narażenia mechaniczne, klimatyczne, elektromagnetyczne itp.) powodują, że sieć i jej elementy muszą spełniać bardzo specyficzne i rygorystyczne wymagania. Szczególnie charakterystyczne jest to w przypadku systemów pola walki, gdzie elementy systemów 5G, opracowane dla sieci komercyjnych, nie są w stanie spełnić wspomnianych wymagań. Reasumując stwierdza się, że technologia 5G nie osiągnęła jeszcze dojrzałości do zastosowań militarnych. Dlatego sformułowano listę niezbędnych prac „uzdatniających" systemy komercyjne do takich zastosowań poprzez eliminację, przedstawionych w prezentacji, tzw. luk technologicznych. Dla przeprowadzenia takich prac jest niezbędna infrastruktura w postaci tesbedów. W prezentacji przedstawiono przykłady testbedów budowanych w różnych krajach i budowanych przez organizacje międzynarodowe. Kierując się takimi doświadczeniami, również w naszym kraju podjęto prace nad opracowaniem militarnego testbedu 5G i w prezentacji omówiono podstawowe założenia przyjęte dla tego projektu.

LITERATURA

[1] 1. IMT Vision – Framework and overall objectives of future development of IMT for 2020 and beyond, Recommendation ITU – R M.2083.
[2] Jingiing Liao, Xinjijan Ou, 5G Military Application Scenarios and Private Networks Architekture, 2020 IEEEInternational Conference on Advances in Electrical Engineering and Computer Applications
[3] Potential of 5G technologies for NATO applications, NCIA Technical Raport, NCIA/TR/2020/NCBO14792/01
[4] 5G technologies for defence, EDA White paper, Version 1.0, January 2021.
[5] MN5G PROJECT , Exploiting the potential of 5G for military application, NITECH: NATO Innovation and Technology – Issue 11, published on Jul 15, 2024.
[6] Technologia 5G w zastosowaniach militarnych, Raport WIŁ-PIB DT/391/2024/5G
[7] https://www.defense.gov/News/Releases/Release/Article/2376743/dod-announces-600-million-for-5g-experimentation-and-testing-at-five-installati/

DOI: 10.15199/59.2025.4.2

Przegląd metod wielodostępu nieortogonalnego i możliwości ich zastosowania w systemach 5G i 6G

Overview of non-orthogonal multiple access methods and their possible applications in 5G and 6G systems

Prof. dr hab. inż. Krzysztof Wesołowski, Instytut Radiokomunikacji, Politechnika Poznańska
krzysztof.wesolowski@put.poznan.pl

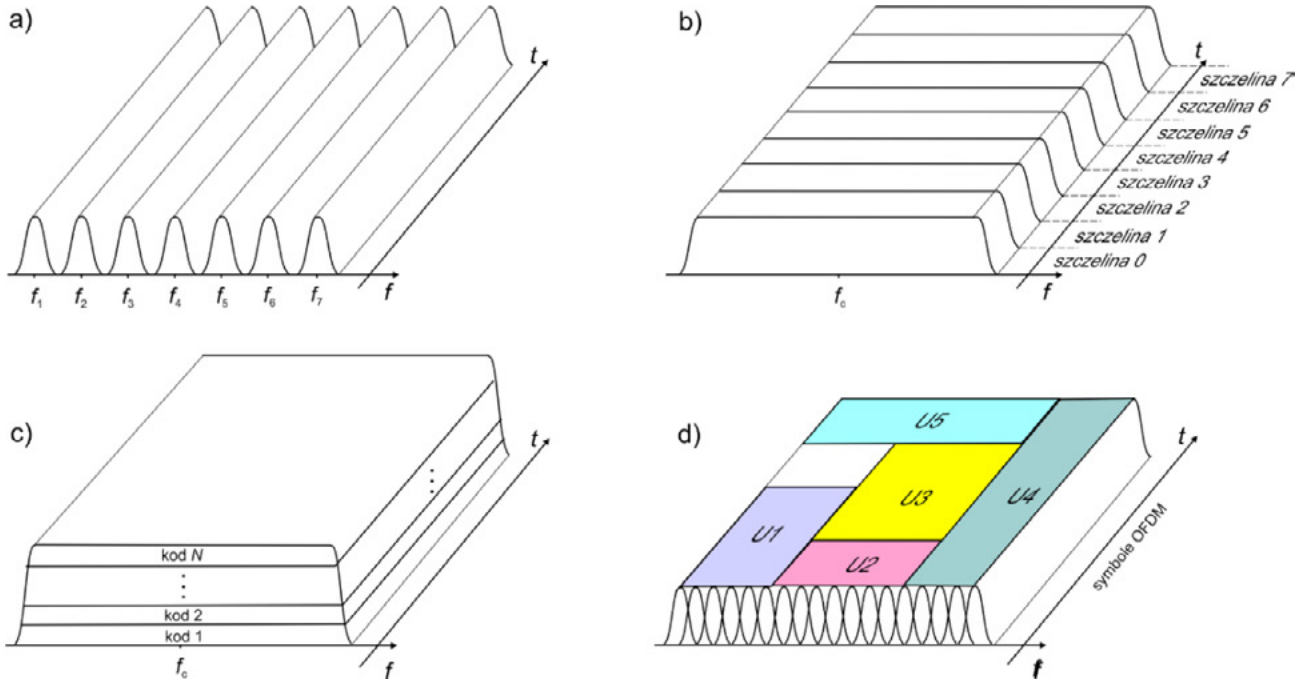
STRESZCZENIE: W artykule dokonano przeglądu metod wielodostępu nieortogonalnego, w szczególności wielodostępu NOMA realizowanego w dziedzinie mocy. Po ogólnym podsumowaniu metod wielodostępu ortogonalnego OMA stosowanych w kolejnych generacjach komunikacji bezprzewodowej, wymieniono podstawowe metody wielodostępu NOMA w dziedzinie kodowej a następnie wyjaśniono zasadę działania systemu NOMA w dziedzinie mocy. Dokonano podstawowej analizy i porównania granicznej przepływności z wielodostępem OMA. Z kolei przedstawiono zaproponowane uprzednio przez autorów ulepszenie procesu kompensacji interferencji w łączu w górę systemu z wielodostępem NOMA w dziedzinie mocy. Następnie przeanalizowano podstawowe własności i zasady działania wielodostępu NOMA w przypadku transmisji wieloantenowej MISO i na podstawie literatury pokazano rozwiązania konkurencyjne, takie jak system wielu użytkowników z liniowym prekodowaniem (MU-LP) oraz system RSMA z rozdzielaniem transmitowanych wiadomości i częściową kompensacją interferencji. W podsumowaniu stwierdzono możliwość zastosowania wielodostępu NOMA w masowej transmisji między maszynami (IoT).

SŁOWA KLUCZOWE: wielodostęp ortogonalny OMA, wielodostęp nieortogonalny NOMA, interferencja, kompensator interferencji, kodowanie wstępne, systemy wieloantenowe MIMO

The paper reviews non-orthogonal multiple access methods, in particular NOMA multiple access implemented in the power domain. After a general summary of orthogonal multiple access methods (OMA) used in subsequent generations of wireless communication, the basic NOMA multiple access methods in the code domain are listed, and then the operating principle of the NOMA system in the power domain is explained. Basic analysis and comparison of achievable rates with OMA multiple access are made. In turn, an improvement of the uplink interference compensation process of the system with NOMA in the power domain, previously proposed by the authors, is presented. Then, the basic properties and operating principles of the NOMA multiple access in the case of multi-antenna MISO transmission are analyzed and, based on the literature, competitive solutions are shown, such as the multi-user system with linear precoding (MU-LP) and the RSMA system with rate splitting and partial interference compensation. The paper ends with the conclusion about the possibility of using NOMA multiple access in massive machine-to-machine (IoT) communication.

KEYWORDS: orthogonal multiple access OMA, non-orthogonal multiple access NOMA, interference, interference cancellation, precoding, MIMO multiple antenna systems

Jednym z najważniejszych czynników mających wpływ na pojemność systemu radiokomunikacji ruchomej jest metoda dostępu do przydzielonego mu pasma częstotliwości i jego szerokość. Pasmo elektromagnetyczne jest zasobem szczególnie cennym i ograniczonym, które powinno być używane przez abonentów równocześnie korzystających z systemu radiokomunikacyjnego w sposób szczególnie efektywny. W tym celu wraz z rozwojem kolejnych generacji systemów radiokomunikacji ruchomej stosowano coraz bardziej efektywne metody wielodostępu do pasma, a w konsekwencji, zasobów systemu. Zasadnicza większość metod wielodostępu bazuje na zasadzie ortogonalności: w dziedzinie częstotliwości (wielodostęp z podziałem częstotliwości FDMA (*Frequency Division Multiple Access*)), w dziedzinie czasu (wielodostęp z podziałem czasu na szczeliny czasowe TDMA (*Time Division Multiple Access*)), czy w dziedzinie kodów (wielodostęp z zastosowaniem modulacji określonej liczby funkcji wzajemnie ortogonalnych – najczęściej w postaci sekwencji rozpraszających CDMA (*Code Division Multiple Access*)). Quasi-ortogonalność występuje również w wielodostępie przestrzennym (*Space Division*



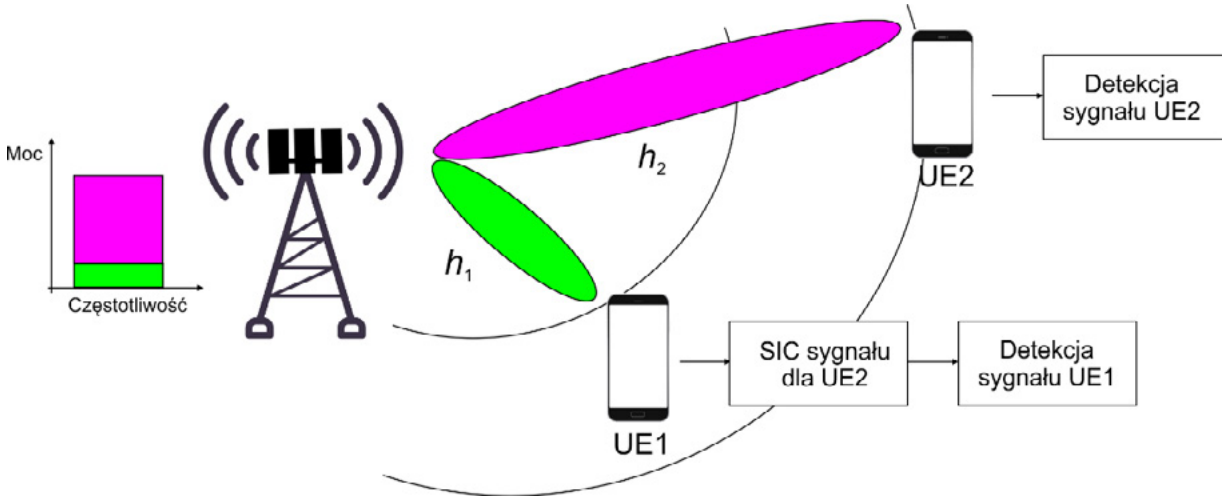
» *Rys. 1. Metody wielodostępu ortogonalnego: a) FDMA, b) TDMA, c) CDMA, d) OFDMA*
» *Fig. 1. Orthogonal methods of multiple access: a) FDMA, b) TDMA, c) CDMA, d) OFDMA*

Multiple Access – SDMA). Często stosuje się rozwiązania hybrydowe, np. TDMA/FDMA, jak np. w systemie GSM i jego pochodnych. Jedna z wersji standardu systemów UMTS przewidywała zastosowanie wielodostępu CDMA/TDMA. Kolejną, dominującą obecnie w systemach komórkowych metodą wielodostępu jest metoda OFDMA (*Orthogonal Frequency Division Multiple Access*) oparta na podziale wzajemnie ortogonalnych w zdefiniowanych interwałach czasowych podnośnych transmisji wielotonowej, w której podstawą jest zastosowanie transmisji OFDM (*Orthogonal Frequency Division Multiplexing*). W metodzie tej dane transmitowane są z wykorzystaniem modulacji podnośnych gęsto rozmieszczonych na osi częstotliwości. Warto dodać, że metody wielodostępu FDMA, TDMA, TDMA/FDMA czy OFDMA cechują się zerową interferencją wewnątrz komórki, w której są zastosowane. W przypadku wielodostępu CDMA interferencje wewnątrzkomórkowe niestety pojawiają się w odbiornikach, ponieważ zastosowane sekwencje rozpraszające nie są wzajemnie ortogonalne względem własnych i innych opóźnionych wersji sekwencji a nadawany sygnał z rozpraszaniem i wielodostępem CDMA jest odbierany po przejściu przez wzajemnie opóźnione ścieżki propagacji kanału wielodrogowego. Na *rys. 1* przedstawiono ilustrację wyżej wymienionych metod wielodostępu. Należą one do kategorii wielodostępu ortogonalnego OMA (*Orthogonal Multiple Access*), w których konkretne zasoby radiowe są w dyspozycji pojedynczego użytkownika.

Chcąc zwiększyć pojemność systemu radiokomunikacyjnego rozważono rezygnację z zasady ortogonalności i umożliwiono korzystanie z określonego zasobu radiowego przez więcej niż jednego użytkownika. Idea ta wynika z nowych rodzajów usług w systemach piątej generacji (5G) i systemach mających po nich nastąpić. Jedną z nich jest konieczność obsługi w ramach danej komórki bardzo dużej liczby terminali o zróżnicowanych szybkościach transmisji i różnych wymaganiach dotyczących dopuszczalnych opóźnień. Liczba realizowanych połączeń może przekraczać liczbę jednostek zasobów pozostających w dyspozycji systemu. W przypadku masowej komunikacji między urządzeniami technicznymi (*massive Machine Type Communications*) przewidywanej w systemach piątej generacji (5G), wiele połączeń odbywa się z terminali przekazujących dane sporadycznie i charakteryzujących się niskimi szybkościami danych. Przydzielanie takim połączeniom całych bloków zasobów fizycznych (PRB – *Physical Resource Block*) w wielodostępie OFDMA jest nieuzasadnione ze względu na użytkowanie dostępnych zasobów. W rzeczywistości bloki PRB mogą być potencjalnie wykorzystywane przez więcej niż jednego użytkownika.

Podstawowe zasady działania wielodostępu NOMA

Powyższe rozważania uzasadniają rozpatrzenie wielodostępu nieortogonalnego NOMA (*Non-Orthogonal Multiple Access*) [1]. Może on być zrealizowany w różnych dziedzinach, w szczególności w dziedzinie kodów oraz w dziedzinie mocy [2]. Przykładami wielodostępu NOMA w dziedzinie kodów są wielodostęp z kodowaniem kratowym TCMA (*Trellis-Coded Multiple Access*) [3], wielodostęp IDMA (*Interleave Division Multiple Access*) [4], wielodostęp LDS-CDMA (*Low-Density Signature CDMA*) [5], wielodostęp MUSA (*Multi-User Shared Access*) [6], czy też PDMA (*Pattern-Division Multiple Access*) [7].



» *Rys. 2. Ilustracja zasady działania wielodostępu NOMA w dziedzinie mocy dla dwóch terminali w łączu w dół*
» *Fig. 2. Illustration of the NOMA power domain multiple access principle in the downlink for two terminals*

W pracy [8] z 2017 roku zrelacjonowano propozycję zastosowania wielodostępu NOMA w dziedzinie mocy w systemie LTE-Advanced w wersji MUST (*Multi-User Superposition Transmission*) w łączu w dół. Zasadniczą ideą wielodostępu NOMA w dziedzinie mocy jest możliwość obsługi wielu użytkowników korzystających z tego samego zasobu za pomocą technik kodowania z superpozycją (*Superposition Coding*) w nadajniku i kompensacji w odbiorniku interferencji pochodzących od sygnałów wysyłanych do kolejnych użytkowników SIC (*Successive Interference Cancellation*). Kluczowym aspektem w realizacji tej metody jest odpowiednie przypisanie mocy w stacji bazowej poszczególnym sygnałom przeznaczonym dla użytkowników/terminali korzystających z tego samego zasobu, co wiąże się z własnościami propagacyjnymi ich kanałów transmisyjnych pomiędzy stacją bazową a terminalem. Sygnał łączny wysyłany przez stację bazową jest superpozycją sygnałów przeznaczonych dla poszczególnych terminali. Na *rys. 2* przedstawiono ideę wielodostępu NOMA w dziedzinie mocy dla dwóch terminali (UE1 i UE2) korzystających z tych samych zasobów. Można ją łatwo rozszerzyć na sytuację, w której z tych samych zasobów w łączu w dół korzysta *K* terminali.

Na *rys. 2* kanał ze stacji bazowej do terminala UE1 charakteryzuje się mniejszym tłumieniem niż w przypadku kanału ze stacji bazowej do terminala UE2. Oznaczając współczynniki kanałów odpowiednio jako i możemy więc przyjąć, że . Superpozycja obu sygnałów wysłanych przez stację bazową przez kanał o współczynniku jest traktowana przez terminal UE2 jako sygnał przeznaczony dla niego z dodatkowym szumem stworzonym przez sygnał wysyłany do terminala UE1 i odbierany również przez UE2. Z kolei terminal UE1 otrzymujący superpozycję obu sygnałów przez kanał najpierw dokonuje detekcji sygnału przeznaczonego dla UE1, następnie znając swój kanał syntetyzuje składową tego sygnału w odebranym sygnale łącznym i odejmuje ją od niego. Po kompensacji tego sygnału dokonuje ostatecznie detekcji sygnału przeznaczonego dla niego. Jak wnioskujemy, konieczna jest znajomość współczynników kanału, aby możliwa była dokładna kompensacja składowej przeznaczonej dla innego użytkownika, co z kolei umożliwi detekcję sygnału pożądanego. Warto również zauważyć, że poziomy mocy obu składowych sygnału łącznego powinny się od siebie znacząco różnić, aby można było w odbiornikach obu terminali traktować jedną ze składowych podlegających w nadajniku superpozycji jako szum i aby usuwanie interferencji było skuteczne. Tak więc powstaje problem efektywnego doboru par lub zbiorów o większej liczebności terminali, jeśli *K* terminali korzysta z tych samych zasobów i uporządkowanie ich względem zastosowanej mocy.

W literaturze można znaleźć analizę przepustowości takiego systemu wielodostępu w porównaniu z systemem OMA [9]. Najczęściej zakłada się w niej, że stacja bazowa emituje kanał rozsiewczy, rozumiany jako taki, którego sygnał dociera do wszystkich terminali znajdujących się w zasięgu tej stacji. Chcąc porównać przepustowość systemu z wielodostępem OMA i NOMA oblicza się przepustowość doświadczaną przez oba terminale. Jak wiadomo z teorii informacji, przepustowość uzyskuje się zakładając gaussowski charakter sygnałów nadawanych. Uzyskane rezultaty należy traktować więc jako górną granicę przepływności kanałów transmitujących sygnały o innych własnościach statystycznych. Dokonajmy takiej analizy dla dwóch terminali działających w trybie OMA lub NOMA. W prostym wywodzie podążamy za analizą przedstawioną w pracy [9].

Załóżmy, że w przypadku wielodostępu OMA zastosowano go w formie TDMA. Zatem ułamek τ czasu jednostkowego $(0 \le t \le 1)$ jest zapewniony dla transmisji do terminala UE1, zaś pozostała część czasu $\bar{\tau} = 1 - \tau$ jest wykorzystywana w transmisji do terminala UE2. Niech moc nadawana przez stację bazową ma wartość *P*, zaś moc szumu będzie znormalizowana do jedności. Stosunek mocy sygnału do mocy szumu dla terminali UE1 i UE2 będzie wtedy określony zależnością:

$$\gamma_i = |h_i|^2 P \quad (i = 1,2) \tag{1}$$

Zatem znormalizowana przepustowość obu kanałów (przepustowość przypadająca na 1 Hz) wynosi odpowiednio $R_1 = \tau C(\gamma_1)$ oraz $R_2 = \bar{\tau} C(\gamma_2)$, gdzie

$$C(x) = \frac{1}{2}\log_2(1+x) \tag{2}$$

W przypadku wielodostępu OMA w trybie FDMA sytuacja jest analogiczna. Podziałowi podlega bowiem dostęp do szerokości całego pasma pozostającego w dyspozycji systemu.

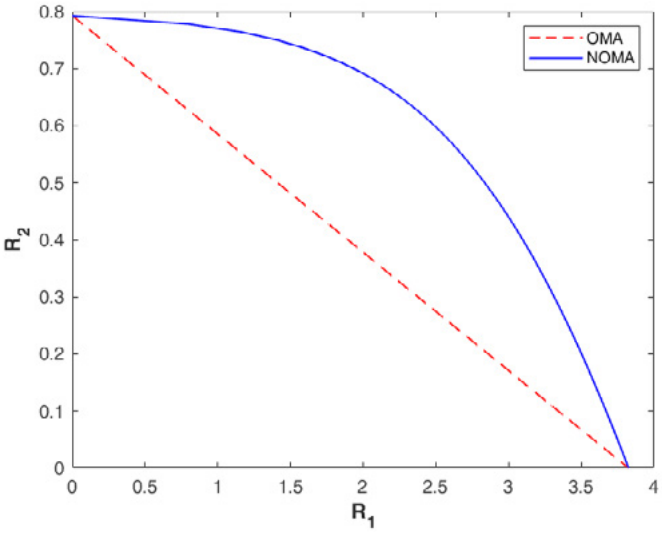
Przeanalizujmy obecnie znormalizowaną przepustowość dla wielodostępu NOMA. Z powodu zastosowanej superpozycji sygnałów moc stacji bazowej jest dzielona pomiędzy sygnały przeznaczone dla obu terminali. Niech sygnał dla terminala UE1 ma przydzieloną moc αP, $(0 \le \alpha \le 1)$ zaś sygnał przeznaczony dla terminala UE2 ma moc Zakładając idealne działanie kompensatora interferencji w terminalu UE1, przepustowość, którą on osiąga jest określona wzorem

$$R_1 = C(\alpha\gamma_1) \tag{3}$$

Terminal UE2 doświadcza zaś dodatkowego szumu w postaci odbieranego sygnału przeznaczonego dla terminala UE1, zatem przepustowość w jego przypadku wynosi

$$R_2 = C\left(\frac{\bar{\alpha}\gamma_2}{\alpha\gamma_2 + 1}\right) \tag{4}$$

Na *rys. 3* przedstawiono wykres obszarów przepustowości dla wielodostępu OMA i NOMA w zależności od podziału czasu (dla wielodostępu OMA) i odpowiednio mocy (dla wielodostępu NOMA). Jak widać, NOMA charakteryzuje się obszarem zapewniającym (co najmniej asymptotycznie) większą przepustowość łączną dla obu terminali. Podobnie jak w pracy [9], rysunek wykonano dla wartości współczynników kanału spełniających zależność $|h_1| = 10|h_2| = \sqrt{5}$ oraz założono wielkość mocy nadawanej P = 40. Zasadniczy wniosek płynący z wykresów na *rys. 3* jest taki, że sumaryczna asymptotyczna przepustowość obsługiwanych terminali jest w przypadku NOMA większa niż w przypadku wielodostępu OMA, zaś staje się równa wtedy, gdy rozdział mocy jest taki, że moc stacji bazowej jest przydzielona wyłącznie transmisji do jednego z terminali.



» *Rys. 3. Wykres osiągalnych obszarów przepustowości dla wielodostępu OMA i NOMA (parametry kanałów i mocy w tekście)*
» *Fig. 3. Plot of achievable rate areas for OMA and NOMA multiple access (channel and power parameters in text)*

Problem kompensacji interferencji w wielodostępie NOMA

Problemem mającym znaczenie praktyczne jest realizacja kompensatora interferencji pochodzącej od silniejszego sygnału przeznaczonego dla innego użytkownika. Ilustruje to sytuacja terminala UE1 na *rys. 2.* Standardowym rozwiązaniem jest detekcja sygnału silniejszego przy traktowaniu sygnału słabszego jako szumu. Jeśli moc sygnału słabszego jest zbyt duża, ma ona negatywny wpływ na prawdopodobieństwo błędu detekcji sygnału silniejszego, a w konsekwencji na skuteczność jego kompensacji w sygnale łącznym. To z kolei wpływa na prawdopodobieństwo błędu detekcji sygnału słabszego otrzymanego po niedoskonałej kompensacji sygnału silniejszego. Bardzo istotny jest odpowiedni dobór terminali korzystających z tych samych zasobów zapewniających znaczącą różnicę w poziomach pomiędzy sygnałem silniejszym i słabszym. Aby zapewnić lepsze działanie kompensatora interferencji, autor niniejszej pracy wraz z doktorantką zaproponowali ulepszone działanie kompensatora interferencji [10]. Rozpatrzono transmisję OFDM w łączu w górę. Zasoby są współdzielone przez dwa terminale nadające sygnały do punktu dostępowego (*rys. 4*) przez kanały scharakteryzowane odpowiednio za pomocą wektorów współczynników $\boldsymbol{h}_1$ oraz $\boldsymbol{h}_2$ opisujących kanały wielodrogowe w dziedzinie czasu. Zakłada się, że oba terminale są na tyle wzajemnie zsynchronizowane, że w odbiorniku możliwe jest znalezienie wspólnego przedziału ortogonalności dla obu sygnałów OFDM docierających do niego. Formalny model systemu transmisyjnego przedstawiono na *rys. 5*.

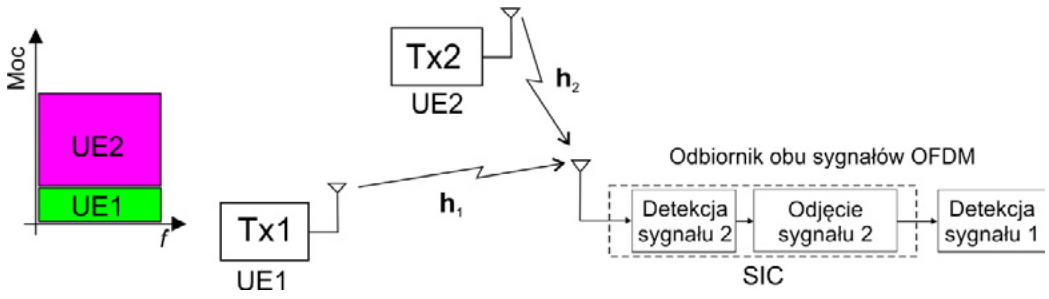
Zauważmy, że skoro odbiornik OFDM sygnału łącznego jest w stanie wyznaczyć dla obu jego składników wspólny przedział ortogonalności symboli OFDM, to jest możliwe wyznaczenie sygnałów na poszczególnych podnośnych za pomocą bloku FFT. Sygnały te oznaczone jako Y_k są wtedy określone dla każdej podnośnej za pomocą równań:

$$Y_k = H_k^{(1)} X_k^{(1)} + H_k^{(2)} X_k^{(2)} + N_k \tag{5}$$

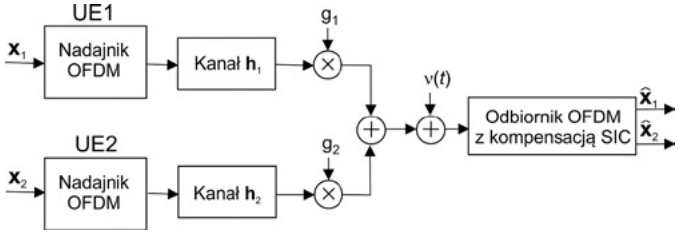
gdzie $X_k^{(1)}, X_k^{(2)}$ są symbolami QPSK/QAM transmitowanymi z obu nadajników na *k*-tej podnośnej, są współczynnikami kanału na tej podnośnej $H_k^{(1)}, H_k^{(2)}$ określonymi na podstawie operacji

$$\mathbf{H}^{(i)} = \left[H_0^{(i)}, H_1^{(i)}, \ldots, H_{N-1}^{(i)}\right] = FFT\{g_i \boldsymbol{h}_i\}, \quad i = 1,2 \tag{6}$$

zaś N_k jest próbką szumu na wyjściu korelatora FFT *k*-tej podnośnej. Współczynniki oraz pozwalają na przyporządkowanie mocy sygnałom pochodzącym od obu terminali. W pracy [10] zaproponowano, aby zamiast standardowego działania kompensatora interferencji dokonano wstępnej detekcji łącznej obu symboli danych $\bar{X}_k^{(1)}, \bar{X}_k^{(2)}$, która zgodnie z zasadą maksymalnej wiarygodności, zakładając biały



» *Rys. 4. Model transmisji w łączu w górę z wielodostępem NOMA*
» *Fig. 4. NOMA multiple access uplink transmission model*



» *Rys. 5. Model systemu transmisji OFDM z wielodostępem NOMA*
» *Fig. 5. NOMA multiple access OFDM transmission system model*

szum gaussowski, może być zrealizowana za pomocą algorytmu:

$$\left(\bar{X}_k^{(1)}, \bar{X}_k^{(2)}\right) = \arg \min_{X_k^{(1)}, X_k^{(2)}} \left| Y_k - \hat{H}_k^{(1)} X_k^{(1)} - \hat{H}_k^{(2)} X_k^{(2)} \right|^2, \quad k = 0,1,\ldots,N-1 \quad (7)$$

Celem wyznaczenia wstępnych decyzji dotyczących symboli danych jest określenie symbolu danych dla kanału słabszego $\bar{X}_k^{(2)}$, aby polepszyć jakość próbek, na podstawie której dokonuje się detekcji sygnału silniejszego za pomocą wzorów:

$$\tilde{X}_k^{(1)} = dec\left(\frac{Y_k - \hat{H}_k^{(2)} \bar{X}_k^{(2)}}{\hat{H}_k^{(1)}}\right), \quad k = 0,1,\ldots,N-1 \quad (8)$$

Sygnał ten jest z kolei zastosowany do kompensacji jego wpływu w sygnale łącznym a następnie dokonania ostatecznej detekcji sygnału słabszego według zależności

$$\tilde{X}_k^{(2)} = dec\left(\frac{Y_k - \hat{H}_k^{(1)} \tilde{X}_k^{(1)}}{\hat{H}_k^{(2)}}\right), \quad k = 0,1,\ldots,N-1 \quad (9)$$

Wyniki symulacyjne przedstawione w [10] wskazują na istotną poprawę działania kompensatora interferencji dzięki zastosowanemu ulepszeniu. Dalsze polepszenie jakości można uzyskać przez wprzęgnięcie do procesu decyzyjnego dekodowania kodu kanałowego zastosowanego w łączu dla sygnału silniejszego a następnie wygenerowania na podstawie zadecydowanych bitów informacyjnych słowa kodowego zastosowanego do powtórnego odwzorowania w symbole QPSK/QAM dla sygnału silniejszego użytych ostatecznie w kompensacji wpływu sygnału silniejszego.

Wielodostęp NOMA w pracach 3GPP

Wielodostęp ortogonalny NOMA był przedmiotem zainteresowania nie tylko środowiska akademickiego, lecz również wielu ośrodków naukowo-badawczych pracujących dla producentów sprzętu radiokomunikacyjnego, co uwidoczniło się w wykonanych badaniach relacjonowanych w ramach spotkań konsorcjum 3GPP (*3rd Generation Partnership Project*). Najnowszy dokument 3GPP oznaczony numerem TR 38.812 poświęcony wielodostępowi NOMA i podsumowujący prace w tej dziedzinie został opublikowany w ramach Wydania 16 standardów i raportów technicznych 3GPP [11] w roku 2018. Zawiera on odwołania do 72 pozycji w tym w większości raportów przedłożonych na spotkaniach grupy roboczej 3GPP RAN1 zajmującej się radiowymi sieciami dostępowymi, w tym wielodostępem NOMA przez kluczowe firmy sektora telekomunikacyjnego, np. Nokię, Ericssona, Huawei, ZTE, Samsunga i wiele innych. W dokumencie [11] skoncentrowano się na zastosowaniu wielodostępu NOMA w łączu w górę, szczególnie z zastosowaniem wielodostępu w dziedzinie kodów i rozpatrywano go dla wszystkich głównych rodzajów usług w systemie 5G, tj. wzbogaconej mobilnej transmisji szerokopasmowej eMBB (*enhanced Mobile Broadband*), transmisji o bardzo wysokiej niezawodności i małych opóźnieniach URLLC (*Ultra-Reliable Low-Latency Communications*) oraz masowej komunikacji między maszynami mMTC (*massive Machine-Type Communications*).

Wielodostęp NOMA pozostaje w sferze zainteresowań w ramach badań nad przyszłymi systemami 6G, w szczególności możliwości zastosowania uczenia maszynowego i sztucznej inteligencji, technik wieloantenowych MIMO (*Multiple Input Multiple Output*) oraz MEC (*Mobile Edge Computing*). Nie uzyskał jak dotąd statusu standaryzacji w systemach 5G. Jedną z ważnych przyczyn tego stanu jest złożoność procesu kolejnej kompensacji interferencji SIC, gdy ze wspólnego zasobu korzysta więcej niż dwóch użytkowników, co powoduje trudności w realizacji praktycznej w niskobudżetowych terminalach. Wspomnianym już problemem jest też skuteczność procesu SIC, bowiem, wystąpienie błędu na jednym poziomie tego procesu pociąga za sobą propagację błędów na niższych poziomach. Kolejnym problemem jest efektywny proces optymalizacji doboru mocy sygnałów nadawanych do różnych terminali i wybór ich we wspólnym przydziale zasobów. Potencjalne działanie wielodostępu NOMA z systemami 5G determinuje również konieczność współdziałania z istniejącą metodą wielodostępu OFDMA.

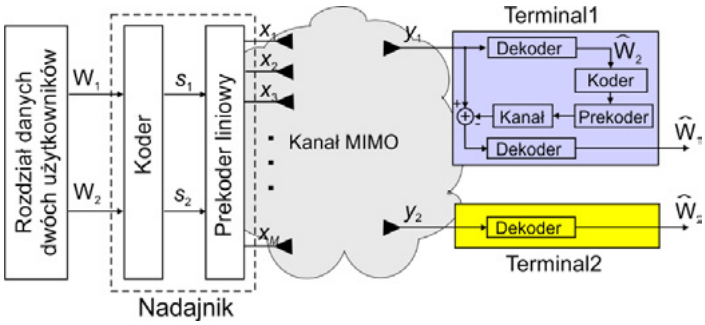
Wielodostęp nieortogonalny w sieciach wieloantenowych

Jak dotąd rozpatrywaliśmy system NOMA z podziałem mocy koncentrując się na sytuacji, w której nadajniki i odbiorniki korzystają w zasadzie z pojedynczych anten. Dodatkowo porównywano efektywność wielodostępu NOMA z wielodostępem ortogonalnym OMA. W 2021 roku

opublikowano jednak znaczący artykuł [12], w którym autorzy rozważają efektywność wielodostępu NOMA w sieciach wieloantenowych. Dodatkowo zwracają uwagę, że należy rozpatrywać jakość działania wielodostępu NOMA nie tylko względem OMA, ale również względem innych istniejących technik wielodostępu.

Załóżmy, że stacja bazowa przesyła do terminali 1 i 2 wiadomości odpowiednio i Stacja bazowa jak i terminale dysponują wspólną książką kodową wykorzystywaną w koderze i dekoderach terminali. W wyniku kodowania wytwarzane są sygnały odpowiednio oraz Są one następnie poddane liniowemu kodowaniu wstępnemu (prekodowaniu) za pomocą wektorów $\mathbf{p}_1$ oraz $\mathbf{p}_2$ o wymiarach 1*xM (M jest liczbą anten stacji bazowej), tak więc wygenerowany przez* nadajnik sygnał ma postać

$$\mathbf{x} = \mathbf{p}_1 s_1 + \mathbf{p}_2 s_2 \tag{10}$$



» *Rys. 6. Konfiguracja systemu z wielodostępem MISO NOMA w kanale w dół dla dwóch użytkowników przy założeniu, że terminal 2 dokonuje detekcji jako pierwszy (według [12])*
» *Fig. 6. Configuration of the MISO NOMA downlink multiple access system for two users assuming that terminal 2 detects first (according to [12])*

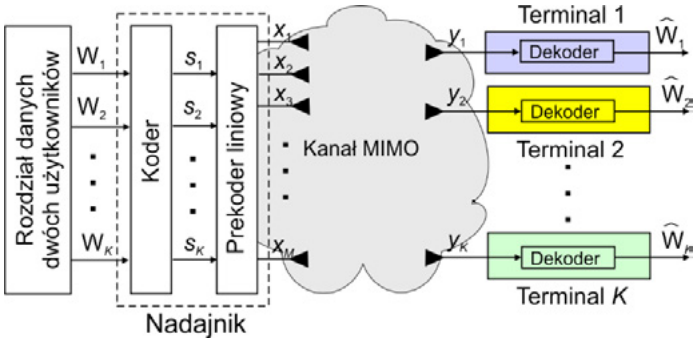
Wektory prekodowania są tak dobrane, że $P_1 = \|\mathbf{p}_k\|^2$, k = 1,2, przy czym moc sygnału nadawanego P spełnia warunek $P_1+P_2 \le P$ Sygnał na wejściu k-tego odbiornika jest określony zależnością:

$$y_k = \mathbf{h}_k^H \mathbf{x} + n_k \tag{11}$$

gdzie $\mathbf{h}_k$ jest wektorem współczynników kanału pomiędzy M antenami nadajnika a odbiornikiem k-tego terminala. Operacja $(.)^H$ oznacza transpozycję hermitowską wektora. Niniejszy schemat działania wielodostępu MISO-NOMA można łatwo uogólnić, gdy transmisja odbywa się do K terminali. W [12] autorzy udowadniają, że maksymalna sumaryczna przepływność w systemie MISO-NOMA nie jest większa niż przepływność dla wielodostępu OMA.

Wielodostęp MISO-NOMA ma jednak konkurencję w postaci innych rozwiązań rozpatrywanych w [12]. Pierwszym z nich jest układ transmisji do wielu użytkowników z liniowymi prekoderami MU-LP (*Multi-User Linear Precoding*). Schemat takiego układu dla transmisji do K terminali pokazano na *rys. 7*. Dla przypadku K = 2 terminale dekodują swoje sygnały s_1 i s_2 traktując resztkową interferencję jako szum. Moc nadajnika może zostać rozłożona równomiernie pomiędzy każdy z sygnałów a wektory prekodowania mogą zostać dobrane np. na zasadzie formowania promieni z wymuszeniem zerowej interferencji ZFBF (*Zero-Forcing Beamforming*). Wtedy bowiem dobrane są one według zasady:

$$\mathbf{h}_1^H \mathbf{p}_2 = \mathbf{h}_2^H \mathbf{p}_1 = 0 \tag{12}$$

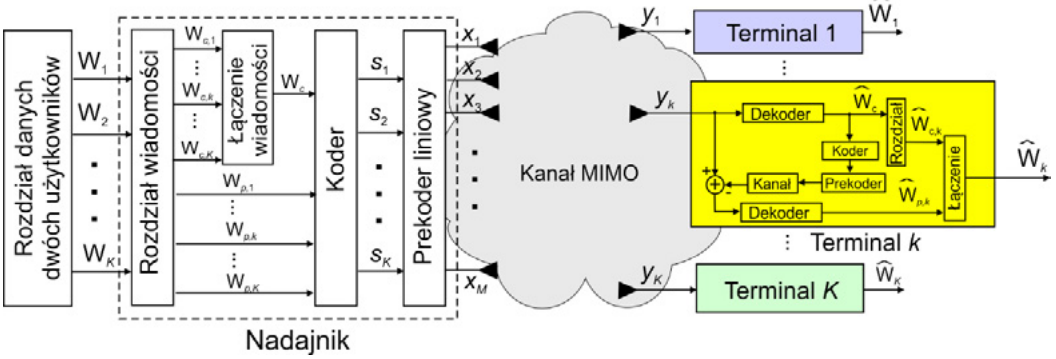


» *Rys. 7. Transmisja MU-LP dla K terminali (według [12])*
» *Fig. 7. MU-LP transmission to K terminals (according to [12])*

Trzecim rozwiązaniem konkurencyjnym wobec MISO-NOMA jest wielodostęp RSMA (*Rate-Splitting Multiple Access*) [13]. Schemat takiego układu pokazano na *rys. 8*.

W przypadku wielodostępu RSMA każda wiadomość W_k przeznaczona do k-tego terminala jest rozdzielona na dwa składniki, tzw. wspólną część $W_{c,k}$ i część prywatną $W_{p,k}$, tj.

$$W_k = W_{c,k} + W_{p,k} \tag{13}$$



» *Fig. 8. Schemat wielodostępu RSMA dla K terminali (według [12]). Każdy z terminali ma schemat blokowy taki jak pokazany na rysunku dla terminala k-tego*
» *Fig. 8. RSMA multiple access scheme for K terminals (according to [12]). Each terminal has a block diagram as shown in the figure for the k-th terminal*

» *Tab. 1. Porównanie podstawowych własności wielodostępu RSMA, NOMA i OMA*
» *Tabl. 1. Comparison of basic properties of RSMA, NOMA and OMA multiple access*

Cecha	RSMA	NOMA	OMA
Rodzaj wielodostępu	częściowo nieortogonalny	w pełni nieortogonalny	w pełni ortogonalny
Zarządzanie interferencją	częściowe dekodowanie, częściowe traktowanie sygnałów jak szum	SIC (silni użytkownicy dekodują słabszych)	unikanie w pełni interferencji
Odporność na niedokładność stanu kanału CSI	wysoka	niska	wysoka
Elastyczność	duża	średnia	mała
Stopień komplikacji	średni	wysoki	niski

Wspólne części wszystkich wiadomości są łączone w jedną wiadomość wspólną W_c. Wiadomość ta jest następnie zakodowania w postaci ciągu wspólnych symboli s_c, Ciąg ten będzie dekodowalny przez wszystkie terminale. Części prywatne wiadomości są indywidualnie kodowanie w postaci symboli s_k, $k = 1,..., K$ Symbole te są liniowo prekodowane, tak więc sygnał nadawany ma postać:

$$\mathbf{x} = \mathbf{p}_c s_c + \sum_{k=1}^{K} \mathbf{p}_k s_k \tag{14}$$

Zakładając, że poszczególne sygnały $s_c, s_1,..., s_K$ są wzajemnie nieskorelowane i mają jednostkową moc, moc nadawana przez stację bazowa iest ograniczona w postaci $P_c + \sum_{k=1}^{K} P_K \leq P$, przy czym $P_c = \|\mathbf{p}_c\|^2$ oraz $P_k = \|\mathbf{p}_k\|^2$. W każdym z terminali najpierw są dekodowane symbole s_c traktując interferencje od prywatnych sygnałów jako szum, dając w rezultacie wspólną łączną część wiadomości $\widehat{W}_c$. Wiadomość ta jest z kolei ponownie zakodowana i poddana prekodowaniu a wynik tego działania jest odjęty od otrzymanego sygnału (por. rys. 8). Dany k-ty terminal dekoduje wtedy jego prywatny sygnał s_k traktując pozostałe sygnały prywatne zawarte w nadanym sygnale (14) jako szum. Na tej podstawie wyznaczona jest prywatna część wiadomości $\widehat{W}_{p,k}$. Ostatecznie syntetyzowana jest wiadomość odebrana $\widehat{W}_k, k = 1, \ldots, K$ zgodnie ze wzorem (13). Opisana architektura wielodostępu RSMA jest opisywana jako architektura jednowarstwowa, ponieważ syntetyzowany jest wyłącznie jeden strumień wspólny sygnałów a w każdym terminalu realizowany jest jeden proces kompensacji interferencji SIC.

Głębsza analiza opisanej powyżej metody wielodostępu wykracza poza ograniczone ramy tego artykułu. Zainteresowanym czytelnikom zaleca się studia przeglądowego artykułu [13]. Autorzy krytycznej względem wielodostępu NOMA pracy [12] stwierdzają, że RSMA cechuje się co najmniej taką przepływnością jak MISO-NOMA, ale w większości wyższą i zawiera znacznie mniej kłopotliwych realizacyjnie układów kompensacji interferencji SIC. Jak dotąd wielodostęp RSMA nie wszedł do etapu standaryzacji 3GPP, choć jest poważnym kandydatem dla systemów 6G. Jest też dalej jednym z gorących tematów badawczych. Przeprowadzone badania pozwalają na sformułowanie tabeli porównawczej dla wielodostępu OMA, NOMA i RSMA przedstawionej poniżej.

Pomimo krytyki wielodostępu NOMA, dalej prowadzone są prace nad tym rodzajem wielodostępu. W szczególności rozpatruje się jego zastosowanie w sieciach Internetu Rzeczy (IoT). Przykładem raportu z takich badań jest praca [14], w której rozpatruje się zastosowanie wielodostępu NOMA w sieciach z terminalami nie wykorzystującymi własnych źródeł energii (BackCom – *Backscatter Communication*).

PODSUMOWANIE

Pomimo faktu, że wielodostęp nieortogonalny NOMA nie został włączony do standardów 3GPP dotyczących systemów 5G New Radio i pomimo jego krytyki związanej z jego ograniczeniami teoretycznymi i implementacyjnymi w systemach wieloantenowych oraz konkurencją innych metod wielodostępu, szczególnie RSMA, budzi on dalej spore zainteresowanie. Wydaje się, że ma on szansę na realizację w łączu w górę w systemach masowej komunikacji między maszynami, w tym w sieciach Internetu Rzeczy.

LITERATURA

[1] Saito, Y., Kishiyama, Y., Benjebbour, A., Nakamura, T., Li, A., Higuchi, K., Non-Orthogonal Multiple Access (NOMA) for Cellular Future Radio Access, *2013 IEEE 77th Veh. Technol. Conf. (VTC Spring)*, 1-5.

[2] Liu, Y., Qin, Z., Elkashlan, M., Ding, Z., Nallanathan, A., Hanzo, L., Non-Orthogonal Multiple Access for 5G and Beyond, *Proceedings of the IEEE*, (2017) 105, nr 12, 2347-2381.

[3] Brannstrom, F., Aulin, T.M., Rasmussen L.K., Iterative Detectors for Trellis-Code Multiple-Access, *IEEE Trans. Commun.*, (2002), 50, nr 9, 1478–1485.

[4] Liu, L., Tong, J., Ping, L., Analysis and Optimization of CDMA Systems with Chip-Level Interleavers, *IEEE J. Sel. Areas Commun.*, (2006), 24, nr 1, 141–150.

[5] Hoshyar, R., Wathan, F.P., Tafazolli, R., Novel Low-Density Signature for Synchronous CDMA Systems over AWGN Channel, *IEEE Trans. Signal Process.*, (2008), 56, nr 4, 1616–1626.

[6] Yuan, Z., Yu, G., Li, W., Yuan, Y., Wang, X., Xu, J., Multi-User Shared Access for Internet of Things," *Proc. IEEE Veh. Technol. Conf. (VTC-Spring)*, (2016), 1–5.

[7] Chen, S., Ren, B., Gao, Q., Kang, S., Sun, S., Niu, K., Pattern Division Multiple Access—A Novel Nonorthogonal Multiple Access for Fifth-Generation Radio Networks, *IEEE Trans. Veh. Technol.*, (2016), 66, nr 4, 3185–3196.

[8] Ding, Z., Liu, Y., Choi, J., Sun, Q., Elkashlan, M., I, Ch.-L., Poor, H.V., Application of Non-Orthogonal Multiple Access in LTE and 5G Networks, *IEEE Commun. Mag.*, (2017) 55, nr 2, 185–191.

[9] Vaezi, M., Schober, R., Ding, Z., Poor, H.V., Non-Orthogonal Multiple Access: Common Myths and Critical Questions, *IEEE Wireless Comm.*, (2019), October, 174-180.

[10] Ghazi, H.S., Wesołowski, K.W., Improved Detection in Successive Interference Cancellation NOMA OFDM Receiver, *IEEE Access*, (2019), 7, 103325–103335.

[11] 3GPP TR 38.812, 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on Non-Orthogonal Multiple Access (NOMA) for NR (Release 16), V16.0.0 (2018-12).

[12] Clerckx, B., Mao, Y., Schober, R., Jorswieck, E.A., Love, D.J., Yuan, J., Hanzo, L., Li, G.Y., Larsson, E.G., Caire, G., In NOMA Efficient in Multi-Antenna Networks? A Critical Look at Next Generation Multiple Access Techniques. *IEEE Open Journal of Communication Society*, (2021), 2, 1310-1342.

[13] Mao, Y., Dizdar, O., Clerckx, B., Schober, R., Popovski, P., Poor, H.V., Rate-Splitting Multiple Access: Fundamentals, Survey, and Future Research Trends, *IEEE Commun. Surveys & Tutorials*, (2022), 24(4), 2073-2126.

[14] Mondal, S., Bepari, D., Chandra A., Singh, K., Li. Ch.-P., Ding Z., A Comprehensive Survey on NOMA-Based Backscatter Communication for IoT Applications, artykuł zaakceptowany do publikacji w *IEEE Internet of Things Journal*, DOI 10.1109/JIOT.2025.3548581.

DOI: 10.15199/59.2025.4.3

Modele numeryczne ciała człowieka umożliwiające analizę i projektowanie nowoczesnych systemów łączności bezprzewodowej

Numerical human body models for the analysis and design of modern wireless communication systems

dr hab. inż. Łukasz Januszkiewicz prof. PŁ, Politechnika Łódzka, Instytut Elektroniki
lukasz.januszkiewicz@p.lodz.pl

STRESZCZENIE: W artykule przedstawione zostały modele numeryczne ciała człowieka umożliwiające przeprowadzenie symulacji komputerowych nowoczesnych systemów łączności bezprzewodowej. Omówione zostały ograniczenia modeli pełnofalowych, w przypadku zastosowania ich do analizy systemów wykorzystujących wielkie częstotliwości. Następnie zaprezentowane zostały modele służące do analizy efektu przesłonięcia łącza bezprzewodowego ciałem człowieka, wykorzystujące jednorodną teorię dyfrakcji. Omówione zostały również modele pozwalające wyznaczyć gęstość mocy fali elektromagnetycznej we wnętrzu ciała. Skupiono się na modelach bazujących na śledzeniu promieni oraz modelach wykorzystujących techniki głębokiego uczenia (Deep-Learning) i sztuczne sieci neuronowe. Wskazano również kierunki rozwoju modeli ciała człowieka w kontekście systemów 6 generacji.

SŁOWA KLUCZOWE: metody numeryczne, łączność bezprzewodowa, elektromagnetyzm obliczeniowy, sieci WBAN, modele ciała człowieka

ABSTRACT: This article presents numerical human body models that enable computer simulations of modern wireless communication systems. Limitations of full-wave models when applied to the analysis of systems utilizing high frequencies are discussed. Subsequently, models for analyzing the human body shadowing effect on wireless links, employing the uniform theory of diffraction, are introduced. Additionally, models for determining the power density of electromagnetic wave within the body are discussed. The focus is on ray-tracing based models and models utilizing deep-learning techniques and artificial neural networks. Directions for the development of human body models in the context of 6th generation systems are also indicated.

KEYWORDS: numerical methods, wireless communication, computational electromagnetics, WBAN networks, human body models

Wraz z dynamicznym rozwojem technologii bezprzewodowych coraz większego znaczenia nabiera możliwość modelowania interakcji fal elektromagnetycznych z ludzkim ciałem. Ze względu na wykorzystywanie przez ludzi w ich bezpośrednim otoczeniu wielu systemów pozwalających na bezprzewodową transmisję danych, takich jak smartfony, smartwatche, słuchawki bezprzewodowe etc. ich projektowanie wymaga uwzględnienia wpływu ciała człowieka. Konieczne jest w tym przypadku przeprowadzenie dokładnej analizy wpływu tkanki biologicznej na propagację fali elektromagnetycznej oraz wyznaczenie ilości mocy pochłoniętej przez ciało. W tym kontekście modele numeryczne ciała człowieka stały się nieodzownym narzędziem inżynierów, umożliwiającym wirtualną walidację rozwiązań przed ich fizyczną implementacją.

Projektowanie systemów i urządzeń bezprzewodowych odbywa się obecnie z wykorzystaniem symulacji komputerowych bazujących na metodach numerycznych. Metody pełnofalowe, bazujące bezpośrednio na równaniach Maxwella są obecnie najdokładniejszymi metodami pozwalającymi na analizę struktur elektromagnetycznych. Należą do nich takie metody jak metoda różnic skończonych w dziedzinie czasu FDTD (ang. *Finite-Difference Time-Domain*), czy metoda elementów skończonych FEM (ang. *Finite Element Method*). Dzięki nim możliwe jest badanie

złożonych struktur biologicznych obejmujące analizę rozkładu pola EM, oszacowanie tłumienia fali wprowadzonego przez ciało oraz określenie ilości energii pochłoniętej przez tkanki. Ze względu na wysoką złożoność obliczeniową, ich zastosowanie bywa ograniczone w przypadku symulacji dużych obiektów, takich jak np. wnętrza budynków lub pojazdy. Uwzględnienie obecności ludzi w takich obiektach w symulacjach wykorzystujących metody pełnofalowe często nie jest możliwe ze względu na wymagania dotyczące pamięci komputera.

W przypadku symulacji obiektów wielokrotnie większych od długości fali EM, stosowane są metody przybliżające zjawisko rozchodzenia się fal elektromagnetycznych za pomocą śledzenia promieni (ang. *ray-tracing*) i analizy ich interakcji z badanymi obiektami. Metody tego rodzaju charakteryzują się większą wydajnością obliczeniową jednak w ich przypadku konieczne jest zastosowanie bardzo uproszczonych modeli ciała człowieka. Dzięki zastosowaniu tych modeli możliwe jest analizowanie systemów bezprzewodowych w środowiskach, w których sygnał bezprzewodowy oddziałuje jednocześnie z wieloma osobami, meblami, ścianami i innymi przeszkodami, przy zachowaniu akceptowalnego czasu obliczeń.

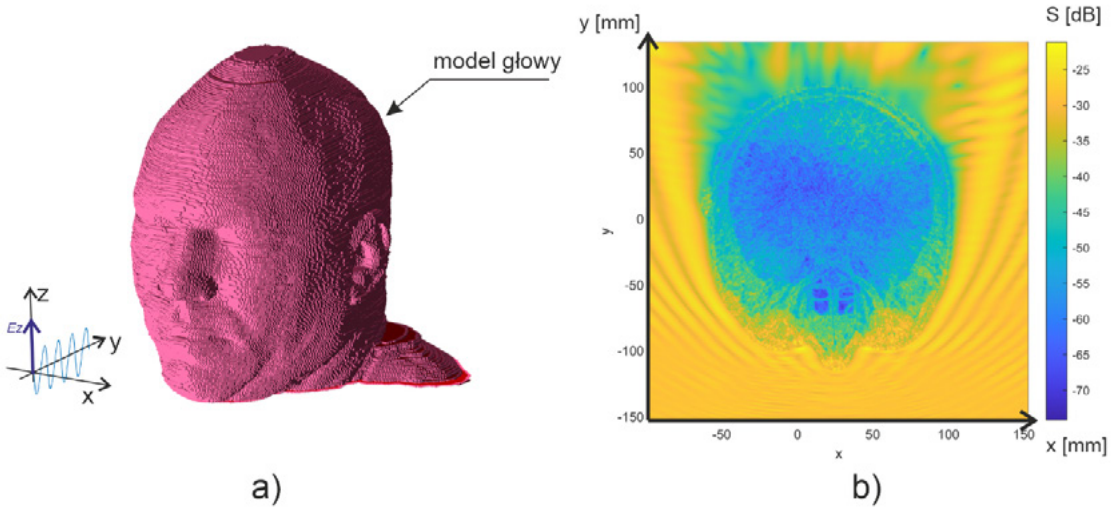
W niniejszym artykule omówiono wybrane rodzaje modeli numerycznych ciała człowieka, które są przydatne w projektowaniu systemów bezprzewodowych, ze szczególnym uwzględnieniem modeli pełnofalowych oraz modeli opartych na śledzeniu promieni. Omówiono również aktualne wyzwania związane z modelowaniem ciała w systemach wykorzystujących wielkie częstotliwości oraz przyszłe kierunki rozwoju tej dziedziny, w tym wykorzystanie uczenia maszynowego do optymalizacji procesów symulacyjnych.

MODELE NUMERYCZNE CIAŁA CZŁOWIEKA W SYMULACJACH PEŁNOFALOWYCH

Ciało człowieka zbudowane jest tkanek o zróżnicowanych właściwościach elektrycznych i skompilowanej geometrii. Dokładna analiza interakcji fali elektromagnetycznej z tak złożonym obiektem wymaga zastosowania pełnofalowych metod numerycznych, które rozwiązują równania Maxwella dla zdyskretyzowanego modelu odwzorowującego właściwości geometryczne i elektryczne obiektu. W tych metodach możliwe jest zastosowanie niejednorodnych modeli anatomicznych, które precyzyjnie odwzorowują wewnętrzną budowę ciała człowieka lub jego fragmentów, uwzględniając parametry elektryczne różnych tkanek. W zależności od potrzebnej dokładności i charakteru symulacji stosuje się różnorodne podejścia do konstrukcji modeli, począwszy od dokładnych odwzorowań opartych na obrazowaniu medycznym, a skończywszy na uproszczonych reprezentacjach wykorzystujących proste bryły geometryczne.

Najbardziej realistyczne modele anatomiczne tworzone są na podstawie danych z tomografii komputerowej (CT) [1] lub rezonansu magnetycznego (MRI) [2]. Przykładem takich modeli są bazy danych opracowane w ramach Visible Human Project [3], czy też modele opublikowane przez IT'IS Foundation, takie jak Duke, Ella czy Thelonious, które uwzględniają zróżnicowanie populacyjne pod względem wieku i budowy ciała [4]. Modele te różnią się liczbą odwzorowanych tkanek oraz zastosowaną rozdzielczością dyskretyzacji. Przykładem anatomicznego modelu wielotkankowego ciała jest model dostępny w programie Remcom XFdtd [5], wykorzystującym metodę FDTD. Model ten, opracowany przez Hershey Institute na podstawie skanu MRI męskiego ciała odwzorowuje 40 tkanek. Jest on zbudowany z prostopadłościennych komórek elementarnych (wokseli). Można wykorzystać różne warianty tego modelu, z wokselami o krawędzi o długościach od 1 mm do 10 mm. Jego wariant zbudowany z wokseli 1mm został przedstawiony na *rys. 1* wraz z przykładowymi wynikami symulacji powierzchniowej gęstości mocy dla częstotliwości 3,5 GHz. Dzięki programowi Varipose firmy Remcom, model całego ciała może być układany w różnych pozach wynikających z ułożenia przestrzennego kończyn.

W przypadku symulacji wymagających mniejszej dokładności, możliwe jest zastosowanie w metodach pełnofalowych uproszczonych modeli ciała. Pierwszy poziom uproszczenia polega na zachowaniu kształtu i wymiarów ciała przy jednoczesnym zastąpieniu tkanek materiałem jednorodnym o właściwościach elektrycznych odpowiada-



» *Rys.1. Wielotkankowy model głowy dostępny w programie XFdtd (a) oraz uzyskane z jego wykorzystaniem wyniki symulacji powierzchniowej gęstości mocy fali EM o częstotliwości 3,5 GHz na płaszczyźnie x-y, na wysokości oczu (b)*
» *Fig. 1. Multi-tissue head model available in XFdtd software (a) and the simulation results obtained using it for the surface power density of an EM wave at 3.5 GHz on the x-y plane, at eye level (b)*

jących uśrednionym parametrom tkanek [6]. Takie modele numeryczne maja swoje odpowiedniki fizyczne w postaci modeli typu SAM (ang. *Specific Antropomorphic Manequin*) dzięki czemu weryfikacja pomiarowa symulacji jest łatwiejsza niż w przypadku modeli dokładnych, dla których odpowiednik fizyczny jest niedostępny. Jeszcze większe uproszczenie stanowią modele parametryczne, w których części ciała aproksymowane są prostymi bryłami geometrycznymi, takimi jak cylindry czy elipsoidy, o właściwościach zbliżonych do rzeczywistych tkanek. Takie model są szczególnie przydane w procesie automatycznej optymalizacji elementów i systemów ponieważ można je łatwo generować w kolejnych iteracjach procesu [7].

» *Tab. 1. Parametry symulacji z wykorzystaniem modelu całego ciała człowieka w metodzie FDTD*
» *Tabl.1. Simulation Parameters Using the Whole-Body Human Model in the FDTD Method*

częstotliwość [GHz]	**1**	**2,5**	**10**	**28**	**100**
długość fali w próżni [mm]	300	120	30	10,71	3
przenikalność elektryczna względna krwi	61	58	45	24	8,3
długość fali we krwi [mm]	38,41	15,76	4,47	2,19	1,04
maksymalna długość krawędzi woksela [mm]	3,84	1,58	0,45	0,22	0,10
liczba wokseli potrzebna do odwzorowania całego ciała [miliony komórek]	13	122	4 388	41 138	372 061
wymagana pamięć komputera do przeprowadzenia symulacji [GB]	0,58	4,3	128	1 200	10 400

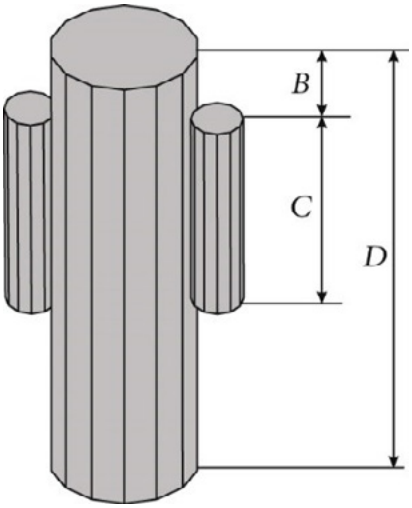
Kluczowym aspektem w symulacjach pełnofalowych jest odpowiednia dyskretyzacja przestrzeni obliczeniowej, która dotyczy również modeli ciała. Przy ustalonych wymiarach modelu ciała, wielkość komórek elementarnych potrzebnych do jego odwzorowania wpływa na ich liczbę, która określa wielkość pamięci komputera wymaganą do przeprowadzenia symulacji. Wielkość komórek elementarnych zależy od metody numerycznej oraz właściwości modelowanego obiektu. W metodzie FDTD wymagane jest zastosowanie wokseli, których długość krawędzi jest co najmniej dziesięciokrotnie mniejsza od długości najkrótszej fali w symulowanym obiekcie. W *tabeli 1* zebrano dla kilku przykładowych częstotliwości parametry symulacji metodą FDTD całego ciała. Dla uproszczenia przyjęto rozmiar woksela wynikający z długości fali we krwi, której parametry podano za [8]. Wyniki ilustrują sposób w jaki wielkość woksela wpływa na wymagania dotyczące pamięci w tej metodzie. Oprócz prostych zależności wynikających z długości fali i efektu skrócenia fali w tkankach, podczas doboru wielkości komórki elementarnej należy uwzględnić również szczególnie wymagające pod względem dyskretyzacji cienkie warstwy graniczne, takie jak skóra, gdzie konieczne może być zastosowanie rozdzielczości rzędu 0,1 mm nawet dla mniejszych częstotliwości. Dla metody FDTD typowe symulacje modelu ciała ludzkiego przy częstotliwościach rzędu 1-2 GHz wymagają od 10 do 10 GB pamięci operacyjnej, przy czym czas obliczeń może wynosić od kilkunastu minut do kilku godzin w zależności od dostępnej mocy obliczeniowej. W zakresie wysokich częstotliwości (powyżej 10 GHz), gdzie długość fali w tkance spada do kilku milimetrów, konieczne staje się stosowanie bardzo małych wokseli, co znacznie zwiększa zapotrzebowanie na zasoby obliczeniowe, wymagając niekiedy setek gigabajtów pamięci RAM i wyspecjalizowanych klastrów obliczeniowych.

Wybór odpowiedniego modelu anatomicznego i metody numerycznej w symulacjach pełnofalowych zależy od konkretnego zastosowania, wymaganej dokładności oraz dostępnych zasobów obliczeniowych. Metoda FDTD jest bardzo często wybierana ze względu na swoją uniwersalność i względną prostotę implementacji. Dzięki zastosowaniu dyskretyzacji o lokalnie dużym poziomie zagęszczenia, Metoda FEM oferuje lepsze odwzorowanie skomplikowanych kształtów, szczególnie w przypadku analiz wymagających wysokiej precyzji lokalnej. Istotnym ograniczeniem tych modeli są bardzo duże wymagania dotyczące dostępnej pamięci komputera zwłaszcza w symulacjach oddziaływania fal o wysokich częstotliwościach.

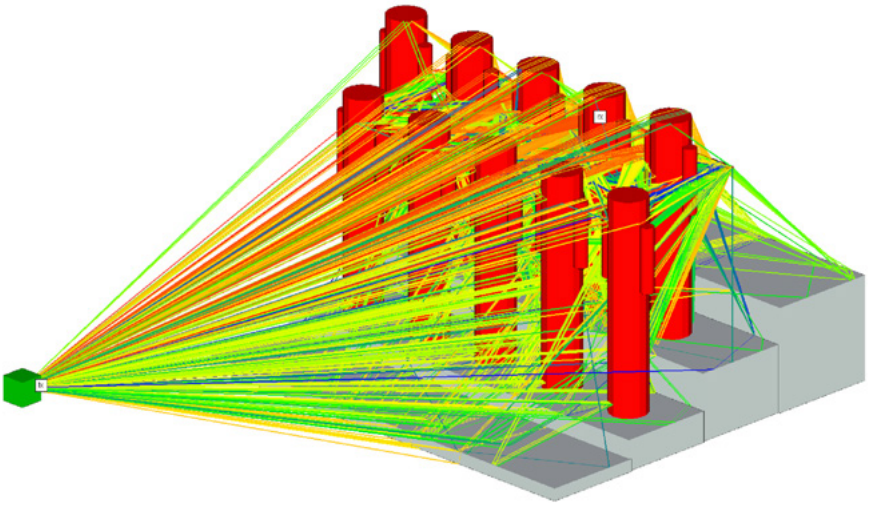
MODELE UMOŻLIWIAJĄCE ANALIZĘ EFEKTU PRZYSŁONIĘCIA ŁĄCZA

W przypadku analizy systemów bezprzewodowych, w których uwzględniane jest oddziaływanie fal z dużymi obiektami takimi jak np. budynki, nie stosuje się metod pełnofalowych, ze względu na duże wymagania dotyczące pamięci komputera. Tych ograniczeń pozbawione są programy działające na zasadzie śledzenia wiązek lub śledzenia promieni. Przykładem takiego programu jest Remcom XGtd działający na zasadzie śledzenia promieni [9]. Wykorzystuje on metodę wystrzeliwania promieni z nadajników, a następnie dokonuje ich analizy w modelowanym obiekcie. Promienie, które docierają do odbiornika oddziałują z obiektami, a program umożliwia symulację różnych interakcji promieni z modelowanymi obiektami. Możliwe jest uwzględnienie odbicia promieni od powierzchni obiektów oraz ich dyfrakcji na krawędzi elementów, jak również przenikanie promieni przez obiekty. W tym programie wykorzystywana jest jednorodna teoria dyfrakcji (ang. *Uniform Theory of Diffraction* UTD) do wyznaczania wartości pola elektrycznego wzdłuż promienia.

Metoda UTD została z powodzeniem zastosowana do modelowania efektu przysłonięcia łącza bezprzewodowego przez ciało człowieka. Zastosowano w niej proste elementarne obiekty aby odwzorować ciało człowieka. Model umożliwiający symulacje wpływu ciała na łącze bezprzewodowe pracujące w paśmie 3,5 GHz za pomocą metody śledzenia promieni przedstawiono w [10]. W tym przypadku ciało człowieka zostało odwzorowane za pomocą graniastosłupów



» *Rys. 2. Model ciała człowieka w metodzie śledzenia promieni na pasmo 3,5 GHz. Wymiary: B = 0,195 m, C = 0,53 m, D = 1,8 m*
» *Fig. 2. Human body model in the ray-tracing method for the 3.5 GHz band. Dimensions: B = 0.195 m, C = 0.53 m, D = 1.8 m*



» *Rys.3. Układ modeli odwzorowując ludzi na fragmencie trybuny stadionu oraz przebieg promieni użytych do wyznaczenia mocy odebranej w obszarze przysłoniętym*
» *Fig. 3. Arrangement of models representing people on a section of a stadium grandstand and the path of rays used to determine the received power in the shadowed area*

foremnych, co pokazano na *rys. 2.* Model umożliwia symulację efektu przysłonięcia ciała dla pasma średniego 5G. Dzięki jego dużej wydajności obliczeniowej możliwe było przeprowadzenie symulacji tego efektu z uwzględnieniem 10 osób znajdujących się na wycinku trybuny stadionu. Na *rys. 3.* przedstawiono układ modeli odwzorowując ludzi oraz przebieg promieni użytych do wyznaczenia mocy odebranej w obszarze przysłoniętym. Czas symulacji wyniósł jedynie 5 minut a wymagana pamięć komputera nie była większa niż 1 GB.

W publikacji [11] opisano podobny model dla pasma 2,4 GHz. W tym przypadku umożliwił on przeprowadzenie symulacji przysłonięcia łącza z uwzględnieniem wnętrza pomieszczenia oraz 4 osób, które w nim przebywały. Również w tym przypadku zasoby obliczeniowe potrzebne do przeprowadzenia symulacji były zdecydowanie mniejsze niż w przypadku modeli stosowanych w metodzie FDTD. Oba modele te zostały zweryfikowane poprzez porównanie z referencyjnymi symulacjami opartymi na modelach wielotkankowych ciała człowieka realizowanych metodą FDTD.

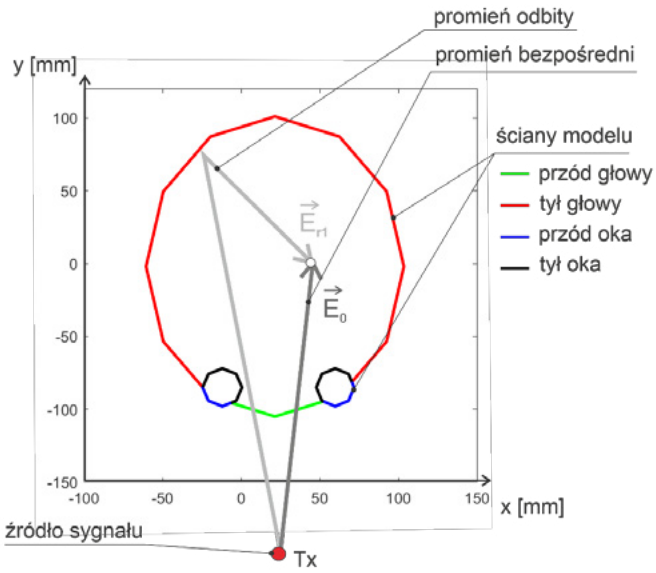
Jak to wykazały opisane wyżej badania, zastosowanie modelowania ciała człowieka za pomocą graniastosłupów w połączeniu ze śledzeniem promieni i jednolitą teorią dyfrakcji (UTD) stanowi bardzo efektywne podejście do przewidywania efektu przesłonięcia łącza bezprzewodowego. Kluczową zaletą tego podejścia jest jego niezwykła wydajność obliczeniowa, która pozwala na symulację złożonych scenariuszy z wieloma osobami obecnymi w budynku. Jest to szczególnie istotne w kontekście rosnącej gęstości użytkowników i urządzeń bezprzewodowych, zarówno w istniejących sieciach, jak i w nowo wdrażanych systemach (takich jak 5G i 6G), zwłaszcza w pasmach milimetrowych.

MODELE UMOŻLIWIAJĄCE WYZNACZENIE GĘSTOŚCI MOCY WE WNĘTRZU CIAŁA

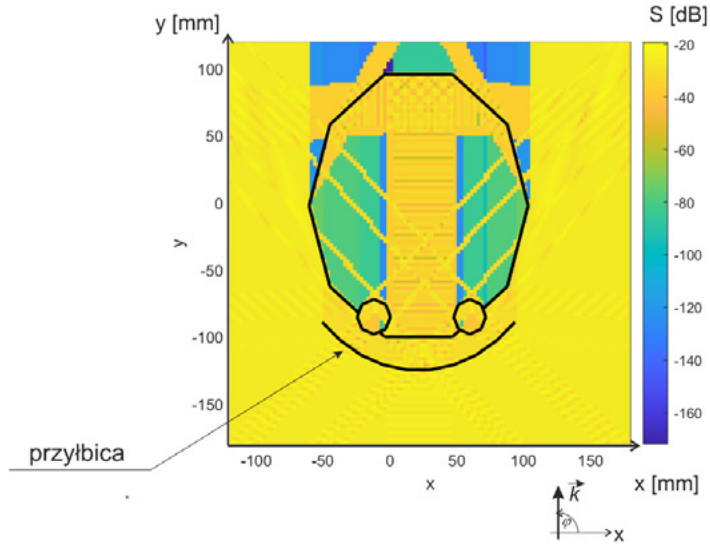
Wyznaczanie gęstości mocy fali elektromagnetycznej w ciele człowieka pozwala na zbadanie poziomu ekspozycji na źródła fal elektromagnetycznych. Dzięki temu możliwe jest zapewnienie bezpieczeństwa użytkownikom urządzeń emitujących promieniowanie, oraz optymalizacja urządzeń i systemów pod katem mocy koniecznej do uzyskania skutecznej transmisji. Typowo stosowane są w tym celu metody pełnofalowe i opisane wyżej modele ciała. Poniżej przedstawiono alternatywne podejście bazujące na modelach bazujących na śledzeniu promieni oraz badania wykorzystujące w tym celu sztuczne sieci neuronowe.

Uproszczony model wielościenny głowy

W ramach badań przedstawionych w [12] opracowany został dwuwymiarowy model numeryczny do symulacji gęstości mocy fali elektromagnetycznej w obszarze głowy i oczu, działający na częstotliwości 26 GHz. Wykorzystano w tym przypadku metodę śledzenia promieni (ray-tracing), która modeluje oddziaływanie fali z ciałem poprzez odwzorowanie tłumienia promieni przy przejściu przez przeszkodę. Uwzględnione zostało również jednokrotne odbicie od ścian. Uproszczony model reprezentuje przekrój poprzeczny głowy na wysokości środka gałek ocznych. Głowa i oczy są odwzorowane za pomocą ścian ułożonych na elipsach, przybliżonych wielokątami – głowa za pomocą 12 ścian, a gałka oczna za pomocą 8 ścian, co pokazano na Rys. 4 . Ściany modelu charakteryzują się określonymi wartościami tłumienia (A) i współczynnika odbicia (Γ). W modelu uwzględniono cztery różne wartości tłumienia ścian dla różnych obszarów: przodu głowy ($A1$), tyłu głowy ($A2$), przodu oka ($A3$) i tyłu



» *Rys. 4. Model wielościenny głowy*
» *Fig. 4. Multi-wall head model*



» *Rys. 5. Przykładowe wyniki symulacji głowy osłoniętej przyłbicą, która została poddana ekspozycji na falę o częstotliwości 26 GHz*
» *Fig. 5. Example simulation results for a head with a face shield, exposed to a 26 GHz wave*

oka (*A4*), a także odpowiadające im współczynniki odbicia (*Γ1,Γ2,Γ3,Γ4*).

Parametry modelu, czyli wartości tłumienia ścian i ich współczynniki odbicia, zostały zidentyfikowane za pomocą algorytmu automatycznej optymalizacji EStra w wariancie najniższego rzędu. Funkcją celu do minimalizacji była maksymalna różnica między średnimi i maksymalnymi wartościami gęstości mocy uzyskanymi z modelu referencyjnego i modelu uproszczonego. Dane referencyjne do optymalizacji pochodziły z symulacji elektromagnetycznych przeprowadzonych metodą FDTD z wykorzystaniem heterogenicznego modelu ciała człowieka. Na podstawie tych symulacji obliczono średnie i maksymalne wartości gęstości mocy dla obszaru głowy i oczu.

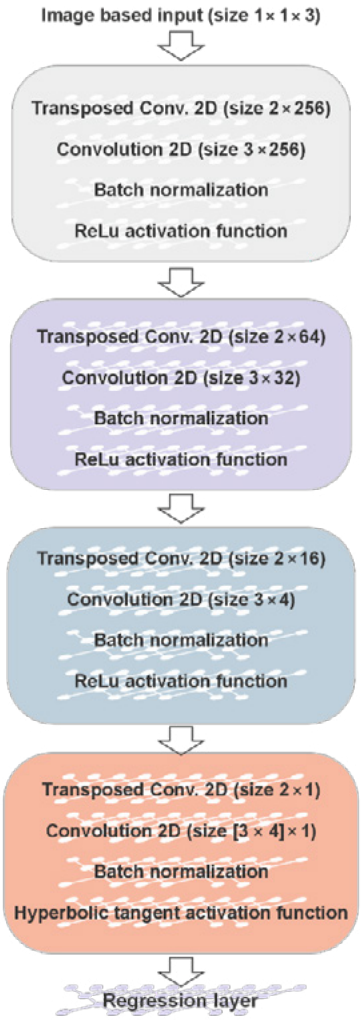
Model wielościenny umożliwia dobre przybliżenie wartości średniej i maksymalnej gęstości mocy fali elektromagnetycznej w obszarze głowy i gałki ocznej. Uzyskane wyniki różnią się od wyników symulacji FDTD o nie więcej niż 0,6 dB. Największa różnica (0,59 dB) występuje w przypadku wartości średniej dla obszaru prawego oka. Dodatkowo, model ten charakteryzuje się znacznie krótszym czasem obliczeń i mniejszymi wymaganiami pamięciowymi w porównaniu do metody FDTD. Symulacja 6390 punktów obliczeniowych w uproszczonym modelu zajmuje około 48 sekund, wykorzystując 1,9 GB pamięci RAM. Dla porównania, symulacja FDTD trwała 9 godzin i 32 minuty i wymagała 39,6 GB pamięci. Oznacza to, że czas obliczeń jest ponad 700-krotnie krótszy w przypadku modelu uproszczonego.

Zastosowaniem tego modelu może być analiza wpływu ochron osobistych na poziom mocy we wnętrzu ciała w zakresach wielkich częstotliwości. Na *rys. 5* przedstawiono przykładowe wyniki symulacji głowy osłoniętej przyłbicą, która została poddana ekspozycji na fale o częstotliwości 26 GHz. Zastosowanie tego modelu pozwoliło oszacować poziom redukcji gęstości mocy w obszarze gałek ocznych wprowadzony przez osłonę.

MODEL WYKORZYSTUJĄCY TECHNIKI GŁĘBOKIEGO UCZENIA

Rozwijane obecnie szeroko metody uczenia maszynowego znalazły również zastosowanie do modelowania zagadnień elektromagnetycznych w ciele człowieka. Przykładem mogą być konwolucyjne sieci neuronowe (CNN), które znajdują zastosowanie do modelowania ilości energii fali EM pochłoniętej przez ciało człowieka [13].

W [14] przedstawiono koncepcję wirtualnego sensora, umożliwiającego modelowanie gęstości mocy w przekroju poprzecznym ludzkiej głowy na wysokości oczu dla częstotliwości 3,5 GHz. Termin „wirtualny" odnosi się w tym kontekście do modelu wykorzystującego konwolucyjną sieć neuronową zamiast pełnofalowych symulacji komputerowych. Zamiast szacować rozkład gęstości mocy w całym ciele, sensor ten umożliwia estymację maksymalnego i średniego poziomu gęstości mocy dla obszarów głowy i oczu. Uzasadnieniem dla takiego podejścia jest kluczowe znaczenie tych parametrów w ocenie bezpieczeństwa osób eksponowanych na fale elektromagnetyczne. W badaniach wykorzystano konwolucyjną sieć neuronową trenowaną od podstaw, składającą się z 18 warstw. Jej architektura (przedstawiona na Rys. 6) opiera się na bloku złożonym z czterech warstw – warstwy transponowanej konwolucji, warstwy konwolucyjnej, warstwy normalizacji wsadowej oraz funkcji aktywacji – powtórzonym czterokrotnie. W pierwszych trzech blokach zastosowano funkcję aktywacji ReLU, pozwalająca ograniczyć efekt przetrenowania sieci. Ostatni blok wykorzystuje funkcję tangens hiperboliczny (tanh), generującą wartości w zakresie od -1 do 1. Pomimo wyższych kosztów obliczeniowych, tanh umożliwia



» *Rys. 6. Struktura sztucznej sieci neuronowej wykorzystanej do modelowania powierzchniowej gęstości mocy fali EM we wnętrzu głowy [14]*
» *Fig. 6. Structure of the artificial neural network used to model the surface power density of an EM wave inside the head [14]*

generowanie silnych gradientów, co przyspiesza aktualizację wag CNN, a jej symetryczność wokół zera sprzyja szybszej konwergencji.

Mimo iż sieć została wytrenowana na wynikach uzyskanych metodą różnic skończonych w dziedzinie czasu (FDTD), po zakończeniu procesu uczenia jest ona zdolna do autonomicznej estymacji parametrów opisujących poziom gęstości mocy. Umożliwia to bardzo szybkie oszacowanie dla dowolnego kierunku fali padającej, co stanowi znaczące przyspieszenie w porównaniu z klasycznym podejściem wykorzystującym FDTD. Wyniki wstępnych badań w tym obszarze są obiecujące i wskazują na możliwość efektywnego wykorzystania konwolucyjnych sieci neuronowych do modelowania złożonych zjawisk, takich jak pochłanianie energii przez ludzkie ciało. Pomimo dużych kosztów obliczeniowych związanych z procesem gromadzenia danych treningowych, docelowa sieć umożliwia znaczące skrócenie czasu niezbędnego do estymacji parametrów dotyczących rozkładu mocy w obszarach głowy i oczu.

KIERUNKI ROZWOJU MODELI CIAŁA CZŁOWIEKA W KONTEKŚCIE SYSTEMÓW 6 GENERACJI

Nowoczesne systemy łączności bezprzewodowej muszą sprostać wykładniczemu przyrostowi przesyłanych danych w kolejnych latach, co wymusza pilną potrzebę rozwoju zaawansowanej infrastruktury cyfrowej. Sieci szóstej generacji (6G), z przewidywanym komercyjnym wdrożeniem około roku 2030, stanowią odpowiedź na te wyzwania. Technologia 6G ma umożliwić generowanie Zettabajtów danych przez ludzi, pojazdy autonomiczne, roboty i drony, zapewniając wszechobecną łączność. Kluczowe innowacje w tym zakresie obejmują rozszerzenie koncepcji zintegrowanego wykrywania i komunikacji (*Integrated Sensing and Communications*) oraz zastosowania sztucznej inteligencji [15].

Podstawą tych innowacji będzie wykorzystanie wyższych częstotliwości, przekraczających kilkadziesiąt GHz, w tym pasm milimetrowych i terahercowych. Szerokie i dostępne spektrum w tych pasmach umożliwi osiągnięcie wysokich przepustowości i małych opóźnień, niezbędnych do realizacji zaawansowanych scenariuszy komunikacyjnych. Ponadto, specyficzne właściwości propagacyjne fal na tych częstotliwościach, takie jak zwiększone tłumienie i ograniczony zasięg, będą sprzyjać precyzyjnemu wykrywaniu obiektów, ich lokalizacji oraz mapowaniu otoczenia, integrując funkcje komunikacyjne z percepcyjnymi [16].

Projektowanie innowacyjnych systemów 6G, integrujących transmisję i wykrywanie, będzie wymagało opracowania nowych modeli ciała ludzkiego. Tradycyjne modele muszą zostać dostosowane do wyższych pasm częstotliwości, uwzględniając właściwości tkanek, a nawet odzieży [17]. Ponadto zmianie ulegną scenariusze ekspozycji ciała na źródło promieniowania. Ze względu na ograniczenie zasięgów rozważane będą częściej lokalne scenariusze komunikacji obejmujące swoim zasięgiem jedno pomieszczenie lub wnętrze pojazdu. Wpłynie to w szczególności na konfigurację modeli uproszczonych ciała, które opracowywane są przy założeniu określonych parametrów ekspozycji.

Ze względu na rozwój technik wykrywania i badania obiektów (w tym ciała człowieka) w systemach 6G, nowe modele muszą wiernie odwzorowywać cechy fizjologiczne ciała, które mogą być monitorowane i wykrywane w tych systemach. W szczególności, kluczowe jest opracowanie dynamicznych modeli, zdolnych do symulowania takich parametrów jak tętno i oddychanie, co jest niezbędne do realizacji funkcji zintegrowanego wykrywania ludzi i ich stanów fizjologicznych w kontekście systemów 6G.

PODSUMOWANIE I WNIOSKI

Numeryczne modele ciała człowieka od dawna są wykorzystywane do projektowania i analizy systemów łączności bezprzewodowej, z których korzystamy na co dzień. Mimo ich ugruntowanej pozycji i dostępności w większości pakietów oprogramowania do symulacji elektromagnetycznych, ograniczenia modeli pełnofalowych wciąż inspirują naukowców do poszukiwania alternatywnych rozwiązań.

Szczególnie obiecujące są obecnie modele oparte na sztucznych sieciach neuronowych, których dynamiczny rozwój w ostatnich latach otwiera nowe perspektywy aplikacyjne. W kontekście nadchodzących systemów 6G kluczowe jest rozszerzenie zakresu częstotliwości, dla których te modele są opracowywane.

Wyzwaniem, zwłaszcza dla najwyższych zakresów częstotliwości, pozostaje wydajność obliczeniowa i zapotrzebowanie na pamięć komputerową. Jednakże, z uwagi na oczekiwany w najbliższych latach rozwój komputerów kwantowych, istnieje realna szansa na przełamanie barier związanych z ogromnymi nakładami obliczeniowymi niezbędnymi do symulacji oddziaływania fal o najwyższych częstotliwościach z ludzkim ciałem.

LITERATURA

[1] A. Takei, K. Murotani, S. -I. Sugimoto, M. Ogino and H. Kawai, „High-Accuracy Electromagnetic Field Simulation Using Numerical Human Body Models," in *IEEE Transactions on Magnetics*, vol. 52, no. 3, pp. 1-4, March 2016, Art no. 7402704, doi: 10.1109/TMAG.2015.2479467

[2] S. N. Makarov *et al.*, „Virtual Human Models for Electromagnetic Studies and Their Applications," in *IEEE Reviews in Biomedical Engineering*, vol. 10, pp. 95-121, 2017, doi: 10.1109/RBME.2017.2722420

[3] M. J. Ackerman, „The Visible Human Project: From Body to Bits," in *IEEE Pulse*, vol. 8, no. 4, pp. 39-41, July-Aug. 2017, doi: 10.1109/MPUL.2017.2701221.

[4] A. Christ i in., „The Virtual Family—development of surface-based anatomical models of two adults and two children for dosimetric simulations", Physics in Medicine & Biology, t. 55, nr 2, s. N23, grudz. 2009, doi: 10.1088/0031-9155/55/2/N01.

[5] N. Homsup, J. Breakall, „Application of XFDTD and FEKO program to the analysis of planar antennas", 2010 10th International Symposium on Communications and Information Technologies, Tokyo, Japan, 19–21 October 2010; pp. 646–650. https://doi.org/10.1109/ISCIT.2010.5665070.

[6] S. Wang, Z. Song, H. Li, G. Guo, i X. Xi, „Numerical simulation and analysis of effects of individual differences on the field distribution in the human brain with electromagnetic pulses", Scientific Reports, t. 11, nr 1, s. 16504, sie. 2021, doi: 10.1038/s41598-021-96059-3.

[7] Ł. Januszkiewicz, S. Hausman and P. D. Barba, „Human Body Modelling for Wireless Body Area Network Optimization," *2020 14th European Conference on Antennas and Propagation (EuCAP)*, Copenhagen, Denmark, 2020, pp. 1-5, doi: 10.23919/EuCAP48036.2020.9136084.

[8] https://itis.swiss/virtual-population/tissue-properties/database/

[9] https://www.remcom.com/resources/brochures/xgtd-overview

[10] Ł. Januszkiewicz, „Model for Ray-Based UTD Simulations of the Human Body Shadowing Effect in 5G Wireless Systems," International Journal of Antennas and Propagation, tom 2018, 2018

[11] Ł. Januszkiewicz, „Analysis of Human Body Shadowing Effect on Wireless Sensor Networks Operating in the 2.4 GHz Band", Sensors, t. 18, nr 10, 2018, doi: 10.3390/s18103412.

[12] J. Kawecki, Ł. Januszkiewicz, P. Di Barba: Model dwuwymiarowy do symulacji gęstości mocy w obszarze głowy i oczu dla częstotliwości 26 GHz. Artykuł w Przegląd Telekomunikacyjny - Wiadomości Telekomunikacyjne, , nr 4, str. 285-288. 2023. Tryb dostępu: https://sigma-not.pl/publikacja-144791-2023-4.html. DOI: 10.15199/59.2023.4.63.

[13] S. Dodge, N. Fontana, M. E. Mognaschi, E. Canicattì, i S. Barmada, „A Deep Learning Based Prediction of Specific Absorption Rate Hot-Spots Induced by Broadband Electromagnetic Devices", IET Science, Measurement & Technology, t. 19, nr 1, s. e70009, sty. 2025, doi: 10.1049/smt2.70009.

[14] P. Di Barba, Ł. Januszkiewicz, J. Kawecki, i M. E. Mognaschi, „Electromagnetic Wave Absorption in the Human Head: A Virtual Sensor Based on a Deep-Learning Model", Sensors, t. 23, nr 6, 2023, doi: 10.3390/s23063131.

[15] K. Trichias, A. Kaloxylos and C. Willcock, „6G Global Landscape: A Comparative Analysis of 6G Targets and Technological Trends," *2024 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, Antwerp, Belgium, 2024, pp. 1-6, doi: 10.1109/EuCNC/6GSummit60053.2024.10597064.

[16] W. Yamada *et al.*, „Pioneering New Frequency bands towards 6G Mobile Communication Systems," *2022 IEEE International Workshop on Electromagnetics: Applications and Student Innovation Competition (iWEM)*, Narashino, Japan, 2022, pp. 95-96, doi: 10.1109/iWEM52897.2022.9993485.

[17] T. Nagaoka, „Estimation of Electromagnetic Absorption Characteristics Using a Human Model with Anatomical Structures at Frequencies above 100 GHz," 2023 International Conference on Electromagnetics in Advanced Applications (ICEAA), Venice, Italy, 2023, pp. 461-461, doi: 10.1109/ICEAA57318.2023.10297688.

DOI: 10.15199/59.2025.4.4

KOMPRESJA WIZJI DLA MASZYN

VIDEO COMPRESSION FOR MACHINES

prof. dr hab. inż. Marek Domański, Politechnika Poznańska, Instytut Telekomunikacji Multimedialnej
(marek.domanski@put.poznan.pl)

STRESZCZENIE: Metody kompresji wizji są opracowywane tak, by przy zadanej małej bitowej szybkości transmisji dawało się w odbiornikach uzyskiwać możliwie dobrą jakość zdekodowanych obrazów. Dotychczas miarą tej jakości była uśredniona ocena grupy widzów. Jednak w ostatniej dekadzie coraz częstsze są zastosowania, w których obraz jest poddawany kompresji przed wykonaniem różnych operacji widzenia maszynowego. Powstaje więc pytanie, czy dotychczasowe metody kompresji zoptymalizowane pod względem subiektywnie ocenianej przez ludzi jakości zdekodowanych obrazów są także optymalne dla kompresji wizji przeznaczanej do wykorzystania w systemach widzenia maszynowego. Odpowiedź na to pytanie okazuje się być negatywna i dlatego w ostatnim dziesięcioleciu prowadzi się intensywne badania nad nowymi metodami kompresji odpowiednimi dla wizji używanej przez systemy widzenia maszynowego, czyli nad metodami kompresji wizji dla maszyn (*video compression/coding for machines*).
Praca skrótowo opisuje rezultaty badań dotyczących metod kompresji wizji dla maszyn. W szczególności podano najważniejsze cechy kodeka MPEG MVC, który został ujęty w projekcie części drugiej normy ISO/IEC MPEG-AI.

SŁOWA KLUCZOWE: kompresja wizji, widzenie maszynowe, MPEG, normy

ABSTRACT: Until recently, video coding standards were developed from the perspective of a human observer; that is, the main goal was to achieve a high subjective quality of the decoded video. This quality was measured by mean opinion score. However, over the past decade, the development of semi- and fully autonomous vehicles, AI-based video analysis, intelligent video surveillance, and video-based control in many new areas has led to a sharp increase in the amount of video data shared between computers. In these cases, direct human consumption of the decoded video is not the primary application. These factors have sparked research interest in video coding where the decoded data serves as input for machine vision tasks.
This paper briefly summarizes the results of research on video coding for machines. Specifically, it briefly describes the upcoming MPEG standard for machine video coding as part two of the ISO/IEC MPEG-AI standard.

KEYWORDS: video compression, machine vision, MPEG, standards

Wynalazek efektywnej kompresji wizji (obrazów ruchomych, ang. video) zrewolucjonizował technikę i umożliwił powstanie wielu nieznanych wcześniej usług i systemów, począwszy od telewizji cyfrowej poprzez cyfrowy zapis filmów, skomplikowane systemy dozoru wizyjnego i strumieniowanie wizji w internecie, aż po systemy wirtualnej rzeczywistości. Dotychczas opracowano wiele metod kompresji wizji: MPEG-2 [1,2], AVC (H.264) [3,4], HEVC (H.265) [5,6], VVC (H.266) [7,8] i inne. Wspomniane metody były opracowywane tak, by przy pewnej bitowej szybkości transmisji dawało się w odbiornikach uzyskiwać zdekodowaną wizję o możliwie dobrej jakości postrzeganej przez widzów. Miarą tej jakości była uśredniona ocena uzyskiwana od grupy widzów. Takie podejście do kompresji wizji odpowiadało potrzebom telewizji i innych usług wspomnianych wyżej.

W ostatniej dekadzie coraz częstsze są jednak zastosowania, w których obraz jest poddawany kompresji przed wykonaniem różnych operacji widzenia maszynowego. Powstaje więc pytanie, czy dotychczasowe metody kompresji zoptymalizowane pod względem subiektywnie ocenianej przez ludzi jakości zdekodowanej wizji są także optymalne dla kompresji wizji przeznaczanej do wykorzystania w systemach widzenia maszynowego. Odpowiedź na to pytanie okazuje się być negatywna i dlatego od kilkunastu lat prowadzi się badania nad nowymi metodami kompresji odpowiednimi dla wizji używanej przez systemy widzenia maszynowego, czyli nad metodami kompresji wizji dla maszyn (ang. *Video Compression/Coding for Machines* – VCM).

W tej pracy skrótowo zostaną omówione podstawowe zagadnienia związane z metodami kompresji wizji dla maszyn i ich normalizacją w Międzynarodowej Organizacji Normalizacyjnej (ISO).

PRZYKŁADY ZASTOSOWAŃ KOMPRESJI WIZJI DLA MASZYN

Grupa ekspertów MPEG (*Moving Pictures Experts Group*) działająca w ramach ISO/IEC JTC1/SC29 przygotowanie norm dotyczących kompresji wizji dla maszyn poprzedziła analizą potencjalnych zastosowań takiej kompresji [9]. Wstępna analiza potencjalnych rozwiązań doprowadziła do wniosku, że istnieją dwie podstawowe konfiguracje systemów związanych z kompresją wizji dla maszyn. W systemach pierwszej grupy (*rys. 1*) wizja jest poddawana kompresji, a w dekoderze odtwarza się wizję stanowiącą obiekt działania systemów widzenia maszynowego. Równolegle z analizą maszynową zdekodowanej wizji możliwe jest też jej oglądanie przez ludzi.

W systemach drugiej grupy już w koderze przeprowadza się analizę wizji prowadzącą do ekstrakcji cech, a właściwej kompresji poddaje się niekoniecznie całą wizję, ale często tylko wyznaczone cechy. Zdekodowane cechy są po stronie odbiorczej przekazywane do systemu widzenia maszynowego (***rys. 2***). Techniki odpowiadające drugiej grupie są na razie mniej rozwinięte i dlatego w dalszej części artykułu nasza uwaga będzie skupiona na systemach pierwszej grupy.

Przeprowadzona analiza [9] wykazała bardzo dużą liczbę bardzo różnorodnych potencjalnych zastosowań. Jako potencjalne obszary zastosowań wymieniono:

- systemy dozoru wizyjnego,
- inteligentne systemy transportowe, w tym komunikacja V2V (*vehicle to vehicle*) oraz I2V (*infrastructure to vehicle*),
- inteligentne miasta,
- inteligentny przemysł,
- inteligentne treści wizualne,
- elektronika powszechnego użytku (ang. *consumer electronics*).

Z wymienionymi przykładowymi obszarami powiązano typowe zadania widzenia maszynowego realizowane po stronie odbiorczej systemu, w tym między innymi:

- wykrywanie obiektów powiązane z określaniem klasy obiektu, identyfikatora obiektu i jego lokalizacji,
- segmentację obrazów,
- poprawę jakości obrazów,
- wyszukiwanie obrazów,
- wyszukiwanie i rozpoznawanie zdarzeń,
- wykrywanie anomalii,
- przetwarzanie wizji wielowidokowej.

Naszkicowana wyżej różnorodność zastosowań i zadań okazała się istotną trudnością w rozwoju technik kodowania wizji dla maszyn, w tym dla procesów normalizacji w tym zakresie.

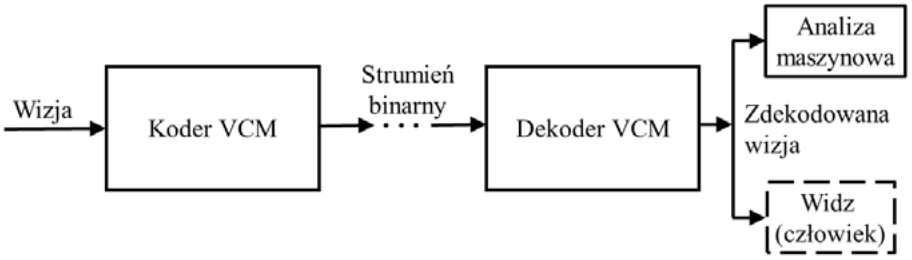
OCENA EFEKTYWNOŚCI METOD KOMPRESJI DLA MASZYN

Klasyczne metody kodowania wizji są oceniane przede wszystkim pod kątem efektywności kompresji rozumianej jako uzyskanie możliwie wysokiej jakości obrazu zdekodowanego ze strumienia binarnego o możliwie małej szybkości bitowej. Jakość mierzy się przede wszystkim subiektywną miarą wyznaczaną średnią ocen grupy obserwatorów (ang. *Mean Opinion Score* – MOS). Ze względu na uciążliwość i wysokie koszty takiej oceny jakości powszechnie stosuje się prostsze do wyznaczenia miary takie, jak PSNR (ang. *Peak Signal To Noise Ratio*) lub bardziej zaawansowane miary jak SSIM (*Structural Similarity Index Measure*) [10] oraz VMAF (*Video Multimethod Assessment Fusion*) [11].

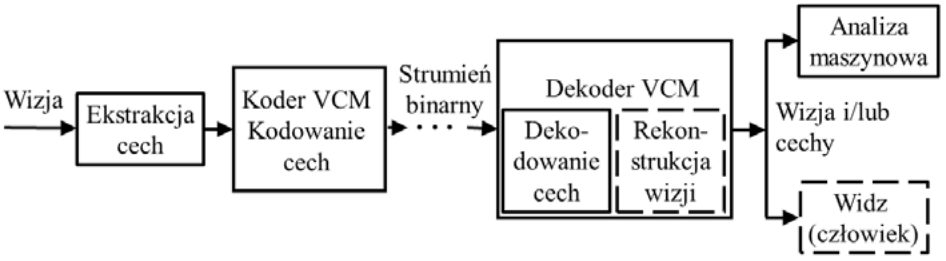
Nawet przy ograniczeniu analizy do układu z *rys. 1*, zdefiniowanie metod oceny efektywności kompresji dla maszyn stwarza istotne trudności wynikające ze zróżnicowania metod przetwarzania zdekodowanej wizji. W grupie ekspertów MPEG przeprowadzono długie dyskusje, które ostatecznie doprowadziły do kompromisowych ustaleń mówiących, że podczas wyboru metod stanowiących podstawę przyszłej normy badana będzie efektywność kompresji tylko dla detekcji obiektów, segmentacji obiektów oraz śledzenia obiektów [12].

Dla detekcji i segmentacji obiektów przyjęto uśrednioną przeciętną precyzję mAP (*Mean Average Precision*) [12].

Dla danej klasy obiektów liczby decyzji prawdziwie pozytywnych (ang. *true positive*) oraz fałszywie pozytywnych (ang. *false positive*) są określane w odniesieniu do wielkości części wspólnych obiektów IoU (ang. *Intersection over Union*), przekraczających wartość progową T_{IoU} określoną



» *Rys. 1. Schemat blokowy systemu, w którym zadania widzenia maszynowego są wykonywane na zdekodowanej wizji*
» *Fig. 1. Block scheme of a system using decoded video for machine vision tasks*



» *Rys. 2. Schemat blokowy systemu, w którym zadania widzenia maszynowego są wykonywane na zdekodowanych cechach*
» *Fig. 2. Block scheme of a system using decoded features for machine vision tasks*

dla tej klasy obiektów. Wspomniane wartości prawdziwe i fałszywe dotyczą wyjścia sieci neuronowej służącej do analizy wizji zdekodowanej w odbiorniku.

Wartość precyzji określa się jako:

$$precision(T_{IoU}) = \frac{TP(T_{IoU})}{TP(T_{IoU})+FP(T_{IoU})} . \quad (1)$$

Dla poszczególnych klas obiektów wyznacza się przeciętne wartości precyzji *AP* (ang. *Average Precision*). Uśredniona przeciętna precyzja *mAP* (ang. *Mean Average Precision*) jest wartością *AP* uśrednioną po wszystkich klasach obiektów w zakresie ustalonych wartości progowych dla IoU.

Dla pomiarów efektywności śledzenia obiektów przyjęto [12] dokładność śledzenia wielu obiektów *MOTA* (ang. Multiple Object Tracking Accuracy) [13].

$$MOTA = 1 - \frac{\sum_t (FN_t + FP_t + mme_t)}{\sum_t g_t} \quad (2)$$

gdzie FN_t, FP_t, mme_t and g_t są odpowiednio liczbami decyzji fałszywie negatywnych oraz fałszywie pozytywnych, a także liczbą błędów niedopasowania i rzeczywistą liczbą obiektów w chwili .

W grupie ekspertów MPEG podczas prac zmierzających do ustanowienia normy dla systemów zgodnych ze schematem z *rys. 1* przyjęto, że wartości *mAP* i *MOTA* będą wyznaczane dla ujednoliconych implementacji systemów widzenia maszynowego przetwarzających zdekodowaną wizję w sztucznych sieciach neuronowych opartych na strukturze sieci Detectron [14, 15].

Oczywiście wszystkie badania efektywności kompresji wymagały zdefiniowania zbiorów obrazów i sekwencji wizyjnych, co zostało przedstawione na przykład w dokumencie [12].

ROZWÓJ METOD KOMPRESJI WIZJI DLA MASZYN

W ostatnim dwudziestoleciu nastąpił szybki rozwój internetu rzeczy, analizy wizji wykorzystującej sztuczną inteligencję, inteligentnych systemów dozoru wizyjnego oraz inteligentnych systemów sterownia. Stymulowało to gwałtowny wzrost ilości danych, w tym danych wizyjnych, wymienianych pomiędzy komputerami. Wymienione zjawiska pobudziły rozwój badań dotyczących kompresji wizji w zastosowaniach, w których obserwacja zdekodowanej wizji przez człowieka nie jest już najważniejszym zadaniem. W ostatnim dwudziestoleciu zostało zaproponowanych wiele odpowiednich metod. Wśród pionierskich metod można wymienić te przedstawione na przykład w pracach [16-21].

W drugiej dekadzie XXI wieku powszechnie uznano znaczenie problemu kompresji wizji dla maszyn, a termin „kodowanie/kompresja wizji dla maszyn" wszedł do szerokiego stosowania. Ponieważ telekomunikacja wizualna zawsze wymaga normalizacji, szybko dostrzeżono tę potrzebę także dla kompresji wizji dla maszyn. Ta potrzeba została też dostrzeżona przez grupę ekspertów MPEG, która już w roku 2019 utworzyła „grupę ad hoc" o nazwie „Video Coding for Machines", której przewodniczącymi początkowo byli Yang Zhang (China Telecom) oraz Patric Dong (*Gerfalcon Technologies*) [22]. Zadaniem wspomnianego zespołu było przygotowanie materiałów i metod niezbędnych do dalszych prac, prowadzących do ustalenia techniki, która mogłaby zostać ujęta w przyszłej normie kompresji wizji dla maszyn. Przyjęto, że najpierw zostaną podjęte badania związane z systemami z *rys. 1*, a dopiero później prace dotyczące systemów z *rys. 2*.

W roku 2022 MPEG ogłosił wezwanie do zgłaszania propozycji technik kompresji wizji dla maszyn (*Call for Proposals on Video Coding for Machines*) [23, 24]. Wezwanie było ograniczone do propozycji systemów o strukturze pokazanej na *rys. 1*. W odpowiedzi na to wezwanie wpłynęło 14 propozycji pochodzących od różnych zespołów badawczych z całego świata, w tym od zespołu z Politechniki Poznańskiej kierowanego przez prof. Marka Domańskiego i dr. hab. Olgierda Stankiewicza. Nadesłane propozycje były dosyć zróżnicowane pod względem technicznym, jednak większość z nich wykorzystywała najbardziej obecnie zaawansowaną ogólną technikę kompresji wizji, czyli technikę VVC [7, 8]. Również propozycja Politechniki Poznańskiej wykorzystywała technikę VVC uzupełnioną o różne nowe narzędzia opracowane w Poznaniu.

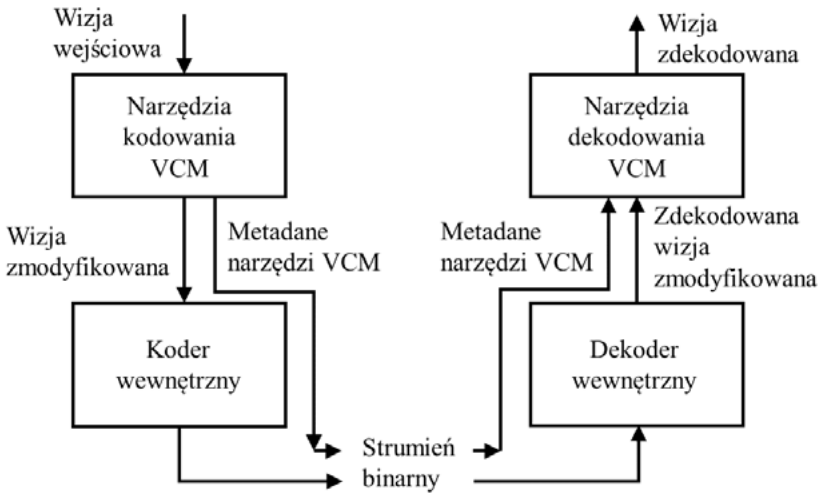
Ponieważ żadna ze zgłoszonych technik nie wyróżniła się zdecydowanie, postanowiono, że do eksperymentów będzie się używać oprogramowania modelowego zawierającego techniki zaproponowane przez różne ośrodki, w tym te pochodzące z Politechniki Poznańskiej. Oprogramowanie modelowe realizuje technikę, która w najbliższej przyszłości ma stanowić normę, co może nastąpić po pozytywnej ocenie w organizacji ISO.

Od pewnego czasu podgrupa zajmująca się przyszłą normą kompresji wizji dla maszyn jest kierowana przez Shan Liu (Tencent) oraz Chrisa Rosewarne (Canon). Koordynatorem oprogramowania VCM-RS (*Video Coding for Machines – Reference Software*), a więc zarządzającym konfiguracjami różnych technik jest Olgierd Stankiewicz (Politechnika Poznańska). Obecnie dla kompresji wizji dla maszyn (według schematu z *rys. 1*) gotowy jest projekt tekstu normy [25], który wkrótce ma się stać drugą częścią międzynarodowej normy MPEG-AI (ISO/IEC 23888). Ta technika bardzo skrótowo zostanie omówiona w następnej części pracy.

Prace badawcze mające na celu ustanowienie norm dla technik nadających się do kodowania wizji dla maszyn są podejmowane także poza grupą MPEG. Związany z grupą MPEG zespół JVET (*Joint Video Experts Team*) stanowiący wspólną grupę roboczą ISO i ITU (Międzynarodowy Związek Telekomunikacyjny) prowadzi prace mające na celu uwzględnienie kodowania wizji dla maszyn w przyszłej ogólnej normie kompresji wizji, która ma wejść w życie za kilka lat i oferować szybkości bitowe dwukrotnie mniejsze niż kodeki zgodne z obecnie najnowszą normą ISO/IEC/ITU, czyli VVC.

Także działająca w ramach ISO grupa robocza JPEG (ISO/IEC JTC1/SC29/WG1) prowadzi prace dotyczące kompresji obrazów statycznych dla maszyn.

Równolegle w grupie MPEG są prowadzone badania mające doprowadzić do opracowania techniki kompre-



» *Rys. 3. Schemat blokowy kodeka MPEG VCM*
» *Fig. 3. Block scheme of MPEG VCM codec*

sji działającej w systemach pokazanych na schemacie z *rys. 2*. W roku 2023 ogłoszono wezwanie do zgłaszania propozycji odpowiednich technik [26]. W odpowiedzi uzyskano pewną liczbę propozycji technicznych i obecnie trwa faza współpracy mająca zakończyć się zaproponowaniem odpowiedniej normy, dla której już opracowano wstępne szkice tekstu [27].

KODEK MPEG VCM

Jak już wspomniano, w grupie ekspertów MPEG przeprowadzono w latach 2022-25 badania mające na celu opracowanie techniki kompresji wizji dla maszyn. Dla systemów działających w strukturze z *rys. 1* ten etap został już zasadzie zakończony publikacją projektu normy [25], która będzie określać składnię i sposób dekodowania strumienia binarnego. Techniczny opis kodeka można znaleźć w dokumencie [28].

Architektura kodeka MPEG MVC (*rys. 3*) wykorzystuje „kodek wewnętrzny", który w zasadzie może być dowolnym kodekiem, nawet takim, który dopiero zostanie opracowany w przyszłości. W trakcie badań realizowanych przez grupę MPEG wykorzystywano kodek VVC [7, 8].

Zaznaczone na *rys. 3* narzędzia kodowania MVC modyfikują kodowaną wizję w taki sposób, by minimalizować szybkość bitową transmisji na wyjściu kodera wewnętrznego, ale by jednocześnie zapewnić dobrą efektywność zadań widzenia maszynowego wykonywanych z wykorzystaniem zdekodowanej wizji uzyskiwanej z narzędzi dekodera VCM.

Narzędzia kodera VCM obejmują między innymi:

- przepróbkowanie w dziedzinie czasu,
- przepróbkowanie przestrzenne i retargeting,
- upraszczanie tła,
- przetwarzanie w regionach zainteresowania,
- dekoloryzację.

W dekoderze przeprowadza się operacje analogiczne do operacji w koderze.

Na koniec można podkreślić bardzo aktywny udział grupy z Politechniki Poznańskiej w opracowaniu tego kodeka. Ten udział dotyczył przede wszystkim upraszczania tła [29], narzędzi przetwarzania w regionach zainteresowania [30] oraz koloryzacji [31].

PODSUMOWANIE

Przedstawiony przegląd wspiera tezę, że kompresja wizji dla maszyn stanowi bardzo ważny obszar badań. Należy się spodziewać, że wyniki badań będą w coraz większym stopniu wykorzystywane w rozmaitych usługach i w przemyśle. W ciągu następnych niewielu lat tematyka kompresji wizji dla maszyn prawdopodobnie będzie musiała znaleźć swoje miejsce w nauczaniu przyszłych inżynierów telekomunikacji, teleinformatyki i informatyki.

LITERATURA

[1] ISO/IEC 13818-1 | ITU-T Rec. H.262, Generic coding of moving pictures and associated audio information: Video.
[2] Mitchell J., Pennebaker W., Fogg Ch., LeGall D.: *MPEG video compression standard*, Kluwer, Dordrecht, 1996.
[3] ISO/IEC IS 14496-10 | ITU-T Rec. H.264, Advanced video coding.
[4] Richardson I.: *The H.264 advanced video compression standard*, 2nd ed., John Wiley & Sons, Chichester, 2010.
[5] ISO/IEC IS 23008-2 | ITU-T Rec. H.265, High efficiency video coding.
[6] Wien M.: *High Efficiency Video Coding – Coding Tools and Specification*, Springer, Berlin, Heidelberg, 2014.
[7] ISO/IEC IS 23090-3 / ITU-T Recommendation H.266, Versatile video coding.
[8] Saldanha M,, Sanchez G., Marcon C., Agostini L., *Versatile Video Coding (VVC)*, Springer, Berlin, Heidelberg, 2022.
[9] ISO/IEC JTC1/SC29/WG2, Doc. N161, Use cases and requirements for Video Coding for Machines, January 2022.
[10] Wang Zhou, Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., *Image quality assessment: from error visibility to structural similarity*, in IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, April 2004.
[11] Rassool, R., VMAF reproducibility: *Validating a perceptual practical video quality metric*, in Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB), Cagliari, Italy, Jun. 2017, pp. 1–2.
[12] ISO/IEC JTC1/SC29/WG2, Doc. N379, Common Test Conditions for Video Coding for Machines, Geneva, July 2023.
[13] Bernardin, K., Elbs, A., , Stiefelhagen, A., *Multiple Object Tracking Performance Metrics and Evaluation in a Smart Room Environment*, in Sixth IEEE International Workshop on Visual Surveillance in Conjunction with ECCV, May 2008.
[14] Wang, Z., Zheng, L. et al., *Towards real-time multi-object tracking*, in European Conf. Computer Vision (ECCV), 2020, pp. 107-122.
[15] Wu, Y., Kirillov, A., Massa, F., et al. *Detectron2*, https://github.com/facebookresearch/detectron2.
[16] Duan, L., Liu, J. , Yang, W., Huang, T. , Gao, W. , *Video Coding for Machines: A Paradigm of Collaborative Compression and Intelligent*

Analytics, in IEEE Transactions on Image Processing, vol. 29, pp. 8680-8695, 2020.
[17] Fischer, K., Brand, F., Herglotz, C., Kaup, A., *Video Coding for Machines with Feature-Based Rate-Distortion Optimization*, 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP), 2020.
[18] Galteri, L., Bertini, M., Seidenari, L., Del Bimbo, A., *Video compression for object detection algorithms*, in Proc. International Conference on Pattern Recognition (ICPR), Aug. 2018, pp. 3007–3012.
[19] Choi, H., Bajic, I. V., *High efficiency compression for object detection*, in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr. 2018, pp. 1792–1796.
[20] Chao, J., Steinbach, E., *Keypoint encoding for improved feature extraction from compressed video at low bitrates*, in IEEE Trans. Multimedia, vol. 18, no. 1, pp. 25–39, Jan. 2016.
[21] Ma, S., Zhang, X., Wang, S., Zhang, X., Jia, C., Wang, S., *Joint feature and texture coding: Toward smart video representation via frontend intelligence*, IEEE Trans. Circuits Syst. Video Technol., vol. 29, no. 10, pp. 3095–3105, Oct. 2019.
[22] ISO/IEC JTC1/SC29/WG11, Doc. N1850, Conclusions of 127th meeting, Gothenburg, July 2019.
[23] ISO/IEC JTC1/SC29/WG2, Doc. N191, Call for Proposals on Video Coding for Machines, April 2022.
[24] ISO/IEC JTC1/SC29/WG2, Doc. N193, Evaluation Framework for Video Coding for Machines, April 2022.
[25] ISO/IEC JTC1/SC29/WG4, Doc. N662, Text of ISO/IEC CD 23888-2 Video coding for machines, May 2025.
[26] ISO/IEC JTC1/SC29/WG2, Doc. N282, Call for Proposals on Feature Compression for Video Coding for Machines, April 2023.
[27] ISO/IEC JTC1/SC29/WG4, Doc. N669, PWD of feature coding for machines, May 2025.
[28] ISO/IEC JTC1/SC29/WG4, Doc. N664, Algorithm description of tools in VCM reference software, June 2025.
[29] Różek, S., Stankiewicz, O., Maćkowiak, S., Domański, M., *Video coding for machines using object analysis and standard video codecs*, in 2023 IEEE International Conference on Visual Communications and Image Processing (VCIP), Jeju, Republic of Korea, December 2023.
[30] Stankiewicz, O., Grajek, T., Maćkowiak, S., Stankowski, J., Różek, S., Lorkiewicz, M., Wawrzyniak, M., Domański, M., *Region-of-interest-based video coding for machines*, in 2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Niagara Falls, ON, Canada, 2024.
[31] Lorkiewicz, M., Różek, S., Stankiewicz, O., Grajek, T., Maćkowiak, S., Domański, M., *Video coding for machines with neural-network-based chroma synthesis*, in IEEE Access, Early access, 2025.

DOI: 10.15199/59.2025.4.5

GENERATYWNA SZTUCZNA INTELIGENCJA A STUDIOWANIE TELEINFORMATYKI

GENERATIVE ARTIFICIAL INTELLIGENCE AND STUDYING INFORMATION AND COMMUNICATION TECHNOLOGY

dr hab. inż. Artur Janicki, prof. PW, Zakład Cyberbezpieczeństwa, Instytut Telekomunikacji i Cyberbezpieczeństwa, Wydział Elektroniki i Technik Informacyjnych, Politechnika Warszawska
Artur.Janicki@pw.edu.pl

STRESZCZENIE: Niniejszy referat prezentuje wybrane korzyści i zagrożenia płynące z korzystania z metod generatywnej sztucznej inteligencji (GenAI) przez studentów na kierunkach teleinformatycznych i pokrewnych. Przedstawiono najczęstsze zastosowania metod GenAI przez studentów, a także zagrożenia płynące z nadużywania narzędzi GenAI podczas studiowania. Zaprezentowano także fragmenty wyników prac projektu ANANAS, którego celem jest ułatwienie analizy dokumentów pod kątem wykrywania tekstów generowanych przez GenAI.

SŁOWA KLUCZOWE: dydaktyka akademicka, generatywna sztuczna inteligencja, przetwarzanie języka naturalnego stylometria, wielkie modele językowe

ABSTRACT: This paper presents selected benefits and risks of using generative artificial intelligence (GenAI) methods by students of information and communication technology and related fields. It presents the most common applications of GenAI methods by students, as well as the risks of overusing GenAI tools during studies. It also presents fragments of the results of the ANANAS project, which aims to facilitate document analysis in terms of detecting texts generated by GenAI.

KEYWORDS: academic teaching, generative artificial intelligence, large language models, natural language processing, stylometry

Generatywna sztuczna inteligencja (ang. *Generative Artificial Intelligence*, GenAI) rozwija się w ogromnym tempie. Algorytmy i narzędzia stosujące GenAI penetrują coraz więcej obszarów naszego życia. W naturalny sposób także środowisko akademickie korzysta z metod GenAI, zarówno jeśli chodzi o naukowców i wykładowców, jak i o samych studentów. W tym referacie skupię się właśnie na tym drugim obszarze.

Metody GenAI należy uznać za narzędzia podwójnego zastosowania. Powszechnie znane są pozytywne aspekty korzystania ze sztucznej inteligencji, takie jak możliwość szybkiego przeprowadzania analiz, łatwego generowania treści tekstowych (przy użyciu wielkich modeli językowych: ang. *large language models,* LLM) oraz multimedialnych: dźwięku, obrazów czy filmów. Z drugiej strony metody GenAI niosą szereg zagrożeń. Należy do nich szerzenie dezinformacji, w tym np. zagrażających zdrowiu i bezpieczeństwu publicznemu, tworzenie deepfake'ów, przeprowadzanie celowanych ataków w inżynierii społecznej czy manipulacje polityczne [4, 5, 12]. Także środowisko akademickie jest narażone na zagrożenia, związane m.in. z plagiatami [3].

Niniejszy referat ma na celu pokazanie wybranych korzyści i zagrożeń płynących z korzystania z metod GenAI przez studentów na kierunkach teleinformatycznych i pokrewnych. W Rozdziale 2 omówię pokrótce najczęstsze zastosowania metod GenAI przez studentów. W Rozdziale 3 przejdę do zagrożeń, a w Rozdziale 4 zaprezentuję fragmenty wyników naszych prac, które dotyczą wykrywania tekstów tworzonych przez GenAI. Rozdział 5 podsumuje referat.

KORZYŚCI Z WYKORZYSTANIA GENERATYWNEGO AI PODCZAS STUDIOWANIA TELEINFORMATYKI

Korzyści z wykorzystania narzędzi GenAI są powszechnie znane, nic więc dziwnego, że chętnie z nich korzystają także studenci teleinformatyki. Algorytmy GenAI pomagają przy złożonym wyszukiwaniu informacji, wyjaśniają pojęcia i np. zasady działania różnych algorytmów czy protokołów wykorzystywanych w teleinformatyce. Pomagają przy edycji dokumentów: korygują przygotowane teksty, tłumaczą materiały z różnych języków. Wyszukują publikacje naukowe i podsumowują ich rezultaty. Streszczają duże dokumenty, generują wykresy i zestawienia. Świet-

nie się spisują w zadaniach programistycznych: można je wykorzystać do generowania kodu albo do wyszukiwania i naprawiania błędów w naszym kodzie. Tworzą zaawansowane grafiki oraz animacje, co można wykorzystać np. w prezentacji na seminarium.

Tabela 1 przedstawia subiektywny wybór narzędzi GenAI wraz z ich potencjalnym zastosowaniem podczas studiów na kierunku teleinformatyka lub pokrewnym. Zbiór ten nie jest oczywiście kompletny, co więcej, nowe narzędzia, a także kolejne wersje istniejących pojawiają się na rynku na bieżąco.

» *Tab. 1. Przykładowe narzędzia GenAI i ich potencjalne zastosowania w kontekście studiowania teleinformatyki (opracowanie własne na podstawie źródeł internetowych)*
» *Tabl. 1. Examples of GenAI tools and their potential applications in the context of studying ICT (own study based on Internet sources)*

Narzędzia	Zastosowanie
ChatGPT (OpenAI), Claude (Anthropic), DeepSeek, Gemini (Google), LLaMA (Meta)	Wyjaśnianie pojęć, tłumaczenie działania algorytmów, protokołów, objaśnianie wzorów, analiza danych, wyszukiwanie informacji, streszczanie dokumentów itd.
GitHub Copilot (Microsoft), AlphaCode (Google), Ghostwriter (Replit), Code- Whisperer (AWS)	Tworzenie, analiza i debugowanie kodu
Gemini 2.5 Deep Research, Scite, Elicit, SciSpace, Scholar GPT	Przeszukiwanie i analiza publikacji naukowych
DeepL, Google Translate, Azure AI Translator	Tłumaczenia tekstów
Grammarly, Writefull	Korekta językowa
PLLuM	Sprawdzanie polskich regulacji prawnych
DALL·E 3 (OpenAI), Midjourney, Stable Diffusion	Generowanie obrazów

ZAGROŻENIA PŁYNĄCE Z KORZYSTANIA Z GENERATYWNEGO AI PRZEZ STUDENTÓW TELEINFORMATYKI

Narzędzia GenAI, jak wiele innych, niosą ze sobą też różnego rodzaju zagrożenia, także gdy wykorzystuje się je podczas studiowania teleinformatyki. Po pierwsze, wiarygodność informacji zwracanych przez modele AI jest trudna do określenia. Modele LLM zazwyczaj są uczone na wielkich zbiorach pozyskanych z zasobów Internetu, które zawierają zarówno wiele wartościowych treści, jak i treści o niskiej jakości, a także informacje najzwyczajniej nieprawdziwe. Dlatego też informacje zwracane przez te modele są obarczone pewnym błędem, o czym od razu na wstępie ostrzegają ich dostawcy, tacy jak OpenAI czy autorzy modelu PLLuM. W niektórych sytuacjach modele GenAI również halucynują, czyli zwracają nieprawdziwe informacje, niemające pokrycia w rzeczywistości.

Po drugie, nadmierne korzystanie z narzędzi GenAI np. podczas zadań programistycznych, szczególnie na początku nauki, prowadzi do tego, że studenci nie mają okazji nabrać wprawy w posługiwaniu się językami programowania. Jest to sytuacja analogiczna do takiej, kiedy uczymy się języka obcego, ale zamiast próbować go używać w naturalnych sytuacjach, korzystamy z wirtualnego asystenta z funkcją tłumaczenia maszynowego. Cel krótkoterminowy osiągamy, bo udaje nam się porozumieć, natomiast cel długoterminowy – opanowanie języka obcego – oddala się.

Kolejnym zagrożeniem z obszaru studiów jest to, że potężne narzędzia takie jak algorytmy GenAI kuszą, by korzystać z nich do szybkiego generowania treści prac zaliczeniowych. Ułatwia to pozornie życie studentom, ale powoduje szereg poważnych implikacji:

- Studenci nie zdobywają umiejętności, które są wymagane przez program studiów danego kierunku, w tym wypadku teleinformatyki.
- Dydaktycy marnują czas, sprawdzając tekst wygenerowany przez GenAI, czyli oceniają pracę modelu LLM, a nie pracę studentów.
- Mimo że dydaktycy angażują swój czas, to i tak nie mają możliwości weryfikacji osiągnięcia efektów kształcenia u studentów, bo nie oni są twórcami pracy zaliczeniowej.

Jest to więc poważny problem, któremu poświęcę resztę referatu.

CZY PRACE GENEROWANE PRZEZ AI RÓŻNIĄ SIĘ OD TYCH PISANYCH PRZEZ STUDENTÓW?

Od kilku lat trwają prace, które mają na celu wykrywanie treści, w tym tekstowych, generowanych przez AI. Niektórzy wychodzą z założenia, że każda treść stworzona przez GenAI powinna zostać oznaczona przy pomocy znakowania wodnego. Inni badacze sugerują, że takie treści powinny być zarejestrowane w (zapewne gigantycznej) bazie danych, by umożliwić wyszukiwanie semantyczne podobnych tekstów w bazie danych z historycznymi wynikami modeli LLM. Część badaczy proponuje wykorzystanie innych modeli LLM (z dostrajaniem, ang. *fine-tuning* lub bez) do wykrywania generowanych tekstów [13]. Inna grupa prac proponuje ich detekcję na podstawie pewnych charakterystycznych cech. Na przykład badacze zauważyli, że teksty generowane przez modele LLM często cechuje brak różnorodności składniowej i leksykalnej [8], powtarzalność [10], a także brak spójności i jasnego celu [1].

Istnieją także narzędzia komercyjne, które w założeniu mają służyć do wykrywania generowanych tekstów. Należą do nich takie aplikacje jak GPTZero, ZeroGPT, Copyleaks AI Content Detector, Giant Language Model Test Room (GLTR) czy Winston AISaplingDetecting- AI.com. Niestety są to zwykle rozwiązania zamknięte, a także nieznana jest ich dokładność. W dodatku większość tych narzędzi jest przeznaczona tylko dla języka angielskiego.

Ponieważ brakuje zbliżonych narzędzi dla języka polskiego, na Politechnice Warszawskiej (PW) rozpoczęliśmy projekt, który ma na celu pomoc w analizie prac studenckich. Skupiamy się na krótszych formach, takich jak raporty z projektów czy sprawozdania z laboratoriów. Projekt nosi

nazwę ANANAS: „System do ANAlizy dokumeNtów wspomagający wykrywAnie tekStów wygenerowanych przez modele językowe". Jego celem jest opracowanie oraz wstępne wdrożenie systemu, który służyłby do analizy dokumentów tekstowych w celu ułatwienia wykrycia, czy ich treść została wygenerowana przy pomocy modeli językowych. W ramach tego projektu prowadzimy między innymi prace, które mają na celu znalezienie cech dystynktywnych, które ułatwią odróżnienie prac pisanych przez człowieka od tych wygenerowanych przez GenAI. Na początku prowadziliśmy badania na tekstach pochodzących z wiadomości email [6, 7], obecnie skupiamy się na pracach studenckich. Wkolejnych podrozdziałach omówię wybrane działania z tego obszaru.

Zgromadzony materiał tekstowy

W ramach projektu ANANAS zgromadziliśmy zbiór 3700 prac studenckich, które powstały na Politechnice Warszawskiej przed końcem 2022 r., kiedy to na rynku pojawiła się usługa ChatGPT, szybko zyskując popularność. Założyliśmy, że z dużym prawdopodobieństwem prace pochodzące sprzed tego czasu zostały napisane przez człowieka. Większość prac studenckich pochodziła z Wydziału Elektroniki i Technik Informacyjnych i została napisana przez osoby studiujące kierunek telekomunikacja, cyberbezpieczeństwo, informatyka i pokrewne.

Następnie dokumenty te zostały poddane przetwarzaniu wstępnemu, które polegało na anonimizacji tekstu, usunięciu tabel i rysunków, usunięciu tekstów niepolskich (np. angielskich, kodu maszynowego), usunięciu spisów treści i bibliografii, a także odrzuceniu dokumentów krótszych niż 1500 znaków, dla których analiza nie byłaby miarodajna. W ten sposób uzyskaliśmy zbiór 2400 prac napisanych (z dużym prawdopodobieństwem własnoręcznie) przez studentki i studentów PW.

Dokumenty te zostały podzielone na bloki po 500 tokenów, a następnie dla każdego bloku przy pomocy modelu GPT-4o utworzyliśmy kilkuzdaniowe streszczenia. Na podstawie tych streszczeń model LLM generował rozwinięcia o zbliżonej długości do oryginalnego tekstu. W ten sposób utworzyliśmy lustrzaną bazę dokumentów wygenerowanych przy pomocy GenAI. Zbiory te posłużyły do badań porównawczych i do eksperymentów z klasyfikacją.

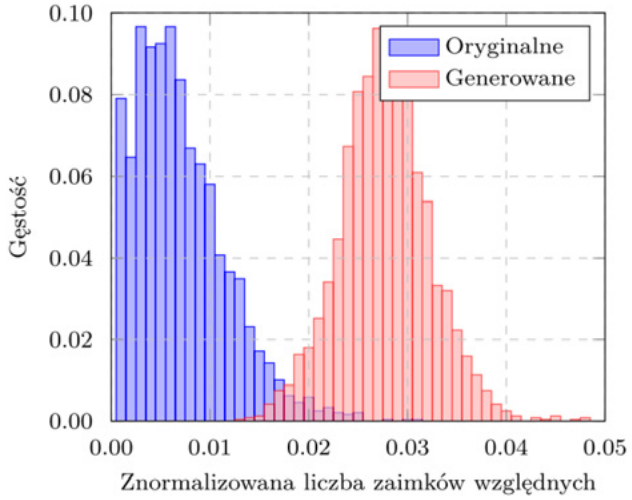
Ekstrakcja cech

Ponieważ zdecydowaliśmy się na stosowanie metody opartej na cechach wyekstrahowanych z tekstu, dla każdego dokumentu, oryginalnego lub wygenerowanego przez AI, wyliczyliśmy 220 cech. Ich skrócona lista wygląda następująco:

- Perpleksja (ang. *perplexity*), czyli nieokreśloność: to miara niepewności wartości kolejnego tokenu [9].
- Wybuchowość (ang. *burstiness*), która mierzy rozkład i wzorce występowania słów w tekście [2].
- Rozkład długości słów: średnia, odchylenie standardowe i wariancja długości słów w znakach.
- Rozkład długości zdań: średnia, odchylenie standardowe i wariancja długości zdań, mierzone zarówno w słowach, jak i znakach.
- Statystyki dokumentu: całkowita liczba zdań, słów i znaków.
- Błędy: 26 cech obejmujących liczbę błędów ortograficznych/ literowych, gramatycznych, błędów typograficznych (np. podwójne spacje, brak spacji po znaku interpunkcyjnym), znormalizowane względem długości tekstu mierzonej w znakach.
- Cechy stylometryczne: 172 cechy uzyskane za pomocą biblioteki StyloMetrix [11], wg następujących kategorii: formy gramatyczne, fleksja, składnia, interpunkcja, leksykalne, psycholingwistyczne, deskryptywne, graficzne.
- Cechy interpunkcyjne: liczba znaków interpunkcyjnych znormalizowana względem liczby zdań oraz względem liczby znaków (gęstość interpunkcyjna).
- Cechy związane z emocjami: łączna liczba emotikonów, znaków zapytania, wykrzykników oraz przypadków podwójnych znaków zapytania lub wykrzykników.

4.3 Porównanie wybranych cech

Jako jeden z celów projektu założyliśmy sobie znalezienie cech, które zapewnią najlepszą detekcję dokumentów generowanych przez GenAI. Dodatkowo próbujemy znaleźć cechy, które będą silnie skorelowane z podziałem na klasę dokumentów oryginalnych i klasę dokumentów generowanych przez modele LLM. W tym celu mierzyliśmy wartość informacji wzajemnej oraz współczynnik korelacji. Wybrane wyniki są zaprezentowane i opisane poniżej.



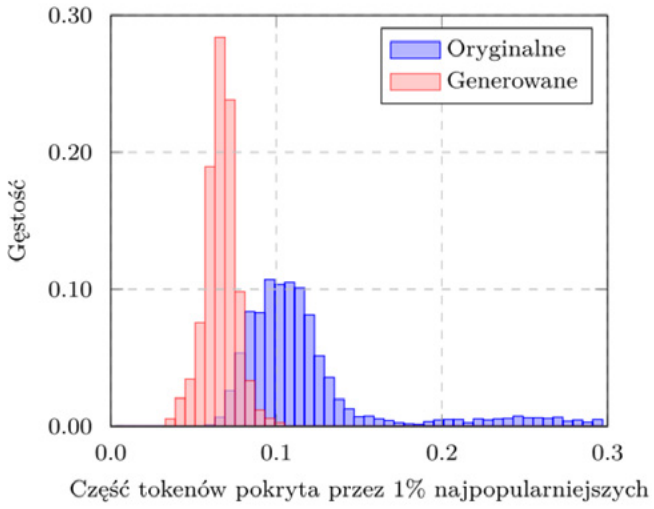
» *Rys. 1. Porównanie rozkładu liczby zaimków względnych dla dokumentów oryginalnych (tworzonych przez człowieka) i generowanych przez model LLM*
» *Fig. 1. Comparison of the distribution of the number of relative pronouns for original (human-written) and LLM-generated documents*

Okazało się, że cechą najbardziej dystynktywną jest liczba zaimków względnych (np. który, jaka, której) znormalizowana względem liczby tokenów.

Rysunek 1 pokazuje wyraźnie, że teksty wygenerowane przez modele LLM wykazują znacznie większą frekwencję tych części mowy. Na jednym z czołowych miejsc znalazła się też cecha mówiąca o bogactwie słownictwa używanego w badanych dokumentach (Rysunek 2). Okazało się, że w dokumentach generowanych udział 1% najczęstszych tokenów nie przekracza 10% całkowitej liczby tokenów

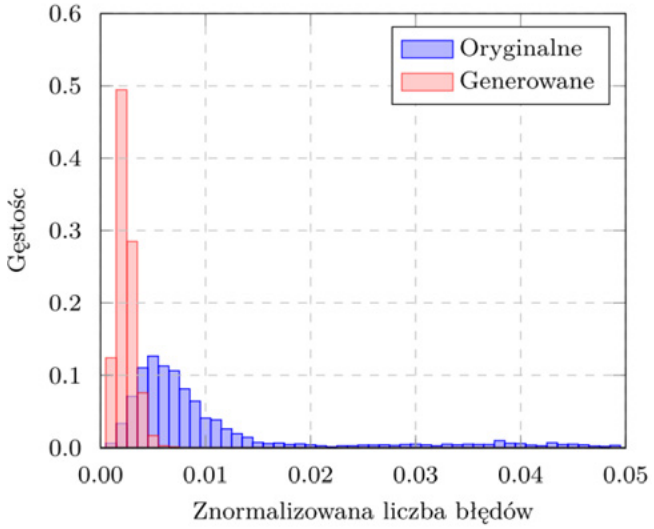
w tekście, podczas gdy dla dokumentów pisanych przez studentów udział ten wahał się od 5% do nawet ok. 30%.

Kolejną cechą dystynktywną okazała się cecha związana z liczbą błędów w dokumentach. Rysunek 3 pokazuje, że dokumenty generowane zawierają bardzo niski odsetek błędów: poniżej 0,005 błędu na znak. Wśród prac studenckich rozrzut tych wartości był znacznie większy: od prac bezbłędnych do 0,05 błędu na znak lub więcej.



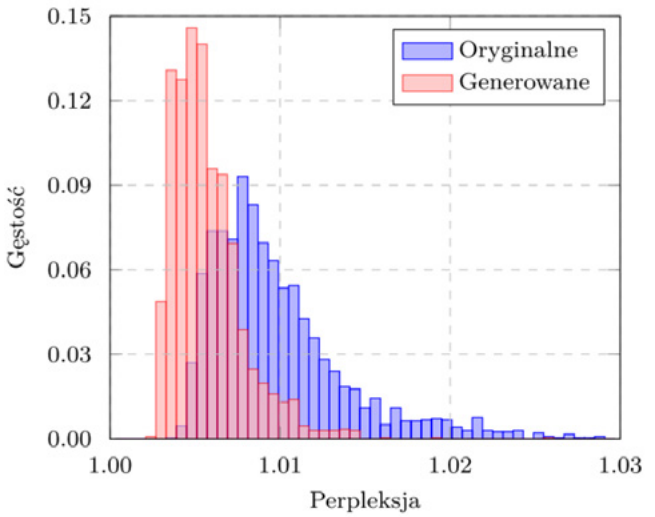
» *Rys. 2. Porównanie rozkładu udziału tokenów zawierających 1% najczęstszych słów dla dokumentów oryginalnych i generowanych*
» *Fig. 2. Comparison of the distribution of tokens containing the 1% most frequent words for original and generated documents*

Wyniki te należy interpretować w ten sposób, że niska stopa błędów w dokumencie nie niesie żadnych informacji o oryginalności dokumentu: może to być zarówno tekst zwrócony przez GenAI, jak i praca starannie napisana przez studentkę. Natomiast podwyższona stopa błędów może sugerować oryginalność tekstu, o ile oczywiście ktoś celowo nie powstawiał błędów do tekstu wygenerowanego przez GenAI.



» *Rys. 3. Porównanie rozkładu liczby błędów dla tekstów oryginalnych i generowanych*
» *Fig. 3. Comparison of the error count distribution for original and generated texts*

Cecha związana z perpleksją znalazła się na dalszym miejscu rankingu względem niesionej informacji, jednak może być ona również pomocna przy odróżnianiu tekstów oryginalnych od stworzonych przez GenAI. Rysunek 4 pokazuje, że znaczna część dokumentów oryginalnych i generowanych może mieć podobne wartości perpleksji. Jednak podwyższone wartości, tj. powyżej 1,015, mogą sugerować to, że dokument został napisany przez człowieka. Wysoka perpleksja oznacza w tym wypadku, że model LLM użyty do jej wyznaczenia (w tym wypadku GPT-2) był częściej „zaskakiwany" wartościami kolejnych tokenów.



» *Rys. 4. Porównanie rozkładu wartości perpleksji dla dokumentów oryginalnych (tworzonych przez człowieka) i generowanych przez generatywny model LLM*
» *Fig. 4. Comparison of the distribution of perplexity values for original documents (created by a human) and those generated by the generative LLM model*

Powyżej zostały umieszczone tylko wybrane cechy, których rozkłady wydały się nam najbardziej interesujące i obiecujące. Prace nad analizą wyekstrahowanych cech, także dla dokumentów wygenerowanych przez różne modele LLM, a także prace nad automatyczną detekcją dokumentów tworzonych przez GenAI nadal trwają. Wkrótce można się więc spodziewać kolejnych wyników.

PODSUMOWANIE

Metody GenAI mają ogromny potencjał i z pewnością mogą pomóc podczas studiowania teleinformatyki, a także oczywiście innych kierunków. Na pewno należy jednak korzystać z tych narzędzi z należytą uwagą, krytycznie odnosząc się do wyników zwracanych przez modele LLM.

Trzeba też pamiętać, że narzędzia te mogą być nadużywane, jeśli np. zostaną użyte do wygenerowania prac zaliczeniowych. Nie jesteśmy tutaj jednak bezbronni, gdyż opracowywane są narzędzia, takie jak system ANANAS, które mają na celu ułatwienie wykrywania dokumentów generowanych przez GenAI.

Narzędzia GenAI to także wyzwania dla dydaktyków: niektóre zadania trzeba przeformułować w ten sposób, by nie dało się ich rozwiązać poprzez wpisanie prostego promptu w modelu LLM i skopiowanie wyniku. Należy też

mieć nadzieję, że studentkom i studentom po prostu zależy na zdobyciu wiedzy i umiejętności przewidzianych w programie studiów teleinformatyki. Wtedy GenAI nie będzie drogą na skróty do zdobycia dyplomu, a stanie się kolejnym narzędziem w ręku inżyniera, o naprawdę imponujących możliwościach.

PODZIĘKOWANIA

Artykuł powstał w ramach projektu ANANAS – „System do ANAlizy dokumeNtów wspomagający wykrywAnie tekStów wygenerowanych przez modele językowe", finansowanego ze środków Politechniki Warszawskiej w ramach Programu Inicjatywa Doskonałości – Uczelnia Badawcza (IDUB). Bardzo dziękuję członkom zespołu ANANAS, z których efektów prac korzystałem podczas przygotowywania tego referatu: inż. Miłoszowi Kutyle, mgr. inż. Pawłowi Gryce, dr. hab. Kacprowi Gradoniowi i dr. inż. Markowi Kozłowskiemu.

LITERATURA

[1] Brown, Tom B. 2020. "Language models are few-shot learners." arXiv preprint arXiv:2005.14165.

[2] Chakraborty, Megha, S.M Towhidul Islam Tonmoy, S M Mehedi Zaman, Shreya Gautam, Tanay Kumar, Krish Sharma, Niyar Barman, et al. 2023. "Counter Turing Test (CT2): AI-Generated Text Detection is Not as Easy as You May Think - Introducing AI Detectability Index (ADI)." In Proc. Conference on Empirical Methods in Natural Language Processing, edited by Houda Bouamor, Juan Pino, and Kalika Bali, 2206–2239. Singapore: AACL. https://doi.org/10.18653/v1/2023.emnlp-main.136.

[3] Dien, Joseph. 2023. "Editorial: Generative artificial intelligence as a plagiarism problem." Biological Psychology 181: 108621. issn: 0301-0511. https://doi.org/https://doi.org/10.1016/j.biopsycho.2023.108621.

[4] Europol. 2023. "ChatGPT: the impact of large language models on law enforcement" (Luxembourg). https://doi.org/10.2813/255453.

[5] Gradoń, Kacper T. 2024. "Generative artificial intelligence and medical disinformation." British Medical Journal, no. 384. https://doi.org/10.1136/bmj.q579.

[6] Gryka, Paweł, Kacper Gradoń, Marek Kozłowski, Miłosz Kutyła, and Artur Janicki. 2024. "Detection of AI-Generated Emails – A Case Study." In Proc. 19th International Conference on Availability, Reliability and Security (ARES 2024), 1–8. 19 141. Vienna, Austria.

[7] Gryka, Paweł, Kacper Gradoń, Marek Kozłowski, Miłosz Kutyła, and Artur Janicki. 2024. "Impact of Spelling and Editing Correctness on Detection of LLM-Generated Emails." In Proc. 19th Conference on Computer Science and Intelligence Systems (FedCSIS 2024), 603–608. Belgrade, Serbia: IEEE.

[8] Holtzman, Ari, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. "The Curious Case of Neural Text Degeneration." In Proc. International Conference on Learning Representations (ICLR 2020).

[9] Hugging Face. 2025. Perplexity of fixed-length models. Https://huggingface.co/docs/transformers/perplexity, Accessed on March 12, 2025. https:// huggingface. co/docs/transformers/perplexity.

[10] Jiang, Shaojie, Thomas Wolf, Christof Monz, and Maarten de Rijke. 2020. "TLDR: token loss dynamic re-weighting for reducing repetitive utterance ge/neration." arXiv preprint arXiv:2003.11963.

[11] Okulska, Inez, Daria Stetsenko, Anna Kołos, Agnieszka Karlińska, Kinga Głąbińska, and Adam Nowakowski. 2023. "StyloMetrix: An Open-Source Multilingual Tool for Representing Stylometric Vectors." arXiv preprint arXiv:2309.12810.

[12] Williams, HJ, and McCulloch, C. 2023. Truth Decay and National Security: Intersections, Insights, and Questions for Future Research. Santa Monica, CA, USA. https://www.rand.org/pubs/perspectives/PEA112- 2.html.

[13] Yang, Xianjun, Liangming Pan, Xuandong Zhao, Haifeng Chen, Linda Petzold, William Yang Wang, and Wei Cheng. 2023. "A survey on detection of LLMsgenerated content." arXiv preprint arXiv:2310.15654.

DOI: 10.15199/59.2025.4.6

„SECURITY-BY-DESIGN" - BEZPIECZEŃSTWO SYSTEMOWE ICT - W PRZEPISACH, NORMACH I PRAKTYCE

"SECURITY-BY-DESIGN" – SYSTEMIC ICT SECURITY – IN REGULATIONS, STANDARDS, AND PRACTICE

mgr inż. Paweł K. Kostkiewicz, NASK – Państwowy Instytut Badawczy
pawel.kostkiewicz@nask.pl

STRESZCZENIE: W niniejszym artykule przedstawiono koncepcję „Security-by-design" jako podstawową zasadę planowania, projektowania, testowania, zamawiania i oceny zgodności rozwiązań ICT. Rozwiązania te obejmują produkty cyfrowe – zarówno sprzętowe, jak i programowe – a także systemy cyfrowe, usługi ICT i procesy informatyczne. Kluczowe tematy obejmują modelowanie zagrożeń, specyfikację wymagań bezpieczeństwa, założenia projektowe, zapewnienie jakości produktu końcowego oraz pełny cykl życia produktu – od wstępnego projektu do końca eksploatacji. Artykuł przedstawia źródła powstania koncepcji, aktualny stan wiedzy oraz założenia do wdrożenia w oparciu o systemowe podejścia w oparciu o ramy normatywne, takie jak: ISO/IEC/IEEE 15288 Systems and software engineering — System life cycle processes, ISO/IEC 27034 Information technology - Security techniques - Application security, ISO/IEC 15408 Common Criteria for Information Technology Security Evaluation, and ETSI EN 303 645 Cyber Security for Consumer IoT: Baseline Requirements.
SŁOWA KLUCZOWE: security-by-design, cyberbezpieczeństwo, normy i standardy, modelowanie zagrożeń, cykl życia oprogramowania, zarządzanie podatnościami, ocena zgodności, kultura bezpieczeństwa

ABSTRACT: This article presents the concept of "Security-by-design" as a fundamental principle in the planning, design, testing, procurement, and conformity assessment of ICT solutions. These solutions include digital products—both hardware and software—as well as digital systems, ICT services, and IT processes.
Key topics covered include threat modeling, security requirements specification, design assumptions, end-product quality assurance, and the full product lifecycle—from initial design to end-of-life.
The article presents the origins of the concept, the current state of knowledge, and an outline of the concept of a systematic approach to the secure design and verification of ICT solutions based on normative framework and standards such as: ISO/IEC/IEEE 15288 Systems and software engineering — System life cycle processes, ISO/IEC 27034 Information technology - Security techniques - Application security, ISO/IEC 15408 Common Criteria for Information Technology Security Evaluation, and ETSI EN 303 645 Cyber Security for Consumer IoT: Baseline Requirements.
.KEYWORDS: Security-by-design, cybersecurity, standards and regulations, threat modelling, software development lifecycle, vulnerability management, conformity assessment.

CZYM JEST SECURITY-BY-DESIGN?

Pojęcie abstrakcyjne, specyfikacja techniczna, norma międzynarodowa, zestaw dobrych praktyk, hasło marketingowe a może postulat polityczny? Wszystko jednocześnie, a nawet… więcej.

Termin „Security-by-design" ma znaczenie nie tylko w dziedzinach nauk technicznych, gdzie od dawna występuje w zbliżonych formach, ale także w obszarze polityki i prawodawstwa w różnych krajach i regionach. Warto zacząć od próby jak najdokładniejszego zdefiniowania tej koncepcji, aby umożliwić dalszą analizę jej zastosowań.

Zacznijmy od wyidealizowanej, ambitnej definicji: Security-by-design – to tworzenie oprogramowania i sprzętu IT, infrastruktury sieciowej i usług cyfrowych, które są bezpieczne od podstaw [INVI]. Jednocześnie „Security-by-design" to podejście, które ma na celu zwiększenie cyberodporności państwa i instytucji publicznych oraz poprawę zaufania między organizacjami. Oba podejścia można postrzegać - łącznie - jako synergiczny, samonapędzający się mechanizm, obejmujący podaż (perspektywa „wytwórcy/dostawcy") i popyt (perspektywa „użytkownika/nabywcy") . W idealnym scenariuszu siły te – choć napędzane przeciwstawnymi motywacjami – przyczyniałyby się do optymalizacji bezpieczeństwa produktów i usług ICT poprzez funkcjonowanie w ramach wspólnie uzgodnionych standardów, wdrożonych i akceptowanych przez wszystkich interesariuszy.. Rzeczywistość jest jednak bardziej gorzka.

Spróbujmy nakreślić tło, prekursorów i ewolucję koncepcji od jej wczesnych początków:

» *Tab. 1 Ewolucja koncepcji „Security-by-desing" od lat 70. XX w. do dzisiaj*
» *Tabl. 1 Evolution of the "Security-by-design" concept from the 1970s to today*

Rok	Publikacja	Wpływ na koncepcję „bezpieczeństwa w fazie projektowania"
Lata 1970te	Raport Ware'a. Studium Andersona.	Raport „Kontrola bezpieczeństwa systemów komputerowych" sporządzony przez grupę zadaniową Defense Science Board. Opracowanie „Computer Security Technology Planning Study".
Lata 1980te	Kryteria oceny zaufanych systemów komputerowych (Orange Book lub TCSEC).	Pierwsze formalne rozważania na temat bezpieczeństwa systemów komputerowych w USA. Wprowadzenie pojęć takich jak „trusted computing base" czy „assurance levels". Podstawa dla późniejszych koncepcji bezpiecznego projektowania systemów komputerowych.
Lata 1990.	ISO 7498-2 i ITU-T X.800 BS 7799 (UK) ISO/IEC 17799 ISO 15408 (Common Criteria)	Norma ISO 7498-2 (1989), jednocześnie zalecenie ITU-T X.800 (1991), które definiuje ogólną architekturę związaną z bezpieczeństwem. Brytyjski standard (BS7799) zarządzania bezpieczeństwem informacji, a następnie (formalnie przyjęty w 2000 r.) międzynarodowy standard (ISO 17799) - prekursor ISO/IEC 27001. W latach 1996-1999 prowadzono prace nad opracowaniem i weryfikacją użyteczności standardu Common Criteria, a także międzynarodowego standardu ISO15408, który m.in. wprowadza pojęcie pakietu uzasadnienia zaufania EAL4 oznaczającego przedmiot oceny „metodycznie zaprojektowany, przetestowany i zweryfikowany".
2002 r.	Microsoft Security Development Lifecycle (SDL) ISO/IEC/IEEE 15288 Procesy cyklu życia systemu.	Formalizacja podejścia zorientowanego na bezpieczeństwo w cyklu rozwoju (życia) oprogramowania. Od tego momentu udokumentowane podejście do bezpiecznego tworzenia kodu zaczęło pojawiać się wśród producentów oprogramowania i sprzętu. Również rok, w którym opublikowano międzynarodowy standard ISO/IEC/IEEE 15288 System Life Cycle Processes.
2005 r.	NIST SP 800-53	Pierwsze wydanie NIST SP 800-53 Recommended Security Controls for Federal Information Systems (Zalecane zabezpieczenia dla federalnych systemów informatycznych)
2011 r.	ISO 27034	Publikacja międzynarodowej standardu ISO/IEC 27034 Information technology - Security normy techniques - Application security, która kładzie podwaliny pod systematyczne podejście do bezpieczeństwa aplikacji i, szerzej, systemów teleinformatycznych.
2018 r.	SAFECode	Organizacja non-profit SAFECode wydała Fundamental Praktyki bezpiecznego tworzenia oprogramowania.
2019 r.	Książka ***Secure by Design*** (Johnsson, Deogun, Sawano).	Popularna publikacja promująca praktyczne podejście do projektowania bezpiecznego kodu. Termin „Secure by Design" staje się powszechnie znany poza wąskim kręgiem specjalistów.
2019–2023	Unijny akt o cyberbezpieczeństwie (2019). Unijny akt o cyberodporności (2022). (UK) Product Security and Telecommunications Infrastructure regulation (2023) (USA) Secure-by-design pledge (2023)	W ostatnich latach obserwujemy ustanawianie regulacji w różnych jurysdykcjach – w tym w Stanach Zjednoczonych, Wielkiej Brytanii oraz Unii Europejskiej – które mają na celu promowanie (USA) lub wręcz wymuszanie (Wielka Brytania, UE) stosowania koncepcji Security-by-Design przez kluczowych uczestników rynku, takich jak producenci, importerzy, integratorzy, nabywcy oraz jednostki oceny zgodności.
2023 r.	„Principles and approaches for Secure by Design software"	Wspólny dokument z zachętą do współdziałania sygnowany przez agencje ds. cyberbezpieczeństwa z Europy, Ameryki i Azji: CISA, NSA, FBI, ACSC, CCCS, CERT NZ, NCSC-NZ, NCSC-UK, BSI, NCSC-NL, NCSC-NO, NÚKIB, INCD, KISA, NISC-JP, JPCERT/CC, CSA, CSIRTAMERICAS

Zarysowana w *Tabeli 1* perspektywa wydaje się optymistyczna. Od ponad pięćdziesięciu lat naukowe i przemysłowe zainteresowanie bezpieczeństwem systemów teleinformatycznych stale rośnie. Liczba inicjatyw dotycząca produkcji bezpiecznego sprzętu i oprogramowania jest prawdopodobnie wystarczająca, aby zaspokoić zapotrzebowanie rynku. Innymi słowy, osoby odpowiedzialne za zamówienia, wdrażanie lub rozwój produktów cyfrowych i kodu źródłowego mogą znaleźć odpowiednie narzędzia formalne i techniczne w celu zapewnienia bezpieczeństwa od podstaw. Miniony okres można podzielić na następujące główne fazy:

- Faza I - Podstawy teoretyczne (wczesne lata 50. do późnych lat 70.)

W tym okresie dominowały badania naukowe i koncepcyjne, często o charakterze futurystycznym i abstrakcyjnym. - które stanowiły intelektualne fundamenty dla późniejszych działań w zakresie bezpieczeństwa. Symbolicznym zwieńczeniem tej fazy jest raport Security Controls for Computer Systems, który po raz pierwszy przedstawił bezpieczeństwo komputerowe jako odrębną, złożoną i wielowarstwową dziedzinę.

- Faza II – Eksperymentalny i komercyjny rozwój systemów (lata 80.)

Lata 80. przyniosły dynamiczny rozwój systemów informatycznych zarówno dla sektora prywatnego, jak i przedsiębiorstw. Pojawiły się pierwsze próby standaryzacji na poziomie branżowym i regionalnym. Kwestie bezpieczeństwa w sektorach telekomunikacji i IT pozostawały jednak na dalszym planie, ustępując miejsca rozwojowi funkcjonalnemu i ekspansji rynkowej.

- Faza III – Strukturyzacja i segmentacja (lata 90.)

W tej dekadzie nastąpił gwałtowny rozwój produktów konsumenckich oraz integracja systemów informatycznych różnych dostawców, które coraz częściej współdziałały za pośrednictwem sieci telekomunikacyjnych. Korporacje ICT konkurowały o dominację w wybranych segmentach rynku, kształtując własne standardy, co jednocześnie przyczyniło się do intensyfikacji regionalnych i globalnych wysiłków standaryzacyjnych. W tym okresie opublikowano kluczowe dokumenty normatywne, takie jak Common Criteria, BS7799 czy zalecenia ITU-T X.8xx. Choć termin security-by-design nie był jeszcze powszechnie stosowany, coraz większą popularność zyskiwały pojęcia takie jak „zasady bezpieczeństwa" czy „architektura bezpieczeństwa systemów i sieci", głównie w środowiskach eksperckich.

- Faza IV – Szeroka dostępność i rosnąca świadomość (2000–2019)

W tym okresie wykształcił się odrębny segment rynku ICT skoncentrowany na bezpieczeństwie. Termin cyberbezpieczeństwo zyskał na znaczeniu, rozwijały się międzynarodowe dobrowolne porozumienia międzynarodowe (np. CCRA, Global Platform), które mają na celu harmonizację zasad bezpieczeństwa w kontekście globalnie dostępnych produktów i systemów połączonych przez Internet. Terminy secure-by-design oraz secure-by-default zaczęły pojawiać się w szerszym dyskursie, choć najczęściej w ujęciu deklaratywnym.

- Faza V – Powszechne wdrożenie i akceptacja (2020 – obecnie)

Pandemia COVID-19 zainicjowała istotną zmianę w postrzeganiu środowiska cyfrowego – z narzędzia wspierającego przekształciło się ono w krytyczny fundament funkcjonowania gospodarki. W odpowiedzi na rosnące zagrożenia, na całym świecie zaczęto wdrażać środki regulacyjne mające na celu wymuszenie stosowania zasad bezpieczeństwa przez wszystkich uczestników rynku. Termin security-by-design został wprowadzony do ram prawnych, co znacząco zwiększyło jego rozpoznawalność. Jednocześnie istnieje ryzyko, że stanie się on jedynie modnym hasłem, pozbawionym rzeczywistej treści. Agencje takie jak CISA odegrały kluczową rolę w popularyzacji tego podejścia, nadając mu globalny rozgłos.[LWRF].

Niniejszy artykuł stanowi o próbie wzmocnienia koncepcji, która – choć zakotwiczona w solidnych podstawach teoretycznych i zweryfikowana w praktyce rynkowej – obecnie, mimo uznania jej przez ekspertów za kluczowy element współczesnych strategii bezpieczeństwa, stoi w obliczu ryzyka trywializacji i redukcji do pustego sloganu. Celem autora jest zatem nie tylko przypomnienie o jej rzeczywistym znaczeniu, ale również wskazanie na konieczność jej świadomego i odpowiedzialnego wdrażania w kontekście dynamicznie zmieniającego się krajobrazu technologicznego i rosnącej niepewności geopolitycznej.

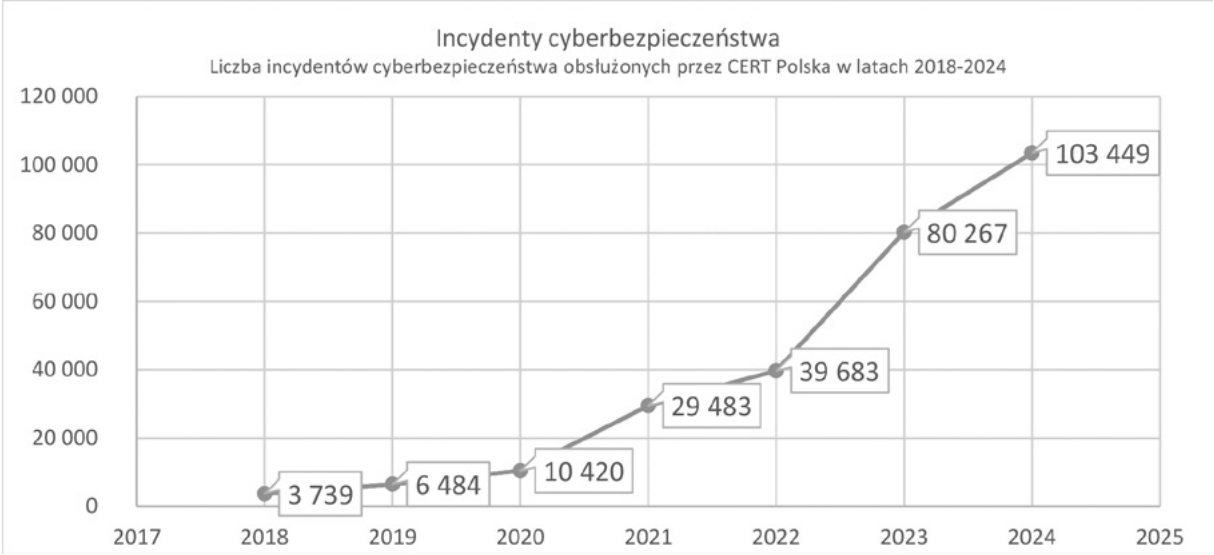
(DLACZEGO) POTRZEBUJEMY „SECRUITY BY DESIGN"?

Czytając w rekomendacji ITU-T X.800 z 1991 r. stwierdzenie, że funkcje bezpieczeństwa zwykle zwiększają koszt systemu i mogą utrudniać jego użytkowanie [X.800] - można być zaskoczonym, jak zdecydowanie pozostaje ono aktualne. Vide szeroko promowane i używane uwierzytelnianie wieloskładnikowe (MFA), którego znaczenia nikt obecnie nie kwestionuje, znajdując jednocześnie niewielu zwolenników wśród użytkowników.

Przed zaprojektowaniem bezpiecznego systemu należy zidentyfikować konkretne zagrożenia (ang. threat), przed którymi wymagana jest ochrona (zabezpieczenie). Proces ten znany jest jako „modelowanie zagrożeń" (ang. threat modeling). Nacisk w modelowaniu powinien być położony na zapewnienie, że wszystkie istotne zagrożenia zostały wzięte pod uwagę i będą mogły zostać odpowiednio odzwierciedlone w specyfikacji bezpieczeństwa systemu. [ANDRE] System jest podatny na wiele sposobów, ale tylko niektóre z nich można wykorzystać, ponieważ atakujący nie ma wystarczających możliwości lub ponieważ spodziewany rezultat nie uzasadni wysiłku i ryzyka wykrycia. [X.800] Po przeprowadzeniu modelowania możliwe jest rozpoczęcie opracowywania projektu bezpieczeństwa [...] systemu [ANDRE]. Powyższe zostało zapisane w 1972 roku [ANDRE], powtórzone w 1991 roku [X.800] i pozostaje tak samo aktualne jak wtedy, ale niestety - rzadko stosowane jest w praktyce. Co gorsza, narzędzia dostępne dla atakujących w porównaniu z rokiem 1972, czy 1991 są nieporównywalnie skuteczniejsze i łatwiejsze w użyciu. Cyberprzestępcy nieustannie szukają nowych sposobów atakowania, łamiąc zabezpieczenia za pomocą technik jailbreakingu i narzędzia „dark AI" w celu wzmocnienia skuteczności ataków [KELA]. Sztuczna inteligencja przekształca krajobraz cyberzagrożeń, rozszerzając pole rażenia (ang. ***attack Surface***), zwiększając siłę nacisku i stale doskonaląc możliwości [Paul Chichester, NCSC]. Sztuczna inteligencja prawie na pewno zwiększy ilość i wpływ cyberataków w ciągu najbliższych [...] lat. Jednak realny wpływ na cyberbezpieczeństwo będzie różny. Rzeczywistość, z którą mamy obecnie do czynienia, jest taka, że skala incydentów rośnie wraz z każdą kolejną publikacją zespołów reagowania (CERT, CSIRT). Przykładowo, CERT Polska (jeden z trzech krajowych zespołów CSiRT odpowiedzialnych za reagowanie na incydenty na rynku konsumenckim, edukacyjnym, samorządowym itp.) odnotowuje kolejny rok z rosnącą liczbą incydentów (patrz wykres poniżej).

Na potrzeby tego artykułu oszacowano liczbę incydentów potencjalnie związanych z lukami (podatnościami) w oprogramowaniu lub sprzęcie. Chociaż odsetek takich incydentów w stosunku do całkowitej liczby obsługiwanych spraw pozostaje stosunkowo niski, jednak liczba bezwzględna - 2336 incydentów - jest jednak znacząca.

Wydaje się, że było wystarczająco dużo czasu i motywacji, aby urzeczywistnić wymagania zaleceń ITU-T z lat dziewięćdziesiątych. Technologia rozwinęła się w tym okresie znacznie, co z pewnością miało istotny wpływ na osiągnięcie celu. Paradygmat ponownego wykorzystania kodu, który stał się rodzajem pragmatycznego dogmatu, dał znaczący impuls do lawinowego wzrostu liczby aplikacji. Duża dostępność rozwiązań społecznościowych (open-source) i narzędzi do automatyzacji tworzenia kodu to kolejne czynniki, które doprowadziły do gwałtownego rozwoju. Otwarte



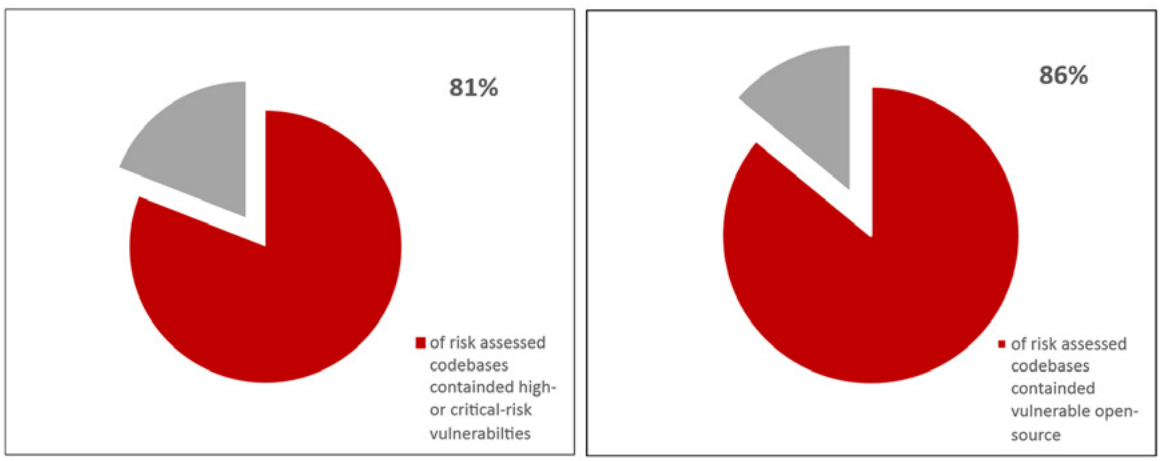
» *Rys. 1. Liczba incydentów cyberbezpieczeństwa obsłużonych przez CERT Polska (2018-2024)*
» *Fig. 1. Number of cybersecurity incidents handled by CERT Polska (2018-2024).*

» *Tab. 2 Rozkład typów incydentów cyberbezpieczeństwa (2024, CERT Polska)*
» *Tabl. 2 Cybersecurity incident type distribution (2024, CERT Polska)*

Typ incydentu cyberbezpieczeństwa	Wszyscy	Proc	Powiązane
Oszustwa komputerowe	97 995	94,73%	
Szkodliwe oprogramowanie	1891	1,83%	
Podatne usługi	1634	1,58%	**1634**
Obraźliwe i nielegalne treści	775	0,75%	
Włamania	447	0,43%	**447**
Dostępność zasobów	426	0,41%	
Próby włamań	179	0,17%	**179**
Atak na bezpieczeństwo informacji	62	0,06%	**62**
Gromadzenie informacji	26	0,03%	
Inne	14	0,01%	**14**
Razem	103 449		**2336**
		Powiązane (%)	2,3%

oprogramowanie zasila nieomal każdy ekosystem nowoczesnego oprogramowania - w tym znaczną część gospodarki cyfrowej [GHB].

Raport dotyczący oceny ryzyka dla poszczególnych sektorów rynku [BLDCK] z perspektywy podatności na zagrożenia, wskazuje, że wysokim poziom ryzyka charakteryzuje się rynek aplikacji mobilnych (100% przeskanowanych repozytoriów kodu z tego sektora zawierało podatności wysokiego ryzyka), sektor sprzętu komputerowego i półprzewodników (87%) oraz oprogramowania dla sektory edukacyjnego (EdTech) i oprogramowania dla przedsiębiorstw/SaaS (86%). Nie powinno dziwić (ale cieszy!), że najniższy odsetek podatności wysokiego ryzyka tj. 60% zawierają kody źródłowe w branży energetycznej/czystych technologii (spośród wszystkich 16 sektorów). Ogólnie rzecz biorąc, ponad 50% baz kodu każdej z branż zawierało podatności wysokiego ryzyka, co jest, można by rzec, naocznym zaprzeczeniem zasad Security-by-design i oczekiwań dotyczących udostępniania na rynku oprogramowania bez znanych podatności.

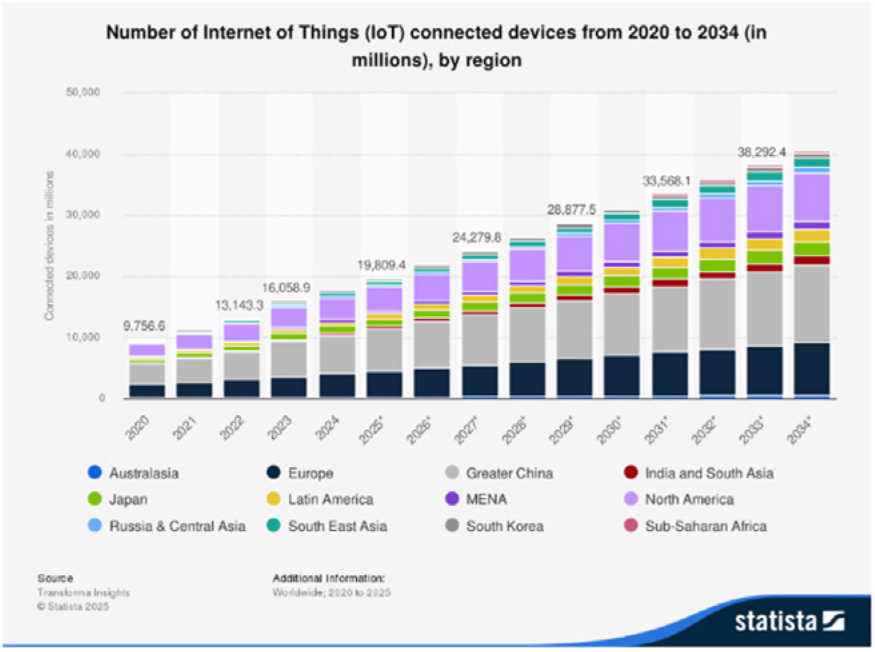


» *Rys. 2. Dane z raportu BlackDuck na temat podatnych na ataki repozytoriów kodu open-source*
» *Fig. 2. BlackDuck report data on open-source code databases vulnerablility.*

Analiza repozytoriów kodu źródłowego *(codebases)* – wyimek raportu BlackDuck:

- 1 658 projektów przeskanowanych przez audyty
- 97% repozytoriów zawierało oprogramowanie open source
- 70% repozytoriów ma swoje źródło w oprogramowaniu open source.

Rynek ICT w ciągu ostatniego półwiecza rozwijał się nieprzerwanie, dynamicznie i nadal rośnie. Przekłada się to nie tylko na rosnącą liczbę produktów cyfrowych, systemów i usług IT, wytwarzanych, udostępnianych, sprzedawanych i wdrażanych, ale także na rosnącą liczbę starzejących się, ale wciąż aktywnych (często - podłączonych do Internetu) urządzeń i oprogramowania.



» *Rys. 3. Liczba i prognoza wzrostu urządzeń IoT na świecie (źródło: Statista)*
» *Fig. 3. Number and growth forecast of IoT devices worldwide (source: Statista)*

Tak więc, pomimo doskonałości w teorii bezpieczeństwa, rzeczywistość wskazuje na istny zalew kodu z podatnościami i jednoczesne umiarkowane (niemniej rosnące) zainteresowanie zasadami bezpieczeństwa. Większa liczba programistów (deweloperów) używa automatyzacji do włączania kodu zależnego (zapożyczonego). Według statystyk GitHub w 2023 r. deweloperzy open-source o 60% częściej niż 2022r. aktualizowali biblioteki podatne na ataki. Jeszcze bardziej interesujący jest rosnący trend zabezpieczania gałęzi kodu open-source. Chronione gałęzie (ang. *secure branch*) dają opiekunom więcej sposobów na zapewnienie bezpieczeństwa ich projektów i [...] 60% najpopularniejszych projektów open source z nich korzysta. [GITHUB]. Można w tym miejscu dostrzec – choć nie w pełni przesądzony – związek z przepisami zawartymi w Akcie o Cyberodporności [CRA], które nakładają na dostawców rozwiązań informatycznych, w tym opartych na otwartym kodzie źródłowym, obowiązek zapewnienia bezpieczeństwa całego produktu. W praktyce może to oznaczać konieczność wzięcia odpowiedzialności również za komponenty zapożyczone, co budzi – nota bene - istotne kontrowersje w środowisku open-source (patrz np.: *Unaware and Uncertain: Is the Open Source Community Prepared for the New Regulatory Reality of the Cyber Resilience Act? – Linux Foundation*).

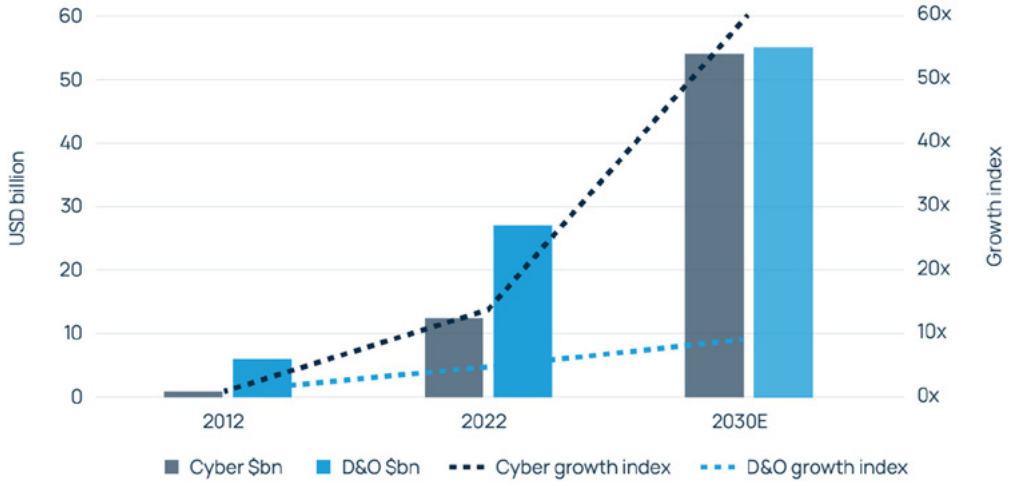
Perfekcyjne bezpieczeństwo techniczne, podobnie jak idealne bezpieczeństwo fizyczne, nie jest możliwe. Celem powinno być zatem sprawienie, aby koszt ataku był wystarczająco wysoki, aby zmniejszyć ryzyko do akceptowalnego poziomu. Obecnie należałoby uzupełnić to zalecenie - przy powszechnej dostępności urządzeń cyfrowych i niemożności zbudowania równie powszechnych kompetencji społecznych w zakresie bezpieczeństwa cyfrowego - o zalecenie, aby produkty cyfrowe minimalizowały prawdopodobieństwo przeprowadzenia ataku. Może to przełożyć się na rodzaj ograniczenia wolności, tj. odmowę wykonania usługi/funkcji w sferze cyfrowej, jeśli może ona zagrozić, na przykład, poufności danych lub ujawnieniu tożsamości.

Produkty i usługi wchodzą na rynek i utrzymują się na nim z podatnościami i lukami w zabezpieczeniach, które mogą zostać wykorzystane przez złośliwe podmioty. Użytkownikom oferowane są podatne na zagrożenia produkty i usługi, a oni sami borykają się z brakiem specjalistycznej wiedzy do zarządzania ryzykiem. Szacunkowa skala problemu jest gigantyczna, dość powiedzieć, że ponad 3,8 miliarda urządzeń IoT jest podłączonych w Europie, a 16 miliardów na całym świecie (źródło: Statista, 2024). Pamiętając przy tym, że 72% przeskanowanej repozytoriów kodu (dla sektora IoT) zawiera podatności wysokiego ryzyka. Nie wspominając o innych sektorach i branżach, w których ekspozycja na wysokie ryzyko jest jeszcze wyższa sięgając 100% zeskanowanych baz kodów ([BLDCK], 2025).

W miarę, jak rynek inteligentnych urządzeń rozszerza się o produkty takie jak autonomiczne pojazdy i zaawansowane domowe sterowniki energetyczne, podatności mogą - i będą - stwarzać ryzyko systemowe dla społeczeństwa i gospodarki. Od 2022 roku (inwazja Rosji na Ukrainę na pełną skalę) wiemy już, że „stan cyfrowej podatności" - nie potencjalnie, ale faktycznie - ułatwia obcą ingerencję w cyberprzestrzeń.

Nie ma jednego rozwiązania, które wyeliminowałoby utrzymujące się zagrożenie ze strony złośliwych cyberprzestępców wykorzystujących luki w technologiach, a produkty, które są „bezpieczne z założenia", będą nadal narażone na wystąpienie podatności; jednak duża ich liczba wynika ze stosunkowo niewielkiego podzbioru przyczyn źródłowych. [SBDP]

Warto tutaj nadmienić, że środki nietechniczne, takie jak ochrona ubezpieczeniowa, mogą być opłacalną alternatywą dla kosztownych technicznych środków bezpieczeństwa [X.800]. W 2023 r. globalny rynek ubezpieczeń był wart 16,6 mld USD. Co więcej, prognoza wzrostu do 2030 r.



» *Rys. 4. Prognozy dot. ubezpieczeń związanych z cyberprzestrzenią do 2020 r. (źródło: HOLDEN)*
» *Fig. 4. Cyber-related insurance projections by 2030 (source: HOLDEN)*

sięga 50 mld USD [Holden], co nie tylko daje wyobrażenie o wielkości rynku, ale – co być może ważniejsze – dotyczy wyceny ryzyka zarówno przez firmy ubezpieczeniowe, jak i ubezpieczających się przedsiębiorców. W tym kontekście warto wspomnieć o rosnących obowiązkach regulacyjnych w krajach Unii Europejskiej i Wielkiej Brytanii zarówno dla producentów (CRA), jak i użytkowników instytucjonalnych (NIS2). Rozwój rynku ubezpieczeń dodaje opcję wyceny odszkodowania do szacowania kosztów zabezpieczeń. Należy pamiętać, że odpowiedzialność deliktowa za produkty cyfrowe jest czynnikiem, który z czasem będzie zyskiwał na znaczeniu. Będzie to zapewne kolejny czynnik wywierający presję na realne wdrożenie pryncypiów Security-by-design. Szczególną uwagę należy tu zwrócić na nowe przepisy europejskie, w szczególności dyrektywę Parlamentu Europejskiego i Rady (UE) 2024/2853 z dnia 23 października 2024 r. w sprawie odpowiedzialności za produkty wadliwe, która stanowi, że przy ocenie wadliwości produktu należy wziąć pod uwagę wszystkie okoliczności, w tym odpowiednie wymogi dotyczące bezpieczeństwa produktu, w tym istotne dla bezpieczeństwa wymogi dotyczące cyberbezpieczeństwa.

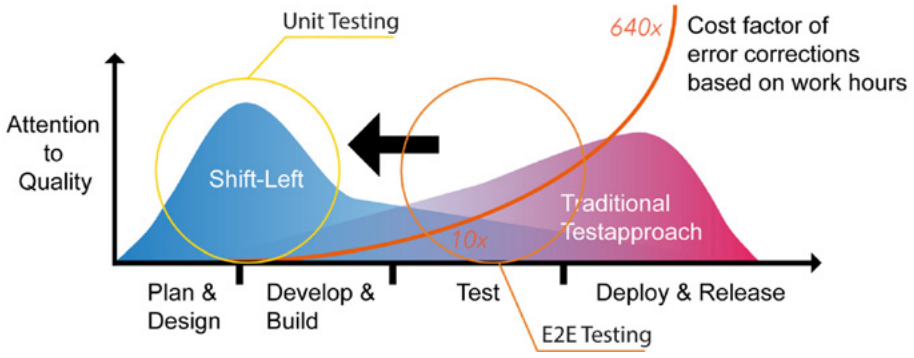
Powyższe rozważania mają dać pogląd na wieloaspektowe tło motywujące do wdrażania koncepcji Security-by-design. Warto w tym miejscu przedstawić definicję konsensusu CISA i wielu agencji ds. cyberbezpieczeństwa na całym świecie: „Security by design" oznacza, że produkty technologiczne są budowane w sposób, który **rozsądnie chroni** przed cyberprzestępcami próbującymi uzyskać dostęp do urządzeń, danych i infrastruktury sieciowej. Producenci produktów cyfrowych powinni przeprowadzić ocenę ryzyka w celu zidentyfikowania znanych zagrożeń dla systemów krytycznych, a następnie zaplanować adekwatne zabezpieczenia w projektach produktów, które uwzględniają zmieniający się krajobraz cyberzagrożeń. [SBDP] Innymi słowy, rozsądna ochrona oznacza, że ktoś będzie musiał za nią zapłacić, czy to w składce ubezpieczeniowej, czy w cenie produktu/usługi, środki ochrony muszą zostać dobrane odpowiednio pod względem technicznym i ekonomicznym.

JAK (WDROŻYĆ) SECURITY BY DESIGN?

Analizują wdrażanie podejścia „Security by design", przedstawiciele firm takich jak Microsoft, Google, CrowdStrike i Schneider Electric podkreślili rozmaite strategie podejścia do integracji bezpieczeństwa w całym cyklu życia produktów. Microsoft wskazuje (swój) Security Development Lifecycle jako przykład ramowy, zauważając, że specyficzne wymagania branżowe - takie jak PCI DSS w płatnościach lub HIPAA w opiece zdrowotnej - mogą kształtować praktyki bezpieczeństwa zależnie od rodzaju projektu. Google podkreśla wykorzystanie wewnętrznych zespołów testujących zabezpieczenia (red-teaming), testów fuzzingowych i analizy raportów o podatnościach przy realizacji projektów. CrowdStrike opowieda się za perspektywą opartą na analizie ryzyka, podczas gdy Schneider Electric skupił się na zgodności ze normami technicznymi i ocenach bezpieczeństwa. Zarówno CISA, jak i Google wspomnieli o wykorzystaniu metod formalnych – formalnych metod projektowania i analizy oprogramowania - jako uzupełniających w praktyce programistycznej. Urzędnicy CISA podkreślili, że bezpieczeństwo nie jest binarne, a koszt wdrożenia środków bezpieczeństwa jest różny zależnie od sposobu podejścia do realizacji projektu. Na przykład przyjęcie języków programowania bezpiecznych dla obsługi pamięci może być droższe niż skonfigurowanie bezpiecznych baz danych na wczesnym etapie rozwoju. Rozmówcy zauważyli również, że strategie wdrażania różnią się w zależności od tego, czy produkt jest wewnętrzny czy zewnętrzny, skierowany do konsumentów czy przedsiębiorstw, lub dostosowany do nowych lub istniejących klientów.

Wielu ekspertów ostrzega, że termin „Security-by-design" może być mylący, jeśli będzie interpretowany wąsko, jako odnoszący się tylko do fazy projektowania, a nie **do całego cyklu życia** - w tym analizy potrzeb, projektowania, tworzenia, wdrażania, konserwacji i wycofywania z rynku. Jest to bardzo ważne spostrzeżenie, ponieważ sugeruje, że skupienie się wyłącznie na projektowaniu może sprawić, że inne kwestie związane z bezpieczeństwem pozostaną nierozwiązane. [LWFR]

Programiści powinni otrzymać wskazówki dotyczące tego, jakich funkcji należy unika, a jakich bezpiecznych odpowiedników w ramach standardów kodowania używać. Ponadto należy wdrożyć narzędzia, które pomogą zidentyfikować i zweryfikować wykorzystanie niebezpiecznych funkcji. Wiele narzędzi do analizy statycznej i lintingu zapewnia mechanizm identyfikacji błędów w czasie kompilacji, a niektóre integrują się z IDE (*Integrated Development Environments*), aby dostarczać programistom wskazówki



» *Rys. 5. Znaczenie ekonomiczne strategii wykonywania testów na wcześniejszym etapie wytwarzania (źródło: Lightrun)*
» *Fig. 5. Shift-Left testing strategy - financial impact (source: Lightrun)*

w czasie pisania kodu. Ponadto wartość mają proste i dostępne rozwiązania, jak np.. plik nagłówkowy banned.h firmy Microsoft, które po włączeniu do projektu spowoduje, że użycie niebezpiecznych funkcji wygeneruje błędy kompilatora. Tworzenie kodu wolnego od niebezpiecznych funkcji jest zatem wykonalne. [SAFEcode]

IBM kładzie nacisk na kompetencje zespołu i strategiczne myślenie o architekturze rozwiązań. „Secure by design" to zasada, zgodnie z którą praktyki projektowania, rozwoju i dostarczania projektu mają być realizowane przez wykwalifikowany i doświadczony zespół. Zasady projektowania zabezpieczeń, kierują praktykami myślenia architektonicznego. Postuluje się wykorzystanie twórczego podejścia projektowego (*enterprise desing thinink*), aby skupić się na oczekiwaniach interesariuszy związanych z ryzykiem, zgodnością i bezpieczeństwem, a co dotyczyć ma zarówno użytkowników wewnętrznych, jak i zewnętrznych. Proces projektowania jest zorientowany na architekturę bezpieczeństwa docelowego rozwiązania. [IBM]

W zgodnej opinii podejście wytwarzania zorientowane na bezpieczeństwo jest kosztowne, czasochłonne i wymaga zaangażowania wysokiej klasy ekspertów. [SLA]

Chociaż środowisko naukowe i przemysł opowiadają się za wdrożeniem koncepcji „Security-by-design", w rzeczywistości wielu projektantów i architektów systemów pracuje w środowiskach o niskim poziomie znajomości norm bezpieczeństwa, rzadko lub nigdy nie włącza mechanizmów bezpieczeństwa do tworzonego systemu, ani nie uważa włączenia problematyki bezpieczeństwa do procesu projektowania za przyjemne, ekscytujące lub interesujące. [BLHD]

Można spekulować, że czynnikiem decydującym o powodzeniu wdrożenia koncepcji „Security-by-desing" będzie aspekt finansowy. To, czy testowanie bezpieczeństwa będzie miało tak duży wpływ, jak wcześnie wykonywane testy funkcjonalne (patrz rysunek 3), nie zostało jeszcze udowodnione.

Istnieje pojęcia tzw. „przesunięcia bezpieczeństwa w lewo" (*security shift-left*), które można interpretować jako strategiczny krok w kierunku wdrażania zasad Security-by-Design. Choć koncepcja ta znajduje coraz szersze uznanie, jej skuteczna realizacja na poziomie przedsiębiorstwa stanowi istotne wyzwanie. Wymaga bowiem nie tylko zmian proceduralnych, ale przede wszystkim transformacji kulturowej obejmującej programistów, testerów, zespoły ds. bezpieczeństwa oraz inne jednostki organizacyjne zaangażowane w wytwarzanie. Co więcej, wdrożenie tego podejścia pociąga za sobą konieczność redefinicji strategii organizacyjnych na wszystkich poziomach zarządzania. Warto zwrócić uwagę na kilka kluczowych aspektów, które zwiększają szanse powodzenie tego procesu:

1. Automatyzacja: musi obejmować automatyczną kompilację, kontrole bezpieczeństwa i testowanie (co oznacza włączenie testów bezpieczeństwa do praktyki CI/DC.
2. Narzędzia do testowania zorientowane na bezpieczeństwo: statyczne testowanie bezpieczeństwa aplikacji (SAST), dynamiczne testowanie bezpieczeństwa aplikacji (DAST), analiza składni (SCA) itp.
3. Modelowanie zagrożeń: narzędzie/metodyka może się różnić (STRIDE, OCTAVE, PASTA, DREAD, NIST 800-30), jednak jej użycie będzie oznaczać, że wszystkie zespoły w projekcie będą koncentrować uwagę na bezpieczeństwie.
4. Przyjęcie zasad bezpieczeństwa w cyklu wytwarzania (SDLC): co wspiera identyfikowanie problemów na wczesnym etapie rozwoju [FORTI].

NORMATYWNE PODEJŚCIE DO KONCEPCJI SECURITY-BY-DESIGN

Przedstawiony powyżej kontekst i zarys problematyki, pozwala stwierdzić, że wdrożenie zasad „Security by Design" jako powszechnej praktyki organizacyjnej stanowi złożone i wielowymiarowe wyzwanie. Zakres problemu wykracza daleko poza kwestie czysto techniczne – obejmuje zarządzanie zespołami i procesami, dotyczy pozyskiwania sprzętu i oprogramowania z zapewnieniem bezpieczeństwa w łańcuchu dostaw, wymaga wysokich kompetencje indywidualne i zespołowych, jest wyzwaniem dla zamawiających, dla a także dla dostawców chcących wykazać zgodność z obowiązującymi regulacjami i standardami branżowymi.

Oczekiwania są wysokie, tj. produkty cyfrowe mają być bezpieczne. Ludzie powinni czuć się chronieni przez bezpieczną i odporną gospodarkę cyfrową, [...] mając pewność, że cyberbezpieczeństwo będzie egzekwowane przez tych, którzy są najbardziej zdolni do zarządzania nim na każdej warstwie technologicznego łańcucha dostaw. Konsumenci i firmy skorzystają na powszechnym przyjęciu norm cyberbezpieczeństwa na rynkach technologii i oprogramowania. Kupując urządzenia cyfrowe lub oprogramowanie, [...] technologia będzie zabezpieczona przed cyberatakami i nie będzie miała luk w zabezpieczeniach – podobne deklaracje możemy przeczytać w wielu manifestach politycznych np. australijskiej strategii cyberbezpieczeństwa [AUCS], czy preambule do europejskiego Aktu o cyberodporności.

Można postawić hipotezę, że istnieją dobrze ugruntowane normy międzynarodowe oraz specyfikacje – zarówno formalnie, jak społecznościowo rozwijane i przyjmowane –

które mogą wspierać przedsiębiorstwa i organizacje w skutecznym wdrażaniu zasad „Security by Design". Normy te, odpowiednio zaadaptowane do kontekstu organizacyjnego, mogą stanowić fundament dla budowy systemowego podejścia do bezpieczeństwa, integrując wymagania techniczne, proceduralne i kompetencyjne.

Powszechne przyjęcie i wdrożenie pryncypiów „Security-by-design" jest możliwe w oparciu o istniejące normy.

Poniższe *tabele* (Tab. 4. *Zestawienie norm mających zastosowanie dla wdrażania koncepcji „Security-by-design" w odniesieniu do produktów cyfrowych* i Tab. 5. *Znaczenie wybranych norm dla interesariuszy ekosystemu „Security-by-design"*) zawierają niewyczerpujący przegląd ustalonych i powszechnie uznanych standardów, reprezentujących obecny stan wiedzy. Oczekuje się, że standardy te, odpowiednio przekazane wykwalifikowanemu personelowi - zapewniające zgodność między wymaganiami zadania a doświadczeniem zespołu - przyniosą zamierzone rezultaty.

Decyzja dotycząca wyboru ram normatywnych jest tylko częściowo uznaniowa i nie zależy wyłącznie od preferencji biznesowych. Na przykład zalecenia wydane przez niemiecki Federalny Urząd ds. Bezpieczeństwa Informacji (BSI) wyraźnie stwierdzają, że „producent MUSI przestrzegać najlepszych praktyk w zakresie bezpiecznego cyklu życia (oprogramowania), takich jak OWASP SAMM, BSI TR-03185 i ISO/IEC 27034.

Uzupełnieniem do norm i specyfikacji wysokiego poziomu (system/proces) są normy specyficznie adresujące zagadnienia bezpieczeństwa produktów informatycznych.

Złożony ekosystem rynku ICT musi być w pełni zaangażowany, aby osiągnąć cel, jakim

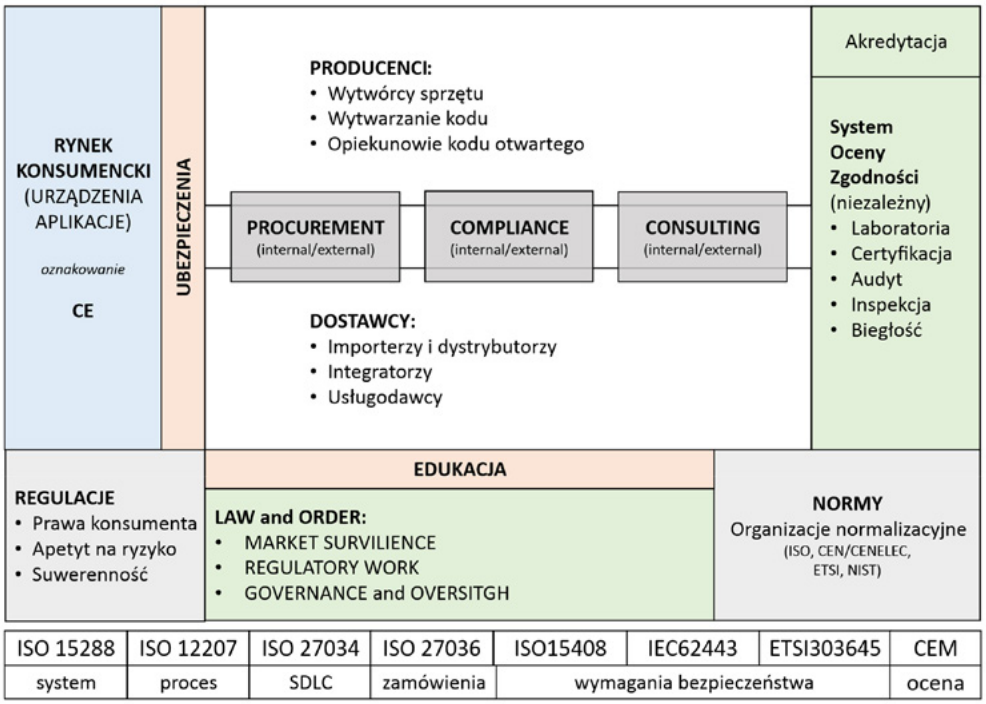
» *Tab. 3. Zestawienie wybranych norm i specyfikacji mających znaczenie dla koncepcji "Security-by-design"*
» *Tabl. 3. Composition of existing "Security-by-design" related standards and frameworks*

Nazwa	Główny cel	Przeznaczenie	Poziom
ISO/IEC/IEEE 15288	Organizacja cyklu życia systemu	Inżynierowie systemów, integratorzy	Inżynieria na poziomie systemu
ISO/IEC 12207	Organizacja cyklu życia oprogramowania	Programiści, kierownicy projektów	Proces tworzenia oprogramowania
ISO/IEC 27000 seria norm	Zarządzanie bezpieczeństwem informacji	Administracja, sektory regulowane, przedsiębiorstwa	Organizacja/system
ISO/IEC 27034	Integracja bezpieczeństwa z cyklem życia oprogramowania	Architekci bezpieczeństwa, zespoły wytwórcze	Organizacja/proces
ISO/IEC 27036	Zarządzanie bezpieczeństwem w łańcuchu dostaw ICT	Administracja, sektory regulowane	Organizacja/proces
NIST 800-53	Ogólne bezpieczeństwo i zgodność	Administracja, sektory regulowane	System/proces/aplikacja
NIST SP 800-218 (SSDF)	Bezpieczne praktyki tworzenia oprogramowania	Programiści, zespoły ds. bezpieczeństwa, specjaliści ds. zgodności	Cykl rozwoju i polityka organizacyjna
OWASP SAMM	Doskonalenie procesów i dojrzałości	Architekci bezpieczeństwa	Strategia wysokiego poziomu
OWASP ASVS	Weryfikacja oprogramowania	Programiści, inżynierowie, architekci	Poziom kodu/konfiguracji
CMMI dla rozwoju (CMMI-DEV)	Doskonalenie procesów i dojrzałości	Organizacje dążące do optymalizacji procesów	Poziom organizacyjny i poziom projektu
Własny* SDLC	Bezpieczne praktyki tworzenia oprogramowania	Architekci, zespoły wytwórcze	Proces

** Microsoft, Oracle, IBM*

» *Tab. 4. Zestawienie wybranych norm mających znaczenie dla koncepcji „Security-by-design" w odniesieniu do produktów cyfrowych*
» *Tabl. 4. Composition of existing Security-by-design related standards for product security*

Nazwa	Tytuł	Rodzaj produktu	Przeznaczenie
Wspólne kryteria (ISO/IEC 15408) Wersja 3.1 Rev. 5 (2017) Wersja 2022	Wspólne kryteria oceny bezpieczeństwa technologii informatycznych	Produkty IT (sprzęt, oprogramowanie, systemy wbudowane) - wszystkie rodzaje	Testy laboratoryjne, ocena i certyfikacja
CEM Wersja 3.1 Wersja 2022	Wspólna metodologia oceny bezpieczeństwa IT	Produkty IT (sprzęt, oprogramowanie, systemy wbudowane) - wszystkie rodzaje	Testy laboratoryjne, ocena i certyfikacja
IEC 62443-4-2 IEC 62443-4-2:2019	Techniczne wymogi bezpieczeństwa dla komponentów IACS	Elementy systemu automatyki przemysłowej i sterowania (IACS)	Ocena i certyfikacja
ETSI EN 303 645 Wersja 3.1.3	Cyberbezpieczeństwo dla konsumenckiego internetu rzeczy	Podłączone produkty konsumenckie (IoT)	Testy fabryczne, Możliwe: badania laboratoryjne i certyfikacja
ETSI TS 103 701 Wersja 2.1.1	Ocena cyberbezpieczeństwa dla konsumenckiego IoT - ocena zgodności	Podłączone produkty konsumenckie (IoT	Laboratoria testowe, producenci poszukujący zgodności z normą EN 303 645
ETSI TS 103 848 Wersja 1.1.1	Cyberbezpieczeństwo inteligentnych liczników - zastosowanie normy EN 303 645	Inteligentne liczniki	Laboratoria testowe, producenci poszukujący zgodności z normą EN 303 645



» *Rys. 6. Struktura interesariuszy ekosystemu Security-by-design*
» *Fig. 6. Synthetic structure of stakeholders in the "Security-by-design" ecosystem*

» Tab. 5. Znaczenie wybranych norm dla interesariuszy ekosystemu "Security-by-design"
» *Tabl. 5. Structured matrix of the relevance and usability of selected standards for stakeholders in the "Security-by-design" ecosystem*

Nazwa/Użyteczność	producenci	dostawcy/Integratorzy	klienci	regulatory	zespoły wytwórcze	menedżerowie ryzyka	menedżerowie ds. zgodności	usługi zaopatrzenia	usługi doradcze	system oceny zgodności	system edukacji	organizacje normalizacyjne
ISO/IEC/IEEE 15288	T	T		T			T		T	T	T	T
ISO/IEC 12207	T	T		T	T		T		T	T	T	T
ISO/IEC 27034	T	T	C	T	T	T	T	T	T	T	T	T
ISO/IEC 27000 (seria)	T	T	T	T	T	T	T	T	T	T	T	T
ISO/IEC 27036		T	T			T		T	T		T	T
NIST 800-53	T			T	T	T	T				T	T
NIST SP 800-218 (SSDF)	T			T	T	T	T*				T	T
OWASP SAMM	T				T	T	T		T		T	T
OWASP ASVS	T				T	T	T		T		T	T
OWASP LLM	T				T	T	T				T	T
CMMI-DEV	T				T		T				T	T
Common Criteria ISO/IEC 15408	T	T*	T*	T		T	T	y		T	T	T
CEM (ISO/IEC 18045)	T					T	T*			T	T	T
IEC 62443-4-2	T	T*	T*	T		T	T			T	T	T
ETSI TS 303 645	T	T*	T*	T		T	T			T	T	T
ETSI TS 103 701	T						T*			T	T	T

jest solidne powszechna implementacja „Security--by-design".

Zainteresowane strony obejmują:
- klienci,
- ustawodawcy i regulatorzy rynku,
- menedżerowie ryzyka,
- menedżerowie ds. zgodności,
- usługi zaopatrzenia (zamawiający),
- usługi doradcze,
- system oceny zgodności,
- system edukacji i nie tylko.

Syntetyczna struktura interesariuszy zaangażowanych we wdrażanie koncepcji Security-by-design została przedstawiona na *rysunku 4*.

Potrzeby i obowiązki każdego interesariusza muszą zostać przeanalizowane, odpowiednio odnotowane, zrozumiane, zweryfikowane, oszacowane i uwzględnione. Zwięzła ilustracja oczekiwań i potencjalnego wkładu wybranych interesariuszy - takich jak organizacje normalizacyjne - może pomóc w wyjaśnieniu zakresu i złożoności zagadnienia. Podczas gdy opracowywanie dokumentów normatywnych jest oczywistym i powszechnie oczekiwanym obowiązkiem, nie jest to bynajmniej jedyny z nich. Podczas gdy zagrożenia i czynniki środowiskowe zmieniają się, podatności i luki w zabezpieczeniach istnieją przez cały okres eksploatacji systemu lub protokołu, chyba że zostaną podjęte konkretne kroki w celu ich wyeliminowania. Instytucje normalizacyjne mają zarówno odpowiedzialność, jak i wyjątkową zdolność do eliminowania podatności, które mogą być nieodłącz-

nie związane ze specyfikacjami, takimi jak architektury, ramy i protokoły. (ITUSEC)

Podsumowanie znaczenia wybranych dokumentów normatywnych dla każdej z grup interesariuszy zebrano w *Tabela 5. Znaczenie wybranych norm dla interesariuszy ekosystemu „Security-by-design".*

SECURITY-BY-DESIGN - KTO ODPOWIADA ZA WDROŻENIE?

Seizo Onoe, Dyrektor Międzynarodowego Związku Telekomunikacyjnego (ITU), w przedmowie do najnowszego wydania Podręcznika Bezpieczeństwa ITU-T twierdzi, że wraz ze wzrostem liczby i różnorodności technologii informacyjno-komunikacyjnych (ICT) stają się one niezbędne w biznesie i życiu codziennym, dla konsumentów w każdym sektorze przemysłu, który przekształca się za pomocą ICT. Wzywa on do ustanowienia „**kultury Security--by-design**", zwracając szczególną uwagę na bezpieczeństwo w całym cyklu życia technologii informacyjno-komunikacyjnych. Z technicznego punktu widzenia to wezwanie może nie wydawać się rewolucyjne, ale wymiar społeczny **budowania kultury** - musi być postrzegany jako główne przedsięwzięcie na nadchodzące lata. W tym miejscu warto docenić wiedzę i intuicję prof. Eugene'a Howarda Spafforda, rozprawiającego się o mitami cyberbezpieczeństwa [CyberMyths]:

1. Cyberbezpieczeństwo to nie tylko ochrona komputerów i sieci.
2. Cyberbezpieczeństwo obejmuje komputery, jednak cyberbezpieczeństwo **dotyczy przede wszystkim ludzi**.
3. Komputery czasem działają nieprawidłowo a ludzie popełniają błędy.
4. **Cyberbezpieczeństwo jest zorientowane na człowieka**.

Pamiętajmy przy tym, że słowa te pochodzą od legendarnego Spafa, który kiedyś stwierdził, że „jedyny system (komputerowy), który jest naprawdę bezpieczny, to taki, który jest wyłączony i odłączony od zasilania, zamknięty w sejfie wyłożonym tytanem, zakopany w betonowym bunkrze i otoczony gazem paraliżującym oraz bardzo dobrze opłacanymi uzbrojonymi strażnikami. A nawet wtedy nie postawiłbym na to swojego życia".

Oto, gdzie jesteśmy w tej chwili: **Security-by-design jest wyzwaniem społecznym, naszą wspólną odpowiedzialnością**, nie częścią kodu źródłowego, ani funkcją maszyny liczącej.

ZARYS KIERUNKÓW DALSZYCH BADAŃ I ROZWOJU.

- **Określenie celów strategicznych: koncepcja „Security-by-design" jako rdzeń cyberbezpieczeństwa.**

Wobec rosnącej złożoności zagrożeń cyfrowych oraz coraz większej zależności państwa od infrastruktury informacyjnej, zasadne wydaje się systemowe ujęcie zasad „Security-by-Design" w ramach krajowej strategii cyberbezpieczeństwa. Dotychczasowe podejścia, koncentrujące się głównie na reagowaniu na incydenty i wzmacnianiu odporności operacyjnej, powinny zostać uzupełnione o proaktywne mechanizmy kontraktowania i projektowania bezpiecznych rozwiązań ICT. Wskazane jest zbudowanie modelu ekonometrycznego umożliwiającego optymalizacje tempa i zakresu podejmowanych decyzji regulacyjnych.

- **Tworzenie zachęt rynkowych: uczynienie z bezpieczeństwa przewagi konkurencyjnej**

Implementacja „Security-by-design" musi być nagradzana, a nie tylko wymagane. Rząd i duże przedsiębiorstwa powinny włączyć zachęty do zamówień i oceny łańcucha dostaw. Obejmować to może opracowanie modelowych kryteriów oceny w przetargach publicznych, preferencyjne traktowanie w łańcuchach dostaw (publicznych i prywatnych). Pod rozwagę należy wziąć stosowanie ulg podatkowych i programy wsparcia.

- **Wzmocnienie systemu zamówień: budowanie kompetencji i kryteriów ceny.**

Zaopatrzenie jest pierwszą linią obrony. Wymogi bezpieczeństwa dot. produktów i usług IT muszą zostać uwzględnione w specyfikacjach przetargowych, a specjaliści ds. zamówień publicznych muszą zostać przeszkoleni w zakresie ich oceny. Ocena ryzyka i modelowanie zagrożeń powinno być częścią fazy inicjowania zamówień, a decydenci muszą być przygotowani do wspierania wyborów opartych na bezpieczeństwie. Można rozważyć przygotowanie rekomendacji branżowych oraz wzorcowych specyfikacji do zakupów, jak również wykorzystanie mechanizmu przedkomercyjnych i innowacyjnych zamówień publicznych.

- **Wykorzystanie ekosystemu oceny zgodności: budowanie zaufania poprzez dobrowolne wykazanie zgodności.**

Niezależna ocena wzmacnia zaufanie w łańcuchu dostaw. Wskazane jest, aby rozważyć i opracować modelowe rozwiązania korzystające z pełnego spektrum narzędzi oceny zgodności – badań i testowania, inspekcji, audytu, oceny biegłości, certyfikacji, akredytacji - aby stale monitorować poprawność działania zasad „Security by design".

- **Regulacja z myślą o ludzkich możliwościach: dostosowanie strategii do realiów poznawczych**

Wdrożenie Security-by-design odniesie sukces lub porażkę w zależności od tego, jak dobrze uwzględni czynnik ludzki. Strategie wdrożeniowe muszą być dostosowane do modeli mentalnych, obciążenia poznawczego i praktycznych możliwości wszystkich interesariuszy - od programistów po decydentów. Wskazane jest opracowanie materiałów dydaktycznych, programów kształcenia (zawodowego i formalnego), włączenia odpowiednio wybranych zagadnień do systemu edukacji.

Zarysowane powyżej obszary wskazują na potrzebę interdyscyplinarnych badań nad operacjonalizacją koncepcji Security-by-Design w kontekście krajowych i regionalnych strategii cyberbezpieczeństwa. Szczególnego znaczenia nabierają tu: rozwój modeli decyzyjnych wspierających polityki publiczne, analiza skuteczności mechanizmów rynkowych i regulacyjnych, badania nad percepcją i adaptacją zasad bezpieczeństwa wśród różnych grup zawodowych, a także ewaluacja narzędzi oceny zgodności w dynamicznie zmieniającym się środowisku technologicznym.

BIBLIOGRAFIA

[1] **ETSI** (2024). *ETSI EN 303 645 V3.1.3: Cyber Security for Consumer Internet of Things: Baseline Requirements*. European Telecommunications Standards Institute. Link: https://www.etsi.org/deliver/etsi_en/303600_303699/303645/03.01.03_60/en_303645v030103p.pdf

[2] **International Telecommunication Union** (dalej **ITU-T**) (2004). *ITU-T X.805: Security architecture for systems providing end-to-end communications*. ITU. Link: https://www.itu.int/rec/T-REC-X.805.

[3] **Bundesamt für Sicherheit in der Informationstechnik** [dalej **BSI**] (2024) Technical Guideline TR-03183: *Cyber Resilience Requirements for Manufacturers and Products, Part 1: General Requirements*. Link: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Publications/TechGuidelines/TR03183/BSI-TR-03183-1-0_9_0.pdf .

[4] **BSI** (2024) Technical Guideline BSI TR-03185 Secure Software Lifecycle. Link: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Publications/TechGuidelines/TR03185/BSI-TR-03185.pdf .

[5] **NIST** (2024) SP 800-218A *Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile. Link: https://csrc.nist.gov/pubs/sp/800/218/a/final .*

[6] **NIST** (2022) SP 800-218 *Secure Software Development Framework (SSDF) Version 1.1: Recommendations for Mitigating the Risk of Software Vulnerabilities. Link: https://csrc.nist.gov/pubs/sp/800/218/final .*

[7] **Digital Policy Office (Hong Kong)** (2024). *Practice Guide for Security by Design* [PDF]. Link: https://www.govcert.gov.hk/doc/PG%20for%20Security%20by%20Design_EN.pdf.

[8] IEEE Cybersecurity (2025) *Building Code for the Internet of Things.* Link: https://ieeecs-media.computer.org/media/technical-activities/CYBSI/docs/Building_Code_IoT_online.pdf.

[9] **UK Government** (2025) *Secure by Design Principles*. National Cyber Security Centre (NCSC).Link: https://www.security.gov.uk/policy-and-guidance/secure-by-design/principles/.

[10] **Ivanti** (2025). *What is Secure by Design? Terminology, Benefits & Guidelines*. Link: https://www.ivanti.com/glossary/secure-by-design.

[11] **10Guards.** (2025. *Secure by Design: From Concept to Cybersecurity Imperative*. Link: https://10guards.com/secure-by-design.

[12] **Bright Defense.** (2025). *Secure By Design Guide*. Link: https://learn.microsoft.com/en-us/windows/security/book/.

[13] **Microsoft Corporation** (2023). *Windows 11 Security Book: Powerful security by design* [PDF]. Link: https://www.microsoft.com/content/dam/microsoft/final/en-us/microsoft-brand/documents/MSFT-Windows-11-Security-guide-RWMvI1.pdf .

[14] **Microsoft Corporation** (2025) *Microsoft Security Development Lifecycle Practices* Link: https://www.microsoft.com/en-us/securityengineering/sdl/practices.

[15] **Oracle Corporation** (2025) Oracle Corporate Security Practices Link: https://www.oracle.com/contracts/docs/corporate-security-practices-4490843.pdf.

[16] **IBM** (2018). Security in Development The IBM Secure Engineering Framework Link: https://www.redbooks.ibm.com/redpapers/pdfs/redp4641.pdf .

[17] [IBM] **IBM** (2025) *IBM Well-Architected Framework* Link: https://www.ibm.com/architectures/well-architected/security .

[18] **Mozilla Foundation.** (2024). *Mozilla Security Principles*. Link: https://infosec.mozilla.org/fundamentals/security_principles.html

[19] **Infosys Limited.** (2020). *Security by design*. Link: Security by design

[20] **CISA** (2023) *Secure By Design Pledge*. Cybersecurity and Infrastructure Security Agency. Link: https://www.cisa.gov/secureby-design/pledge

[21] **European Commision** (2019). *Cybersecurity Act*. Link: https://eur-lex.europa.eu/eli/reg/2019/881

[22] **NIST** (2020). *NIST Special Publication 800-160 Vol. 1: Systems Security Engineering. Considerations for a Multidisciplinary Approach in the Engineering of Trustworthy Secure Systems*; Link: https://doi.org/10.6028/NIST.SP.800-160v1

[23] **OWASP Foundation**(2021) *OWASP Top 10* Link: https://owasp.org/Top10/

[24] **CSA (Cloud Security Alliance).** (2023). *Cloud Controls Matrix (CCM)*. Link: https://cloudsecurityalliance.org/research/ccm/

[25] **European Commission** (2024). *Cyber Resilience Act (CRA) Regulation (EU) 2024/2847*. Link: https://digital-strategy.ec.europa.eu/en/library/cyber-resilience-act

[26] **Israel National Cyber Directorate** - *Supply Chain Guide.* Link: https://www.gov.il/he/pages/supply_chain_guide

[27] [SBDP] **CISA, NSA, FBI, ACSC, CCCS, CERT NZ, NCSC-NZ, NCSC-UK, BSI, NCSC-NL, NCSC-NO, NÚKIB, INCD, KISA, NISC-JP, JPCERT/CC, CSA, CSIRTAMERICAS** (2023) *Shifting the Balance of Cybersecurity Risk (Principles and Approaches for Secure by Design Software)*. Link: https://www.cisa.gov/sites/default/files/2023-10/Shifting-the-Balance-of-Cybersecurity-Risk-Principles-and-Approaches-for-Secure-by-design-Software.pdf.

[28] **Michael Benis** (2023). *Unlocking Cyber Resilience: INCD Cyber Supply Chain Methodology Explained*. Link: https://www.linkedin.com/pulse/unlocking-cyber-resilience-incd-supply-chain-explained-michael-benis-kyjnf/

[29] **Serge Drolet** (2024) *Évoluer avec l'approche Secure by design*. Link: https://owasp.org/www-chapter-quebec-city/assets/presentations/OWASPSecureByDesign20241017.pdf + 2024/10/17 - Secure by Design et DevSecops | OWASP Foundation

[30] **ISO** ISO 31700-1:2023 *Consumer protection — Privacy by design for consumer goods and services — Part 1: High-level requirements*. https://www.iso.org/obp/ui/en/#iso:std:iso:31700:-1:ed-1:v1:en

[31] **Madhu Murty** (2025) *The Hidden Cost of Ignoring Quality Early: Why Shift Left Still Hasn't Shifted Minds*. Link: https://www.linkedin.com/pulse/hidden-cost-ignoring-quality-early-why-shift-left-still-madhu-murty-lpeke/

[32] **Fortinet** (2025) What Is Shift Left Security? Link: https://www.fortinet.com/resources/cyberglossary/shift-left-security

[33] **Invi Grid** (2025) *Difference Between Security by Design and Shift Left.* Link: https://www.invigrid.com/post/difference-between-security-by-design-and-shift-left

[34] **NCSC** (2024) GovAssure announcements and newsletters - UK Government Security - Beta

[35] ISO (20211) *ISO/IEC 27034-1:2011 Information technology — Security techniques — Application security Part 1: Overview and concepts*

[36] [SLA] **Casola V**. and others. *A novel Security-by-design methodology: Modeling and assessing security by SLAs with a quantitative approach*

[37] [CyberMyths] **Spafford E.H., Metcalf L., Dykstra J.**.(2023) Cybersecurity Myths and Misconceptions: Avoiding the Hazards and Pitfalls that Derail Us.

[38] [BHLD] **Peretz Renana Arizon, Hadar Irit, and Gil Luria** (2022) The Importance of Security Is in the Eye of the Beholder: Cultural, Organizational, and Personal Factors Affecting the Implementation of Security by Design
[39] [SEB-Fuchs] **A. Fuchs and C. Rudolph** „Security Engineering Based on Structured Formal Reasoning," (2012) ASE/IEEE International Conference on BioMedical Computing (BioMedCom).
[40] NIST SP 800-37, Risk Management Framework for Information Systems and Organizations—A System Life Cycle Approach for Security and Privacy
[41] NIST SP 800-53, Security and Privacy Controls for Information Systems and Organizations.
[42] [HOWDEN] https://www.howdengroup.com/sites/corporate.howdenprod.com/files/2023-07/9100%20Cyber%20Report%20June%2023%20v04.pdf
[43] [AUCS] **ACSC** (2023) Australian Cybers Security Strategy https://www.homeaffairs.gov.au/cyber-security-subsite/files/2023-cyber-security-strategy.pdf
[44] [KELA] **KELA** (2025) ***AI Threat Report How cybercriminals are weaponizing AI technology.*** Link: https://info.ke-la.com/hubfs/Reports/KELA%20Report%20-%202025%20AI%20Threat%20Report.pdf
[45] [LWFR] E.Lostri, J.Sherman (2024) ***"Security by Design" in Practice: Assessing Concepts, Definitions, and Approaches*** Link: https://s3.documentcloud.org/documents/25049674/sbd_lostrisherman_final.pdf
[46] [ANDER] (1972) ***Computer Security Technology Planning Study***. Link: https://apps.dtic.mil/sti/tr/pdf/AD0772806.pdf
[47] [GITHUB] Daigle K., GitHub Blog (2025) Octoverse: The state of open source and rise of AI in 2023. Link: https://github.blog/news-insights/research/the-state-of-open-source-and-ai
[48] [BLDCK] BlackDuck (2025) 2025 Open Source Security and Risk Analysis Report. Link: https://www.blackduck.com/content/dam/black-duck/en-us/reports/rep-ossra.pdf
[49] [X.800] **ITU-T** (1991) X.800: Security Architecture for Open Systems Interconnection for CCITT Applications
[50] [ITUSEC] **ITU-T** (2024) ***Handbook ITU-T SEC-MANUAL (09/2024) - Security in telśecommunications and information technology (8th edition)*** Link: https://www.itu.int/epublications/publication/itu-t-sec-manual-2024-09-security-in-telecommunications-and-information-technology-8th-edition

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

ANALIZA WPŁYWU STRUKTURY SYGNAŁÓW RADIOWYCH NA SKUTECZNOŚĆ GŁĘBOKIEJ SIECI NEURONOWEJ

ANALYSIS OF THE IMPACT OF RADIO SIGNAL STRUCTURE ON THE EFFECTIVENESS OF A DEEP NEURAL NETWORK

Alicja Olejniczak [1]; Jarosław Sadowski [2]

[1] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, alicja.olejniczak@pg.edu.pl
[2] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, jarsadow1@pg.edu.pl

DOI: 10.15199/59.2025.4.7

Streszczenie: Niniejszy artykuł ma na celu podjęcie próby określenia sposobu definiowania danych wejściowych w postaci sygnałów radiowych w celu optymalizacji procesu uczenia i poprawienia końcowej efektywności sieci neuronowej. Prezentowane badania podejmują zagadnienie odbioru sygnałów radiowych z użyciem głębokiej splotowej sieci neuronowej oraz przedstawiają znaczenie jej parametrów w kontekście detekcji sygnałów GMSK, będącej alternatywą dla klasycznego detektora MLSE.

Abstract: This article aims to explore how input data in the form of radio signals can be defined in order to optimize the training process and improve the final performance of a neural network. The presented research addresses the problem of radio signal reception using a deep convolutional neural network and highlights the importance of its parameters in the context of detecting GMSK signals, which serve as an alternative to the classical MLSE detector.

Słowa kluczowe: CNN, głębokie uczenie, GMSK, uczenie maszynowe

Keywords: CNN, deep learning, GMSK, machine learning

1. WSTĘP

Dynamiczny rozwój narzędzi z obszaru sztucznej inteligencji spowodował ich szerokie wykorzystanie w wielu dziedzinach, również w ujęciu badań naukowych. Nie inaczej jest w przypadku szeroko rozumianej radiokomunikacji. Z jednej strony, modele uczenia maszynowego mogą być przydatne w przypadku analizy i modelowania problemów nieliniowych czy takich, dla których nie istnieją modele matematyczne lub są wysoce niedokładne. Z drugiej jednak, sposób doboru danych uczących czy nawet ich przetwarzania wewnątrz sieci neuronowej nie jest ustandaryzowany i często bazuje na metodyce wywodzącej się z obszaru przetwarzania obrazów, która niekoniecznie będzie optymalna w przypadku odmiennych problemów badawczych.

Zagadnienie odbioru sygnałów radiowych z użyciem modeli neuronowych zostało podjęte w licznych pracach naukowych, z których niektóre pomijały nawet aspekt specyfiki przetwarzania sygnałów w ramach warstwy fizycznej zgodnie z przyjętym formatem transmisji i symulowały działanie części: nadawczej i odbiorczej połączonych w postaci autoenkodera, bazującego na nieznanej konstelacji wytworzonej w wyniku trenowania obu sieci.

Takie podejście zaproponowano w [7], a stosunkowo płytkie sieci (2 warstwy ukryte) nadajnika i odbiornika trenowane i testowane były jedynie dla kanału AWGN, symulowanego również w postaci warstwy neuronowej. Analiza porównawcza końcowej skuteczności modeli bazowała na metryce BER (ang. *bit error rate*) i uwzględniła dobór algorytmów optymalizacyjnych.

Model autoenkodera zaprezentowano również w [6]. W artykule zdefiniowano zagadnienie detekcji jako problem klasyfikacyjny, do którego rozwiązania wykorzystano 2-warstwowe sieci FNN w części nadawczej i odbiorczej (formowanie siatki zasobów OFDM zrealizowane zostało bez użycia sieci). Badania zakładały znajomość idealnego CSI (ang. *channel state information*), a oprócz metryki BLER (ang. *block error rate*) autorzy zaprezentowali również wykresy konstelacji wytworzonych w ramach prezentowanego autoenkodera.

Koncepcję odbiornika DFNN służącego predykcji symbolowej (zamiast bitowej) zaprezentowano w [4]. Autorzy przebadali różne typy modulacji a końcowe wyniki ocenili na podstawie symbolowej stopy błędów. Należy jednak podkreślić, że podejście to może być nieskuteczne w przypadku prób odwzorowania zależności fazowych sygnału.

W [3, 5] autorzy zaprezentowali detektory IM-OFDM (ang. *index modulation orthogonal frequency-division multiplexing*) bazujące na modelach CNN (ang. *convolutional neural network*) i DFNN (ang. *deep feedforward neural network*). Co ciekawe, w [5] dane uczące analizowane były w postaci biegunowej (moduł i faza), w celu poprawienia skuteczności detekcji nieaktywnych podnośnych, podczas gdy w [3] zaproponowano analizę próbek w postaci liczb zepolonych.

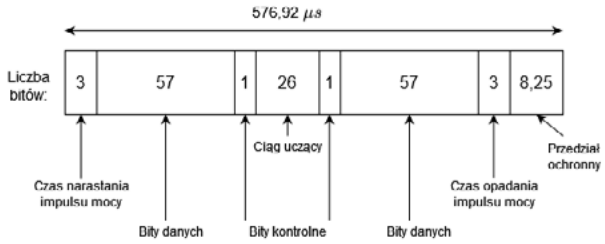
Przedstawiony przegląd literatury wskazuje przede wszystkim skupienie zainteresowania badawczego wokół analizy sygnałów OFDM oraz brak spójności wobec wykorzystania sztucznej inteligencji do przetwarzania sygnałów radiowych. Struktury sieci oraz sposób reprezentacji i interpretacji wektorów wejściowych i wyjściowych jest

różny, próżno też szukać uzasadnienia przyjętej metodyki badawczej.

Mając na uwadze, że istniejące biblioteki dedykowane projektowaniu modeli uczenia maszynowego ML (ang. *machine learning*) nie obsługują lub też nie są zoptymalizowane pod kątem przetwarzania liczb zespolonych, zrozumiałym wydaje się przetwarzanie sygnałów radiowych w postaci wektorów osobno reprezentujących część rzeczywistą oraz urojoną i takie podejście prezentowane jest w przeważającej części prac naukowych. Alternatywnym sposobem reprezentacji sygnałów jest postać biegunowa, w której próbki wyrażone są osobno w postaci modułu i fazy [5]. Niestety, żaden z wymienionych artykułów nie podjął się uzasadnienia przyjętego sposobu analizy danych, ani też nie zaprezentował porównania skuteczności algorytmów w zależności od przyjętego podejścia. Celem niniejszego artkułu jest uzupełnienie tej niejednoznaczności oraz próba zdefiniowania ogólnych rekomendacji dot. przetwarzania sygnałów radiowych w ramach modeli ML.

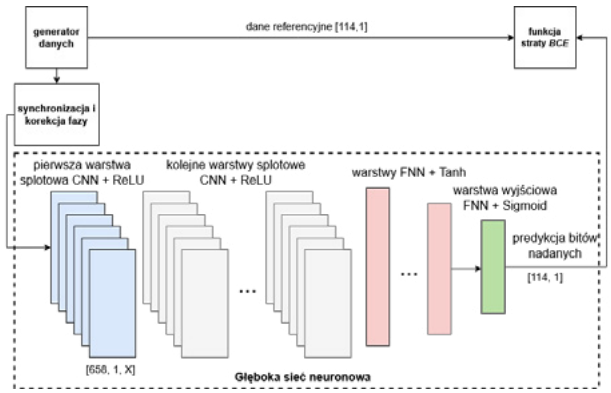
2. METODYKA BADAWCZA

Poniższa analiza jest kontynuacją prac obejmujących detekcję sygnałów z pojedynczą nośną i modulacją GMSK (ang. *gaussian minimum shift keying*), zgodnych ze standardem 3GPP w ramach GSM (ang. *Global system for mobile communication*) i EC-GSM-IoT (ang. *extended coverage GSM Internet of things*), prezentowanych w [8, 9]. Poza prezentacją przyjętej struktury pakietów (rys. 1), sposób przetwarzania sygnałów w ramach warstwy fizycznej nie będzie opisywany bardziej szczegółowo.



Rysunek 1: Struktura pakietu TCH/F (ang. *traffic channel/fullrate)*

W celu uwzględnienia zależności czasowych pomiędzy kolejnymi symbolami transmitowanych sygnałów GMSK w przeprowadzonych badaniach analizowano strukturę splotowej sieci neuronowej CNN, ze stosunkowo dużym rozmiarem filtra (zależnie od konfiguracji obejmującym od kilku do kilkudziesięciu kolejnych próbek sygnału) realizującego operację splotu na przetwarzanych danych. Ponadto, w symulacjach przyjęto profil kanału EPA (ang. *extended pedestrian A*) [1]. Przed przekazaniem danych do sieci pakiety odebrane były wstępnie synchronizowane, a w toku prac konieczne okazało się również dodanie bloku korekcji fazy [9]. Ogólną koncepcję realizowanej detekcji przedstawiono na rys. 2



Rysunek 2: Koncepcja głębokiego detektora GMSK

W tym miejscu należy zaznaczyć, że w celu przebadania skuteczności modeli ML w zależności od sposobu reprezentacji danych wejściowych przyjęto dwa warianty analizy. W rezultacie przetwarzano pojedynczy pakiet, na który składał się wektor 2-wymiarowy (oznaczony jako X na rys. 2) zawierający 658 próbek w układzie kartezjańskim (część rzeczywista i urojona) lub biegunowym (moduł i faza wyrażona w postaci kątowej). Testowane modele sieci neuronowych oprócz części splotowej składały się również z warstw w pełni połączonych FC (ang. *fully connected*) w ramach części FNN. Sam proces uczenia każdego z modeli bazował na określeniu funkcji straty BCE (ang. *binary cross-entrophy loss*) pomiędzy wektorem wyjściowym sieci, zawierającym dane zdemodulowane z przetwarzanego pakietu, a wektorem referencyjnym (114 bitów nadanych) przekazywanym z generatora danych. Funkcję BCE można opisać następującym wzorem:

$$BCE = -\frac{1}{m}\sum_{i=1}^{m}\left(y_i \log \hat{y}_i \log(1-\hat{y}_i)\right) \qquad (1)$$

gdzie y_i oraz $\hat{y}_i$ reprezentują i-te wartości wektorów: odniesienia i predykcji. Z uwagi na rolę projektowanych modeli, oprócz funkcji straty podczas procesu uczenia analizowano również bitową stopę błędów BER, obliczaną osobno dla zbiórów: treningowego (szeroki zakres E_b/N_0) i walidacyjnego ($E_b/N_0 = 7$ dB).

Ewaluacja modeli odbyła się poprzez połączenie tzw. grid search w przypadku poszukiwania optymalnej struktury sieci (liczba i rodzaj warstw) oraz random search, w którym dobór hiperparamterów i rozmiary: zbioru treningowego, batcha oraz filtru był przeprowadzany losowo dla każdego z badanych modeli. Liczba iteracji była dobierana zależnie od złożoności badanego modelu, a poziom szumu pakietów odebranych był losowany z przyjętego przedziału E_b/N_0. Proces puszukiwania był monitorowany przy pomocy oprogramowania Weights & Biases [2]. Przyjęte zakresy parametrów usystematyzowano w formie tabeli 1:

Tabela 1: Zakresy parametrów modeli ML dla fazy treningowej

Parametr	*Zakres*
wsp. uczenia	[0,0001; 0,001]
rozmiar zbioru uczącego	[100000; 500000]
liczba iteracji	{100; 150; 200}
batch_size	[100, 1000]
kernel	[2; 24]
stride	{1; 2; 3; 4; 8}
dropout	[0,1; 0,5]
E_b/N_0	7-18 dB

Należy mieć na uwadze, że końcowa skuteczność sieci zależy nie tylko od wybranej struktury, ale i inicjalizowanych losowo: parametrów, hiperparametrów i zbioru treningowego, dlatego też w ramach ewaluacji modeli neuronowych dla każdej z architektur fazę treningową wywoływano 5-krotnie, a końcowe wyniki prezentowane w następnym rozdziale będą się odnosić do modeli o najwyższej efektywności.

3. WYNIKI

W pierwszej części prezentowanych wyników niezbędne jest podsumowanie fazy treningowej modeli dla dwóch wariantów struktury danych wejściowych: w układzie kartezjańskim (K) i w układzie biegunowym z kątową reprezentacją fazy (B). Struktury sieci, rozmiary zbiorów treningowych T_{size}, wartości funkcji straty BCE oraz wartości bitowej stopy błędów BER_{val} dla zbioru walidacyjnego ($E_b/N_0 = 7$ dB) podsumowane zostały w tabeli 2:

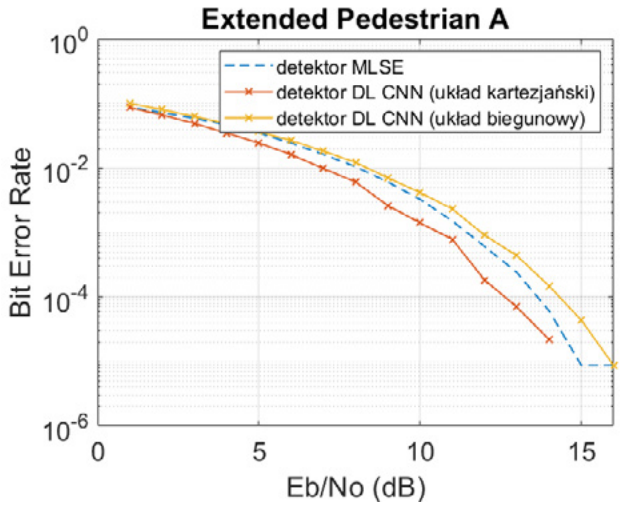
Tabela 2: Podsumowanie fazy treningowej badanych modeli ML

	K	*B*
T_{size}	390015	443370
CNN	3	3
FC	2	3
L_{drop}	1	2
wsp. uczenia	0,0005	0,0003
batch_size	569	804
kernel	7	3
BCE	0,0127	0,0190
BER_{val}	0,0103	0,0154

Parametr CNN reprezentuje liczbę warstw splotowych, FC liczbę warstw FNN, a L_{drop} liczbę warstw dropout. Zauważalny jest spadek skuteczności modelu w przypadku wariantu B, należy jednak zaznaczyć, że wartość bitowej stopy błędów BER_{val} była jedynie parametrem poglądowym, wyznaczonym dla ograniczonego zbioru danych, a ogólna ocena obu wariantów wyrażona zostanie na podstawie charakterystyk szumowych BER.

3.1 Charakterystyki bitowej stopy błędów

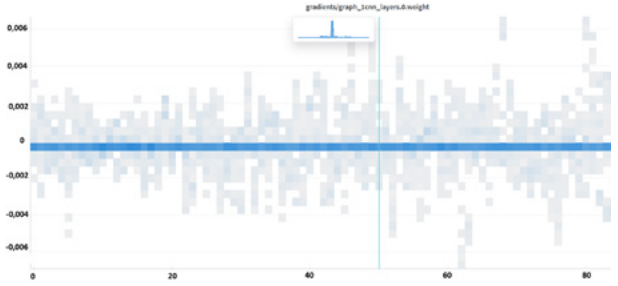
Wykresy charakterystyk bitowej stopy błędu dla dwóch najlepszych wyodrębnionych modeli przedstawiono na rys. 3 wraz z krzywą referencyjną klasycznego detektora MLSE (ang. *maximum likelihood sequence estimation*), który oznaczono linią przerywaną.



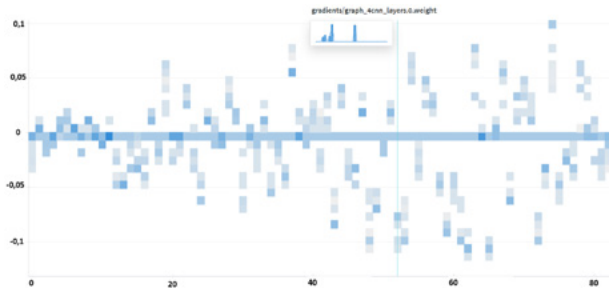
Rysunek 3: Charakterystyki BER dla testowanych modeli

Można zauważyć że wyższa skuteczność sieci CNN przetwarzającej próbki w układzie kartezjańskim, zauważalna podczas fazy treningowej, pozostała zachowana również w przypadku analizy zbioru danych testowych, a różnica pomiędzy testowanymi wariantami wyniosła prawie 2 dB.

Pogorszenie skuteczności może wynikać z przeskoków fazy, jakie występują w przypadku reprezentacji kątowej. Analiza gradientów, tj. wektorów zawierających pochodne funkcji straty dla wszystkich parametrów wewnętrznych sieci zdaje się potwierdzać to przypuszczenie. Na rys. 4-5 przedstawiono przykładowe wykresy gradientu, dostępne w oprogramowaniu W&B dla współczynników wagowych pojedynczej warstwy splotowej w obu badanych modelach (K) i (B) dla każdej iteracji (oś odciętych) w ramach fazy treningowej. W przypadku przetwarzania danych w układzie kartezjańskim zauważyć można skupioną linię niewielkich wartości gradientu utrzymującą się przez wszystkie iteracje fazy uczenia. Z kolei w przypadku reprezentacji biegunowej, rozróżnić można systematyczne rozrzuty skupionych wartości gradientu, co może sugerować odchyłkę parametrów sieci, powtarzającą się przez cały czas trwania treningu.



Rysunek 4: Wartości gradientu warstwy splotowej w trakcie fazy uczenia ze zbiorem danych (K)



Rysunek 5: Wartości gradientu warstwy splotowej w trakcie fazy uczenia ze zbiorem danych (B)

Należy również wspomnieć, że różnica pomiędzy detektorem CNN (K) a optymalnym detektorem MLSE, wynika wyłącznie z nieidealnego procesu estymacji współczynników kanału, realizowanego w przypadku klasycznego odbioru radiowego.

3.2 Parametry CNN w ujęciu radiokomunikacyjnym

Celem niniejszego artykułu jest zdefiniowanie sposobu przetwarzania danych w sieci CNN, który będzie skuteczny w ujęciu analizy sygnałów radiowych, choć niekoniecznie zachowujący spójność z mechanizmami wykorzystywanymi choćby w przetwarzaniu obrazów. Stąd też, poza przeprowadzeniem badań obejmujących dobór struktury danych wejściowych sieci, postanowiono przeanalizować również parametry opisujące działanie modeli CNN. Oprócz rozmiaru filtru, w ramach którego realizowana jest operacja splotu, wyróżnić można jeszcze dwa dodatkowe parametry dotyczące przetwarzania CNN: stride oraz dilation.

Pierwszy z parametrów definiuje krok z jakim przesuwany jest filtr na płaszczyźnie elementów/danych warstwy splotowej. Tym samym zwiększenie kroku do wartości przekraczającej rozmiar filtru może skutkować całkowitym pominięciem pojedynczych próbek lub grup próbek. Wyniki prezentujące wpływ zwiększenia odstępów pomiędzy kolejnymi obszarami obliczenia splotu na końcową skuteczność sieci neuronowej zaprezentowano w tabeli 3.

Tabela 3: Skuteczność modeli ML w zależności od zwiększenia parametru kroku

stride	*BCE*	BER_{val}
1	0,0127	0,0103
2	0,1264	0,1204
3	0,2271	0,3753
4	0,2418	0,4246

Analiza wartości BER oraz *BCE* wskazuje na spadek skuteczności detekcji GMSK proporcjonalny do rosnącego kroku pomiędzy filtrowanymi obszarami danych. Co istotne, odstęp odpowiadający pojedynczemu symbolowi (4 próbki sygnału) spowodował całkowity brak zbieżności trenowanych modeli.

Drugim z podstawowych parametrów CNN jest dilation i determinuje on ziarnistość filtru - tj. które z elementów objętych filtrem zostaną wyodrębnione do obliczenia splotu. W przypadku analizy sygnałów z pojedynczą nośną, byłoby to równoznaczne z usunięciem części rzeczywistej lub urojonej próbek sygnału - dlatego też modyfikacja tego parametru w przypadku problemu badawczego zdefiniowanego w ramach niniejszego artykułu byłaby bezzasadna. Nie można jednak wykluczyć, że parametr ten znalazłby uzasadnienie w przypadku analizy sygnałów rozmieszczonych w siatce zasobów, np. OFDM.

4. PODSUMOWANIE

Zaprezentowane wyniki pokazały, że wykorzystanie narzędzi ML może stanowić alternatywę dla klasycznych algorytmów odbioru radiowego, jednak powinno być poparte zrozumieniem analizowanego problemu i struktury przetwarzanych danych. Czesto nieprzemyślane stosowanie upowszechnionych rozwiązań z obszaru modelowania sieci neuronowych może okazać się niewłaściwe dla podejmowanego zagadnienia, co w niektórych przypadkach będzie skutkować drastycznym pogorszeniem skuteczności projektowanych modeli.

LITERATURA

[1] 3GPP, TS 136 116. 2015. “LTE Evolved Universal Terrestrial Radio Access (E- UTRA) Relay Radio Transmission and Reception.”

[2] Biewald, Lukas. 2020. “Experiment Tracking with Weights and Biases”

[3] Chen, X. Liu, M. Gui, G. Adebisi, B. Gacanin, H. Sari, H. 2021. “Complex deep neural network based intelligent signal detection methods for OFDM-IM systems” Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit): 90–94.

[4] Ha, C.-B. Song, H.-K. 2018. “Signal detection scheme based on adaptive ensemble deep learning model” IEEE Access 6: 21342–21349.

[5] Luong, T. V., Ko, Y., Vien, N.A., Nguyen D.H.N., Matthaiou, M. 2019. “Deep learning-based detector for OFDM-IM” IEEE Wireless Communications Letters 8 (4): 1159–1162.

[6] Luong, T. V., Ko, Y. Matthaiou, M. Vien, N.A. Le, M.-T. Ngo, V.-D. 2021. “Deep learning-aided multicarrier systems” IEEE ansactions on Wireless Communications 20 (3): 2109-2119.

[7] Morocho-Cayamcela, M.E. Njoku, J.N. Park, J. Lim, W. 2020. “Learning to communicate with autoencoders: Rethinking wireless systems with deep learning” International Conference on Artificial Intelligence in Information and Communication: 308–311.

[8] Olejniczak, Alicja. Cwalina, Krzysztof K. Sadowski, Jarosław. Stefański, Jacek. 2023. “Analiza skuteczności detekcji sygnałów GMSK w kanale z propagacją wielodrogową przy użyciu metod głębokiego uczenia”. Konferencja Radiokomunikacji i Teleinformatyki.

[9] Olejniczak, Alicja. Sadowski, Jarosław. 2023. “Głębokie uczenie do korekcji fazy sygnałów GMSK w rzeczywistym środowisku wewnątrzbudynkowym”. Konferencja Radiokomunikacji i Teleinformatyki.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

HIERARCHICZNE, NIEHIERARCHICZNE I GĘSTOŚCIOWE ALGORYTMY ANALIZY SKUPIEŃ W GRUPOWANIU SYGNAŁÓW RADAROWYCH

HIERARCHICAL, NON-HIERARCHICAL AND DENSITY-BASED CLUSTER ANALYSIS ALGORITHMS IN A RADAR SIGNALS GROUPING

Łukasz Rybak[1]; Janusz Dudczyk[2]; Jakub Olszewski[3]; Piotr Mazur[4]

[1] Wojskowa Akademia Techniczna, Warszawa, lukasz.rybak@wat.edu.pl
[2] Wojskowa Akademia Techniczna, Warszawa, janusz.dudczyk@wat.edu.pl
[3] Wojskowa Akademia Techniczna, Warszawa, jakub.olszewski01@student.wat.edu.pl
[4] Wojskowa Akademia Techniczna, Warszawa, piotr.mazur03@student.wat.edu.pl

DOI: 10.15199/59.2025.4.8

Streszczenie**: Celem artykułu było określenie właściwości znanych algorytmów klasteryzacji pod kątem ich aplikacyjności w grupowaniu sygnałów radarowych. Użyte metody należały do zbiorów podejść: hierarchicznych, niehierarchicznych i gęstościowych. Zastosowano znormalizowany trójwymiarowy zbiór danych impulsów pochodzących z trzech typów źródeł emisji. Uzyskane wyniki ujawniły wpływ różnych strategii grupowania impulsów na liczbę tworzonych klastrów oraz ich kształt w trójwymiarowej przestrzeni cech.**

Abstract**: A purpose of article was to determine the properties of known clustering algorithms in the terms of their applicability in a clustering radar signals. The methods used belonged to sets of approaches: hierarchical, non-hierarchical and density-based. A normalized three-dimensional dataset of pulses from three types of emission sources was used. The results revealed an influence of different pulse clustering strategies on a number of clusters formed and their shape in the three-dimensional feature space.**

Słowa kluczowe**: grupowanie gęstościowe, grupowanie hierarchiczne, grupowanie niehierarchiczne, rozpoznanie radioelektroniczne, sygnały radarowe**

Keywords**: density-based clustering, electronic intelligence, hierarchical clustering, non-hierarchical clustering, radar signals**

1. WSTĘP

Każda z istniejących gałęzi uczenia maszynowego ML (ang. *Machine Learning*) jest charakteryzowana poprzez immanentne cechy szczególne, które pozwalają na ich jednoznaczną wzajemną separację. Grupowanie danych stanowi jeden z obszarów tematycznych, będący przedmiotem zainteresowania naukowców, pochylających się nad eksploracją danych z zastosowaniem metod uczenia nienadzorowanego UL (ang. *Unsupervised Learning*) [14]. Cecha, która w sposób wyraźny odróżnia grupę algorytmów uczenia maszynowego typu UL, zawarta jest w rodzaju danych, które są przez te metody przetwarzane. W tej materii kluczowym aspektem jest zastosowanie danych, którym wcześniej nie zostały przypisane etykiety. Innymi słowy, są to dane, którym nie nadano zewnętrznego znaczenia, a występują w nich jedynie natywne zależności, ukryte w rozkładach poszczególnych atrybutów i korelacjach zachodzących pomiędzy tymi deskryptorami. Celem zastosowania opisywanych podejść jest ekstrakcja wzorców, które są ukryte w danych podanych na wejście przedmiotowych metod [10]. Uszczegóławiając, cel implementacji opisywanych algorytmów można postrzegać jako tworzenie w heterogenicznej przestrzeni cech maksymalnie homogenicznych grup elementów lub selekcję relewantnych atrybutów zbioru danych.

Ze względu na strategię stosowaną w celu utworzenia grup obiektów, algorytmy analizy skupień dzielone są na kilka klas. Podejścia, które na podstawie badania podobieństwa pomiędzy elementami zbioru danych tworzą hierarchię klastrów, klasyfikowane są do zestawu - którego nazwa nawiązuje do filozofii działania jego mechanizmów - grupowania hierarchicznego HC (ang. *Hierarchical Clustering*) [4]. Popularnym w tym katalogu jest algorytm AGNES (ang. *AGglomerative NESting hierarchical clustering*). Druga z opisywanych w tym artykule gałęzi drzewa metod analizy skupień nosi nazwę grupowania niehierarchicznego NHC (ang. *Non-Hierarchical Clustering*) [7]. W przeprowadzonych badaniach, z katalogu NHC, analizie poddano algorytm *k*-średnich (ang. *k-means*). Ostatnim z rozpatrywanych w niniejszym artykule zbiorów metod grupowania jest zestaw narzędzi, bazujących na gęstości przestrzeni cech, kryjący się pod akronimem DC (ang. *Denisty Clustering*). Istotną kwestią procesu tworzenia klastrów, z zastosowaniem strategii z grupy DC, jest ich kształt. Powstałe, na bazie gęstych fragmentów przestrzeni cech, grupy obiektów bazy danych, we wspomnianej przestrzeni cech często mogą być aproksymowane wielokątami/bryłami wklęsłymi. Z gałęzi DC pod uwagę wzięto algorytm DBSCAN (ang. *Density-Based Spatial Clustering of Applications with Noise*) [5] i OPTICS (ang. *Ordering Points To Identify the Clustering Structure*) [1].

W głównym nurcie myślowym sygnały radarowe kojarzone są z aplikacjami wojskowymi, co jest zbieżne z genezą akronimu RADAR (ang. *RAdio Detection And Ranging*) sięgającą roku 1940. Wówczas marynarka wojenna Stanów Zjednoczonych wprowadziła do obiegu przedmiotowy skrót. Systemy radiolokacyjne poprzez

przetwarzanie sygnałów radarowych są w stanie realizować zdefiniowany im cel. Jest nim pozyskiwanie informacji o obiektach odbijających, wyemitowane przez różne źródła promieniowania, zaburzenia pola elektromagnetycznego. Zakres informacji o elementach, stanowiących cele i znajdujących się w otaczającym system radarowy środowisku, w szczególności dotyczy ich położenia [2]. Spoglądając na podział działalności wywiadowczej prowadzonej w środowisku sygnałów elektromagnetycznych SIGINT (ang. *SIGnal INTelligence*) i biorąc pod uwagę, że opisywane sygnały radarowe nie są stosowane do przenoszenia komunikatów, tematyka niniejszego artykułu osadzona jest w dziedzinie rozpoznania elektronicznego ELINT (ang. *ELectronic INTelligence*) [6]. W kontekście bogactwa źródeł promieniowania elektromagnetycznego na współczesnym polu walki, aktualnym i istotnym wyzwaniem dla systemów walki radioelektronicznej jest proces rozplatania sygnałów radarowych.

Autorzy przedmiotowego artykułu zagadnienie, wskazane w ostatnim zdaniu poprzedniego akapitu skorelowali z problematyką analizy skupień, którą przybliżyli we wcześniejszych paragrafach dokumentu. Pokłosiem było zdefiniowanie celu prac badawczych, który został skoncentrowany na określeniu właściwości algorytmów AGNES, *k*-means, DBSCAN i OPTICS pod kątem ich aplikacyjności w procesie grupowania sygnałów radarowych. W drodze do jego osiągnięcia zastosowano metodę analizy porównawczej. Przedmiotem niniejszych prac analitycznych były zobrazowania przestrzeni cech, zawierających skupienia utworzone z użyciem wymienionych metod. Na ich podstawie określono właściwości wybranych podejść oraz wykazano, że najlepsze rezultaty w procesie grupowania sygnałów radarowych uzyskały algorytmy klasteryzacji gęstościowej.

2. PRZEGLĄD LITERATURY

Mimo, iż zastosowaniu algorytmów grupowania przyświeca ten sam cel, to różne algorytmy analizy skupień, w zależności od grupy metod, którą reprezentują: hierarchiczną, niehierarchiczną lub gęstościową, charakteryzują się różnymi strategiami tworzenia klastrów. Z tego często wynikają istotne różnice w rezultatach ich działania. W kolejnych podrozdziałach zostały w sposób syntetyczny scharakteryzowane wybrane podejścia analizy skupień.

2.1. AGNES

Jedną z metod grupowania hierarchicznego jest metoda aglomeracyjna (*ang. AGglomerative NESting*). Zasada działania tego algorytmu polega na traktowaniu każdego elementu zbioru jako pojedynczego klastra. Następnie na podstawie najmniejszej odległości pomiędzy grupami tworzone są ich pary, co jest realizowane iteracyjnie do momentu utworzenia jednego skupiska zawierającego wszystkie rozpatrywane elementy [7, 8, 16]. Aglomeracyjne algorytmy grupowania danych są stosowane w różnych gałęziach analizy. Dowodem ich aplikacyjności jest użycie AGNES w quasi-optymalizacji zrównoważonego zużycia energii w bezprzewodowych sieciach czujników (*ang. Wireless Sensor Network*), czego realizacja została przedstawiona w [18].

2.2. *k*-means

Algorytm *k*-średnich (*ang. k-means*) jest jednym z najczęściej stosowanych podejść grupowania niehierarchicznego. Jego logika działania opiera się na dzieleniu zbioru danych na podobne grupy z wykorzystaniem matematycznej miary odległości od punktu centralnego. Często stosowaną przestrzenią metryczną jest ta, z zadaną metryką euklidesową. Proces przetwarzania danych jest poprzedzony zdefiniowaniem wartości parametru *k*, który określa liczbę klastrów, na jakie dane mają zostać podzielone [9, 13]. Algorytm *k*-means znajduje zastosowanie w wielu różnych dziedzinach, w tym w elektronice. Przykładem tego jest użycie przedmiotowej metody do identyfikacji urządzeń elektronicznych w inteligentnych budynkach [17]. W rezultacie, otrzymano wyniki klasyfikacji, które umożliwiły przypisanie urządzeń do odpowiednich grup na podstawie ich zużycia energii.

2.3. DBSCAN

DBSCAN (*ang. Density-Based Spatial Clustering of Applications with Noise*) jest algorytmem grupowania danych opartym na analizie gęstości obszarów przestrzeni cech. Metoda dzieli obszar o określonej gęstości na klastry, oddzielając je od regionów o mniejszej gęstości, skutecznie wyodrębniając grupy elementów o dowolnych kształtach i eliminując przy tym wartości odstające. DBSCAN wymaga dwóch parametrów wejściowych. Pierwszym jest maksymalny promień sąsiedztwa *eps*, oznaczający minimalną odległość, dzielącą dwie obserwacje w przestrzeni cech. Drugim wymaganym parametrem jest wartość minimalnej liczby obserwacji *minPts*, które mogą tworzyć klaster. Dobór odpowiednich wartości powyższych parametrów jest zależny od wykorzystywanego zbioru danych i ma znaczący wpływ na wyniki końcowe [5, 12, 15]. Algorytm znajduje zastosowanie w wielu obszarach, a jeden z praktycznych przykładów został przedstawiony w [12]. Autorzy przywołanego artykułu użyli DBSCAN do grupowania danych związanych z jakością powietrza w miastach. Rezultatem tych prac było wyodrębnienie obszarów o zróżnicowanej jakości powietrza, identyfikując przy tym regiony o wysokim poziomie zanieczyszczeń.

2.4. OPTICS

Kolejnym reprezentantem algorytmów bazujących na gęstości jest często stosowany OPTICS (*ang. Ordering Points To Identify the Clustering Structure*) [1]. W niektórych pozycjach literaturowych [3] określany jest jako rozszerzenie DBSCAN [5]. OPTICS w przeciwieństwie do DBSCAN zachowuje strukturę hierarchiczną klastrów przy zmiennym promieniu sąsiedztwa, co czyni go znacznie skuteczniejszym modelem przy analizie dużych zbiorów danych [1]. Podobnie jak DBSCAN, OPTICS wymaga zdefiniowania wartości dwóch parametrów wejściowych: *eps* i *minPts*. OPTICS znalazł zastosowanie w analizie danych z sensora LiDAR (ang. *Light Detection and Ranging*) przy mierzeniu wysokości obszaru leśnego w danym regionie. W artykule [19] przedstawiono rezultaty metody bazującej na OPTICS w aspekcie redukcji szumów w danych obrazowych, co umożliwiło bardziej dokładne przeprowadzenie analizy wysokości lasów, przy eliminacji wartości odstających, niż w przypadku użycia podejścia DBSCAN.

3. PRZEBIEG BADAŃ

W zrealizowanych badaniach wybrane i opisane w poprzednim rozdziale algorytmy zostały zaimplementowane w języku Python z zastosowana biblioteki *scikit-learn* [19]. W eksperymentach użyto zbiór danych rzeczywistych poddany wstępnemu przetwarzaniu opisujący trzy atrybuty impulsów sygnałów radarowych pochodzących z trzech emiterów różnego typu. Wymiarami analizowanej przestrzeni cech były częstotliwość nośna FREQ, okres powtarzania impulsu PRI oraz czas trwania PW. Baza danych została poddana procesowi normalizacji, który polegał na zmapowaniu zakresu wartości każdej z cech, z zakresu oryginalnego na przedział [0-1].

4. REZULTATY

Badania były skoncentrowane na analizie porównawczej zobrazowań przestrzeni cech, zawierających grupy utworzone z użyciem, opisanych w rozdziale drugim artykułu, metod analizy skupień. Wybrane algorytmy sparametryzowano zgodnie z wartościami z Tab. 1. Zastosowane wartości parametrów zostały ustalone w drodze przeprowadzenia kilku iteracji testów.

Tab. 1. Parametryzacja zastosowanych algorytmów

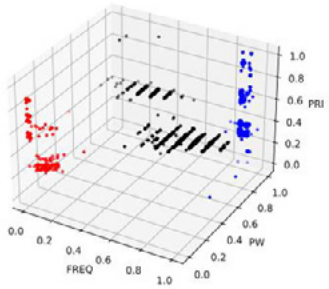
Algorytm	*Parametr*	*Wartość*
AGNES	*n_clusters*	3
k-means	*k*	3
DBSCAN	*eps*	0,25
	minPts	100
OPTICS	*eps*	0,45
	minPts	400

Wyniki działania algorytmów określane miarą jakości (*ang. silhouette_score*) przedstawiono w Tab. 2.

Tab. 2. Rezultaty klasteryzacji

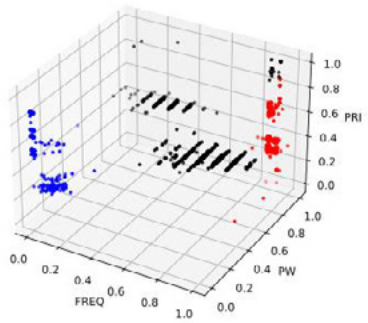
Algorytm	*silhouette_score*
AGNES	0,60
k-means	0,59
DBSCAN	0,59
OPTICS	0,70

Ponadto, poniżej zamieszczone zostały wizualizacje znormalizowanej, trójwymiarowej przestrzeni cech. Aglomeracyjna analiza skupień utworzyła trzy wyraźne grupy oznaczone kolorami: czerwonym, czarnym i niebieskim. Nie zostały zidentyfikowane punkty szumu [Rys. 1.].



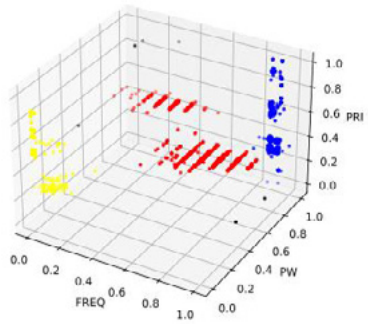
Rys. 1. Rezultat uzyskany algorytmem AGNES
Źródło: Opracowanie własne

Algorytm *k*-średnich, zgodnie ze zdefiniowaną wartością parametru $k = 3$, również utworzył trzy klastry punktów. Jednak dostrzegalna jest niejednorodność skupienia impulsów w górnym zakresie wartości wszystkich cech. Punkty szumu nie zostały oznaczone [Rys. 2.].



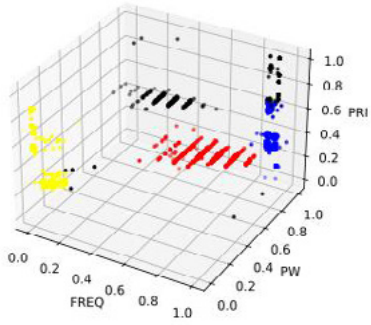
Rys. 2. Rezultat uzyskany algorytmem k-means
Źródło: Opracowanie własne

Algorytm DBSCAN, bez parametryzacji w aspekcie liczby tworzonych klastrów, zidentyfikował trzy wyraźne skupienia (żółte, czerwone, niebieskie), oznaczając wartości odstające w postaci punktów czarnych [Rys. 3.].



Rys. 3. Rezultat uzyskany algorytmem DBSCAN
Źródło: Opracowanie własne

Algorytm OPTICS, pozbawiony definicji docelowej liczby grup, wykrył trzy skupienia. Ponadto odseparował on wiele impulsów szumu (punkty czarne) [Rys. 4.].



Rys. 4. Rezultat uzyskany algorytmem OPTICS
Źródło: Opracowanie własne

Analiza danych przedstawionych w Tab. 2, prezentujących jakość użytych algorytmów oraz wizualizacje [Rys. 1-4] wykazały, iż dla wykorzystywanego zbioru danych, najwyższy wynik uzyskał jeden z algorytmów bazujących na gęstości – OPTICS, który osiągnął wartość miary *silhouette_score* na poziomie 0,70.

5. PODSUMOWANIE

W artykule w sposób syntetyczny opisano hierarchiczne, niehierarchiczne i gęstościowe algorytmy analizy skupień w kontekście ich potencjału do grupowania sygnałów radarowych. Dokonano implementacji poszczególnych metod z użyciem biblioteki *scikit-learn* dla języka programowania Python i przedstawiono rezultaty w postaci zobrazowań znormalizowanych trójwymiarowych przestrzeni cech. Uzyskane wyniki ukazały właściwości poszczególnych algorytmów.

Podejścia bazujące na gęstości poprawnie zidentyfikowały liczbę grup impulsów, która odpowiadała rzeczywistej liczbie źródeł emisji, z których one pochodzą. Zakładając, że w trakcie prowadzenia działań operacyjnych nie jest znana *a priori* liczba emiterów, algorytmy niewymagające definiowania liczby klastrów, do której należą metody gęstościowe, wykazują potencjał aplikacyjny w zagadnieniu rozplatania sygnałów radarowych. Analiza otrzymanych rezultatów pozwala również na sformułowanie wniosku, że metody bazujące na gęstości charakteryzuje potencjał aplikacyjny w filtracji danych ELINT.

Podsumowując, zagadnienie stosowania algorytmów analizy skupień w kontekście przetwarzania danych dotyczących impulsów radarowych stanowi badawczy konglomerat, który jest warty dalszych prac. Otrzymane rezultaty otworzyły kierunek dalszych badań. Potencjalna ścieżka rozwoju obejmuje zastosowanie innych algorytmów grupowania, w szczególności z katalogu podejść bazujących na gęstości lub im podobnych. Istotny wkład w stan wiedzy może zostać uzyskany przy użyciu bardziej rozbudowanych i zróżnicowanych zbiorów danych.

LITERATURA

[1] Ankerst Mihael, Breunig Markus M., Kriegel Hans-Peter, Sander Jörg. 1999. "OPTICS: Ordering Points to Identify the Clustering Structure". *ACM SIGMOD International Conference on Management of Data*, 28 (2) : 49-60.

[2] Bhatta Niraj Prasad, Geethapriya M. 2016. „RADAR and its Applications". *International Conference on Advanced Communication Control and Computing Technologies*, 9 (28) : 1-9.

[3] Drab Klaudia. 2015. „Rozprawa doktorska pt. Metody grupowania danych i ich wybrane modyfikacje dedykowane eksploracji danych eksperymentalnych". *Wydział Matematyki, Fizyki i Chemii, Uniwersytet Śląski*.

[4] Dudczyk Janusz. 2016. „Radar Emission Sources Identification Based on Hierarchical Agglomerative Clustering for Large Data Sets". *Journal of Sensors*, 2016 (1) : 1-9.

[5] Ester Martin, Kriegel Hans-Peter, Sander Jörg, Xu, Xiaowei. 1996. „A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*: 226-231.

[6] Gautam Vibhor, Shishodia Vikalp. 2022. „The E-Intelligence System". *arXiv: Electrical Engineering and Systems Science*.

[7] Gülağız Fidan Kaya, Şahin Suhap. 2017. „Comparison of Hierarchical and Non-Hierarchical Clustering Algorithms". *International Journal of Computer Engineering and Information Technology*, 9 (1) : 6–14.

[8] Kaufman Leonard, Rousseeuw Peter J. 1990. "Agglomerative Nesting (Program AGNES)". *Finding Groups in Data: An Introduction to Cluster Analysis*: 199-252.

[9] Kavlakoglu Eda, Winland Vanna. 2024. „What is k-means clustering?". *IBM.* Online: https://www.ibm.com/think/topics/k-means-clustering [Dostęp: 16.05.2025].

[10] Naeem Samreen, Ali Aqib, Anam Sania, Ahmed Munawar. 2023. "An Unsupervised Machine Learning Algorithms: Comprehensive Review". *International Journal of Computing and Digital Systems*, 13 (1) : 911-921.

[11] Pedregosa Fabian, Varoquaux Gaël, Gramfort Alexandre, Michel Vincent, Thirion Bertrand, Grisel Olivier, Blondel Mathieu, Prettenhofer Peter, Weiss Ron, Dubourg Vincent, Vanderplas Jake, Passos Alexandre, Cournapeau David, Brucher Matthieu, Perrot Matthieu, Duchesnay Édouard. 2011. "Scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research*. 12 (85) : 2825-2830.

[12] Shetty Chetan, Shedole Seema, Sowmya B. J., Nandalike Rajesh, Shivashankar Supreeth, Dayananda P., Rohith S., Vishwanath, Y., Ranjan Rajeev i Goud Venugopal. 2024. "A machine learning approach for environmental assessment on air quality and mitigation strategy," *Journal of Engineering*, 2024 (1) : 1-18.

[13] Singh Satya Prakash, Yadav Asmita. 2013. „Study of K-Means and Enhanced K-Means Clustering Algorithm". *International Journal of Advanced Research in Computer Science*, 4 (10) : 103–107.

[14] Theodoridis Sergios, Koutroumbas Konstantinos. 2009 (Wydanie IV). „Chapter 11 - Clustering: Basic Concepts". *Pattern Recognition*: 595-625.

[15] Trojanowski Łukasz. 2016. „Praca dyplomowa magisterska pt. Analiza położenia użytkownika sieci geospołecznościowej". *Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej, Katedra Automatyki i Robotyki, Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie.*

[16] Vaghela Rajdipsinh, Iyer Sailesh. 2022. "A Comparative Analysis of Clustering Algorithm". *ECS Transactions*, 107 (1) : 2435-2443.

[17] Zhang Guiqing, Li Yong, Deng Xiaoping. 2020. „K-Means Clustering-Based Electrical Equipment Identification for Smart Building Application". *Information*, 11 (1) : 1-18.

[18] Zhao Zhidong, Xu Kaida, Hui Guohua, Hu Liqin. 2018. "An Anergy-efficient Clustering Routing Protocol for Wireless Sensor Networks based on AGNES with Balanced Energy Consumption Optimization". *Sensors*, 18 (11) : 1-27.

[19] Zhu Xiaoxiao, Nie Sheng, Wang Cheng, Xi Xiaohuan, Wang Jinsong, Li Dong, Zhou Hangyu. 2020. „A Noise Removal Algorithm Based on OPTICS for Photon-Counting LiDAR Data". *IEEE Geoscience and Remote Sensing Letters*. 18 (8) : 1471-1475.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

Koncepcja wielokanałowego systemu wyboru kanału radiowego w oparciu o predykcję zajętości kanałów.

Concept of a Multichannel System for Radio Channel Selection using Channels Occupancy Prediction

Adrian Kubowicz[1]; Jerzy Łopatka[2]

[1] Wojskowa Akademia Techniczna, Instytut Systemów Łączności, Warszawa, adrian.kubowicz@student.wat.edu.pl
[2] Wojskowa Akademia Techniczna, Instytut Systemów Łączności, Warszawa, jerzy.lopatka@wat.edu.pl

DOI: 10.15199/59.2025.4.9

Streszczenie: **W artykule przedstawiono koncepcję wielokanałowego systemu radiokomunikacyjnego posiadającego funkcjonalność predykcji zajętości i wyboru kanału radiowego. Pozwoli to na wybór kanału z odpowiednio długim oknem czasowym w celu dokonania własnej transmisji w ramach dynamicznego dostępu do widma DSA (ang. Dynamic Spectrum Access). Rozwinięto koncepcję użycia sieci głębokiego uczenia typu RNN (ang. Recurrent Neural Network) do predykcji zajętości widma bądź określonego kanału.**

Abstract: **The article presents a concept of a multichannel radiocommunication system with occupancy prediction and radio channel selection functionality. This enables the selection of a channel with a sufficiently long time window for own transmission within the framework of Dynamic Spectrum Access (DSA). The concept further develops the use of deep learning recurrent neural network (RNN) for predicting spectrum or specific channel occupancy.**

Słowa kluczowe: **DSA, Radio, RNN, różnice czasowe z LSTM**

Keywords: **DSA, Radio, RNN, time differential LSTM**

1. WSTĘP

Oportunistyczny dostęp użytkownika wtórnego SU (ang. Secondary User) do kanału może odbywać się według możliwości dostępu do jednego lub do wielu kanałów [1]. W drugim przypadku, SU może mieć możliwość nasłuchu jednego bądź więcej kanałów jednocześnie, co zależy głównie od możliwości platformy sprzętowej, wyposażonej w równoległe tory odbiorcze lub rozwiązania opartego o radio zdefiniowane programowo (ang. Software Defined Radio). Jest to możliwość charakterystyczna dla stacji bazowych przy centralizowanych systemach. Dla zamkniętego systemu komunikacji, np. w zastosowaniach wojskowych i specjalnych, instalacja urządzenia końcowego typu SU z tą funkcjonalnością pozwoliłaby na jednoczesny nasłuch wielu urządzeń użytkowników pierwotnych PU (ang. Primary User), umożliwiając dodanie cech koncepcji radia kognitywnego do istniejącego systemu podlegającemu modernizacji. Wpisuje się to w ogólny trend, ponieważ z jednej strony obecnie trwają prace nad wbudowaniem inteligentnych rozwiązań do czołowej generacji telefonii 6G [2], ale także spotykane są koncepcje rozbudowy istniejących systemów np. 4G/5G. W takim podejściu, dotychczasowe urządzenia w rozważanym przypadku zachowałyby kompatybilność wsteczną. Łącze punkt-punkt pozwoliłoby zarówno na nasłuch i komunikację z urządzaniami PU jeśli takie wsparcie systemu jest wymagane, lub kognitywną transmisję o minimalnej interferencji z PU.

W [3] zaproponowano oportunistyczny dostęp do widma OSA bazujący na wzorach aktywności użytkownika pierwotnego. Ważnym elementem takiego systemu jest moduł predykcji, który musi umożliwiać bieżące uczenie się dynamicznie zmieniającego się wzorca . Ustalenie punktu przesunięcia się wzorca może zostać wykonane za pomocą metod korelacji, niemniej jednak z powodu statystycznego rozrzutu aktywności PU może to nie być możliwe. Zaproponowano zatem zastosowanie głębokiego uczenia do korelacji ciągu wzorca transmisji UP w kanale, tak aby móc dokonać predykcji kontynuacji ciągu według aktualnych pomiarów zajętości kanału. Jeśli odpowiednio długi czas do wystąpienia transmisji UP jest dostępny, SU może dokonać transmisji w danym kanale, w przeciwnym przypadku następuje odczekanie.

Zastosowanie metod uczenia maszynowego ML (ang. Machine Learning) i głębokiego uczenia DL (ang. Deep Learning) w dziedzinie dostępu do widma i ogólnie w systemach radiokomunikacyjnych stanowi logiczną konsekwencję rozwoju tych technologii. Dotyczy to zarówno ML które wyraźnie składają się z etapu ekstrakcji cech z sygnału które podlegają klasyfikacji bądź regresji przez rozwiązania takie jak SVM, KNN, DT etc., jak i sieci głębokich DL takich jak RNN, CNN, Autoenkodery czy Transformery [4]. W literaturze znajduje się wiele ich modyfikacji i adaptacji poprzez połączenia typu CNN-LSTM. Zaleta sieci głębokich ma być możliwość pracy z niejawną ekstrakcją cech, których z powodu złożoności opracowanie jak w ML stanowiło by wyzwanie, ponieważ wiąże wiedzę ekspercką z koniecznością adaptacji modelu [5].

Niemniej ML stanowią lżejszy obliczeniowo odpowiednich sieci głębokich tam gdzie to jest możliwe. Stąd w rozważanym przypadku uznano za zasadne podejście w którym łączy się szybkość algorytmu ekstrakcji cechy którą jest binarna informacja o występowaniu transmisji PU bądź nie, wraz z możliwościami uczenia i wnioskowania sieci głębokich w celu predykcji przebiegu komunikacji w najbliższym czasie.

Środowisko radiowe z powodu efektów wielotorowości, efektu Dopplera i ogólnie obserwowanej zmienno-

ści stanowi wyzwanie w projektowaniu algorytmów dostępu do widma. Należy założyć, iż nawet w idealnym przypadku nadawania przez PU tylko według określonego cyklicznego wzoru, odbiornik powinien obsługiwać rozkład losowy momentu odbioru ramki PU w danym kanale. Wtedy w predykcji wzoru transmisji PU, ograniczone zastosowanie mają znane metody przetwarzania cyfrowego sygnałów takie jak korelacja, czy też DTW(ang. Dynamic Time Warping). Dzięki zdolności uczenia, sieć głęboka poprzez odpowiedni zestaw danych uczących powinna nauczyć się dopuszczalnego rozkładu losowego danych i wygenerować najbardziej prawdopodobną predykcje przebiegu transmisji w czasie. O to założenie oparto koncepcję wielokanałowej predykcji zajętości. Celem jest wykorzystanie historycznych wzorów aktywności PU do predykcji zajętości kanałów w najbliższej przyszłości, kiedy w dalszym ciągu istnieje potrzeba dokonania transmisji przez SU

2. MODEL SYGNAŁU

W docelowym systemie jest M kanałów w których aktywność użytkowników nie jest ze sobą skorelowana. Źródło danych o aktywności użytkowników może zostać przetworzone na różne sposoby, zależnie od platformy sprzętowo programowej. Jeżeli dostępny jest strumień próbek IQ, należy poddać go obróbce cyfrowej, w typowym rozwiązaniu poddając go przetwarzaniu z FFT (ang. Fast Fourier Transform) uzyska się podstawę do detekcji energii sygnału ED (ang. Energy Detection). Wymagany jest próg detekcji, zmienność SNR (ang. Signal Noise Ratio) częściowo może być uwzględniona adaptacyjnie [6]. Stanowi to zagadnienie testowania hipotezy obecności lub braku transmisji PU, zwiększając złożoność metody wypracowania decyzji przy rozszerzeniu o celowe zakłócanie i wpływ efektów toru radiowego. Możliwe jest więc stosowanie np. filtrów dopasowanych do sygnału. Analizowany scenariusz zakłada jednak uproszczony model sygnału, ponieważ dokonując rozbudowany istniejącego systemu o moduł predykcji zajętości i wyboru kanału zakłada się, iż dostępne są już wyjścia np. RSSI (ang. Received Signal Strenght Indicator) czy też binarna informacja o detekcji ramki PU, nawet przy braku możliwości jej pełnej identyfikacji, biorąc pod uwagę metody szyfrowania i zabezpieczania. Jednakże zasadniczą cechą, jaką musi posiadać transmisja w kanale PU aby dokonać predykcji, jest cykliczność komunikacji na poziomie ramek, gdzie każda realizacja przesłanej ramki obarczona jest rozkładem losowym czasu wystąpienia. W literaturze istnieją koncepcje z użyciem DL które przyjmują strumień próbek sygnału traktując go wieloaspektowo, jednakże są one wymagające obliczeniowo zarówno na etapie wnioskowania oraz uczenia.

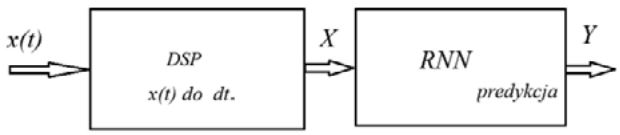
3. PRZETWARZANIE Z RNN

Problemem dla urządzeń końcowych w których istotny jest kompromis pomiędzy kosztem, wielkością fizyczną czy szczególnie poborem energii, jest złożoność obliczeniowa, która im mniejsza tym krótsze będzie opóźnienie powodowane przez etap predykcji dla danej platformy. Dotyczy to także etapu uczenia, jeśli jest to adaptacyjne/wzmacniane uczenie w trakcie działania urządzenia.

Wybrano konfigurację połączenia RNN, wariantu LSTM (ang. Long Short-Term Memory) bądź uproszczonego GRU (ang. Gated Recurrent Unit) ze wstępnym przetwarzaniem cechy - obecności transmisji. Złożoność obliczeniowa proponowanej sieci RNN liniowo zależy od długości wektora wejściowego, bądź ilości przetwarzanych próbek na sekundę dla danych w funkcji czasu. Zasadne jest zatem podzielenie strumienia danych, tak aby zmniejszyć złożoność obliczeniową samego RNN. Na Rys.(1) pokazano związany z tym problem, zmniejszenie częstotliwości próbkowania (czarne strzałki) na wejściu algorytmu spowoduje błąd synchronizacji w czasie i nie rozwiąże kwestii występowania długich przerw w transmisji, których ciągła obsługa przez RNN oznaczałaby pobór energii.

Rys. 1 Wizualizacja problemu próbkowania aktywności PU

Zaproponowano zatem modyfikację, polegającą na podaniu na wejście RNN informacji o czasach występowania transmisji PU (zielone strzałki na Rys.1). Aby umożliwić uczenie wzorców, czasy te przekształcane są na różnice w czasie. Tego rodzaju zaadresowanie problemu znajduje odniesienie w literaturze [7] czy też przekształcenie na serię symboli [8], w ogólnych zastosowaniach. Na Rys.2 przedstawiono schemat blokowy składający się z etapu przetwarzania cyfrowego DSP (ang. Digital Signal Processing) i RNN otrzymującego strumień informacji o aktywności w kanale.



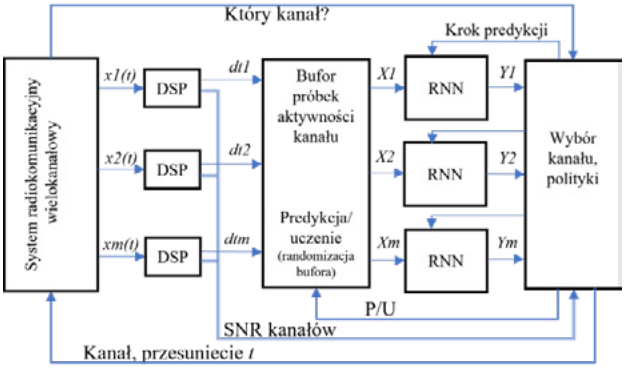
Rys. 2 Schemat przetwarzania z RNN

Wyjście RNN dokonuje predykcji czasu binarnego wystąpienia aktywności w kanale, co stanowi zredukowaną informację na rzecz zmniejszenia złożoności RNN. Z powodu stosowania funkcji *tanh* w RNN, dodatkowo dane do i z RNN muszą być normalizowane do zakresu [-1,1].

Sieci RNN analizują złożoność próbek w czasie, nie są jednak praktyczne tutaj metody douczania, co sprawdzono eksperymentalnie, stąd douczanie do nowego wzorca odbywa się poprzez kolejne uczenie, dzięki zredukowaniu złożoności w proponowanym podejściu.

4. MODUŁ PREDYKCJI

Rozwinięcie koncepcji predykcji z różnicami czasowymi i RNN na moduł wielokanałowy przedstawiono na Rys.3.



Rys. 3 Schemat blokowy modułu z predykcją wielokanałową

Informacja o odbiorze pakietów w danym kanale, RSSI, bądź strumień próbek zostaje poddany obróbce w bloku DSP. Na jego wyjściu znajduje się ciąg wartości oznaczających czasy pomiędzy kolejnymi aktywnościami w kanale. Przeciwnie do wejścia DSP, są to próbki aktywności bez równomiernego rzutowania w czasie. Na tym etapie możliwe jest adaptacyjne odrzucanie aktywności w kanale niezwiązanych z transmisją użytkownika. Bufor próbek dla każdego kanału umożliwia gromadzenie próbek aktywności. Podstawowym celem jest zgromadzenie próbek aktywności na potrzeby wykonania z nich predykcji na żądanie w nieznanym z góry momencie. Drugim celem jest zebranie próbek w celu wykonania uczenia RNN na żądanie z randomizacją próbek, która odbywa się według ograniczonego rozkładu normalnego o określonym odchyleniu standardowym, symulując rozkład losowy aktywności użytkownika kanału i ucząc RNN do jego tolerancji. Blok bufora i randomizacji ma wejście przełączające tryby predykcji i uczenia.

Bloki RNN mają zadane wektory próbek wejściowe *X* i wyjściowe *Y* o określonej długości, dopasowanej do ilości próbek aktywności, ponieważ przy predykcji ze stanu resetu wytrenowanej sieci RNN wymagana jest więcej niż jedna realizacja cyklu aktywności w buforze, jeśli dla oszczędności energii przetwarzania RNN nie są aktualizowane o dane wejściowe każdego kroku DSP. Wyjście *Y* to spodziewana kontynuacja *X*, stanowiąca predykcję którego długość musi być dobrana do wymaganego okna czasu predykcji, stanowiącego tutaj sumę elementów *Y*, uwzględniającego czas predykcji wszystkich kanałów.

Ostatni blok na żądanie inicjuje predykcję kanałów. Wyjścia predykcji *Y* zawierają przewidywane próbki aktywności kanałów, co powoli, wraz z implementacją polityk, na ustalenie i przekazanie do systemu informacji o numerze kanału i przesunięciu w czasie do rozpoczęcia transmisji. Dodatkowym zagadnieniem jest synchronizacja wyniku predykcji i międzyczasów zdarzeń na wyjściu DSP. Podjęcie decyzji o wyborze kanału zależy zatem od wygenerowanej predykcji i wyniku pomiaru SNR bądź innej informacji o aktywności w kanale w chwili zakończenia predykcji, aby odrzucić kanały w których predykcja nie zgadza się z bieżącą sytuacją. Tego rodzaju walidacja może być także dodatkowym mechanizmem, przeprowadzana poprzez cykliczne uruchamianie predykcji w celu sprawdzenia jej jakości, w przypadku istotnego braku pokrycia predykcji z zebraną rzeczywistą kontynuacja ciągu aktywności powinno nastąpić ponowne uczenie dla danego kanału.

5. SYMULACJE

Jako że zaproponowano użycie RNN w celu predykcji zajętości kanałów i możliwości wyboru kanału do transmisji SU z najmniejszym prawdopodobieństwem kolizji z PU, przeprowadzono symulacje z RNN pomijając bloki DSP. Założono pseudolosowe ciągi aktywności w trzech kanałach:

$$t1 = \{407, 820\} \quad (1)$$

$$t2 = \{230, 320, 831\} \quad (2)$$

$$t3 = \{120{,}300{,}380, 450{,}602{,}1100\} \quad (3)$$

które zostały przekształcone na ciąg różnic w czasie:

$$dt1 = \{407, 413\} \quad (4)$$

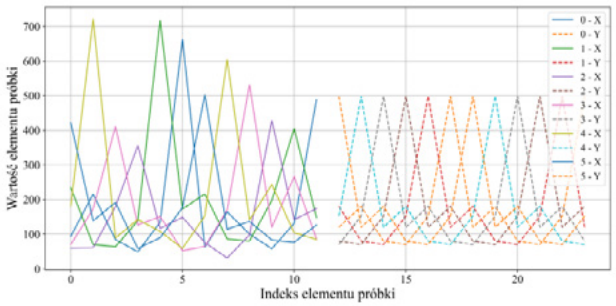
$$dt2 = \{230, 90, 510\} \quad (5)$$

$$dt3 = \{120{,}180{,}80, 70{,}152{,}498\} \quad (6)$$

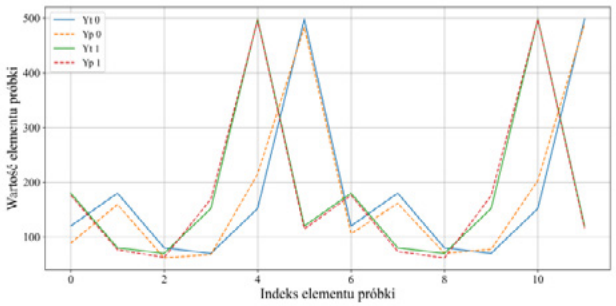
Wartości są w milisekundach, niemniej działanie algorytmu jest niezależne od jednostek. Następnie ustalono sieć RNN – LSTM, korzystając z pakietów Keras i Tensorflow z następującymi ustawieniami: liczba elementów ukrytych H=64, długość wektora wejścia do przeprowadzenia kroku wnioskowania w przód X=12 i długość wyjścia – wektora predykcji Y=12, z zadaniem regresji czyli wyjściową funkcją liniową. Niezbędne są tutaj 3 niezależne zestawy sieci i danych uczących. Następnie przygotowano dane uczące, wektory *dt1-dt3* poddano metodzie przesunięcia kolejnych elementów aby móc nauczyć sieć rozpoznawać wektor z dowolnego punktu startu, co skutkuje wygenerowaną liczbą wierszy próbek uczących równą liczbie elementów wektorów *dt*. Wiersze próbek uczących są rozszerzone o oczekiwaną dla danego przesunięcia ciągu jego kontynuację, czyli ilość elementów każdej z próbek uczących wynosi tutaj $X+Y=24$. Następnie próbki – liczbę wierszy należy zwielokrotnić poprzez randomizację każdego wiersza, założono eksperymentalnie iż konieczne jest 25 krotne randomizowanie każdego wiersza, co oznacza np. dla *dt3 6 x25* wierszy. Randomizacja wielokrotnie zwiększa złożoność danych uczących, ale jest to niezbędne. Ustalone odchylenie standardowe rozkładu normalnego randomizacji $\sigma=0.2$, jest to parametr który powinien być wyznaczany w systemie drogą obserwacji aktywności w kanałach. Należy zaznaczyć, iż randomizacji podlega tylko część wierszy próbek *X, Y* stanowi kopię wartości z *dt*.

Rys.4 przedstawia 6 realizacji próbek uczących z szumem *X* dla danych uczących *dt3*, podobnie dla *dt1* i *dt2*. Z kolei Rys.5 przedstawia efekt nałożenia wyniku predykcji *Y* i wartości oczekiwanych dla dwóch nowych próbek nie uczestniczących w uczeniu, wygenerowanych

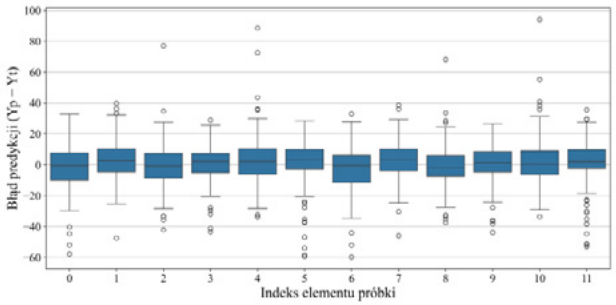
z ww. regułami, pokrywają się one. Rys.6 stanowi potwierdzenie działania RNN dla kompletnych danych testujących o wielkości uczących, czyli 6 *x* 12 *x* 25=1800 próbek (wierszy). Niebieskie prostokąty zawierają rozkład 50% wartości, średnio nie przekraczające zakresu 12.5ms, kółka to pojedyncze odstające wartości, porównywalne do najmniejszych wartości wektora *dt3*, jednakże w najgorszym przypadku jest to 20% największego elementu *dt3*.



Rys. 4 Pierwsze 6 próbek danych uczących dt3, X i Y

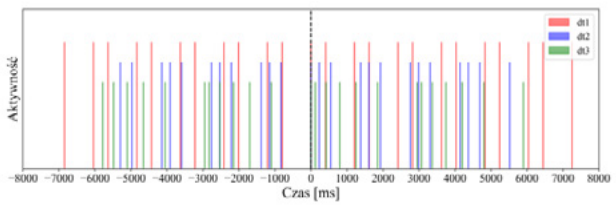


Rys. 5 Porównanie wartości wyniku predykcji (Yp0 Yp1) i oczekiwanych testowych (Yt0 Yt1)



Rys. 6 Statystyczne porównanie wartości wyniku predykcji i oczekiwanych testowych (Y)

Uzyskując w ten sposób potwierdzenie działania LSTM w tym zastosowaniu, przedstawiono na Rys.7 sposób postępowania z próbkami dotychczas zebranymi (*Xm*) i wynikającymi z predykcji (*Ym*), dla określonej realizacji zebranych próbek trzech kanałów.



Rys. 7 Rozkład w czasie aktywności trzech kanałów i predykcji

Próbki są wyrównane względem środka zakresu, jako punkt *0ms*, czyli moment uruchomienia predykcji, w docelowym rozwiązaniu należy podjąć zagadnienie obliczania ich wyrównania oraz uwzględniania czasów aktywności użytkowników.

6. PODSUMOWANIE

Przedstawione podejście polaczenia bloku DSP z RNN umożliwi rozbudowę systemu radiokomunikacyjnego o funkcjonalność wyboru kanału o najmniejszym prawdopodobieństwie kolizji własnej z transmisjami użytkowników pierwotnych kanałów.

Wyniki działania LSTM na testowych próbkach potwierdzają działanie w tym zastosowaniu, niemniej docelowa implementacja wymaga opracowania metod min. synchronizacji, czy też optymalizacji sposobu generowania danych uczących. Dodatkowo należy rozważyć adaptacyjne pomijanie bloków RNN w przypadku wykrycia stosunkowo prostych wzorów aktywności, jak *dt1* i *dt2*. Na dalszym etapie planowane jest zastosowanie metody w rzeczywistym środowisku radiowym, także dla bardziej złożonych wzorów aktywności. Na Rys.7 można odnaleźć kolejne „białe plamy" w aktywności kanałów. Ograniczeniem zastosowania metody są jawnie cykliczne wzory transmisji, bez możliwości uwzględnienia cyklostacjonarności wyższego rzędu, w zamian jest to rozwiązanie o umiarkowanym koszcie obliczeniowym sieci DL, dając możliwość rezygnacji z projektu dostosowanego klasycznego algorytmu predykcji aktywności w kanałach radiowych na rzecz rozwiązania bazującego na uczeniu.

LITERATURA

[1] Gaaloul, F., et al. 2012. „Opportunistic Spectrum Access in Cognitive Radio Based on Channel Switching". *CROWNCOM*, 270–274.

[2] Bai, W., et al. 2025. „Multi-User Opportunistic Spectrum Access via Multi-Head Self-Attention and Multi-Agent DRL". *Sensors* 25 (7): 2025.

[3] Arshid, K., et al. 2020. „Primary User Traffic Pattern Based Opportunistic Spectrum Handoff". *Appl. Sci.* 10 (5): 1674.

[4] Pan, G., et al. 2025. „Spectrum Prediction With Deep 3D Pyramid Vision Transformer Learning". *IEEE Trans. Wireless Commun.* 24 (1): 509–525.

[5] Wong, L. J., A. J. Michaels. 2022. „Transfer Learning for RF Machine Learning: Taxonomy and Survey". *Sensors* 22 (4): 1416.

[6] Arjoune, Y., et al. 2018. „Spectrum Sensing: Enhanced Energy Detection via Noise Measurement". *CCWC*, 828–834.

[7] Zhu, Y., et al. 2017. „What to Do Next: Modeling User Behaviors by Time-LSTM". *IJCAI*, 3602–3608.

[8] Elsworth, S., S. Güttel. 2020. „Time Series Forecasting Using LSTM Networks: A Symbolic Approach". *arXiv*:2003.05672.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

KONTRADYKTORYJNE INTERFERENCJE NEUROINTERFEJSU RADIOWEGO 5G-NR

5G-NR RADIO NEUROINTERFACE ADVERSARIAL INTERFERENCES

Krzysztof K. Cwalina; Piotr Rajchowski; Jarosław Sadowski

Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk
{kkcwalina, piorajch, jarsadow1}@pg.edu.pl

DOI: 10.15199/59.2025.4.10

Streszczenie: W niniejszym artykule przedstawiono charakterystykę oraz badania wrażliwości neurotechnik estymacji kanału radiowego w interfejsie 5G-NR na kontradyktoryjne interferencje generowane z użyciem dwóch metod. Przedstawiono architekturę uzyskanego estymatora kanału radiowego zgodnego ze standardem 5G-NR oraz jego analizę w zakresie zmiennych wartości SNR oraz ISNR w obliczu występujących interferencji zsynchronizowanych z występowaniem sygnałów pilotowych dla kanału współdzielonego PDSCH w łączu w dół.

Abstract: In this article the characteristics and sensitivity studies of the 5G-NR radio channel estimation neurotechnics to adversarial interference generated using two methods are described. The architecture of the obtained 5G-NR radio channel estimator and its analysis in the range of variable SNR and ISNR values in the face of interference synchronized with the occurrence of pilot signals for the downlink PDSCH shared channel are presented.

Słowa kluczowe: 5G-NR, estymacja kanału, głębokie uczenie, interfejs radiowy, sieci neuronowe

Keywords: 5G-NR, channel estimation, deep learning, neural networks, radio interface

1. WSTĘP

W dobie wykorzystania systemów radiokomunikacyjnych bazujących na technice OFDM (Orthogonal Frequency Division Multiplexing), niezwykle istotnym i cały czas aktualnym zagadnieniem jest rozwój warstwy fizycznej interfejsu radiowego i technik odbioru radiowego. Wpływ wielodrogowości, przesunięcia Dopplerowskiego, czasu trwania odpowiedzi impulsowej definiują zmienny w czasie, selektywny częstotliwościowo kanał radiowy, którego wpływ na jakość odbioru powinien być minimalizowany. Dodatkowo, występujące interferencje oraz celowe zakłócenia sygnałów radiowych, sprawiają, iż detekcja, identyfikacja oraz minimalizacja ich wpływu jest cały czas tematem prowadzonych prac badawczych.

Mając na uwadze niniejsze oraz wzrost zainteresowania rozwiązaniami opartymi na uczeniu maszynowym, a w szczególności na głębokich sieciach neuronowych, naturalnie powstały różne formy interfejsu radiowego zawierającego elementy, bloki przetwarzania sygnałów oraz sprzęgniętych torów nadawczych i odbiorczych w formie sieci neuronowych [12]. Stopień ich integracji w warstwie fizycznej warunkuje powstanie tzw. neurointerfejsu radiowego [2], który może przewyższać efektywnością pracy klasyczne rozwiązania [11], także dla interfejsu 5G-NR [10]. Zastosowanie neurointerfejsu radiowego wskazywane jest także jako jeden z podstawowych elementów w sieciach kolejnych generacji opartych na algorytmach sztucznej inteligencji [7].

W takim przypadku, uzasadnionym wydaje się realizowanie badań w zakresie szeroko pojętego zakresu stosowalności neurointerfejsu radiowego, także w zakresie interferencji i zagłuszania. Przy czym, oprócz znanej dotychczas problematyki generowania sygnałów interferujących, stosowanie neurotechnik umożliwia realizację nowych typów ataków, tzw. kontradyktoryjnych (adversarial attacks), które mogą wpływać na efektywność pracy modeli neuronowych [6], w tym także modeli stosowanych jako grupa neurotechnik odbioru radiowego [16].

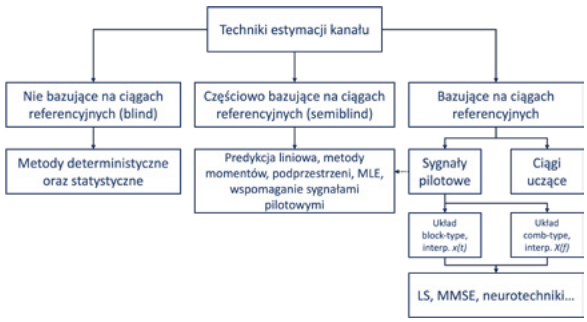
W ramach niniejszych prac opracowano neuronowy estymator kanału radiowego zgodnego ze standardem 5G-NR i przebadano zakres jego stosowalności, w szczególności jego wrażliwość, w obliczu występujących kontradyktoryjnych interferencji sygnałów pilotowych dla kanału współdzielonego łącza w dół zgodnego ze standardem 5G-NR. Podobne badania i zakres tematyczny podejmowany był np. w [3], przy czym niejednoznaczne jest czy badania zrealizowano dla łącza w górę, czy dla łącza w dół, a przedstawione wektory danych wejściowych wskazują na oddzielną analizę siatki zasobów dla części rzeczywistej i urojonej odbieranych sygnałów. W [13] badania skupiono na komunikacji w paśmie fal milimetrowych, stosując także mechanizm atencji dla komunikacji V2X (Vehicle-to-Everything). W [4] badania oparto na frameworku dla systemów 6G dla nieiteracyjnie generowanej interferencji sieci splotowej połączonej z siecią resztkową (Residual). W [9] zastosowano modulację QAM (Quadrature Amplitude Modulation) wraz z układem pilotów block-type dla trzech wariantów generowanych nieiteracyjnych interferencji płytkich sieci neuronowych. Warto także zaznaczyć, iż często zbiory danych uczących były znacząco ograniczone, co wynikać mogło prawdopodobnie ze złożoności zrealizowanych modeli i można zakładać, że ich adaptacyjność w takim przypadku będzie także ograniczona dla innych scenariuszy użycia. W [5] opisano splotowy estymator kanału radiowego dla 11 warstw splotowych i zwe-

ryfikowano zakres jego stosowalności, przy czym nie rozważano w tychże badaniach sygnałów interferujących.

Mając na uwadze wskazane ograniczenia, niniejszy artykuł można uznać za rozwinięcie istniejących już prac w tym obszarze badawczym, który stanowi element dalszych prac opisanych początkowo w [14].

2. CHARAKTERYSTYKA NEUROTECHNIK ODBIORU RADIOWEGO

Jednym z elementów szeroko pojętych technik odbioru radiowego, w tym także neurotechnik, szczególnie w komunikacji radiowej opartej na technice OFDM, jest przetwarzanie sygnałów radiowych celem estymacji kanału radiowego [8, 15]. Techniki te możemy w ogólności podzielić na trzy grupy [10], tj. metody bazujące na ciągach referencyjnych, metody częściowo bazujące na ciągach referencyjnych (semiblind) oraz metody nie bazujące na ciągach referencyjnych (blind) co szczegółowo przedstawiono na rys. 1.



Rys.1. Podział technik estymacji kanału dla interfejsu radiowego opartego na technice OFDM.

W niniejszych pracach skupiono się na interfejsie 5G-NR co naturalnie determinuje grupę technik estymacji kanału radiowego bazującą na ciągach referencyjnych. W przypadku analizowanego łącza w dół, zgodnie ze standardem umieszczane są w siatce zasobów sygnały referencyjne DMRS (Demodulation Reference Signal). Mając na względzie przedstawiony sposób grupowania, można zdefiniować podgrupę określoną jako metody klasyczne, interpretowalne [15], np. najmniejszych kwadratów LS (Least Square), estymatora MMSE (Minimum Mean-Square Error), liniowego estymatora LMMSE (Linear MMSE), czy uproszczonej interpolacji w dziedzinie czasu i częstotliwości [8]. Drugą podgrupę stanowią nieinterpretowalne neurotechniki, oparte na np. adaptacji sieci do przetwarzania obrazów [3], autoenkoderach, realizujących wybrane grupy bloków przetwarzania sygnałów lub większość warstwy fizycznej [12], głębokich sieciach rekurencyjnych [17].

Niewątpliwie, udowodniono dotychczas, iż istnieją zalety stosowania neurotechnik odbioru radiowego, które dla dla wybranych profili kanału radiowego mogą uzyskiwać zbliżone, lub lepsze wyniki estymacji kanału (rozumianej jako minimalizację błędu średniokwadratowego, czy bitowej stopy błędów) względem praktycznych implementacji estymatorów MMSE. Dodatkowo pod względem złożoności obliczeniowej mogą konkurować z implementacjami estymatora MMSE i umożliwiają zrównoleglanie obliczeń i ich efektywną implementację sprzętową.

Oprócz zalet stosowania neurotechnik odbioru radiowego mają one wiele ograniczeń, które póki co wpływają na ich zakres stosowalności w praktycznych rozwiązaniach na poziomie warstwy fizycznej interfejsu radiowego. Wyróżnić można m.in. ograniczoną interpretowalność modeli, ograniczoną adaptacyjność wynikającą np. ze zmienności kanału radiowego, występowania i zmiennych w czasie interferencji czy nawet sposobu organizacji transmisji w warstwie fizycznej, czy też wrażliwość na różnego rodzaju ataki [6] wpływające na efektywność pracy modeli sieci neuronowych.

2.1 Kontradyktoryjne interferencje

Uwzględniając aktualny stan literatury, ograniczony jest zbiór prac badawczych dotyczących podatności sieci neuronowych służących do estymacji kanału radiowego na wszelkiego typu zakłócenia - w tym zamierzone. Zagadnienie kontradyktoryjnych sygnałów, które mogą wpływać na degradację efektywności pracy sieci neuronowych jest zagadnieniem znanym i cały czas badanym w innych specjalizacjach. Naturalnie, coraz popularniejsze stosowanie metod neuronowych w tym także w warstwie fizycznej interfejsu radiowego, determinuje potrzebę realizacji prac biorąc pod uwagę specyfikę ich zastosowań także w warstwie fizycznej interfejsu radiowego.

W ogólności przyjmuje się, iż dany model sieci neuronowej ma modelować pewną wielowymiarową, nieliniową funkcję F

$$F : X \to Y \tag{1}$$

gdzie X to macierz danych wejściowych, a Y to macierz danych wyjściowych. W takim przypadku proces uczenia można w ogólności opisać jako

$$\theta_{(n+1)} = \theta_n - \alpha \cdot \nabla \mathcal{L}(F(\theta_n, X)) \tag{2}$$

gdzie α to współczynnik uczenia, $\mathcal{L}$ to funkcja kosztu, a θ to parametry modelu w kroku n uczenia. Następnie w rzeczywistych warunkach pracy danego modelu dane wejściowe X^* w ogólności mogą być (a w praktyce często będą) inne niż te, które stosowano w procesie uczenia X tj. $X^* \neq X$. Jak wspomniano wcześniej w takim przypadku należy zweryfikować zakres stosowalności danego modelu (np. w zakresie innego profilu kanału radiowego, czy interferencje), który może wpływać negatywnie na efektywność jego pracy w nowych warunkach.

W takim przypadku zasadnym jest rozpatrywanie celowych ataków i wprowadzanie celowych interferencji dla estymatora kanału radiowego spełniających warunek

$$\arg\max_{X^*} \mathcal{L}\left(F(\theta_n, X^*)\right), \tag{3}$$

czyli wspomniane sygnały kontradyktoryjne. Zagadnienie atakowania sieci neuronowych jest przedmiotem aktualnie prowadzonych badań w literaturze, które grupowane są na ataki z pełną wiedzą o modelu (white box), z częściową wiedzą o modelu (grey box), czy z brakiem wiedzy o modelu (black box), których skuteczność zależy silnie od zakresu wiedzy jaki posiadamy o atakowanym modelu [1].

W ramach niniejszych prac analizowana będzie grupa ataków z pełną wiedzą o modelu, co znajduje swoje uzasadnienie także w praktyce, bowiem (podobnie jak w opisywanych pracach) modele wyuczane są na znanych i standaryzowanych profilach kanałów radiowych. Co powoduje, iż można utworzyć cyfrowy odpowiednik neuronowego estymatora kanału i przygotować interferencje kon-

tradyktoryjne, które będą stosowane już na docelowym neurointerfejsie radiowym.

W ramach ataków z pełną wiedzą o modelu można zdefiniować nieiteracyjną interferencję FGSM (Fast Gradient Sign Method Attack)

$$X_{adv} = X + \epsilon \cdot sign(\nabla \mathcal{L}\left(F\left(X, Y\right)\right)), \tag{4}$$

gdzie ϵ jest współczynnikiem amplitudy danej interferencji oraz iteracyjną interferencję BIM (Basic Iterative Method Attack)

$$X_{adv} = X + \beta \cdot sign(\nabla \mathcal{L}\left(F\left(X_{adv}, Y\right)\right)), \tag{5}$$

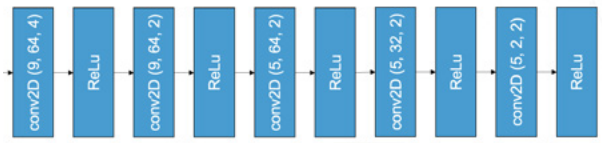
gdzie β jest współczynnikiem kroku zmian amplitudy danej interferencji generowanej iteracyjnie.

Mając na względzie omawianą charakterystykę neurotechnik odbioru radiowego, zdecydowano się zrealizować estymator neuronowy dla interfejsu radiowego 5G-NR i określić m.in. efektywność i wrażliwość proponowanej głębokiej sieci neuronowej na wspomniane interferencje kontradyktoryjne.

3. BADANIA NEUROTECHNIK ODBIORU RADIOWEGO

Na tym etapie prac, zrealizowano splotową, głęboką sieć neuronową (GSN) do estymacji parametrów kanału radiowego. Warto zaznaczyć, iż jednym z ograniczeń realizacji prac jest brak jednoznacznie zdefiniowanych reguł przy projektowaniu neuronowych estymatorów kanału radiowego. Stąd, opracowywana architektura GSN realizowana była poprzez rozbudowywanie kolejnych warstw splotowych bazując na wynikach przedstawianych w [3] przy założeniu, iż dla danego zbioru danych wartości średnie dla kolejnych wartości SNR (Signal-to-Noise Ratio) maleją oraz dane wejściowe stanowią zespolone wartości elementów siatki zasobów w paśmie podstawowym. Na rys. 2 przedstawiono architekturę proponowanej, wynikowej splotowej GSN, przy czym parametry uczenia były następujące:

- współczynnik uczenia: 0,0001;
- maksymalna liczba iteracji: 20;
- liczba danych (realizacji siatek zasobów): 30 000;
- fragmentacja zbioru danych: 64.



Rys.2. Architektura proponowanego estymatora kanału radiowego dla interfejsu radiowego 5G-NR.

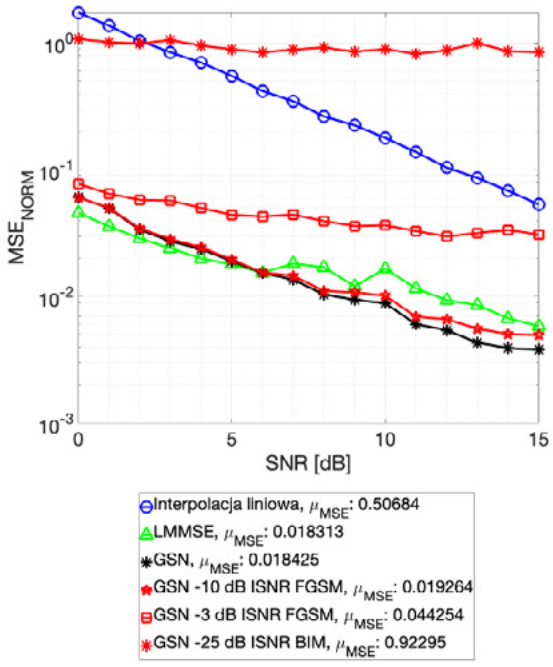
Ze względu na mnogość możliwych konfiguracji interfejsu 5G-NR opierając się na wynikach prezentowanych w [14], przyjęto następujące parametry transmisji dla kanału współdzielonego PDSCH (Physical Data Shared Channel) zgodne ze standardem 3GPP dla 5G-NR:

- profil dla warunków NLOS, z rozkładem zaników Rayleigh'a, 23-odczepowy, znormalizowany względem rozrzutu opóźnienia 50 ns - modelowany jako linia opóźniająca Tapped Delay Line: TDL-A;
- odchyłka dopplerowska = 50 Hz;
- raster częstotliwościowy = 30 kHz
- 51 bloków zasobów RB (Resource Blocks);
- transmisja typu SISO;
- dodatkowo wybrane parametry: MappingType A, modulacja QPSK, DMRS jednosymbolowy wraz z dodatkowym symbolem DMRS;
- częstotliwość próbkowania: 30,72 MHz.

Dodatkowo, na tym etapie badań przyjęto następujące założenia:

- brak błędów na etapie synchronizacji czasowo-częstotliwościowej;
- minimalizowano wartości mocy interferencji przy maksymalizacji błędu estymatora;
- kontradyktoryjne interferencje są generowane wyłącznie w zasobach symboli DMRS.

Na rys. 3 przedstawiono uzyskane wyniki znormalizowanego błędu średniokwadratowego różnicy modułu emulowanej zespolonej wartości oraz wynikowej zespolonej wartości współczynników transmitancji kanału radiowego dla bloków zasobów RB dla wybranych metod estymacji kanału radiowego w interfejsie 5G-NR. Dla porządku przedstawiono także metody klasyczne. Model testowany był dla tego samego profilu kanału dla jakiego był uczony - by ograniczyć wpływ zmiennych parametrów i móc wskazać głównie wpływ interferencji kontradyktoryjnych, a nie zmienności środowiska propagacyjnego.



Rys.3. Wyniki efektywności wybranych metod estymacji kanału radiowego dla interfejsu radiowego 5G-NR.

Zgodnie z oczekiwaniami estymator GSN uzyskał średnie wyniki zbliżone do zrealizowanego estymatora MMSE, przy czym powyżej 5 dB wartości SNR widoczny jest wzrost efektywności jego pracy i mniejsze wartości MSE o 0.004 co ma przełożenie na zysk rzędu 1-2 dB. Wskazane wyniki pogorszyły się o ok. 0,5 dB dla scenariusza generowania nieiteracyjnych interferencji kontradyktoryjnych FGSM o wartości -10 dB ISNR (Interference-to-Signal-to-Noise-Ratio). Zagłuszanie ustrukturyzowaną

interferencją symboli DMRS o wartości -3 dB ISNR wpłynęło na znaczące (powyżej 10 dB dla analizowanych zakresów SNR) pogorszenie efektywności pracy GSN, które można uznać właściwie za brak przydatności tegoż estymatora.

Szczególną uwagę warto zwrócić na wygenerowaną iteracyjnie interferencję BIM, która nawet dla wartości -25 dB ISNR skutecznie spowodowała całkowitą rozbieżność uzyskiwanych wyników z GSN nawet dla wysokich wartości SNR. W tym przypadku (5) przyjęto wartości $\beta = 0,0001$ i maksymalną liczbę iteracji równą 300 celem uzyskania stosownego ISNR. Dodatkowo, zweryfikowano, iż ograniczenie interferencji tylko do połowy sygnałów pilotowych nie wpływa na efektywność ataku dla analizowanych danych. Warto mieć jednak na względzie, iż w rzeczywistych warunkach propagacyjnych - nawet przy znajdującym się źródle sygnału interferującego w bliskim otoczeniu odbiornika - efektywność wskazanych wektorów kontradyktoryjnych będzie w ogólności gorsza. Uzyskane wyniki w takim przypadku można uznawać za górną granicę podatności modeli na ataki z użyciem emisji radiowych, co należałoby zweryfikować na drodze badań pomiarowych i może stanowić element przyszłych prac wraz z weryfikacją metod przeciwdziałaniu tymże atakom [3].

4. PODSUMOWANIE

W ramach niniejszych prac przedstawiono charakterystykę neurotechnik odbioru radiowego i problematykę kontradyktoryjnych interferencji dla neurointerfejsu radiowego. W ramach tego obszaru prac opracowano estymator kanału radiowego dla interfejsu radiowego 5G-NR oparty na głębokich, splotowych sieciach neuronowych, wykazano efektywność jego pracy i podatność na dwa rodzaje kontradyktoryjnych interferencji.

Oczywiście, opracowywane są metody przeciwdziałaniu atakom kontradyktoryjnym, np. poprzez kompresję uzyskanego modelu [3], włączenie próbek kontradyktoryjnych już w trakcie uczenia - co może jednak powodować wzrost złożoności architektury GSN, czy nawet zwielokrotnianie modeli pracujących równolegle. Ze względu na obszerność podejmowanej tematyki, w ramach przyszłych prac można rozważać realizację minimalizacji wpływu kanału i dalszej integracji w warstwie fizycznej 5G-NR, badań w zakresie detekcji i minimalizacji wpływu interferencji dla neuronowych modułów toru odbiorczego oraz weryfikację wyników na podstawie pomiarów.

5. FINANSOWANIE

Przedstawione prace badawcze zrealizowano w ramach projektu „Definiowalny programowo układ do wykrywania i kompensowania wpływu zakłóceń wewnątrz i zewnątrz systemowych oraz transmisji niepożądanych w interfejsie radiowym 4G-LTE, 5G-NR" finansowanego ze środków krajowych przez Narodowe Centrum Badań i Rozwoju w ramach programu Lider XIV, umowa nr LIDER14/0116/2023 z dnia 13.02.2024 r.

LITERATURA

[1] Abusaid O.M., Alaqtaa M. R. 2024. "Toward Security Enhancement for Deep Learning-Based Channel Estimation Model Under Adversarial Attacks in Next-G Wireless Networks'. IEEE 4th International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA).

[2] Cammerer S., Aoudia F.A., Hoydis J., Oeldemann A., Roessler A., Mayer T., Keller A. 2023. "A Neural Receiver for 5G NR Multi-User MIMO". ArXiv.

[3] Catak F. O., Kuzlu M., Catak E., Cali U., Guler O.. 2022. "Defensive Distillation-Based Adversarial Attack Mitigation Method for Channel Estimation Using Deep Learning Models in Next-Generation Wireless Networks'. IEEE Access.

[4] Cedbro, A. N. 2023. "Robust Neural Receiver in Wireless Communication', Praca magisterska, Linkoping University, 2023.

[5] Gupta R., Gupta J. 2022. "Channel Estimation of 5G New Radio Using Convolutional Neural Network". 8th International Conference on Signal Processing and Communication (ICSC). Indie.

[6] Kuznetsov P., Riley E., Ted X., Humza I., Raul P., Noah G., Shannon S. 2021. "Kontradyktoryjne uczenie maszynowe". Napędy i Sterowanie. Wydawnictwo "Druk-Art" SC, nr 7/8.

[7] Letaief K. B., Chen W., Shi Y., Zhang J., Zhang Y.-J. A. 2019. "The Roadmap to 6G: AI Empowered Wireless Networks". IEEE Communications Magazine.

[8] Magiera, Jarosław. 2023. "A Simplistic Downlink Channel Estimation Method for NB-IoT". Applied Sciences, nr 23.

[9] Mendonça M. O. K., Diniz P. S. R., Morales J.M., Frossard P. 2024. "Adversarial Training for Jamming-Robust Channel Estimation in OFDM Systems'. IEEE Open Journal of Signal Processing.

[10] Mousa M. A. S., Taman A. I. A., Hassan A. M., i Zekry A., „Deep Learning Channel Estimation for OFDM 5G Systems with Different Channel Models", Wireless Personal Communications 128, nr 4, 2023.

[11] Olejniczak A., Cwalina K. K., Sadowski J., Stefański J. 2023. "Analiza skuteczności detekcji sygnałów GMSK w kanale z propagacją wielodrogową przy użyciu metod głębokiego uczenia". Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne.

[12] O'Shea T., Hoydis J. 2017. "An Introduction to Deep Learning for the Physical Layer'. IEEE Transactions on Cognitive Communications and Networking.

[13] Qi, X., Liu Y., Ye Y. 2024. "Attention-Enhanced Defensive Distillation Network for Channel Estimation in V2X Mm-Wave Secure Communication". Sensors.

[14] Rajchowski P. 2024. "Examination of 5G NR, LTE, and NB-IoT Radio Interfaces and Their Vulnerabilities to Interference". Journal of Telecommunications and Information Technology.

[15] van de Beek J.-J. 1998. "Synchronization and Channel Estimation in OFDM Systems", Rozprawa doktorska, Szwecja.

[16] Ye, Youjie, Yunfei Chen, and Mingqian Liu. 2022. "Multiuser Adversarial Attack on Deep Learning for OFDM Detection'. IEEE Wireless Communications Letters 11 (12): 2527–31.

[17] Yong L., Hua Y., Dai X., Yao H., Yang X. 2019. "ChanEstNet: A Deep Learning Based Channel Estimation for High-Speed Scenarios". IEEE International Conference on Communications (ICC), Chiny.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

WPŁYW WYMIAROWOŚCI DANYCH ROZPOZNANIA RADIOELEKTRONICZNEGO NA DOKŁADNOŚĆ ALGORYTMÓW KLASYFIKACJI

AN IMPACT OF ELECTRONIC INTELLIGENCE DATA DIMENSIONALITY ON AN ACCURACY OF CLASSIFICATION ALGORITHMS

Piotr Mazur[1], Łukasz Rybak[2], Janusz Dudczyk[3], Jakub Olszewski[4]

[1] Wojskowa Akademia Techniczna, Warszawa, piotr.mazur03@student.wat.edu.pl
[2] Wojskowa Akademia Techniczna, Warszawa, lukasz.rybak@wat.edu.pl
[3] Wojskowa Akademia Techniczna, Warszawa, janusz.dudczyk@wat.edu.pl
[4] Wojskowa Akademia Techniczna, Warszawa, jakub.olszewski01@student.wat.edu.pl

DOI: 10.15199/59.2025.4.11

Streszczenie: **Celem artykułu było określenie wpływu wymiarowości danych rozpoznania radioelektronicznego na dokładność algorytmów klasyfikacji. W badaniach użyto zbiór impulsów, pochodzących z trzech różnych typów źródeł emisji elektromagnetycznej, opisanych zestawem atrybutów: częstotliwość nośna, czas trwania i okres powtarzania. W eksperymentach zastosowano dwa wybrane klasyfikatory: *k* najbliższych sąsiadów i maszynę wektorów nośnych. Wykazano, że wymiarowość danych ELINT wpływa na dokładność algorytmów klasyfikacji.**

Abstract: **The purpose of article was to determine the influence of electronic intelligence data dimensionality on the accuracy of classification algorithms. The study used a set of pulses, coming from three different types of electromagnetic emission sources, described by a set of attributes: carrier frequency, duration and repetition period. Two selected classifiers were used in the experiments: *k* nearest neighbor and support vector machine. The dimensionality of ELINT data has been shown to affect the accuracy of classification algorithms.**

Słowa kluczowe: **ELINT, KNN, rozpoznanie radioelektroniczne, SVM, wymiarowość.**

Keywords: **dimensionality, electronic intelligence, ELINT, KNN, SVM.**

1. WSTĘP

Analizując zobrazowania danych, opisujących dowolny fragment otaczającej rzeczywistości można dostrzec, że wraz ze wzrostem ich wymiarowości trudniejszym staje się podjęcie decyzji dotyczącej informatywności poszczególnych kombinacji przetwarzanych cech. Przedstawiona prawidłowość może być powiązana ze sformułowaniem, pochodzącym ze źródła [11]. W przywołanej pracy [11] stwierdzono, że zadaniami prostymi są wizualizacje jednowymiarowych danych w postaci osi liczbowej, dwuwymiarowych danych w formie płaszczyzny czy danych trójwymiarowych jako przestrzeni. Jednak tworzenie i późniejsze analizowanie zobrazowań danych, których wymiar *n* jest większy niż 3, stanowi zagadnienie nietrywialne.

Istotnym zadaniem realizowanym w ramach prowadzenia rozpoznania radioelektronicznego jest klasyfikacja źródeł promieniowania elektromagnetycznego. Rejestrowana aktywność, w wyniku przetwarzania przez sensor, często reprezentowana jest między innymi w postaci wielowymiarowego wektora cech, opisującego pojedynczy impuls, za pośrednictwem zbioru atrybutów ***PDW*** (ang. *Pulse Description Word*) [3]. W wymienionym katalogu parametrów wspomnianej struktury danych można wyróżnić następujące cechy impulsu: amplitudę AMP (ang. *Amplitude*), częstotliwość nośną RF (ang. *Radio Frequency*), czas trwania PW (ang. *Pulse Width*) oraz okres powtarzania impulsów PRI (ang. *Pulse Repetition Interval*) [10]. Biorąc pod uwagę fakt występowania na współczesnym polu walki bogatej aktywności elektromagnetycznej, obejmującej nie tylko interesujące sygnały radarowe, ale również sygnały, stanowiące szum [5], ważnym aspektem w drodze do efektywnej klasyfikacji impulsów jest zastosowanie referencyjnej bazy danych uczących. Stanowi ona kluczowy czynnik, wpływający na skuteczność algorytmów predykcyjnych, których celem jest określenie źródła emisji poszczególnych impulsów. Wspomnianą bazę danych można utożsamiać z macierzą o wymiarach *n* x |***PDW***|, gdzie *n* symbolizuje liczbę znanych elementów.

Celem prac badawczych opisanych w przedmiotowym artykule było określenie wpływu wymiarowości danych rozpoznania radioelektronicznego na efektywność algorytmów klasyfikacji. Założeniem eksperymentu było zastosowanie dwóch wybranych klasyfikatorów, spośród katalogu popularnych metod oraz dwu- i trójwymiarowych konfiguracji przestrzeni cech, utworzonych na podstawie rzeczywistego zbioru danych rozpoznania radioelektronicznego, którego atrybutami były częstotliwość nośna impulsu, czas jego trwania oraz okres powtarzania. Klasyfikacja realizowana była w trójelementowym zbiorze etykiet: 1, 2, 3.

2. PRZEGLĄD LITERATURY

Analizując zagadnienia teoretyczne z dziedziny sztucznej inteligencji, można dostrzec, że tytułowe algorytmy klasyfikacji wpisują się w gałąź nadzorowanego

uczenia maszynowego. Ważną cechą tych podejść jest fakt stosowania, w procesie uczenia się modelu, zestawu znanych danych treningowych. Jednym z aspektów zbiorów danych jest ich wymiarowość, którą należy postrzegać jako liczbę cech, opisujących obiekty zgromadzone w bazie. Połączenie wymienionych powyżej kwestii kreśli problem badawczy dotyczący wpływu wymiarowości danych na dokładność klasyfikatorów. W dalszej części rozdziału przedstawiono dwa wybrane algorytmy klasyfikacji, które zastosowano w przeprowadzonych badaniach.

2.1 Maszyna wektorów nośnych

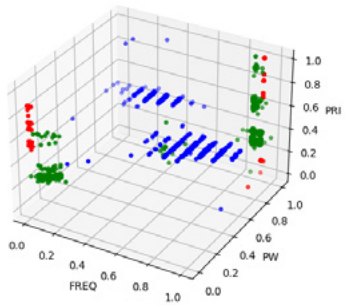
SVM (ang. *Support Vector Machine*) należy do zestawu powszechnie stosowanych metod klasyfikacji danych. Maszyna wektorów nośnych wykazuje zdolność do rozwiązywania problemów predykcyjnych zarówno klas separowalnych liniowo jak i nieliniowo w przestrzeni cech. Opisywana metoda w wyniku przetwarzania danych definiuje granicę decyzyjną w postaci linii lub hiperpłaszczyzny. Wymienione elementy na etapie rozpoznawania stosowane są do określenia etykiet klas analizowanych danych [2]. Najlepszą granicą decyzyjną nazwać można obiekt, dla którego odległość od najbliżej niego zlokalizowanych elementów poszczególnych klas jest jak największa. Istotnym w przypadku niniejszego mechanizmu jest fakt, iż w chwili, gdy rozkład danych nie jest liniowo rozdzielny, koniecznym jest dodanie większej liczby wymiarów przetwarzanej przestrzeni cech [7]. W ramach artykułu [1] autorzy stwierdzili, że SVM radzi sobie lepiej ze zjawiskiem *„klątwy wymiarowości”* danych niż klasyfikator KNN.

2.2 Algorytm *k* najbliższych sąsiadów

Podejście KNN (ang. *k-nearest neighbors*) stanowi jeden z najbardziej popularnych algorytmów klasyfikacji. Na przełomie kilku ostatnich lat opisywana metoda doczekała się wielu modyfikacji, poprawiających jej wydajność. Metoda *k* najbliższych sąsiadów, w oparciu o podobieństwo elementów przestrzeni cech, w pierwszym kroku wyszukuje *k* sąsiadów, położonych w najmniejszej odległości względem nieznanej próbki. W następnym etapie, bazując na liczności poszczególnych etykiet w *k* elementowym zbiorze, algorytm określa klasę, do której przypisywana jest nieznana dotychczas próbka [6]. Algorytm KNN jest obciążony tak zwaną *„klątwą wymiarowości”* [8]. Wraz ze wzrostem liczby wymiarów, przestrzeń cech rozrzedza się coraz bardziej. Oznacza to, że próbki danych są od siebie znacznie oddalone w przestrzeni cech, a znalezienie istotnych, najbliższych sąsiadów staje się trudniejsze. Wzrost wymiarowości powoduje również fakt, iż odległości między punktami stają się coraz bardziej wyrównane, ponieważ atrybuty gwarantujące rozróżnialność stanowią coraz mniejszą część ogólnego wymiaru przestrzeni cech. Ostatnią kwestią jest rosnący koszt obliczeniowy, w szczególności występujący w przetwarzaniu dużych złożonych zbiorów danych [4].

3. REALIZACJA BADAŃ

W badaniach skoncentrowanych na określeniu wpływu wymiarowości danych rozpoznania radioelektronicznego na efektywność algorytmów klasyfikacji zastosowano implementacje wybranych algorytmów, udostępniane w bibliotece języka programowania Python - *scikit-learn* [9]. Użyty zbiór danych, zawierał 3358 wektorów, opisujących pojedyncze impulsy. Jak wspomniano we wstępie, analizie poddano 3 elementów zbiór cech ***PDW***: częstotliwość nośna, czas trwania oraz okres powtarzania. W ramach wstępnego przetwarzania danych, wartości wszystkich cech zostały poddane procesowi normalizacji, w wyniku którego różne zakresy wartości poszczególnych atrybutów zbioru impulsów zostały zmapowane na przedział od 0 do 1. Pierwszy eksperyment został poświęcony klasyfikacji z użyciem różnych dwuwymiarowych przestrzeni cech: PW(RF), PRI(RF) oraz PRI(PW). Druga część badań obejmowała klasyfikację w trójwymiarowej przestrzeni cech. Na Rys. 1. zwizualizowano rozkład impulsów poszczególnych źródeł emisji w znormalizowanej trójwymiarowej przestrzeni cech. Na poniższym zobrazowaniu etykiety zostały zilustrowane w postaci kolorystycznej: kolor czerwony – źródło 1, kolor zielony – źródło 2 oraz kolor niebieski – źródło 3.



Rys. 1. Zobrazowanie danych wejściowych w znormalizowanej trójwymiarowej przestrzeni cech. Źródło: Opracowanie własne

Wybrane algorytmy, opisane w rozdziale drugim, zostały sparametryzowane zgodnie z danymi w Tab. 1.

Tab. 1. Parametryzacja algorytmów klasyfikacji. Źródło: opracowanie własne

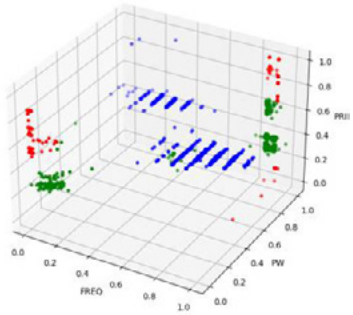
Klasyfikator	*Parametr*	*Wartość*
SVM	*kernel*	rbf
	random_state	42
KNN	*n_neighbors*	50

W przypadku klasyfikatora SVM zdecydowano się na domyślną jego konfigurację w obrębie biblioteki. Klasyfikator KNN został sparametryzowany w oparciu o analizę, przedstawioną w pozycji [12].

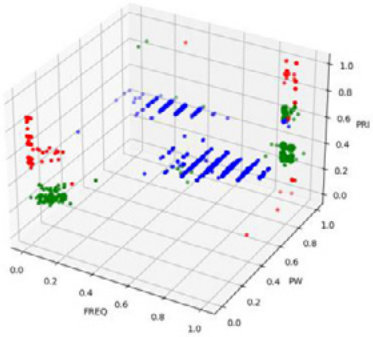
4. REZULTATY BADAŃ

Rezultaty zrealizowanych badań zostały przedstawione w postaci wizualizacji trójwymiarowych przestrzeni cech, zawierających punkty, których kolor reprezentował etykietę źródła emisji rozpoznaną przez klasyfikator. Na Rys. 2.-5. przedstawiono zobrazowania przestrzeni cech z etykietami impulsów przewidzianymi przez metodę SVM. Rys. 6.-9. obrazują etykiety impulsów, rozpoznane przez algorytm *k* najbliższych sąsiadów. Wizualizację referencyjną, przedstawiającą rzeczywiste etykiety poszczególnych impulsów, stanowi Rys. 1., zamieszczony w rozdziale 3. Ponadto dla każdego wykonanego procesu klasyfikacji, został określony, w formie procen-

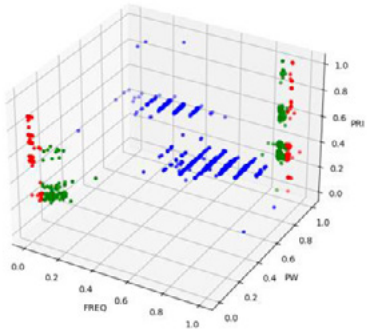
towej, wynik dokładności, stanowiący stosunek liczby poprawnie rozpoznanych etykiet do liczby wszystkich prób klasyfikacji. Wspomniane rezultaty, zostały zawarte w Tab. 2. Na rysunkach 2-9, etykiety zostały przedstawione za pomocą kolorów: czerwony – źródło 1, zielony – źródło 2, niebieski – źródło 3.



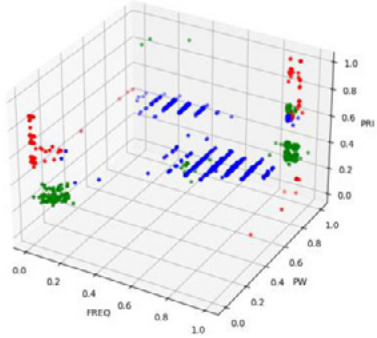
Rys. 2. Efekt klasyfikacji SVM dla zależności dwuwymiarowej PRI od RF. Źródło: Opracowanie własne



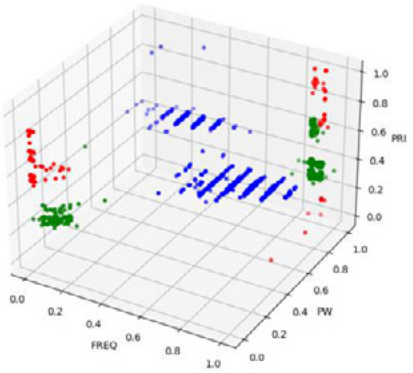
Rys. 3. Efekt klasyfikacji SVM dla zależności dwuwymiarowej PW od RF. Źródło: Opracowanie własne



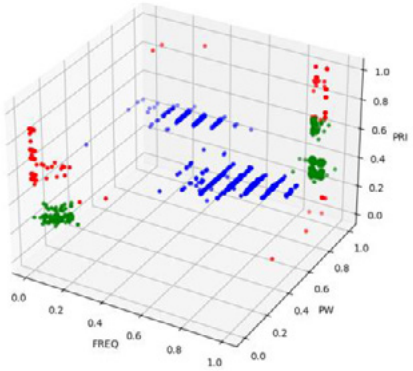
Rys. 4. Efekt klasyfikacji SVM dla zależności dwuwymiarowej PW od PRI. Źródło: Opracowanie własne



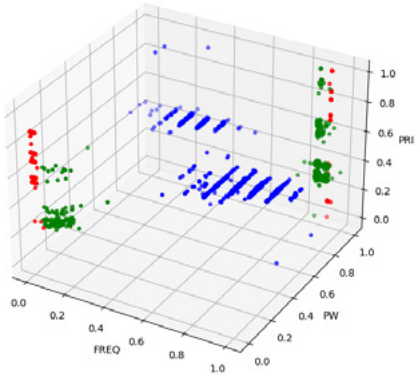
Rys. 5. Efekt klasyfikacji SVM dla zależności trójwymiarowej PRI od PW i RF. Źródło: Opracowanie własne



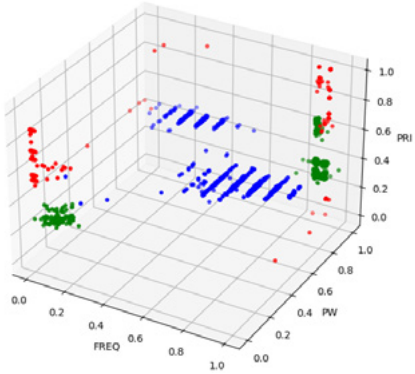
Rys. 6. Efekt klasyfikacji KNN dla zależności dwuwymiarowej PRI od RF. Źródło: Opracowanie własne



Rys. 7. Efekt klasyfikacji KNN dla zależności dwuwymiarowej PW od RF. Źródło: Opracowanie własne



Rys. 8. Efekt klasyfikacji KNN dla zależności dwuwymiarowej PW od PRI. Źródło: Opracowanie własne



Rys. 9. Efekt klasyfikacji KNN dla zależności trójwymiarowej PRI od PW i RF. Źródło: Opracowanie własne

Tab. 2. Wyniki dokładności klasyfikatorów SVM oraz KNN. Źródło: Opracowanie własne

	SVM	KNN
PRI od RF	98,0%	98,1%
PW od RF	96,0%	99,1%
PW od PRI	97,6%	98,4%
PRI od PW i RF	99,2%	98,4%

Na podstawie zobrazowań wyników, przedstawionych na rysunkach od 2 do 9, widocznym jest, iż w zależności od badanych przypadków, wyniki klasyfikacji uzyskiwały różne efekty. Etykiety w przypadku klasyfikatora SVM uzyskały największą zgodność ze stanem faktycznym (Rys. 1.) w przypadku procesu klasyfikacji realizowanego w trójwymiarowej przestrzeni cech (zależności PRI od PW i RF), co potwierdza otrzymany wynik dokładności, wyszczególniony w Tab.2. Algorytm *k* najbliższych sąsiadów rezultat najbardziej zbliżony do stanu rzeczywistego uzyskał w przypadku procesu klasyfikacji realizowanego w dwuwymiarowej przestrzeni cech, utworzonej na bazie parametrów impulsu: PW i RF. We wspomnianym przypadku KNN osiągnął dokładność na poziomie 99,1%.

Biorąc pod uwagę przestrzenie cech, dla których obie metody uzyskały najwyższe wartości miary dokładności można zaobserwować, że cechami impulsu, które wystąpiły w obu zestawach atrybutów bazowych były: częstotliwość nośna i czas trwania.

5. PODSUMOWANIE

Wyniki badań opisanych na łamach tego artykułu pozwoliły na określenie wpływu wymiarowości danych rozpoznania radioelektronicznego na dokładność algorytmów klasyfikacji. Istnieje przesłanka, że wymiarowość danych rozpoznania radioelektronicznego wpływa na dokładność procesu ich klasyfikacji w różny sposób, w zależności zastosowanego zestawu cech impulsów oraz użytych algorytmów.

W przypadku maszyny wektorów nośnych najwyższą dokładność procesu rozpoznawania, wynoszącą 99,2%, uzyskano z zastosowaniem trójwymiarowej przestrzeni cech, składającej się ze zbioru atrybutów ***PDW***: częstotliwości nośnej, czasu trwania i okresu powtarzania. Natomiast stosując algorytm *k* najbliższych sąsiadów największą dokładność, na poziomie 99,1%, osiągnięto w oparciu o przetwarzanie dwuwymiarowej przestrzeni cech, zbudowanej z pary cech impulsu: częstotliwości nośnej i czasu trwania.

W klasyfikacji danych rozpoznania radioelektronicznego, opisujących trzy źródła emisji różnego typu, parametrami impulsu, które za każdy razem stanowiły wymiary składowe najbardziej informatywnej przestrzeni cech były: częstotliwość nośna i czas trwania.

Potencjalny kierunek dalszych badań nad przedmiotową problematyką obejmuje użycie większej liczby klasyfikatorów, zastosowanie szerszego zestawu miar ich jakości oraz przeprowadzenie eksperymentów na zbiorach danych rozpoznania radioelektronicznego, zawierających impulsy pochodzące z większej liczby źródeł emisji.

BIBLIOGRAFIA

[1] Bengio Yoshua, Delalleau Olivier, Le Roux Nicolas. 2005 „The Curse of Dimensionality for Local Kernel Machines," w *Advances in Neural Information Processing Systems*, 1-17.

[2] Demir Kerem Utku, Sener Batiham, Unver Hakki Ozgur. 2022. „Investigation of Dimensionality Reduction Methods for Chatter Detection with SVM," w *19th International Conference on Machine Design and Production (UMTIK)*, 1-17.

[3] Dudczyk Janusz, Kawalec Adam. 2014. „Wykorzystanie cech fraktalnych ekstrahowanych z sygnałów radarowych w procesie specyficznej identyfikacji". *Przegląd Elektrotechniczny*, 11 : 212-215.

[4] GeeksforGeeks, „K-Nearest Neighbors and Curse of Dimensionality," 4 Marzec 2024. https://www.geeksforgeeks.org/k-nearest-neighbors-and-curse-of-dimensionality/.

[5] Han Zhuangzhi, Zhao Zihang, Gao Zhenbin, Xie Hui. 2020. „Radar Signal Sorting Algorithm Based on Pulse Repetition Interval," w *2020 5th International Conference on Electromechanical Control Technology and Transportation (ICECTT)*, 163-166.

[6] Hu Juan, Chen Guangchun, Peng Hong, Wang Jun, Huang Xiangnian, Luo Xiaohui. 2017. „A kNN classifier optimized by P systems," w *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, 432-437.

[7] Jain Parul, Agarwal Ritu. 2023. „Sentiment Analysis using SVM, Twin SVM, and Naïve Bayes," w *2023 5th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, 456-459.

[8] Kouiroukidis Nikolaos, Evangelidis Georgios. 2011. „The Effects of Dimensionality Curse in High Dimensional kNN Search," w *2011 15th Panhellenic Conference on Informatics*, 41-45.

[9] Pedregosa Fabian et al. 2011. „Scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research*, 12(85) : 2825-2830.

[10] Rybak Łukasz, Dudczyk Janusz, Mazur Piotr, 2024. „ALGORYTMY SZTUCZNEJ INTELIGENCJI W PRZETWARZANIU DANYCH ROZPOZNANIA RADIOELEKTRONICZNEGO ELINT". *Przegląd Telekomunikacyjny - Wiadomości Telekomunikacyjne*, 4 : 121-124.

[11] Salem Nema, Hussein Sahar. 2019. „Data dimensional reduction and principal components analysis". *Procedia Computer Science*, 163 : 292-299.

[12] Zhang S, IEEE Senior Member, Li X, Zong M, Zhu X i Wang R, „Efficient kNN Classification With Different Numbers of Nearest Neighbors," *IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS,* tom 29, nr 5, 2018.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

Wykorzystanie uczenia maszynowego do efektywnego wykrywania anomalii oraz predykcji przepustowości w sieciach 5G

Machine learning for efficient anomaly detection and throughput prediction in 5G networks

Piotr Olesiński[1,2]; Michał Panek[2]; Róża Goścień[1];

[1] Department of Systems and Computer Networks, Faculty of Information and Communication Technologies, Wroclaw University of Science and Technology, Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland,
e-mail: piotr.olesinski@pwr.edu.pl
[2] Nokia Solutions and Networks Sp. z o.o., Szybowcowa 2, 54-130 Wroclaw, Poland

DOI: 10.15199/59.2025.4.12

Streszczenie: Artykuł porusza problem użycia uczenia maszynowego do analizy sieci 5G. Definiuje problem przewidywania przepustowości komórki sieci na podstawie wskaźników efektywności sieci. Do weryfikacji używa rzeczywistego zbioru danych, który zawiera liczne anomalie utrudniające dopasowanie modelu predykcyjnego do danych. Jako rozwiązanie, proponuje podejście w którym najpierw identyfikowane i usuwane są anomalie, a następnie budowany jest model predykcyjny. Badania pokazują, iż podejście to znacząco podnosi jakość modeli predykcji.

Abstract: The paper studies machine learning-based 5G network analysis. It defines problem of predicting a cell bandwidth based on key peformance indicators. A dataset form real network is used in the evaluation. It contains anomalies preventing a proper prediction model fit. As a solution, the paper proposes an approach that firstly identifies and removes anomalies and then fits forecasting model. The experiments show that the approach significantly improves the model fit.

Słowa kluczowe: detekcja anomalii, predykcja przepustowości, sieci mobilne, uczenie maszynowe, 5G

Keywords: anomaly detection, throughput prediction, machine learning, mobile networks, 5G

1. WSTĘP

Stale rosnące zainteresowanie użytkowników sieciami telekomunikacyjnymi skutkuje ich bardzo intensywnym rozwojem, szczególnie w kontekście sieci mobilnych. Sieci te są obecnie odpowiedzialne za przekazywanie około 20-30% całkowitego ruchu sieciowego, a wartości te mają prognozy wzrostowe [1]. Rozwój sieci mobilnych i poprawa ich działania są szczególnie istotnym i wymagającym zagadnieniem ze względu na wysoką złożoność i dynamikę/zmienność (w czasie i przestrzeni) tych systemów. Obiecującym rozwiązaniem, wprowadzonym w sieciach 5G, jest ich optymalizacja z wykorzystaniem algorytmów uczenia maszynowego (ang. Machine Learning, ML), umożliwiających prognozowanie parametrów i zjawisk kluczowych dla wydajności i niezawodności tych sieci [1, 2]. Przykładowymi obszarami zainteresowania operatorów mobilnych jest przewidywanie przepustowości sieci, jakości sygnału lub współczynnika tłumienia [3, 4, 5]. Jednak wydajność i niezawodność tych metod mogą być zakłócane przez *anomalie*, które mogą wystąpić w działaniu sieci i powiązanych z siecią pomiarach i obserwacjach (używanych do trenowania modeli ML) [6]. Anomalie te mogą również bezpośrednio wpływać na wydajność i niezawodność sieci. Zatem ich wykrywanie jest kluczowe dla poprawy ogólnego działania sieci.

Liczne badania wykorzystują modele ML do przewidywania parametrów sieci mobilnych [4, 5, 7]. Praca [5] prognozuje średnią przepustowość i opóźnienie, stosując wielowymiarową regresję liniową (ang. Multiple Linear Regression, MLR) i metodę gradientu prostego. W pracy [7] przewiduje się przepustowość za pomocą modelu ARIMA (ang. Auto-Regressive Integrated Moving Average), analizując szeregi czasowe. Praca [4] prognozuje przepustowość, wykorzystując sieć neuronową i relację między konfiguracją komórki a warunkami sieciowymi. Dodatkowo, liczne badania skupiają się również na wykrywaniu anomalii w środowisku/danych sieciowych [8, 9]. Autorzy [8] używają podejść statystycznych do dekompozycji szeregów czasowych, a autorzy [9] stosują metody ML do wykrywania awarii sprzętu. Istniejące badania wskazują na wysoki potencjał tych metod, jednocześnie wymagając dalszych prac rozwojowych.

W niniejszym artykule poruszany jest problem prognozowania przepustowości komórek w sieci 5G na podstawie wektora kluczowych wskaźników wydajności (ang. Key Performance Indicator, KPI), z wykorzystaniem algorytmów ML. Badania prowadzone są na rzeczywistym zbiorze danych, dostarczonym przez firmę Nokia z komercyjnej sieci testowej. Zbiór danych przedstawia zależność między wybranymi wskaźnikami KPI a wynikającą z tego przepustowością komórki sieci 5G. Dane pochodzą z rzeczywistej, testowej sieci operacyjnej i zawierają liczne obserwacje odstające (tj. anomalie), wpływające na możliwość zbudowania wysoce wydajnego i niezawodnego modelu prognostycznego. Stąd też, artykuł skupia się również na problemie wykrywania i usuwania anomalii z danych, dążąc do zwiększenia dokładności przewidywania przepustowości. Praca weryfikuje, czy możliwe jest zaproponowanie wysoce wydajnego modelu predykcyjnego po wyeliminowaniu obserwacji związanych z anomaliami.

Artykuł jest zorganizowany następująco: Sekcja 2 przedstawia definicję problemu. Sekcja 3 opisuje metody

użyte w badaniu. Sekcja 4 prezentuje i omawia wyniki. Na koniec, Sekcja 5 podsumowuje badania.

2. DEFINICJA PROBLEMU

Formalnie skupiamy się na dwóch problemach: (*i*) prognozowaniu przepustowości komórki sieci 5G na podstawie wektora wybranych wskaźników KPI, (*ii*) wykrywaniu i usuwaniu anomalii w zbiorze danych w celu poprawy efektywności prognozowania. Na podstawie danych od operatora, zakładamy że anomalie są odchyleniem od trendu pomiarowego, wynikającym z błędów podczas akwizycji oraz przetwarzania danych lub błędów pomiarowych.

Problem prognozowania definiujemy jako zadanie regresji $w_i = \{KPI_1, KPI_2, ..., KPI_n\} \rightarrow TP_i$, gdzie w_i jest wektorem wskaźników KPI dla komórki sieciowej i, a TP_i jest odpowiadającą mu przepustowością łącza w dół (DL). Dla opisanego problemu, wektor w_i obejmuje wybrane parametry środowiskowe sieci, w tym wykorzystanie bloku zasobów fizycznych (PRB), współczynnik błędów blokowych (BLER), wskaźnik jakości kanału (CQI), liczbę podłączonych użytkowników (CU) oraz wolumen danych w kanale w dół (DL DV). Pierwotne dane poddano standaryzacji ze względu na ich szeroki zakres wartości. Proces ten może prowadzić do wystąpienia wartości ujemnych w zbiorze zmodyfikowanym. Wskaźniki KPI są danymi ciągłymi reprezentującymi wartości w zakresach określonych w Tabeli 1.

Tabela 1: Zakres wartości wskaźników KPI w danych

KPI	*Min*	*Max*
DL TP	-0,83	17,18
PRBU	-0,33	33,72
BLER	-0,77	10,77
CQI	-0,26	56,72
CU	-4,19	1,86
DL DV	-0,67	27,34

Z punktu widzenia modeli ML, zbiór danych to zestaw par składających się z wektora cech w_i oraz odpowiadającej mu wartości przepustowości komórki $TP_i, i \in I$ (wartości mierzone w komercyjnej sieci testowej). Zbiór reprezentuje wierną próbkę danych sieciowych, zawierającą 1169670 rekordów opisujących 3846 komórek. Wartości te są reprezentatywne dla sieci, ponieważ obejmują tylko komórki z co najmniej 200 wpisami i wykluczają wszelkie rekordy o nieznanych (NaN) wartościach.

Zbiór danych zawiera liczne anomalie (obserwacje odstające), które wpływają na możliwość zbudowania wysoce wydajnego i niezawodnego modelu prognozującego. Dlatego drugim problemem jest wykrycie (identyfikacja) anomalii, usunięcie ich ze zbioru danych i sprawdzenie, czy operacja ta poprawia efektywność predykcji.

3. METODY

W tej sekcji omawiamy algorytmy ML wybrane do rozwiązywania zadań predykcji i wykrywania anomalii, a także zastosowane metryki porównawcze. Do predykcji przepustowości wykorzystujemy następujące algorytmy:

- **Wielokrotna Regresja Liniowa, MLR**, statystyczna metoda określania liniowej zależności między cechami a zmienną wynikową [10]. MLR jest metodą prostą w implementacji i interpretacji, o niskiej złożoności obliczeniowej. Stąd wybraliśmy ją do porównania, pomimo nieliniowej zależności przepływności od CQI.
- **Metoda Najmniejszych Odchyleń Bezwzględnych (ang. Least Absolute Deviations, LAD)**, która minimalizuje sumę bezwzględnych różnic między przewidywanymi a rzeczywistymi wartościami. Jej celem jest znalezienie linii lub hiperpłaszczyzny minimalizującej wspomnianą sumę bezwzględnych różnic [11].
- **Regresja z Wzmacnianiem Gradientowym (ang. Gradient Boosting Regression, GBR)**, łączy koncepcje wzmacniania gradientowego i regresji, tworząc solidny i dokładny model predykcyjny. Założenie GBR polega na minimalizacji funkcji straty poprzez iteracyjne dopasowywanie nowych modeli do ujemnego gradientu funkcji straty względem przewidywanych wartości [12].

Do zadania wykrywania anomalii, w niniejszym artykule wykorzystano następujące metody [13]:

- **Lokalny Wskaźnik Odległości (ang. Local Outlier Factor, LOF)**, zakłada że anomalia ma niższą gęstość otoczenia w porównaniu do swoich sąsiadów. LOF uwzględnia konteksty lokalne i różnice w gęstości, co czyni go skutecznym w identyfikowaniu anomalii w zróżnicowanych zbiorach danych. Ten algorytm jest wrażliwy na modyfikacje parametru *nneighbors* dotyczące liczby sąsiadów.
- **Jednoklasowa Maszyna Wektorów Nośnych (ang. One Class Support Vector Machine, OCSVM)**, koncentruje się na znalezieniu hiperpłaszczyzny oddzielającej dane normalne od anomalii. Algorytm zakłada, że większość danych należy do jednej klasy, a anomalie są rzadkie i znacząco różnią się od pozostałych. Pozwala na znalezienie granicy oddzielającej dane anomalne od danych normalnych, co umożliwia ich wykrycie.
- **Las Izolacyjny (ang. Isolation Forest, IF)**, dzieli dane na drzewa decyzyjne, w oparciu o losowy wybór atrybutów i progów, izoluje anomalie, które wymagają mniej podziałów, aby oddzielić się od reszty danych. Dane normalne wymagają więcej podziałów, aby zostać odizolowane. Im mniej podziałów wymaganych, tym bardziej obserwacja jest uważana za anomalię. IF ocenia stopień izolacji dla każdej obserwacji.

Porównując modele predykcyjne, skupiamy się na jakości ich dopasowania do danych (zdolność do rozpoznawania ukrytych wzorców). Wykorzystujemy skorygowany współczynnik R^2 (w skrócie adj R^2) [12], którego wartość nie może przekroczyć 1 i powinna być maksymalizowana.

4. WYNIKI I ANALIZA

Ten rozdział przedstawia wyniki symulacji, które zostały podzielone na pięć kroków: (*i*) dostrojenie modeli predykcyjnych, (*ii*) ocena jakości modeli predykcyjnych dla początkowego zbioru danych, (*iii*) dostrojenie metod wykrywania anomalii, (*iv*) wykrycie i usunięcie anomalii ze zbioru danych, (*v*) ocena jakości modeli predykcyjnych dla zmodyfikowanego zbioru danych. Początkowy zbiór

danych użyty w eksperymentach zawiera dane z 3846 różnych komórek sieciowych. W eksperymentach został podzielony na podzbiory treningowy i testowy w proporcji 70-30. Wszystkie metody regresji i wykrywania anomalii zostały zaimplementowane na poziomie komórki sieciowej przy użyciu dedykowanej biblioteki Python SKLEARN.

Po pierwsze, skupiamy się na strojeniu metod regresji. Dla wybranych losowo 300 komórek testujemy wszystkie możliwe kombinacje wskazanych parametrów, zakładając dla MLR: *n jobs* $\in \{1; 2; 4; 8\}$, dla LADR: $\alpha \in \{0,001; 0,01; 0,1; 1; 10\}$, dla GBR: *learning rate* $\in \{0,01; 0,1; 0,25\}$; *n estimators* $\in \{50; 100; 200\}$; *max depth* $\in \{1, 3, 5\}$. Zgodnie z wynikami uśrednionymi dla testowanych komórek, zalecamy ustawienia domyślne dla MLR i LAD oraz użycie następującej konfiguracji dla GBR: *learning rate* = 0,25, *n estimators* = 50, *max depth* = 1.

Po dostrojeniu metod przechodzimy do oceny ich jakości dla zbioru danych zawierającego anomalie. Trenujemy i dopasowujemy trzy metody regresji dla każdej komórki sieciowej osobno (na zbiorze treningowym), a następnie porównujemy ich wyniki zgodnie z uśrednionym współczynnikiem adj R^2 i liczbą komórek z adj $R^2 > 0$ (biorąc pod uwagę zbiór testowy). Wyniki przedstawiono w Tabeli 2. Analiza wskazuje, że dla znaczącej liczby komórek (od 504 (dla GBR) do 1645 (dla LADR)) jakość predykcji jest bardzo niska (tj. adj $R^2 \leq 0$). Uśrednione wartości adj R^2 są poniżej 0 dla MLR i LADR, podczas gdy dla GBR wartość współczynnika jest dodatnia, ale stosunkowo niska – 0,31. Co więcej, uzyskane wartości adj R^2 znacznie różnią się między różnymi komórkami. Na przykład, model MLR uzyskał wartości adj R^2 w zakresie od -1628 do 0,94, co może sugerować iż dla niektórych komórek nastąpiło bardzo dobre dopasowanie modelu a dla innych dopasowania nie udało się znaleźć (co może być skutkiem licznych anomalii). Wyniki pokazują więc, iż początkowy zbiór danych nie daje możliwości zbudowania efektywnego i niezawodnego modelu predykcyjnego dla każdej komórki.

Tabela 2: Ocena efektywności predykcji dla początkowego zbioru danych

	Liczba komórek dla których		śr. wartość
	adj $R^2 > 0$	*adj* $R^2 \leq 0$	*adj* R^2
MLR	2758	1088	-1,55
LADR	2201	1645	-2,49
GBR	3342	504	0,31

Strojąc metody detekcji anomalii, maksymalizowaliśmy liczbę wykrytych wartości odstających (błąd bezwzględny: wzór 1), oceniając skuteczność poprzez porównanie sumy liczby największych odchyleń klasyfikowanych jako anomalie z faktyczną liczbą anomalii (wzór 2). Aby przeprowadzić właściwe strojenie dla każdej metody, przeprowadzamy testy dla 10 losowo wybranych komórek i wszystkich możliwych kombinacji wartości następujących parametrów (dla każdego parametru pogrubioną czcionką wyróżniamy jego wartość rekomendowaną do dalszych eksperymentów):

- LOF: *n neighbors* $\in \{10; \mathbf{20}; 30; 40\}$.
- OCSVM: *nu* $\in \{0,03; 0,05; \mathbf{0,08}; 0,1\}$.
- IF: *n estimators* $\in \{10; 100; \mathbf{200}\}$; *contamination* $\in \{\mathbf{0,05}; 0,75\}$.

$$\epsilon = |TP^{'} - TP| \tag{1}$$

$$\theta[\%] = \frac{a_{consec}^{max}}{a_{all}} \cdot 100 \tag{2}$$

Gdzie:
ϵ – wartość błędu bezwzględnego; $TP^{'}$ – przewidywana wartość przepustowości; TP – zmierzona (dokładna) wartość przepustowości; θ – efektywność detekcji [%]; a_{consec}^{max} – liczba maksymalnie kolejnych wykrytych anomalii; a_{all} – liczba wszystkich wykrytych anomalii.

Następnie, wykorzystujemy dostrojone metody do wykrywania i usuwania anomalii z oryginalnego zbioru danych, otrzymując w ten sposób zmodyfikowany zbiór danych. Liczba wykrytych anomalii różni się w zależności od komórek sieciowych. Tabela 3 przedstawia minimalną, maksymalną i średnią liczbę anomalii wykrytych przez każdą metodę (dla wszystkich komórek). Największą liczbę anomalii wykryto za pomocą metody LOF. Metoda ta zidentyfikowała od 6 do 51 anomalii, średnio 29,5. Następnie, metoda OCSVM wykryła od 4 do 44 obserwacji odstających, średnio 20,6. Na koniec, metoda IF wykryła najmniejszą liczbę anomalii – od 3 do 17, średnio 14,9.

Tabela 3: Liczba wykrytych anomalii.

	Min	*Max*	*Średnia*
LOF	6	51	29,5
OCSVM	4	44	20,6
IF	3	17	14,9

Na koniec testujemy jakość modeli predykcyjnych dla zmodyfikowanego zbioru danych. Tabela 4 przedstawia uzyskane wyniki biorąc pod uwagę liczbę komórek z adj $R^2 > 0$ i średnią wartością tego współczynnika. Badania wykazują znaczący wzrost obu metryk, co dowodzi że proces wykrywania anomalii może poprawić zadanie predykcji przepustowości w sieci mobilnej. Tabela 5 szczegółowo przedstawia wyniki i prezentuje procentowy wzrost liczby komórek z adj $R^2 > 0$ (w porównaniu do początkowego zbioru danych). Jak można zaobserwować, proces wykrywania i usuwania anomalii pozwolił zwiększyć tę liczbę o maksymalnie 42,7% (dla LADR, LOF).

Pomimo znaczącego wzrostu obu kryteriów, nie wszystkie modele wykazują zadowalającą efektywność w zadaniu predykcji. Spośród trzech metod regresji, tylko metoda GBR pozwoliła na uzyskanie średniego adj $R^2 > 0,5$, co może sugerować akceptowalne dopasowanie modelu do danych (MLR i LADR uzyskały średnie adj $R^2 \leq 0,21$, co nie gwarantuje dobrego dopasowania modelu do danych). Porównując metody wykrywania anomalii, obserwujemy, że najlepsze wyniki uzyskano przy użyciu LOF, która zidentyfikowała największą liczbę anomalii spośród wszystkich podejść. Sugeruje to, że metoda prawidłowo wykryła obserwacje odstające, a proces ich usuwania ujawnił ukryte wzorce.

Podsumowując, najlepsze dopasowanie modelu predykcyjnego uzyskano, gdy LOF został użyty do wykrywania anomalii, a GBR jako model prognozujący.

Tabela 4: Ocena wyników modelu dla wszystkich komórek gdzie adj $R^2 > 0$

-	*MLR*			*LADR*			*GBR*		
	LOF	*OCSVM*	*IF*	*LOF*	*OCSVM*	*IF*	*LOF*	*OCSVM*	*IF*
Liczba komórek ze średnim adj $R^2 > 0$									
zbiór początkowy		2758			2201			3342	
zbiór zmodyfikowany	3307	3307	3295	3141	2765	2987	3762	3709	3738
Średni adj R^2									
zbiór początkowy		-1,55			-2,29			0,31	
zbiór zmodyfikowany	0,21	0,21	0,19	0,18	-0,18	0,12	0,62	0,54	0,56

Tabela 5: Wzrost [%] liczby komórek z adj $R^2 > 0$

Metoda	*LOF*	*OCSVM*	*IF*
MLR	19,9%	19,9%	19,5%
LADR	42,7%	25,6%	35,7%
GBR	12,3%	11%	11,7%

5. WNIOSKI

W niniejszej pracy skupiliśmy się na zastosowaniu algorytmów ML w celu poprawy predykcji przepustowości w komórkach sieci 5G. Po pierwsze, zdefiniowaliśmy problem przewidywania przepustowości łącza w dół na podstawie wektora wskaźników KPI sieci i zaproponowaliśmy jego rozwiązanie za pomocą trzech metod regresji. Do celów ewaluacji wykorzystaliśmy rzeczywisty zbiór danych sieciowych, który zawiera wiele anomalii (obserwacji odstających), które znacząco pogarszają jakość predykcji. W celu zaadresowania tego problemu, zaproponowaliśmy zastosowanie dedykowanych podejść do wykrywania tych anomalii, usunięcia ich z zestawu danych i ponownego wykonania zadania predykcji. Wyniki badań dowodzą, że proces wykrywania i usuwania anomalii znacząco poprawia jakość predykcji przepustowości, a w wielu przypadkach wręcz umożliwia wykonanie tego zadania.

Uzyskane wyniki są niezwykle istotne, ponieważ odkrycie wzorców zachowań komórki sieciowej może umożliwić wdrożenie zaawansowanych technologii optymalizacji i symulacji sieci, na przykład — w zastosowaniach cyfrowych bliźniaków. Dlatego też w naszej przyszłej pracy planujemy znacząco rozszerzyć nasze badania poprzez:

- Eksplorację bardziej efektywnych algorytmów wykrywania anomalii, takich jak modele głębokiego uczenia.
- Implementację wykrywania anomalii w czasie rzeczywistym w celu monitorowania wydajności sieci i proaktywnego identyfikowania oraz rozwiązywania problemów.
- Opracowanie modeli predykcyjnych uwzględniających interakcje między wieloma komórkami w celu uchwycenia bardziej złożonej dynamiki sieci.
- Integrację modeli wykrywania anomalii i predykcji z istniejącymi systemami zarządzania siecią w celu automatyzacji optymalizacji i rozwiązywania problemów.

PODZIĘKOWANIA

Praca finansowana ze środków finansowych Ministerstwa Nauki i Szkolnictwa Wyższego na podstawie umowy Nr DWD/8/0116/2024, program Doktorat Wdrożeniowy edycja VIII, w ramach realizacji projektu "Projekt i analiza aplikacji cyfrowego bliźniaka dla węzła sieci 5G".

LITERATURA

[1] Nokia. 2020. Deepfield Network Intelligence Report Networks in 2020

[2] U. Challita, et al. 2020. When Machine Learning Meets Wireless Cellular Networks: Deployment, Challenges, and Applications. IEEE Communications Magazine 58 (6) : 12-18.

[3] A. Mekrache, et al. 2024. Toward Enhanced Beyond 5G Network Management. IEEE Communications Surveys & Tutorials 26, (4) : 2769-2797.

[4] M. Panek, et al. 2024. Modeling configuration-performance relation in a mobile network: a data-driven approach. Proc. of IEEE PIMRC : pp. 1-6,

[5] N. Mourad and M. Hamdi. 2019. LTE QoS Parameters Prediction Using Multivariate Linear Regression Algorithm. Proc. of ICIN : 145-150.

[6] A. B. Nassif, et al. 2021. Machine Learning for Anomaly Detection: A Systematic Review. IEEE Access 9 : 78658-78700.

[7] L. Dongwon, et al. 2020. Prediction of Network Throughput using ARIMA. Prof. of. ICAIIC : pp. 1-5.

[8] L. Cerdà-Alabern, et al. 2023. Anomaly detection for fault detection in wireless community networks using machine learning. Computer Communications 202 : 191-203.

[9] A. Nediyanchath, et al. 2020. Anomaly Detection in Mobile Networks. Proc. of IEEE WCNCW : 1-5.

[10] J. M. Chiou, et al. 2016. Multivariate functional linear regression and prediction. Journal of Multivariate Analysis 146 : 301-312.

[11] Z. Liu and Y. Ying. 2020. Least absolute deviations estimation for uncertain regression with imprecise observations. Fuzzy Optimization and Decision Making 19 : 1573-2908

[12] D. Alekseeva, et al. 2021. Comparison of Machine Learning Techniques Applied to Traffic Prediction of Real Wireless Network. IEEE Access 9 : 159495-159514.

[13] A. Yahia, et al. 2024. Leveraging Machine Learning for Anomaly Detection Methods in Cryptocurrency: A Data-Driven Study. Proc. of ICOA : 1-5

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

ARCHITEKTURA PRZETWARZANIA BRZEGOWEGO DLA EFEKTYWNEJ ANALIZY WIDEO W SYSTEMACH MONITORINGU WIZYJNEGO

EDGE COMPUTING ARCHITECTURE FOR EFFICIENT VIDEO ANALYSIS IN CCTV

Andrzej Matiolański[1];

[1] AGH Akademia Górniczo-Hutnicza, Kraków
email: andrzej.matiolanski@agh.edu.pl

DOI: 10.15199/59.2025.4.13

Streszczenie: Artykuł analizuje architekturę przetwarzania brzegowego w systemach monitoringu wizyjnego. Przedstawiono jego zalety, takie jak redukcja opóźnień, optymalizacja wykorzystania sieci i poprawa bezpieczeństwa danych. Porównano trzy scenariusze przetwarzania (globalny, regionalny, brzegowy), wykazując istotną przewagę urządzeń brzegowych pod względem szybkości i efektywności energetycznej. Podkreślono również znaczenie wdrożenia mechanizmów bezpieczeństwa oraz perspektywy rozwoju architektur brzegowych.

Abstract: The article analyzes edge computing architecture in video surveillance systems. It presents the advantages of edge computing, such as latency reduction, network resource optimization, and enhanced data security. Three processing scenarios (global, regional, edge) are compared, demonstrating a significant advantage of edge devices in terms of speed and energy efficiency. The importance of implementing security mechanisms and the prospects for further development of edge architectures are also highlighted.

Słowa kluczowe: analiza wideo, monitoring wizyjny, opóźnienia, przetwarzanie brzegowe

Keywords: edge computing, latency, video analysis, video surveillance

1. WSTĘP

Szybki wzrost ilości danych wideo wymaga innowacyjnych metod, które umożliwią efektywne ich przetwarzanie i analizę, szczególnie w kontekście analizy danych wideo. Wraz ze wzrostem zapotrzebowania na analizę wideo w czasie rzeczywistym, systemy muszą sprostać wyzwaniom związanym z obsługą dużych ilości danych generowanych w środowiskach takich jak monitoring miejski czy pojazdy autonomiczne. Przetwarzanie brzegowe oferuje wiarygodne rozwiązanie poprzez dystrybucję zadań analitycznych między chmurę, brzeg i urządzenia końcowe, co pozwala na optymalizację wykorzystania zasobów oraz redukcję opóźnień. Takie podejście nie tylko zwiększa wydajność przetwarzania wideo, ale także rozwiązuje problemy związane z prywatnością dzięki lokalnemu przetwarzaniu danych.

Jak wskazano w pracy [3], rozproszona analityka wideo w systemach współpracujących w modelu chmura-brzeg-terminal umożliwia przełomowe rozwiązania w wielu dziedzinach. Umożliwia wdrożenie nowych funkcjonalności i redukcję kosztów. Z kolei badania [1] szczegółowo opisują postępy w rozpoznawaniu chodu, prezentując efektywne wykorzystanie głębokiego uczenia w ramach rozproszonej struktury. Niniejszy artykuł bada implikacje zastosowania architektury przetwarzania brzegowego zaprojektowanej do efektywnej analizy wideo, podkreślając jej potencjał transformacyjny w różnorodnych zastosowaniach. Obie prace podkreślają znaczenie wykorzystanie przetwarzania brzegowego, które stanowi wartość dodaną do prezentowanych rozwiązań.

Przetwarzanie brzegowe oferuje również możliwość integracji z technologiami uczenia maszynowego, które są wdrażane bezpośrednio na urządzeniach brzegowych. Takie podejście pozwala na realizację zaawansowanych algorytmów analizy obrazu i wykrywania zdarzeń bez konieczności przesyłania dużych ilości danych do scentralizowanych serwerów chmurowych [7]. Na przykład, systemy nadzoru miejskiego mogą wykorzystywać modele głębokiego uczenia do automatycznego wykrywania incydentów, takich jak wypadki drogowe czy zachowania podejrzane, co znacząco poprawia bezpieczeństwo publiczne [6].

Ponadto, integracja podejść hybrydowych, łączących możliwości chmury i brzegu, ułatwia optymalizację narzędzi do archiwizacji i wyszukiwania multimediów. Jak wskazują badania [8], takie rozwiązania wspierają lepiej publiczny dostęp do treści wideo oraz zwiększają ich użyteczność poprzez efektywne zarządzanie zasobami i dynamiczną alokację zadań obliczeniowych. Hybrydowe podejście zmniejsza obciążenie scentralizowanych serwerów chmurowych, jednocześnie optymalizując efektywność energetyczną i zmniejszając zależność od dużej przepustowości sieci [9].

Dodatkowo, rozwój technologii takich jak 5G oraz nowoczesne protokoły komunikacyjne umożliwia jeszcze szybsze przesyłanie danych między urządzeniami brzegowymi a chmurą. Dzięki temu systemy przetwarzania brzegowego mogą sprostać wymaganiom aplikacji czasu rzeczywistego, takich jak transmisje na żywo czy analityka wideo w środowiskach o wysokiej dynamice danych [4]. Wprowadzenie technologii neuromorficznych w urządzeniach brzegowych dodatkowo zwiększy ich efektywność energetyczną, co jest kluczowe dla zastosowań wymagających ciągłego monitorowania (na przykład w monitoringu miejskim) [2].

2. ARCHITEKTURA PRZETWARZANIA BRZEGOWEGO

Ewolucja architektury przetwarzania brzegowego znacząco zwiększyła efektywność analizy wideo w systemach, rozwiązując kluczowe ograniczenia tradycyjnego przetwarzania w chmurze. Wraz ze wzrostem zapotrzebowania na niskie opóźnienia oraz zredukowane zużycie energii, frameworki takie jak Cloud-Edge Collaborative Architecture (CEA) stały się niezbędnym elementem nowoczesnych rozwiązań. CEA oferuje solidną infrastrukturę do przetwarzania strumieni wideo, wspierając efektywne strategie odciążania zadań obliczeniowych online. Dzięki dynamicznej alokacji zasobów, architektura ta optymalizuje doświadczenia użytkownika, szczególnie w kontekście ograniczeń budżetowych oraz redundancji zadań [8].

Ponadto, zaawansowane rozwiązania takie jak CloudEye demonstrują potencjał systemów percepcji wizualnej mobilnej w czasie rzeczywistym. Wykorzystując serwery brzegowe, CloudEye minimalizuje opóźnienia i zużycie zasobów, jednocześnie utrzymując wysoką dokładność detekcji obiektów [9]. Architektura ta umożliwia wydobywanie informacji o treści na poziomie brzegu sieci, co nie tylko redukuje zapotrzebowanie na przepustowość, ale także znacząco zwiększa szybkość przetwarzania danych. Dzięki temu systemy przetwarzania brzegowego stają się bardziej efektywne i wszechstronne, wspierając różnorodne zastosowania analityki wideo.

Przykłady zastosowań obejmują inteligentne miasta, gdzie analiza strumieni wideo pozwala na monitorowanie ruchu drogowego oraz wykrywanie incydentów w czasie rzeczywistym. W sektorze pojazdów autonomicznych architektury brzegowe umożliwiają szybkie podejmowanie decyzji na podstawie danych sensorycznych, co jest kluczowe dla bezpieczeństwa i efektywności transportu [4]. Dodatkowo, integracja z technologiami sztucznej inteligencji pozwala na realizację zaawansowanych funkcji takich jak predykcja zdarzeń czy automatyzacja procesów decyzyjnych.

2.1 Kluczowe komponenty i struktura systemów przetwarzania brzegowego

Systemy przetwarzania brzegowego są definiowane przez ich zdecentralizowaną architekturę, która strategicznie lokalizuje możliwości obliczeniowe bliżej źródła danych. Takie podejście często obejmuje integrację z elementami sieciowymi, takimi jak mikroserwery, jednostki zdalne czy urządzenia IoT. Bliskość przetwarzania danych do ich źródła pozwala na minimalizację opóźnień, zwiększenie efektywności wykorzystania przepustowości sieci oraz realizację analityki w czasie rzeczywistym, co jest kluczowe dla takich zastosowań jak analiza wideo w systemach nadzoru, autonomiczne pojazdy czy inteligentne miasta.

Jednym z najważniejszych aspektów tych systemów jest opracowanie odpowiednich architektur sprzętowych. Właściwe ich dopasowanie i dostosowanie pozwala na optymalizację obsługi danych przy jednoczesnym zmniejszeniu zużycia energii. Dedykowane układy obliczeniowe rozwiązują kluczowe wyzwania związane z wysokim zapotrzebowaniem na energię w tradycyjnych infrastrukturach chmurowych [4]. Dzięki temu systemy przetwarzania brzegowego mogą sprostać wymaganiom aplikacji o dużej intensywności danych, takich jak zaawansowana analityka wideo czy predykcja zdarzeń w czasie rzeczywistym.

Dodatkowo, nowe trendy w mobilnym przetwarzaniu brzegowym wskazują na rosnącą rolę umieszczania możliwości obliczeniowych i przechowywania danych na brzegu sieci mobilnych. Takie podejście wspiera szybsze i bardziej zoptymalizowane usługi dzięki zmniejszeniu opóźnień transmisji oraz dynamicznemu zarządzaniu zasobami [2]. Na przykład, w aplikacjach związanych z transmisją strumieniową wideo, mobilne urządzenia brzegowe mogą realizować lokalne przetwarzanie danych, co prowadzi do poprawy jakości usług oraz redukcji kosztów operacyjnych.

3. KORZYŚCI PRZETWARZANIA BRZEGOWEGO DLA ANALIZY WIDEO

W obliczu rosnącego zapotrzebowania na zasoby obliczeniowe w analizie wideo, przetwarzanie brzegowe stanowi transformacyjne rozwiązanie, skutecznie odpowiadające na wyzwania związane z opóźnieniami i przepustowością. Przetwarzając dane bezpośrednio w miejscu ich generowania – tam, gdzie wideo jest przechwytywane – architektura brzegowa eliminuje opóźnienia charakterystyczne dla transmisji dużych plików do scentralizowanych serwerów chmurowych. Lokalne przetwarzanie umożliwia analizę w czasie rzeczywistym, co ma kluczowe znaczenie dla zastosowań takich jak:

- Inteligentne systemy nadzoru miejskiego (wykrywanie zdarzeń w czasie < 150 ms) [10]
- Pojazdy autonomiczne (przetwarzanie danych z kamer 360° z częstotliwością 60 kl/s) [5]
- Transmisja wydarzeń sportowych na żywo z dynamiczną adaptacją bitrate [11]

Integracja przetwarzania brzegowego z architekturami analitycznymi wprowadza strukturalne usprawnienia w zarządzaniu zasobami. Model *Edge-Cloud Orchestration* pozwala na dynamiczne równoważenie obciążenia między:

- Urządzeniami końcowymi (np. kamery z akceleratorami AI)
- Węzłami brzegowymi (serwery GPU w stacjach bazowych 5G)
- Zasobami chmurowymi (analiza historyczna i trening modeli)

Wdrożenie technologii 5G wzmacnia korzyści architektur brzegowych. Dzięki latencji poniżej 10 ms możliwe jest praktycznie całkowite wyeliminowanie opóźnień w transmisji danych. Przepustowość na poziomie 1–10 Gb/s umożliwia przesyłanie strumieni wideo o bardzo wysokiej jakości, takich jak 8K przy 120 klatkach na sekundę. Dodatkowo, gęstość węzłów sięgająca 100 na kilometr kwadratowy pozwala na znaczące zwiększenie liczby urządzeń pracujących w jednej sieci, co sprzyja rozwojowi zaawansowanych systemów monitoringu i analizy danych w czasie rzeczywistym.

Przykładem skutecznej implementacji jest system EdgeVision wykorzystywany w przemyśle 4.0, gdzie analiza wideo z linii produkcyjnych realizowana jest na brzegu sieci z dokładnością 99.4% przy opóźnieniu poniżej 20 ms [**?**]. Rozwiązanie to redukuje zużycie energii o 40% w porównaniu z architekturą chmurową.

Przetwarzanie brzegowe w systemach monitoringu przenosi część obciążenia z modelu Video Surveillance as a Service (VSaaS) na zasoby lokalne, umożliwiając wykonywanie kluczowych operacji analitycznych bez konieczności przesyłania dużych ilości danych do chmury. Lokalna analiza zmniejsza zapotrzebowanie na przepustowość

sieci, poprawia bezpieczeństwo danych oraz minimalizuje ryzyko związane z ich przesyłem [8].

Dodatkowo, wdrożenie usługi staje się prostsze – wystarczy podłączenie urządzenia brzegowego w infrastrukturze klienta. Model ten redukuje koszty instalacji i pozwala na szybkie skalowanie systemu zgodnie z potrzebami [2]. Urządzenia brzegowe zapewniają także ciągłość działania w przypadku awarii połączenia sieciowego, zwiększając niezawodność systemu [10].

4. BADANIA WYDAJNOŚCI SYSTEMU VSAAS W RÓŻNYCH SCENARIUSZACH

4.1 Metodologia badań

Przeprowadzono symulację trzech scenariuszy analizy wideo CCTV przy użyciu:

- Scenariusz 1: Globalny serwer VSaaS (Kanada) z API, przetwarzający wszystkie klatki.
- Scenariusz 2: Regionalny serwer VSaaS (Polska) z API, przetwarzający wszystkie klatki.
- Scenariusz 3: Urządzenie brzegowe NVIDIA Jetson Nano 4GB z analizą ruchu (przesyłanie tylko klatek z aktywnością).

W każdym z analizowanych scenariuszy szczegółowo określono parametry sprzętowe. Dla scenariusza 1. i 2. testy przeprowadzono na serwerach o parametrach: Intel Xeon Silver 4210, 32 GB RAM, GPU NVIDIA RTX 4060 Ti. W przypadku scenaiusza 3. analizę przeprowadzono na urządzeniu NVIDIA Jetson Nano 4GB (CPU ARM Cortex-A57, 4 GB RAM, GPU 128 CUDA cores). W każdym środowisku zapewniono brak konkurencyjnych procesów obciążających system, a całość zasobów była dostępna wyłącznie dla analizowanej aplikacji.

W analizowanych wariantach istotnie różniły się techniki pobierania i przesyłania klatek z kamer: W scenariuszach serwerowych (globalny i regionalny) do analizy przesyłano pełny strumień wideo. Następnie dokonano analizy aktywności, z potem każda klatka była przetwarzana niezależnie przez serwer VSaaS. W scenariuszu brzegowym urządzenie Jetson Nano realizowało lokalną detekcję ruchu i przesyłało do dalszej analizy jedynie te klatki, w których wykryto aktywność (ok. 28% wszystkich klatek). Dzięki temu znacząco ograniczono zarówno obciążenie sieci, jak i czas przetwarzania, co stanowi realny wskaźnik zysków czasowych i efektywności tego rozwiązania. Różnice te mają wpływ na uzyskane wyniki. Scenariusz brzegowy nie tylko minimalizuje opóźnienia propagacyjne, ale także redukuje liczbę klatek wymagających analizy, co przekłada się na krótszy średni czas przetwarzania i mniejsze zużycie zasobów sieciowych. Zgodnie z rozważaniem teoretycznym najbardziej zasadna jest detekcja aktywności jak najbliżej kamery (omawiany scenaiusz 3), w celu rezygnacji z przesyłania dodatkowych danych.

Parametry symulacji:

- Liczba analizowanych klatek: 10000
- Ilość klatek bezczynności: ∼70%
- Protokół komunikacji: REST API v2.1 TLS

4.2 Analiza statystyczna

W ramach pracy zbadano czasy przetwarzania (generowanych opóźnień) dla trzech instancji usługi VSaaS służacej do analizy obrazu z kamer monitoringu. Dane przedstawiono w tabelach 1, 2 i 3.

Tabela 1: Parametry statystyczne czasu przetwarzania klatki – scenariusz Globalny

Parametr	Wartość
Średni czas [s]	0,99 ± 0,28
Odchylenie standardowe [s]	0,16

Tabela 2: Parametry statystyczne czasu przetwarzania klatki – scenariusz Regionalny

Parametr	Wartość
Średni czas [s]	0,54 ± 0,15
Odchylenie standardowe [s]	0,08

Tabela 3: Parametry statystyczne czasu przetwarzania klatki – scenariusz Brzegowy

Parametr	Wartość
Średni czas [s]	0,19 ± 0,06
Odchylenie standardowe [s]	0,03

4.3 Wnioski szczegółowe

Na podstawie literatury oraz przeprowadzonych testów można sformułować następujące wnioski:

- Urządzenie brzegowe (Jetson Nano) wykazało około 5-krotnie krótszy średni czas przetwarzania klatki w porównaniu do serwera globalnego
- Redukcja odchylenia standardowego o 79.1% w porównaniu do serwera globalnego wskazuje na większą stabilność czasową przetwarzania brzegowego
- Architektura brzegowa przetworzyła tylko 2814 klatek (28.14% wszystkich), redukując:
  - Obciążenie sieciowe
  - Zapotrzebowanie na energię

4.4 Analiza aspektów bezpieczeństwa

Bezpieczeństwo danych i infrastruktury stanowi kluczowy aspekt architektury przetwarzania brzegowego w systemach monitoringu wizyjnego. Urządzenia brzegowe, często rozmieszczone w łatwo dostępnych lokalizacjach, wymagają skutecznych zabezpieczeń zarówno na poziomie sprzętowym, jak i programowym. W celu ochrony poufności i integralności danych zaleca się stosowanie silnego szyfrowania zarówno podczas przechowywania, jak i transmisji. Dostęp do urządzeń powinien być kontrolowany za pomocą mechanizmów uwierzytelniania, takich jak certyfikaty sprzętowe, uwierzytelnianie wieloskładnikowe oraz zarządzanie uprawnieniami użytkowników zgodnie z zasadą najmniejszych uprawnień.

Ważnym elementem jest także fizyczna ochrona urządzeń, obejmująca zabezpieczenia mechaniczne, czujniki manipulacji oraz monitorowanie środowiskowe. Regularne aktualizacje oprogramowania i firmware'u pozwalają na eliminację znanych podatności, a segmentacja sieci oraz monitoring ruchu pomagają w wykrywaniu i blokowaniu prób nieautoryzowanego dostępu. Systemy powinny być wyposażone w mechanizmy wykrywania incydentów i reagowania na nie, takie jak izolacja zainfekowanych urządzeń czy automatyczne raportowanie zdarzeń. Wdrożenie redundancji oraz planów ciągłości działania minimalizuje ryzyko utraty danych i przestojów. Kompleksowe podejście do bezpieczeństwa, obejmujące powyższe elementy, nie tylko chroni system przed zagrożeniami, ale także wspiera zgodność z regulacjami prawnymi i buduje zaufanie do rozwiązań opartych na przetwarzaniu brzegowym [11].

5. PODSUMOWANIE

Podsumowując, badanie architektury przetwarzania brzegowego ujawnia jej znaczący potencjał w usprawnianiu systemów analizy wideo poprzez efektywne zarządzanie zasobami i zredukowane opóźnienia. Poprzez strategiczne wykorzystanie zasobów obliczeniowych na brzegu sieci, systemy mogą sprostać rosnącemu zapotrzebowaniu na przetwarzanie wideo w czasie rzeczywistym, niezbędne dla zastosowań od bezpieczeństwa publicznego po inteligentne miasta. Efektywność tej architektury leży w jej zdolności do zrównoważenia ograniczeń zasobów pomiędzy urządzeniami końcowymi, urządzeniami brzegowymi i systemami chmurowymi, jak wskazuje potrzeba kompleksowych badań w tej dziedzinie [12].

Ponadto, usprawnienia przewidywane wraz z implementacją technologii 5G obiecują dalszą optymalizację prędkości transmisji i przetwarzania danych, czyniąc przetwarzanie brzegowe niezbędnym komponentem dla przyszłych infrastruktur komunikacyjnych [13]. Wraz z postępem badań, adresowanie obecnych luk, szczególnie w niefunkcjonalnych właściwościach i zarządzaniu energią, będzie kluczowe dla realizacji pełnych możliwości analizy wideo opartej na przetwarzaniu brzegowym.

6. PODZIĘKOWANIE

Podziękowanie dla firmy Aiseemo Sp. z o. o. za udostępnienie serwerów obliczeniowych, co pozwoliło przeprowadzenie pomiarów we wszystkich rozważanych scenariuszach testowych.

LITERATURA

[1] Conchari Christian, Guillermo Sahonero-Alvarez, Raul Mollocuaquira, Edgar Salazar. 2024. „Distributed Edge Computing for Appearance-Based Gait Recognition". *IEEE ANDESCON, 2024*. Online. Available: https://www.semanticscholar.org/paper/f5fac3b62ba605b010382af968553d09ad578766. [Dostęp: 2025-04-08].

[2] Fajardo Jose Oscar, Fidel Liberal, Ioannis Giannoulakis, Emmanouil Kafetzakis, Vincenzo Pii, Irena Trajkovska, et al. 2016. „Introducing mobile edge computing capabilities through distributed 5G Cloud Enabled Small Cells". *Springer Science and Business Media LLC*. Online. Available: https://core.ac.uk/download/81572469.pdf. [Dostęp: 2025-04-08].

[3] Gong Linxiao, Hao Yang, Gaoyun Fang, Bobo Ju, Juncen Guo, Xiaoguang Zhu, Xiping Hu, Yan Wang, Peng Sun, Azzedine Boukerche. 2025. „A Survey on Video Analytics in Cloud-Edge-Terminal Collaborative Systems". *ArXiv*. Online. Available: https://arxiv.org/abs/2502.06581. [Dostęp: 2025-04-08].

[4] Krestinskaya Olga, Alex Pappachen James, Leon O. Chua. 2018. „Neuro-memristive Circuits for Edge Computing: A review". *Institute of Electrical and Electronics Engineers (IEEE)*. Online. Available: http://arxiv.org/abs/1807.00962. [Dostęp: 2025-04-08].

[5] Kaleem Zeeshan, Muhammad Yousaf, Aamir Qamar, Ayaz Ahmad, Trung Q. Duong, Wan Choi, Abbas Jamalipour. 2019. „UAV-Empowered Disaster-Resilient Edge Architecture for Delay-Sensitive Communication". Online. Available: http://arxiv.org/abs/1809.09617. [Dostęp: 2025-04-08].

[6] Markov Igor L. 2014. „Limits on Fundamental Limits to Computation". *Springer Science and Business Media LLC*. Online. Available: http://arxiv.org/abs/1408.3821. [Dostęp: 2025-04-08].

[7] Latifi Mohammad. 2024. „Synergizing Could And On-Promises: A Knowledge Management Tool For Laboratorio Di Storia Orale". Online. Available: https://core.ac.uk/download/604157890.pdf. [Dostęp: 2025-04-08].

[8] Yuan Shijing, Yuxin Liu, Song Guo, Jie Li, Hongyang Chen, Chentao Wu, Yang Yang. 2025. „Efficient Online Computing Offloading for Budget-Constrained Cloud-Edge Collaborative Video Streaming Systems". *IEEE Transactions on Cloud Computing*. Online. Available: https://www.semanticscholar.org/paper/d0055dd3552f156b26ab6afab661cb97b66adf45. [Dostęp: 2025-04-08].

[9] Cui Huan, Qing Li, Hanling Wang, Yong Jiang. 2024. „CloudEye: A New Paradigm of Video Analysis System for Mobile Visual Scenarios". *ArXiv*. Online. Available: https://www.semanticscholar.org/paper/73d773565c2c7d84459ba1530abed1fd6ee2edd8. [Dostęp: 2025-04-08].

[10] Toczé Klervie, Simin Nadjm-Tehrani. 2018. „A Taxonomy for Management and Optimization of Multiple Resources in Edge Computing". Online. Available: http://arxiv.org/abs/1801.05610. [Dostęp: 2025-04-08].

[11] Wei Liang Yu, Fan He Xiaofei, Hatcher William, Lu Chao, Lin Jie, Yang Xinyu. 2023. „A Survey on Edge Computing for the Internet of Things". *IEEE Access*, 11: 12156–12185. doi: 10.1109/ACCESS.2023.3245678. [Dostęp: 2025-04-08].

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

IDENTYFIKACJA NIESTACJONARNYCH KANAŁÓW KOMUNIKACYJNYCH

IDENTIFICATION OF NONSTATIONARY COMMUNICATION CHANNELS

Marcin Ciołek

Wydział Elektroniki, Telekomunikacji i Informatyki, Katedra Sygnałów i Systemów, Politechnika Gdańska, Gdańsk, marciole@pg.edu.pl

DOI: 10.15199/59.2025.4.14

Streszczenie: W niniejszym artykule rozważany jest problem identyfikacji niestacjonarnych kanałów komunikacyjnych o rzadkiej strukturze, z użyciem lokalnych funkcji bazowych. Typowe algorytmy wykorzystują regularyzację ℓ_1, która nie zawsze ma rozwiązanie analityczne i wymaga wtedy kosztownego wyszukiwania numerycznego. Proponowany jest szybki regularyzowany lokalny estymator funkcji bazowych, oparty na przybliżonej regularyzacji ℓ_2.

Abstract: In this paper identification of nonstationary communication channels with sparse structure using local basis functions is addressed. While standard approaches rely on ℓ_1 regularization, in cases where no closed-form solution exists, they require a computationally expensive numerical search. A fast regularized local basis functions algorithm based on reweighted ℓ_2 regularization is proposed.

Słowa kluczowe: akustyka podwodna, kanał rzadki, preestymaty, regularyzacja, zmienny w czasie system.

Keywords: preestimate, regularization, sparse channel, time-varying system, underwater acoustics.

1. WSTĘP

Wiele niestacjonarnych kanałów komunikacyjnych (naziemnych, podwodnych) może być dobrze aproksymowanych przez model zmienny w czasie o skończonej odpowiedzi impulsowej (ang. *Finite Impulse Response*, FIR) [10] – [9]

$$y(t) = \sum_{i=1}^{n} \theta_i^*(t) u(t-i+1) + e(t) = \boldsymbol{\theta}^{\mathrm{H}}(t)\boldsymbol{\varphi}(t) + e(t) \quad (1)$$

gdzie $t = \ldots, -1, 0, 1, \ldots$ oznacza dyskretny (znormalizowany) czas, $y(t)$ to zespolony sygnał wyjściowy, $\boldsymbol{\varphi}(t) = [u(t), \ldots, u(t-n+1)]^{\mathrm{T}}$ to wektor regresji złożony z przeszłych próbek zespolonego sygnału wejściowego $u(t)$, $\boldsymbol{\theta}(t) = [\theta_1(t), \ldots, \theta_n(t)]^{\mathrm{T}}$ to wektor zmiennych w czasie współczynników systemu, a $\{e(t)\}$ oznacza szum pomiarowy. Symbol $*$ oznacza sprzężenie zespolone, a „H” — transpozycję hermitowską (sprzężoną zespoloną).

Przykładem niestacjonarnego kanału komunikacyjnego jest model w podwodnym systemie akustycznym typu full-duplex. W takich systemach antena nadawcza i antena odbiorcza działają jednocześnie w tym samym paśmie częstotliwości [7]. W konsekwencji sygnał z nadajnika może trafić z powrotem do odbiornika, powodując interferencje własne. Celem jest wyodrębnienie sygnału $e(t)$ od $y(t)$ w równaniu (1), który zawiera zarówno sygnał odległy, jak i szum pomiarowy, poprzez usunięcie składowej interferencji własnych $\boldsymbol{\theta}^{\mathrm{H}}(t)\boldsymbol{\varphi}(t)$. Interferencje własne są spowodowane odbiciami od otaczających obiektów, takich jak powierzchnia morza, dno morskie, ryby czy statki. Ponieważ sygnał własny $u(t)$, generowany przez antenę nadawczą, jest zawsze znany, problem sprowadza się do śledzenia współczynników odpowiedzi impulsowej kanału interferencji własnych. Taki kanał komunikacyjny wykazuje cechę rzadkości, gdyż tylko niewielka liczba zmieniających się w czasie współczynników różni się od zera [10]. Algorytmy estymacji uwzględniające rzadką strukturę odpowiedzi impulsowej zazwyczaj wykorzystują regularyzację ℓ^1 i należą do rodziny metod LASSO (ang. *Least Absolute Shrinkage and Selection Operator*) [11]. Ponieważ problem minimalizacji nie zawsze ma rozwiązania w postaci analitycznej, konieczne jest zastosowanie wyszukiwania numerycznego co jest istotną wadą w przypadku estymacji realizowanej w trybie przesuwającego się okna.

W niniejszej pracy zaproponowano szybką regularyzowaną lokalną metodę funkcji bazowych (ang. *fast Regularized Local Basis Functions, fRLBF*), wykorzystującą dwa człony regularyzacyjne, mające na celu rozwiązanie przedstawionych problemów. W proponowanym podejściu regularyzację ℓ^1 zastąpiono odpowiednio przeskalowaną regularyzacją ℓ^2, co pozwoliło na uzyskanie rozwiązania w postaci analitycznej.

2. LOKALNA METODA FUNKCJI BAZOWYCH

Przyjmijmy typowe założenia dla systemów komunikacji bezprzewodowej:

(A1) $\{u(t)\}$ jest białym szumem kołowym o zerowej wartości średniej i wariancji σ_u^2.

(A2) $\{e(t)\}$ jest białym szumem kołowym o zerowej wartości średniej i wariancji σ_e^2, niezależnym od $\{u(t)\}$.

(A3) $\{\boldsymbol{\theta}(t)\}$ jest ciągiem niezależnym od $\{u(t)\}$ i $\{e(t)\}$.

Biały szum kołowy to ciąg niezależnych i identycznie rozłożonych zmiennych losowych, których części rzeczywista i urojona są wzajemnie niezależne.

Lokalna metoda identyfikacji za pomocą funkcji bazowych (ang. *Local Basis Functions*, LBF) zakłada, że w lokalnym przedziale analizy $T(t) = [t-k, t+k]$ o długości $K = 2k+1$, wyśrodkowanym na chwili t, parametry systemu mogą być wyrażone jako liniowe kombinacje pewnej liczby niezależnych (rzeczywistych lub zespolonych) funkcji czasu, nazwanych funkcjami bazowymi. W ten sposób uzyskujemy poniższy hipermodel

$$\begin{gathered}\theta_i(t+j) = \sum_{l=1}^{m} f_l^*(j) a_{il}(t) = \mathbf{f}^{\mathrm{H}}(j)\boldsymbol{\alpha}_i(t) \\ i = 1, \ldots, n, \quad \boldsymbol{\alpha}_i(t) = [a_{i1}(t), \ldots, a_{im}(t)]^{\mathrm{T}}\end{gathered} \tag{2}$$

gdzie $\mathbf{f}(j) = [f_1(j), \ldots, f_m(j)]^{\mathrm{T}},\ j \in I_k = [-k, k]$.

Zgodnie z estymacją lokalną, trajektorie parametrów oparte na hipermodelu (2) są estymowane niezależnie dla każdego położenia przedziału analizy $T(t)$, z wykorzystaniem metody przesuwającego się okna. Oznacza to, że hiperparametry systemu a_{il} są traktowane jako stałe w każdym przedziale $[t-k, t+k]$, jednak ich wartości mogą zmieniać się wraz z przesuwaniem się okna analizy, dlatego są one również funkcjami czasu t.

Hipermodel (2) można zapisać w zwartej formie

$$\boldsymbol{\theta}(t+j) = \mathbf{F}(j)\boldsymbol{\alpha}(t), \quad j \in I_k \tag{3}$$

gdzie $\boldsymbol{\alpha}(t) = [\boldsymbol{\alpha}_1^{\mathrm{T}}, \ldots, \boldsymbol{\alpha}_n^{\mathrm{T}}]^{\mathrm{T}}$, $\mathbf{F}(j) = \mathbf{I}_n \otimes \mathbf{f}^{\mathrm{H}}(j)$, a $\otimes$ oznacza iloczyn Kroneckera odpowiednich wektorów/macierzy. Korzystając z równania (3), równanie systemu (1) można przekształcić do postaci

$$y(t+j) = \boldsymbol{\alpha}^{\mathrm{H}}(t)\boldsymbol{\psi}(t,j) + e(t+j), \quad j \in I_k \tag{4}$$

gdzie $\boldsymbol{\psi}(t,j) = \boldsymbol{\varphi}(t+j) \otimes \mathbf{f}(j)$ oznacza uogólniony wektor regresji.

Najczęściej stosowane funkcje bazowe, które umożliwiają rekursywne obliczenia przy przesuwającym się oknie analizy, to potęgi czasu, funkcje cosinusów oraz zespolone funkcje wykładnicze [10]. W niniejszej pracy przyjmujemy rzeczywiste potęgi czasu jako funkcje bazowe (dla ogólności zachowujemy transpozycję sprzężoną). Dla wygody, i bez utraty ogólności przyjmujemy, że wybrane funkcje bazowe są ortonormalne, tj.

$$\sum_{j=-k}^{k} \mathbf{f}(j)\mathbf{f}^{\mathrm{H}}(j) = \mathbf{I}_m, \quad \sum_{j=-k}^{k} \mathbf{F}^{\mathrm{H}}(j)\mathbf{F}(j) = \mathbf{I}_{nm} \tag{5}$$

gdzie $\mathbf{I}_p$ oznacza macierz jednostkową o wymiarze $p \times p$.

Estymator LBF ma postać [5]

$$\widehat{\boldsymbol{\theta}}^{\mathrm{LBF}}(t) = \mathbf{F}_0 \widehat{\boldsymbol{\alpha}}^{\mathrm{LBF}}(t) \tag{6}$$

$$\widehat{\boldsymbol{\alpha}}^{\mathrm{LBF}}(t) = \arg\min_{\boldsymbol{\alpha}} \sum_{j=-k}^{k} |y(t+j) - \boldsymbol{\alpha}^{\mathrm{H}}\boldsymbol{\psi}(t,j)|^2 \tag{7}$$

$$= \left[\sum_{j=-k}^{k} \boldsymbol{\psi}(t,j)\boldsymbol{\psi}(t,j)^{\mathrm{H}}\right]^{-1} \left[\sum_{j=-k}^{k} \boldsymbol{\psi}(t,j) y^*(t+j)\right]$$

gdzie $\mathbf{F}_0 = \mathbf{F}(0) = \mathbf{I}_n \otimes \mathbf{f}_0^{\mathrm{H}}, \mathbf{f}_0 = \mathbf{f}(0)$.

Istotną cechą estymatora LBF jest to, że dokonuje on identyfikacji nieprzyczynowej (ang. non-causal), co zmniejsza opóźnienie estymacji, jednak wiąże się z opóźnieniem decyzyjnym. Liczba estymowanych hiperparametrów nm determinuje minimalną długość okna K, która musi być znacznie większa od nm, aby uniknąć problemów numerycznych. Niestety, technika LBF wymaga odwracania macierzy o wymiarach $nm \times nm$ w każdej chwili t, co stanowi wadę obliczeniową.

3. SZYBKA LOKALNA METODA FUNKCJI BAZOWYCH

Jak pokazano w [6], przy założeniach (A1)–(A3), estymaty LBF $\widehat{\boldsymbol{\alpha}}^{\mathrm{LBF}}(t)$ oraz $\widehat{\boldsymbol{\theta}}^{\mathrm{LBF}}(t)$ mogą być aproksymowane za pomocą następujących, obliczeniowo szybkich formuł (ang. *fast Local Basis Functions*, fLBF)

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}^{\mathrm{fLBF}}(t) &= \arg\min_{\boldsymbol{\alpha}} \sum_{j=-k}^{k} ||\widetilde{\boldsymbol{\theta}}(t+j) - \mathbf{F}(j)\boldsymbol{\alpha}||_2^2 \\ \widehat{\boldsymbol{\theta}}^{\mathrm{fLBF}}(t) &= \mathbf{F}_0 \widehat{\boldsymbol{\alpha}}^{\mathrm{fLBF}}(t) = \sum_{j=-k}^{k} \mathbf{f}_0^{\mathrm{H}} \mathbf{f}(j) \widetilde{\boldsymbol{\theta}}(t+j)\end{aligned} \tag{8}$$

gdzie $\{\widetilde{\boldsymbol{\theta}}(t)\}$ oznacza wektor preestymowanych trajektorii. Te preestymaty można uzyskać poprzez filtrację odwrotną estymat otrzymanych za pomocą algorytmu najmniejszych kwadratów z zapominaniem wykładniczym (ang. *Exponentially Weighted Least Squares*, EWLS):

$$\widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t) = \arg\min_{\boldsymbol{\theta}} \sum_{j=0}^{t-1} \lambda^j |y(t-j) - \boldsymbol{\theta}^{\mathrm{H}}\boldsymbol{\varphi}(t-j)|^2 \tag{9}$$

gdzie $\lambda,\ 0 < \lambda < 1$, oznacza stałą zapominania. Estymaty EWLS można obliczyć za pomocą znanego algorytmu rekursywnego [8]:

$$\begin{aligned}\varepsilon(t) &= y(t) - \left[\widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t-1)\right]^{\mathrm{H}} \boldsymbol{\varphi}(t) \\ \mathbf{k}(t) &= \frac{\mathbf{R}(t-1)\boldsymbol{\varphi}(t)}{\lambda + \boldsymbol{\varphi}^{\mathrm{H}}(t)\mathbf{R}(t-1)\boldsymbol{\varphi}(t)} \\ \widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t) &= \widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t-1) + \mathbf{k}(t)\varepsilon^*(t) \\ \mathbf{R}(t) &= \frac{1}{\lambda}[\mathbf{I}_n - \mathbf{k}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t)]\mathbf{R}(t-1)\end{aligned} \tag{10}$$

z warunkami początkowymi $\widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(0) = 0$ oraz $\mathbf{R}(0) = c\mathbf{I}_n$, gdzie c oznacza dużą dodatnią stałą.

Filtracja odwrotna może być zdefiniowana w następujący sposób

$$\widetilde{\boldsymbol{\theta}}(t) = L_t \widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t) - \lambda L_{t-1} \widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t-1) \tag{11}$$

gdzie $L_t = \sum_{j=0}^{t-1} \lambda^j = \lambda L_{t-1} + 1,\ L_0 = 1$ oznacza efektywną szerokość okna wykładniczego. Dla dużych wartości t, gdy efektywna szerokość okna osiąga stan ustalony $L_\infty = 1/(1-\lambda)$, wzór (11) można zastąpić

$$\widetilde{\boldsymbol{\theta}}(t) = \frac{1}{1-\lambda}\left[\widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t) - \lambda \widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t-1)\right]. \tag{12}$$

Zalecany wybór współczynnika zapominania to $\lambda = \max\{0{,}9; 1 - 2/n\}$.

Przy założeniach (A1)–(A3), preestymatę $\widetilde{\boldsymbol{\theta}}(t)$ można zapisać w postaci

$$\widetilde{\boldsymbol{\theta}}(t) \cong \boldsymbol{\theta}(t) + \mathbf{z}(t), \tag{13}$$

gdzie $\mathbf{z}(t)$ oznacza biały szum o zerowej wartości średniej i dużej macierzy kowariancji $\mathrm{cov}[\mathbf{z}(t)] = \sigma_z^2 \mathbf{I}_n$. W tej sytuacji estymatę fLBF $\widehat{\boldsymbol{\theta}}^{\mathrm{fLBF}}(t)$ można interpretować jako wynik „odszumiania" preestymaty $\widetilde{\boldsymbol{\theta}}(t)$ uzyskany z użyciem podejścia opartego na funkcjach bazowych [6].

4. REGULARYZOWANA METODA FLBF

Oznaczmy przez $||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^{n}|x_i|^2}$ normę ℓ^2 wektora zespolonego $\mathbf{x} = [x_1, \ldots, x_n]^{\mathrm{T}}$, a przez $||\mathbf{x}||_1 = \sum_{i=1}^{n}|x_i|$ — jego normę ℓ^1. Na koniec, niech $||\mathbf{x}||_{\mathbf{A}} = \sqrt{\mathbf{x}^{\mathrm{H}}\mathbf{A}\mathbf{x}}$, gdzie $\mathbf{A}$ oznacza macierz Hermitowską dodatnio określoną o wymiarach $n \times n$.

Zdefiniujmy estymator fLBF z regularyzacją typu LASSO uwzględniający dwa człony regularyzacyjne, jako

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}^{\mathrm{LASSO}}(t) = \arg\min_{\boldsymbol{\alpha}}\{ &\sum_{j=-k}^{k}||\widetilde{\boldsymbol{\theta}}(t+j) - \mathbf{F}(j)\boldsymbol{\alpha}||_2^2 \\ &+\eta||\boldsymbol{\alpha}||_1 + \mu||\mathbf{F}_0\boldsymbol{\alpha}||_1\} \\ \widehat{\boldsymbol{\theta}}^{\mathrm{LASSO}}(t) &= \mathbf{F}_0\widehat{\boldsymbol{\alpha}}^{\mathrm{LASSO}}(t)\end{aligned} \tag{14}$$

gdzie $\eta, \mu > 0$ oznaczają współczynniki regularyzacji. Pierwszy człon regularyzacyjny, $\eta||\boldsymbol{\alpha}||_1$, promuje rzadkość w dziedzinie częstotliwości, poprzez karanie nadmiernej liczby funkcji bazowych użytych do aproksymacji zmiennych w czasie parametrów kanału $\theta_1(t), \ldots, \theta_n(t)$. Drugi człon regularyzacyjny, $\mu||\mathbf{F}_0\boldsymbol{\alpha}||_1 = \mu||\boldsymbol{\theta}(t)||_1$, promuje rzadkość w dziedzinie czasu/opóźnienia, poprzez karanie niezerowych składowych wektora $\boldsymbol{\theta}(t)$. Dzięki uwzględnieniu drugiego członu regularyzacji, schemat (14) stanowi rozwiązanie typu LASSO, zwiększające rzadkość modelu zarówno na poziomie pojedynczych (hiperparametrów), jak i grupowych (parametrów) poziomów składowych [3].

Reweighting to znana technika optymalizacji, polegająca na zmianie wag [1], [2]. Zauważmy, że norma ℓ^1 wektora $\mathbf{x}$ może być wyrażona jako

$$||\mathbf{x}||_1 = \sum_{i=1}^{n}\frac{x_i^* x_i}{|x_i|} = \mathbf{x}^{\mathrm{H}}\mathbf{W}\mathbf{x} = ||\mathbf{x}||_{\mathbf{W}}^2 \tag{15}$$

gdzie $\mathbf{W} = \mathrm{diag}\{|x_1|^{-1}, \ldots, |x_n|^{-1}\}$. Stosując to przekształcenie, otrzymujemy przybliżenia

$$\begin{aligned}||\boldsymbol{\alpha}(t)||_1 &\cong ||\boldsymbol{\alpha}(t)||_{\boldsymbol{\Delta}(t)}^2 \\ \boldsymbol{\Delta}(t) &= \mathrm{bl\ diag}\{\boldsymbol{\Delta}_1(t), \ldots, \boldsymbol{\Delta}_n(t)\} \\ \boldsymbol{\Delta}_i(t) &= \mathrm{diag}\{\delta_{i1}(t), \ldots, \delta_{im}(t)\} \\ \delta_{il}(t) &= |\widehat{a}_{il}^{\mathrm{fLBF}}(t)|^{-1}, i = 1, \ldots, n,\ l = 1, \ldots, m\end{aligned}$$

i

$$\begin{aligned}||\boldsymbol{\theta}(t)||_1 = ||\mathbf{F}_0\boldsymbol{\alpha}(t)||_1 &\cong ||\boldsymbol{\alpha}(t)||_{\mathbf{F}_0^{\mathrm{H}}\boldsymbol{\Gamma}(t)\mathbf{F}_0}^2 \\ \boldsymbol{\Gamma}(t) &= \mathrm{diag}\{\gamma_1(t), \ldots, \gamma_n(t)\} \\ \gamma_i(t) &= |\widehat{\theta}^{\mathrm{fLBF}}(t)|^{-1}, i = 1, \ldots, n.\end{aligned}$$

Powyższe przekształcenia, prowadzą do następującej przybliżonej wersji równania (14), dalej nazywanej estymatorem fRLBF (ang. *fast Regularized Local Basis Functions*)

$$\begin{aligned}&\widehat{\boldsymbol{\alpha}}^{\mathrm{fRLBF}}(t) = \\ &\arg\min_{\boldsymbol{\alpha}}\left\{\sum_{j=-k}^{k}||\widetilde{\boldsymbol{\theta}}(t+j) - \mathbf{F}(j)\boldsymbol{\alpha}||_2^2 + ||\boldsymbol{\alpha}||_{\Sigma(t;\eta,\mu)}^2\right\} \\ &= [\mathbf{I}_{mn} + \boldsymbol{\Sigma}(t;\eta,\mu)]^{-1}\widehat{\boldsymbol{\alpha}}^{\mathrm{fLBF}}(t) \\ &\widehat{\boldsymbol{\theta}}^{\mathrm{fRLBF}}(t) = \mathbf{F}_0\widehat{\boldsymbol{\alpha}}^{\mathrm{fRLBF}}(t)\end{aligned} \tag{16}$$

gdzie

$$\begin{aligned}\boldsymbol{\Sigma}(t;\eta,\mu) &= \eta\boldsymbol{\Delta}(t) + \mu\mathbf{F}_0^{\mathrm{H}}\boldsymbol{\Gamma}(t)\mathbf{F}_0 \\ &= \eta\boldsymbol{\Delta}(t) + \mu\boldsymbol{\Gamma}(t) \otimes (\mathbf{f}_0\mathbf{f}_0^{\mathrm{H}}).\end{aligned} \tag{17}$$

Ostatnie przejście w (17) wynika z tożsamości

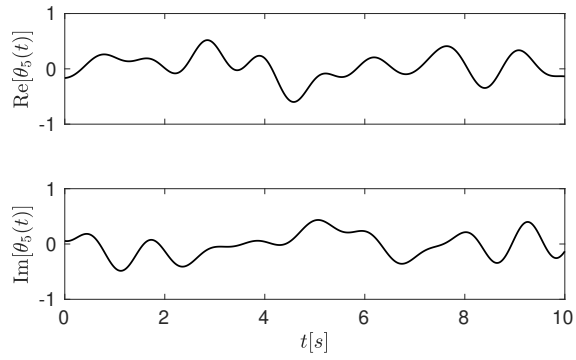
$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{A}\mathbf{C}) \otimes (\mathbf{B}\mathbf{D}) \tag{18}$$

która zachodzi dla iloczynów Kroneckera, pod warunkiem zgodności wszystkich wymiarów. Wykorzystując fakt, że macierz $\boldsymbol{\Sigma}(t;\eta,\mu)$ jest blokowo diagonalna, możliwe jest obliczanie $\widehat{\boldsymbol{\alpha}}^{\mathrm{fRLBF}}(t)$ bez potrzeby odwracania macierzy o rozmiarze $nm \times nm$ w każdej chwili czasu t.

Estymatory typu LASSO odrzucają nieistotne składowe, zastępując je zerami. Proponowany schemat fRLBF przybliża to zachowanie: gdy estymaty fLBF są bardzo małe, wagi $\delta_{il}(t)$ i $\gamma_i(t)$ stają się bardzo duże, skutecznie zmniejszając odpowiadające im estymaty fLBF w kierunku zera. W rezultacie estymator ten promuje rzadkie rozwiązania zarówno w dziedzinie czasu, jak i częstotliwości, przy zachowaniu niskich kosztów obliczeniowych.

5. WYNIKI SYMULACJI

W symulacjach komputerowych wykorzystano model kanału luźno inspirowany akustycznym systemem podwodnym z interferencjami własnymi typu full-duplex, opisanym w pracy [7]. Zgodnie z ich metodologią wygenerowano model kanału o wykładniczej obwiedni, reprezentowany przez filtr FIR o 50 niezależnych współczynnikach zespolonych. Następnie, wprowadzono zmiany do współczynników kanału, aby lepiej odwzorować warunki rzeczywiste: 30% tych współczynników zostało losowo ustawionych na zero, 50% współczynników otrzymało stałe wartości, a 20% współczynników pozostawiono niezmienionych, aby symulować efekty zmieniające się w czasie. Sygnał wejściowy został wygenerowany jako kołowy biały ciąg binarny $u(t) = \pm 1 \pm \mathrm{j}$, natomiast szum pomiarowy modelowano jako kołowy biały szum Gaussowski o wariancji σ_e^2. Sygnały przeskalowano tak, aby uzyskać docelowy stosunek sygnału do szumu równy 10 dB. Typowe trajektorie zmieniających się w czasie współczynników kanału są pokazane na Rysunku 1.



Rysunek 1: Typowe trajektorie zmieniających się w czasie współczynników kanału.

W eksperymencie numerycznym porównano trzy algorytmy: LBF, fLBF i fRLBF. Porównania te przeprowadzono dla wartości parametrów $k = 300$, $m = 5$. W przypadku algorytmu fRLBF przetestowano wszystkie

kombinacje parametrów regularyzacji z zakresu $\{\mu, \eta\} \in \{0; 0{,}1; 0{,}5; 0{,}7\}$. Do oceny skuteczności działania algorytmów wybrano dwie metryki. Pierwszą z nich jest współczynnik eliminacji interferencji własnej (ang. *Self-Interference Cancellation Factor*, SICF) [7], zdefiniowany jako

$$\mathrm{SICF} = \frac{\sum_t |\boldsymbol{\theta}^{\mathrm{H}}(t)\boldsymbol{\varphi}(t)|^2}{\sum_t |[\boldsymbol{\theta}(t) - \widehat{\boldsymbol{\theta}}(t)]^{\mathrm{H}}\boldsymbol{\varphi}(t)|^2}. \tag{19}$$

Druga metryka mierzy znormalizowany błąd średniokwadratowy dopasowania [4], zdefiniowana jako

$$\mathrm{FIT}(t) = 100\left(1 - \left[\frac{\sum_{i=1}^{50} |\theta_i(t) - \widehat{\theta}_i(t)|^2}{\sum_{i=1}^{50} |\theta_i(t) - \bar{\theta}(t)|^2}\right]^{1/2}\right) \tag{20}$$

gdzie $\bar{\theta}(t) = \frac{1}{50}\sum_{i=1}^{50} \theta_i(t)$. Wartość FIT = 100% oznacza idealne dopasowanie między prawdziwą a oszacowaną odpowiedzią impulsową. W eksperymencie symulacyjnym wszystkie wyniki zostały uśrednione przez 20 niezależnych realizacji oraz po czasie (10000 próbek czasowych). Każda realizacja miała inny przebieg zmian współczynników. Ponadto, każda realizacja danych zawierała dodatkowe 1000 próbek czasowych na początku i na końcu, aby wyeliminować problemy brzegowe.

Otrzymane uśrednione wyniki FIT[%]/SICF[dB] dla metod LBF i fLBF wyniosły odpowiednio 92/25,4 i 91,5/24,6. Metoda LBF była lepsza o ΔFIT = 0,5% i ΔSICF = 0,8 dB, ale bardziej kosztowna obliczeniowo. W tabelach 1 i 2 pokazano uśrednione wyniki FIT/SICF przy różnych wartościach współczynników η i μ. Najlepsze wyniki zostały wyróżnione pogrubieniem. Są one lepsze od wyników metody LBF o ΔFIT = 1,2% i ΔSICF = 1,1 dB. Warto zauważyć, że wyniki w tabelach, dla kombinacji współczynników $\eta = 0$ i $\mu = 0$, odpowiadają wynikom metody fLBF. Stosując regularyzację tylko jednego z członów obserwujemy przeważnie początkowy wzrost, a następnie spadek wartości w obu metrykach. W przypadku uwzględnienia obu członów regularyzacyjnych otrzymujemy lepsze wyniki niż wyniki metod LBF i fLBF.

W przyszłych pracach planuje się zastosowanie adaptacyjnego doboru wartości współczynników regularyzacyjnych w dwóch wariantach. Pierwszy scenariusz zakłada, że w danej chwili czasu wybrane współczynniki η i μ będą jednakowe dla wszystkich parametrów modelu systemu. Drugi wariant przewiduje bardziej elastyczne podejście, w którym optymalizacja współczynników regularyzacyjnych będzie prowadzona niezależnie dla każdego parametru. Dostosowanie wartości η i μ do lokalnych warunków może przynieść dalszą poprawę jakości estymacji.

Tabela 1: Uśredniony FIT [%] dla fRLBF przy różnych $\eta\backslash\mu$. Najlepsze wyniki zostały pogrubione.

$\eta\backslash\mu$	0	0,1	0,5	0,7
0	91,5	91,6	92,1	92,3
0,1	93	93,1	**93,2**	**93,2**
0,5	92,6	92,6	92,4	92,4
0,7	91,6	91,5	91,4	91,3

Tabela 2: Uśredniony SICF [dB] dla fRLBF przy różnych $\eta\backslash\mu$. Najlepsze wyniki zostały pogrubione.

$\eta\backslash\mu$	0	0,1	0,5	0,7
0	24,6	24,8	25,4	25,5
0,1	**26,5**	**26,5**	**26,5**	26,4
0,5	24,1	24,1	23,9	23,8
0,7	22,6	22,6	22,5	22,4

6. PODSUMOWANIE

Zaproponowany estymator fRLBF umożliwia identyfikację niestacjonarnych kanałów komunikacyjnych o rzadkiej strukturze odpowiedzi impulsowej, w której tylko nieliczne współczynniki mają istotne wartości. Schemat ten może być stosowany w analizie z użyciem przesuwającego się okna bez potrzeby odwracania macierzy, co jest wymagane w klasycznej metodzie LBF. Przy odpowiednim doborze współczynników regularyzacyjnych estymator fRLBF osiąga lepsze wyniki niż metody LBF i fLBF.

LITERATURA

[1] Bani Mohamed S., Chalmers Bruce L. 1984. „Best approximation in L^{inf} via iterative Hilbert space procedures". Journal of Approximation Theory, 42 (2) : 173–180.

[2] Burrus Charles S., Barreto J. A., Selesnick Ivan W. 1994. „Iterative reweighted least squares design of FIR filters". IEEE Transactions on Signal Processing, 42 : 2926–2936.

[3] Friedman Jerome, Hastie Trevor, Tibshirani Robert. 2010. „A note on the group lasso and sparse group lasso". arXiv: 1001.0736.

[4] Ljung Lennart, Chen Tianshi. 2013. „What can regularization offer for estimation of dynamic systems?". Proc. of the 11th IFAC Workshop on Adaptation and Learning in Control and Signal Processing, Caen, France, 1–8.

[5] Niedźwiecki Maciej, Ciołek Marcin. 2020. „Generalized Savitzky-Golay filters for identification of nonstationary systems". Automatica, 108 (108477).

[6] Niedźwiecki Maciej, Ciołek Marcin, Gańcza Artur. 2020. „A new look at the statistical identification of nonstationary systems," Automatica, 118 (109037) : 1–9.

[7] Shen Lu, Zakharov Yuriy, et al. 2020. „Adaptive filtering for full-duplex UWA systems with time-varying self-interference channel". IEEE Access, 8 : 187590–187604.

[8] Söderström Torsten, Stoica Peter. 1988. System Identification. Prentice-Hall.

[9] Stojanovic Milica, Preisig James. 2009. „Underwater acoustic communication channels: Propagation models and statistical characterization". IEEE Communications Magazine, 47 (1) : 84–89.

[10] Tsatsanis Michail K., Giannaki Georgios B. 1996. „Modeling and equalization of rapidly fading channels". Int. J. Adaptive Contr. Signal Process. 10 (2-3) : 159–176.

[11] Tibshirani Robert. 1996. „Regression shrinkage and selection via the LASSO". J. R. Statist. Soc. B, 58 (1) : 267–288.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

OPROGRAMOWANE Z OTWARTYMI ŹRÓDŁAMI W NAUCZANIU TELEKOMUNIKACJI

THE OPEN SOURCE SOFTWARE IN TEACHING TELECOMMUNICATIONS

Marcin Kucharczyk[1]; Grzegorz Dziwoki[2]; Wojciech Sułek[3]

[1] Politechnika Śląska, Katedra Telekomunikacji i Teleinformatyki, Gliwice, Marcin.Kucharczyk@polsl.pl
[2] Politechnika Śląska, Katedra Telekomunikacji i Teleinformatyki, Gliwice, Grzegorz.Dziwoki@polsl.pl
[3] Politechnika Śląska, Katedra Telekomunikacji i Teleinformatyki, Gliwice, Wojciech.Sulek@polsl.pl

DOI: 10.15199/59.2025.4.15

Streszczenie: **Znajomość teorii telekomunikacji oraz symulacyjnych implementacji algorytmów transmisji danych nie zawsze wystarcza do zrozumienia zjawisk zachodzących w rzeczywistych systemach. W praktyce pojawiają się zakłócenia i problemy synchronizacyjne, często pomijane w klasycznych modelach modulacji i demodulacji. Moduły radia programowalnego (SDR) umożliwiają praktyczną weryfikację działania tych algorytmów, lecz zwiększają koszty dydaktyczne. W celu ich ograniczenia opracowano środowisko laboratoryjne oparte na sprzęcie o niskim koszcie jednostkowym i otwartym oprogramowaniu. Artykuł omawia wdrożenie tego rozwiązania, jego zalety, ograniczenia oraz przykładowe ćwiczenia z zakresu cyfrowych systemów telekomunikacyjnych.**

Abstract: **Theoretical knowledge of telecommunications and simulation-based implementations of transmission algorithms is often insufficient for understanding phenomena in real-world systems. Practical issues such as interference and synchronization errors frequently arise but are typically overlooked in classical models of modulation and demodulation. Software-defined radio (SDR) platforms enable hands-on validation of these algorithms but introduce additional educational costs. To mitigate this, a laboratory environment based on low-unit-cost hardware and open-source software was developed. The article discusses the deployment of this solution, its advantages and limitations, and presents example lab exercises related to digital communication systems.**

Słowa kluczowe: **nauczanie telekomunikacji, oprogramowanie otwartoźródłowe, radio definiowane programowo.**

Keywords: **open source software, software defined radio, telecommunications education.**

1. WSTĘP

Nauczanie podstaw telekomunikacji, tak jak i innych przedmiotów technicznych, wymaga nie tylko przedstawienia podstaw teoretycznych, ale również analizy problemów związanych tematyką przedmiotu w ramach zajęć laboratoryjnych, gdzie implementowane są poznane wcześniej algorytmy. Zrozumie działania zaawansowanych systemów telekomunikacyjnych 5G i nadchodzących 6G, coraz szybszych sieci Wi-Fi, czy systemów rozsiewczych cyfrowej telewizji i radia, wymaga najpierw opanowania podstaw związanych z modulacjami analogowymi i cyfrowymi, filtracją sygnałów, czy przetwarzaniem pakietów.

W roku 2017 została wydana książka „MATLAB i podstawy telekomunikacji", pod redakcją Jacka Izydorczyka [10], napisana przez pracowników ówczesnego Zakładu Telekomunikacji w Instytucie Elektroniki Politechniki Śląskiej, który podczas zmiany struktury organizacyjnej na Uczelni został przekształcony w Katedrę Telekomunikacji i Teleinformatyki. Książka omawia podstawowe algorytmy stosowane w systemach telekomunikacyjnych, zarówno analogowych jak i cyfrowych, oraz ich przykładowe implementacje w języku MATLAB.

Aktualnie Politechnika Śląska posiada pełną licencję CWS (*ang.* Campus Wide Suite) na oprogramowanie MATLAB firmy MathWorks oraz dodatkowe pakiety rozszerzające funkcjonalność, ale w czasie powstawania książki zakres licencji był ograniczony i pozycja powstawała ze świadomością kosztów, jakie zakup licencji niesie. Przykładowy kod powstawał więc tak, żeby można go było uruchomić na wersji podstawowej aplikacji z rozszerzeniem *Signal Processing Toolbox*. Rozszerzenie to można uznać za podstawowe w przypadku przetwarzania sygnałów, nie tylko w telekomunikacji, ponieważ zawiera podstawowe funkcje do analizy widmowej czy narzędzia do projektowania filtrów.

Dzięki ograniczeniu korzystania z pakietów dodatkowych środowiska firmy MathWorks zadania będące integralną częścią wspomnianej książki można realizować bez problemów w oprogramowaniu GNU Octave z rozszerzeniem *signal package*. Jak podają twórcy oprogramowania służy ono do obliczeń numerycznych, które definiowane są w języku o wysokiej zgodności z językiem MATLAB. Oprogramowanie to dostępne jest bezpłatnie. Jest dostępne w postaci źródłowej oraz skompilowanej na wszystkie popularne systemy operacyjne. Licencja nie stawia ograniczeń jak chodzi o korzystanie zarówno dla użytkowników prywatnych, edukacyjnych jak i komercyjnych.

Kolejny krok w rozwoju algorytmów obliczeniowych, a następnie w nauczaniu telekomunikacji to pojawienie się na rynku urządzeń określanych skrótem SDR (*ang.* Software Defined Radio), czyli radia definiowanego programowo. Algorytmy implementowane są w środowisku symulacyjno-obliczeniowym, lecz przetwarzają rzeczywiste sygnały radiowe – pojawia się szum, zakłócenia i zniekształcenia powodowane niedoskonałością sprzętu.

2. RADIO DEFINIOWANE PROGRAMOWO

Rosnąca moc obliczeniowa mikroprocesorów umożliwiła z jednej strony praktyczną implementację algorytmów, które pojawiły się w literaturze już wcześniej, jak FFT czy propagacja przekonań (*ang.* belief propagation) w dekoderze kodu LDPC. A z drugiej spowodowały rozwój algorytmów obliczeniowych używanych w systemach telekomunikacyjnych. Testowanie nowych algorytmów wyłącznie w środowiskach symulacyjnych nie daje jednak pełnego obrazu jeżeli chodzi o ich działanie w rzeczywistości. Taką możliwość daje połączenia systemu obliczeniowego z urządzeniem SDR.

Przy budowie nadajnika wejściem układu SDR jest ciąg próbek, a dokładniej dwa ciągi próbek, odpowiednio dla składowej synfazowej I i składowej kwadraturowej Q, które po przejściu przez przetworniki cyfrowo-analogowe wysyłane są przez układ radiowy. W odbiorniku dwa przetworniki analogowo-cyfrowe próbkują wzajemnie ortogonalne sygnały i przesyłają próbki I oraz Q do systemu obliczeniowego. Próbkowany sygnał zawiera dodatkowe zakłócenia związane z szumem czy charakterystyką kanału radiowego, oraz zniekształcenia związane z brakiem synchronizacji modułów nadawczego i odbiorczego. SDR zapewnia więc interfejs radiowy, a za realizację algorytmów nadajnika lub odbiornika odpowiedzialny jest układ liczący. W najbardziej wydajnych urządzeniach układ ten realizowany jest w postaci odpowiednio skonfigurowanej matrycy FPGA, natomiast przekierowanie obliczeń do programu uruchomionego na komputerze osobistym poszerza krąg użytkowników. Ograniczeniem wydajności w takiej konfiguracji jest moc obliczeniowa komputera oraz szybkość interfejsu przez jaki przesyłane są dane pomiędzy urządzeniami. Wysoką wydajność mają urządzenia z interfejsem Ethernet, PCIe czy USB 3.0. Zastosowanie interfejsu USB 2.0 ogranicza przepustowość, ale obniża koszt urządzenia.

Podążając za informacjami z Wikipedii można zauważyć, że idea jaką realizują obecnie urządzenia SDR pojawiła się w roku 1970, a pierwsze urządzenia to rok 1982 [3]. W kontekście dostępu amatorskiego za początek można uznać rok 2004 [9] kiedy pojawiły się relatywnie tanie urządzenia USRP1 (*ang.* Universal Software Radio Peripheral), a wraz z nimi sterownik UHD (*ang.* USRP Hardware Driver) [5,6], którego kod źródłowy został udostępniony. Efektem tego była i ciągle jest możliwość obsługi tych urządzeń z poziomu różnych aplikacji. USRP1, podobnie jak jego następcy, przykładowo model B100 z interfejsem USB 2.0, zapewniał interfejs do przesyłania danych z komputerem, a za transmisję radiową odpowiedzialny był dodatkowy moduł zawierający mieszacz.

W roku 2014 firma Analog Devices wypuściła na rynek układy AD9361/AD9363/AD9364 [1], które stanowią konfigurowalne programowo układy radiowe zawierające zarówno interfejs komunikacyjny jak i moduł radiowy. Stają się one m.in. podstawą urządzeń USRP z serii B200/B210. W roku 2016 firma Analog Devices zaprezentowała urządzenia ADALM-PLUTO [2], które na bazie tego samego układu co B2xx, ma stanowić tanią platformę edukacyjną umożliwiającą zapoznanie się z podstawami działania systemów SDR.

Produkty Ettusa uznawane są za niedrogie, ale koszt wspomnianych urządzeń to 6-12 tys. zł zależnie od wersji. Natomiast cena ADALM-PLUTO została ustalona na poziomie poniżej 1,5 tys. zł, co czyni te urządzenia bardziej dostępnymi dla indywidualnego użytkownika. W kwestii wyposażenia laboratorium, gdzie potrzeba kilku stanowisk oszczędności są znaczące. Jak chodzi o komunikację z komputerem urządzenie obsługiwane jest przez interfejs IIO [2], źródła sterownika zostały udostępnione i sam sterownik dostępny jest, podobnie jak UHD, na różne systemy operacyjne co umożliwia obsługę z poziomu różnych aplikacji.

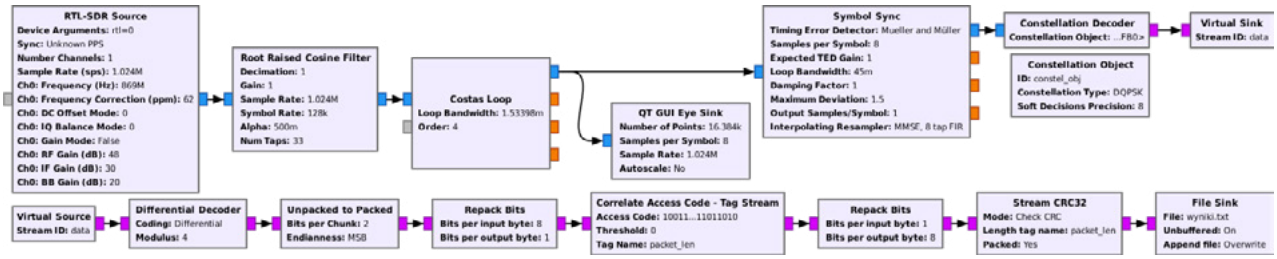
Warto jeszcze wspomnieć o układzie Realtek RTL2832U, który pojawił się na rynku w 2008 roku [11]. Jest to zintegrowany w układzie scalonym odbiornik cyfrowej telewizji DVB-T, jednak bliższa analiza pokazała, że zmieniając sterownik (na ogólny libusb) można bezpośrednio odbierać dane I/Q i przetwarzać je w dowolnym oprogramowaniu. Urządzenia RTL-SDR mogą funkcjonować wyłącznie jako odbiornik i można je kupić w cenie poniżej 100 zł. Kompletując zestaw z ADALM-PLUTO można zbudować stanowisko laboratoryjne, czyli nadajnik i odbiornik SDR, w cenie ok. 1500 zł.

3. OPROGRAMOWANIE DO OBSŁUGI SDR

Poszukując oprogramowania do obsługi modułów SDR, zarówno w trybie nadajnika jak i odbiornika, dobrym rozwiązaniem są produkty komercyjne: MATLAB z modułem Simulink [12] oraz LabVIEW, obydwa oferujące graficzny język programowania. Mimo wysokiego stopnia dopracowania, w praktyce dydaktycznej dostrzegamy ograniczenia związane z kosztami licencyjnymi. Choć w środowisku akademickim dostępne są warianty edukacyjne w obniżonej cenie, próby samodzielnego wdrożenia stanowisk przez studentów (do pracy dyplomowej czy też do użytku amatorskiego) lub uzasadnienie zakupu oprogramowania przez komercyjnego pracodawcę stanowią istotną barierę.

Biorąc pod uwagę powyższe zadecydowano, że laboratorium zostanie uruchomione w oparciu o dostępne na licencji GPL oprogramowanie GNU Radio [4], z otwartymi źródłami, więc teoretycznie z możliwością uruchomienia na wszystkich popularnych systemach operacyjnych. Historia oprogramowania sięga roku 2001, kiedy powstało ono na bazie programu Pspectra z MIT, ale w 2004 roku kod został napisany od nowa, a jednym z twórców, którzy wtedy dołączyli do projektu był Matt Ettus [5,7], czyli twórca USRP, wtedy platformy sprzętowej typu SDR do współpracy z GNU Radio.

Ogólnie GNU Radio jest zestawem bibliotek do cyfrowego przetwarzania sygnałów, a sama aplikacja dba o to, żeby sygnał był efektywnie przesyłany pomiędzy blokami [7]. Interfejs graficzny o nazwie GNU Radio Companion (GRC), który powstał w latach 2006-2007 i został dołączony do pakietu głównego w wersji 3.2. GRC jest generatorem kodu w języku Python dla GNU Radio. Graf przepływu (*ang.* flowgraph) z GRC jest kompilowany, a wynikowy kod tworzy interfejs użytkownika (GUI) dla tworzonej aplikacji oraz połączenia pomiędzy blokami dostępnymi w bibliotekach GNU Radio.



Rys. 1. Przykładowy schemat, graf przepływu, w GNU Radio (odbiornik DQPSK)

Od strony użytkownika mamy więc graficzny interfejs do tworzenia połączeń pomiędzy blokami, który generuje kod efektywnie przetwarzany przez główne biblioteki GNU Radio. Biblioteki te zawierają moduły począwszy standardowych operacji matematycznych i logicznych, przez filtry cyfrowe i analizę Fouriera, bloki do synchronizacji częstotliwościowej i symbolowej, czy detekcji i korekcji błędów, aż po specjalizowane bloki realizujące poszczególne funkcje cyfrowych systemów transmisyjnych jak DVB-T czy IEEE 802.11. Dodatkowo dostępne są bloki do budowy interfejsu użytkownika, czyli zmiany parametrów sygnału w czasie pracy czy wyświetlania wyników obliczeń na ekranie w postaci graficznej: w dziedzinie czasu, częstotliwości, konstelacji, itd.

GRC, podobnie jak inne duże projekty otwartoźródłowe (Open Source), ma dokumentację, której jakość jest różna dla poszczególnych bloków. Dokumentacja na bieżąco jest uzupełniania przez programistów oraz użytkowników, dostępne są również tzw. samouczki (*ang.* tutorials) objaśniające działanie bloków [7,8]. W przypadku braku dobrego opisu użytkownikowi pozostaje analiza źródeł danego bloku celem pełnego zrozumienia jego użycia w schemacie.

Kiedy potrzebnego bloku nie ma w bibliotekach, to oprogramowanie daje możliwość napisania własnego kodu. Nowe bloki można tworzyć zewnętrznie pisząc kod w C++ albo w Pythonie, co wymaga dodatkowych narzędzi do kompilacji. Można też tworzyć bloki typu *Hier Block*, czyli nowe bloki do biblioteki, które wewnątrz złożone są z już istniejących bloków. Wreszcie można tworzyć własne funkcje w Pythonie dodając do schematu *Embedded Python Block*. Dokumentacja opisuje szczegóły użycia tego bloku, czyli wpisania własnego kodu w struktury przetwarzania sygnału GNU Radio. Jedynym wymaganym narzędziem jest edytor tekstowy.

4. PROGRAM LABORATORIUM

Głównym celem stworzenia zestawu zajęć laboratoryjnych powiązanych z modułami SDR było, poza podstawami telekomunikacji, przybliżenie studentom takich zagadnień związanych z transmisją sygnałów jak: zakłócenia (interferencje) innych sygnałów i szumy, regulacja mocy, synchronizacja częstotliwościowa i symbolowa. Wszystkie te elementy można zasymulować, co jednak nie do końca odpowiada zetknięciu się z tym problemami w rzeczywistości. Dodatkowo tani sprzęt w postaci ADALM-PLUTO, a szczególnie RTL-SDR jest przyczyną kolejnych wyzwań po stronie odbiorczej związanych z niedokładnymi i niestabilnymi zegarami taktującymi.

Tworząc program zajęć laboratoryjnych przyjęto, że studenci powinni zapoznać się zarówno z postawami działania układów SDR, jak i mieć możliwość implementacji praktycznych systemów transmisyjnych analogowych i cyfrowych. W przypadku implementacji urządzeń nadawczych transmisja realizowana jest w pasmach ISM, a pracując w kilku sekcjach konieczne jest zadbanie o lokalny przydział pasma radiowego. Tematyka zaproponowanych ćwiczeń jest następująca:

- Podstawy GNU Radio i działania urządzeń SDR: zespolony format sygnału – składowe I oraz Q, synchronizacja nadajnika i odbiornika, filtracja cyfrowa i przesuwanie widma.
- Modulacja amplitudy: AM z detektorem obwiedni i demodulatorem synchronicznym, odtwarzanie fali nośnej w systemie DSB, pętla fazowa PLL.
- Modulacja częstotliwości: nadajnik i odbiornik FM, bloki wbudowane i własne, transmisja stereofoniczna i problemy synchronizacji, detekcja sygnału pilota, przełączanie źródeł sygnału.
- Modulacja OOK: przejście ze świata analogowego do cyfrowego, zamiana ciągu bitów na sygnał analogowy, problemy zajętości pasma i kształtowania impulsów, grupowanie bitów w ramki w nadajniku oraz detekcja nagłówków i odzyskiwanie strumienia danych w odbiorniku.
- Jednotonowe modulacje cyfrowe: kodowanie i dekodowanie konstelacji, synchronizacja częstotliwościowa i symbolowa, detekcja i korekcja błędów, przeplot bitowy.
- Modulator i demodulator OFDM: przykład wykorzystania gotowych bloków dostępnych w GRC.

Na rys. 1 przedstawiono przykładowy schemat, nazywany w GRC grafem przepływu (*ang.* flowgraph) odbiornika cyfrowego sygnału nadawanego z modulacją DQPSK. Na jego podstawie zostaną opisane problemy z jakimi mierzą się studenci podczas zajęć, gdy transmitowane są rzeczywiste sygnały.

Odbiór zaczyna się od skonfigurowania urządzenia odbiorczego (blok *RTL-SDR Source*, który pobiera sygnał z urządzenia zawierającego RTL2832U). Trzeba odpowiednio ustawić parametry takie jak odbierane pasmo (częstotliwość środkowa i próbkowanie), korekcję częstotliwości (współczynnik PPM korygujący generator lokalny SDRa) czy poziom wzmocnienia sygnału.

Filtr RRC przeprowadza dodatkową filtrację dolnoprzepustową sygnału, jednocześnie ogranicza interferencje międzysymbolowe, jeżeli jego parametry są odpowiednio dobrane do podobnego układu w nadajniku. Studenci mogą przetestować jak parametry filtru wpływają

na jakość odbioru. Przy innych technikach modulacji system może być uzupełniony o dodatkowe filtry i układy realizujące przepróbkowanie sygnału.

Ustawianie częstotliwości pracy w urządzeniu SDR nie gwarantuje idealnej synchronizacji odbiornika z nadajnikiem. Powodem jest zarówno rozdzielczość w zakresie ustawiania parametrów jak i niestabilność tych ustawień w czasie. Jest to sytuacja niekorzystnie wpływająca na odbiór ale też taka, którą można określić jako typową w rzeczywistym odbiorniku. Kolejny etap odbioru to synchronizacja częstotliwości (blok *Costas Loop*), gdzie rząd i pasmo pętli należy dostosować do odbieranego sygnału. Przy innych technikach modulacji może tutaj pojawić się pętla fazowa, czyli PLL (*ang.* Phase Locked Loop).

Za blokiem synchronizacji częstotliwościowej w systemie analogowym pojawi się odpowiedni demodulator, który albo będzie dostępny w bibliotece GNU Radio, albo będzie wymagał zbudowania z bloków podstawowych. W systemie cyfrowym dobierając do zastosowanej modulacji odpowiedni algorytm i jego parametry w bloku *Symbol Sync* dokonywana jest synchronizacja symbolowa, czyli wybierane są momenty w których należy próbkować sygnał analogowy. Jeżeli wszystko zostanie poprawnie ustawione i zadziała zgodnie z oczekiwaniami, na wyjściu bloku mamy sytuację, która w literaturze opisywana jest jako „idealna synchronizacja". Następny krok to zamiana analogowego sygnału na cyfrowy i dalsze przetwarzanie dotyczy ciągu bitów.

Od strony „sygnału cyfrowego" GNU Radio daje narzędzia do formowania bitów w ramki w nadajniku, a w odbiorniku, co pokazano na rys. 1, do wyszukiwania początku ramek, czyli ustalonej sekwencji synchronizującej. Dostępne są bloki do kodowania i dekodowania korekcyjnego, detekcji błędów (przykładowo blok *Stream CRC32* z rysunku) czy przeplatania bitów.

Dane transmitowane mogą być pobierane i zapisywane do pliku (jak na rys. 1), ale mogą to być dane z generatorów sygnałów o określonym kształcie albo generatorów losowych. Dostępne są także bloki łączące GNU Radio z systemem audio komputera czy ze stosem sieciowym. W przypadku tych ostatnich po raz kolejny pojawia się problem synchronizacji. Tym razem jest to zsynchronizowanie szybkości transmisji danych przez łącze radiowe z szybkością pracy podsystemu audio czy sieci komputerowej.

Program laboratorium przygotowano tak, żeby zapoznać studentów z jednej strony z metodami modulacji i demodulacji sygnałów, grupowania danych w ramki oraz ich zabezpieczenia przed przekłamaniami. Z drugiej istotne jest pokazanie jak ważne w systemach telekomunikacyjnych są zagadnienia związane z synchronizacją, zarządzaniem pasmem i mocą sygnału, z interferencjami oraz z szumem.

5. PODSUMOWANIE

Pojawienie się na rynku urządzeń realizujących ideę radia definiowanego programowo dało inżynierom narzędzie praktycznej implementacji algorytmów obliczeniowych związanych z telekomunikacją. Kiedy wymagana jest wysoka wydajność, to obliczenia te powinien realizować połączony z SDRem układ FPGA. Na etapie testowania algorytmów, a także w edukacji, co jest tematem niniejszego opracowania, procedury obliczeniowe realizowane mogą być przez oprogramowanie uruchomione na komputerze osobistym.

Łącząc ze sobą tanie urządzenia SDR jakimi są zestawy ADALM-PLUTO z oprogramowaniem GNU Radio Companion zainstalowanym w systemie operacyjnym bazującym na jądrze Linux tworzony jest system nadawczo-odbiorczy, którego główny koszt stanowi sam komputer. Oprogramowanie GRC łączy w sobie prostotę użytkowania (łączenie bloków) ze swobodą budowania algorytmów w związku z możliwością łatwego tworzenia własnych bloków. Niską precyzję i niestabilność układów zegarowych także należy uznać za zaletę, bo dzięki temu studenci mogą zmierzyć się z kolejnymi problemami jakie mogą się pojawić przy implementacji algorytmów cyfrowego przetwarzania sygnałów w rzeczywistych systemach telekomunikacyjnych.

LITERATURA

[1] Analog Devices, Inc. 2014. „AD9361: RF Agile Transceiver – Data Sheet". One Analog Way, Wilmington, U.S.A.

[2] Robin Getz. 2021, „Wiki: ADALM-PLUTO Overview". Online: https://wiki.analog.com/university/tools/pluto

[3] ARRL, The National Association for Amateur Radio. 2017. „Ulrich Rohde, N1UL, Recognized for Pioneering Work on SDR". Online: https://www.arrl.org/news/ulrich-rohde-n1ul-recognized-for-pioneering-work-on-sdr.

[4] Chaudhari Qasim. 2018. „Wireless Communications From the Ground Up – An SDR Perspective". Melbourne, Australia.

[5] Ettus Matt. 2010. „Ettus Research acquired by National Instruments". Ettus Research News. Online: https://lwn.net/Articles/374453/

[6] Ettus Research, A National Instruments Company. 2025. „UHD". Online: https://kb.ettus.com/UHD.

[7] GNU Radio Community. 2024. „Documentaion: Main Page". Online: https://wiki.gnuradio.org/index.php/Main_Page

[8] GNU Radio Community. 2024. „Documentaion: Tutorials". Online: https://wiki.gnuradio.org/index.php?title=Tutorials

[9] Holland Oliver, Bogucka Hanna, Medeisis Arturas. 2015. „Opportunistic Spectrum Sharing and White Space Access : The Practical Reality". John Wiley & Sons, Inc., Hoboken, New Jersey.

[10] Izydorczyk Jacek i inni. 2017. „MATLAB i podstawy telekomunikacji". Helion, Gliwice.

[11] Realtek Semiconductor Corp. 2010. „RTL2832U: DVB-T COFDM Demodulator + USB 2.0 – Datasheet, Rev. 1.4". Hsinchu, Taiwan.

[12] Stewart Robert , Barlee Kenneth, Atkinson Dale, Crockett Louise. 2015. „Software Defined Radio using MATLAB & Simulink and the RTL-SDR". University of Strathclyde, Glasgow, Scotland, UK.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

PROTOKÓŁ KOMUNIKACYJNY W MOBILNYCH SYSTEMACH ŁĄCZNOŚCI PODWODNEJ

THE PROTOCOL FOR UNDERWATER WIRELESS MOBILE COMMUNICATION

Dawid Budnarowski[1]; Agnieszka Czapiewska[2]; Andrzej Łuksza[1]; Ryszard Studański[1]; Łukasz Wojewódka[1]; Andrzej Żak[3]

[1] Uniwersytet Morski w Gdyni, 81-255 Gdynia, d.budnarowski@we.umg.edu.pl, a.luksza@we.umg.edu.pl, r.studanski@we.umg.edu.pl, l.wojewodka@we.umg.edu.pl
[2] Politechnika Gdańska, 80-233 Gdańsk, agnieszka.czapiewska@pg.edu.pl
[3] Akademia Marynarki Wojennej, 81-255 Gdynia, a.zak@amw.gdynia.pl

DOI: 10.15199/59.2025.4.16

Streszczenie: Akustyczny kanał podwodny charakteryzuje się bardzo dużymi opóźnieniami oraz bardzo małą przepływnością. W związku z czym niemożliwe jest proste zaadoptowanie protokołów transmisji znanych z radiokomunikacji. Niniejszy artykuł prezentuje propozycję protokołu bezprzewodowej komunikacji podwodnej przeznaczonego dla systemu składającego się z 8 użytkowników, wykorzystującego modulację MBFSK (Multiple Binary Frequency Shift Keying). W artykule omówiono problemy wynikające z podwodnego medium transmisyjnego.

Abstract: The acoustic underwater channel is characterized by very high latency and very low bit rate. Therefore, it is impossible to simply adopt solutions known from radio communication. This article presents a proposal for a wireless underwater communication protocol designed for a system consisting of 8 users, using MBFSK (Multiple Binary Frequency Shift Keying) modulation. The article discusses problems arising from the underwater transmission medium.

Słowa kluczowe**: bezprzewodowa komunikacja podwodna, MBFSK, protokół komunikacyjny.**

Keywords**: wireless underwater communication, MBFSK, communication protocol.**

1. WSTĘP

Systemy bezprzewodowej łączności podwodnej w większości, jako medium transmisyjne, wykorzystują falę akustyczną. Zaletą tego medium jest możliwość uzyskiwania znaczących zasięgów rzędu kilometrów. Wadą jest bardzo mała prędkość propagacji fali wynosząca ok. 1500 m/s. Dodatkowo wartość tej prędkości ulega znaczącym zmianom w środowisku propagacyjnym w wyniku zmiany zasolenia wody, temperatury czy ciśnienia. Duża zmiana tej prędkości wraz z głębokością powoduje warstwowość ośrodka, która między innymi jest przyczyną wielodrogowości, szczególnie intensywnej w wodach płytkich oraz w akwenach z infrastrukturą hydrotechniczną np. w portach. Prezentowany w niniejszym referacie protokół jest projektowany dla systemu łączności bezprzewodowej, mającego pracować w bardzo trudnych warunkach propagacyjnych: baseny portowe oraz podczas penetracji wraków. Dotychczas realizowane przez autorów pomiary w basenach modelowych na Politechnice Gdańskiej oraz Centrum Techniki Okrętowej (CTO) w Gdańsku pozwoliły wybrać najbardziej odporny typ modulacji do pracy w tych warunkach [2]. W literaturze proponowane są inne rodzaje modulacji dla podwodnych kanałów hydroakustycznych, jednakże przeznaczone są one do pracy w warunkach oceanicznych, gdzie problem wielodrogowości nie jest takim wyzwaniem.

Jedynym praktycznie stosowanym protokołem komunikacji podwodnej jest standard STANAG 1475 [5], powszechnie znany jako JANUS. Stosowana jest tu modulacja binarnego kluczowania częstotliwości z hoppingiem częstotliwościowym (FH-BFSK - Frequency-Hopping Binary Frequency Shift Keying). Dostępnych jest 5 podpasm w zakresie od 4,9 kHz do 32,5 kHz. Uzyskiwany zasięg wynosi do 28 km, praktyczna szybkość transmisji danych użytkowych to od kilku do kilkunastu bitów na sekundę.

W literaturze nieustannie proponowane są nowe protokoły komunikacyjne dla podwodnych sieci bezprzewodowych [3], [6], [7], jednakże projektowane są dla dużych akwenów i relatywnie dużych zasięgów pracy. Ponadto, analiza sposobu działania tych protokołów jest realizowana symulacyjnie w oderwaniu od uwarunkowań warstwy fizycznej. Co najwyżej podane są spodziewane przepływności (150 b/s w [6], 1 kb/s w [3] i 2 kb/s w [7]). Jednakże projektując bezprzewodowy system łączności konieczne jest rozważenie scenariusza zastosowania projektowanego systemu, w jakich warunkach propagacyjnych będzie on pracował i jakie fizyczne ograniczenia te warunki nakładają na projektowany system łączności. Ta analiza dopiero pozwala na wybór odpowiedniej techniki modulacji i jej parametrów. Te decyzje natomiast mają znaczący wpływ na dostępną liczbę bitów w ramce oraz przepływność całego systemu. Długość ramki oraz częstotliwość ich przesyłania ma znaczący wpływ na wybór kodowania kanałowego.

W niniejszym referacie szczegółowo opisano zaproponowany protokół komunikacji podwodnej do pracy w bardzo trudnych warunkach propagacyjnych dla mobilnych węzłów z uwzględnieniem budowy warstwy fizycznej oraz zastosowanego kodowania kanałowego.

2. ZAŁOŻENIA PROJEKTOWE

Projektowany protokół komunikacyjny ma być przeznaczony dla podwodnego bezprzewodowego systemu łączności krótkiego zasięgu, który szacuje się na 150 m. System ma zapewniać łączność 8 użytkownikom. Ponadto, zakłada się, że użytkownicy systemu np. nurkowie lub autonomiczne pojazdy podwodne (AUV - Autonomous Underwater Vehicle) mogą się przemieszczać z wzajemną prędkością nawet 5 m/s. Przyjęto również, że transmisja może być realizowana jednocześnie tylko w jednym kierunku, a łączność w danej chwili jest realizowana pomiędzy dwoma węzłami. Podczas wyznaczania parametrów czasowych protokołu przyjęto, że prędkość propagacji fali wyniesie 1477 m/s. Taką prędkość propagacji fali zmierzono w basenie w CTO.

W warstwie fizycznej zdecydowano się na użycie modulacji MBFSK (Multiple Binary Frequency Shift Keying) opisanej w [1] i [4]. Dotychczasowe badania wykazały, że modulacja ta bardzo dobrze sprawdza się w wyjątkowo trudnych warunkach propagacyjnych wód płytkich oraz podczas wzajemnego ruchu urządzeń nadawczo-odbiorczych. W modulacji tej jednocześnie na wielu ortogonalnych podnośnych przesyłane są bity. Zbiór takich podnośnych wysyłanych przez czas T_{sym} nazywany będzie dalej symbolem MBFSK. W modulacji tej jakość odbioru silnie zależy od odstępu pomiędzy podnośnymi. Im większy jest obserwowany efekt Dopplera podczas transmisji tym większy powinien być odstęp pomiędzy podnośnymi. Niestety, im większy jest ten odstęp tym mniejsza jest liczba dostępnych podnośnych do transmisji danych przy założeniu stałej szerokości wykorzystywanego pasma. Założono, że transmisja jest realizowana na częstotliwości środkowej 100 kHz i szerokości pasma 100 kHz. W projektowanym protokole przyjęto, że parametry modulacji będą adoptowane do panujących warunków w kanale propagacyjnym. Przed każdym symbolem MBFSK przesyłany jest sygnał chirp służący do synchronizacji czasowej.

Ponadto w kanale hydroakustycznym obserwuje się bardzo długie czasy trwania odpowiedzi impulsowej, rzędu setek ms. W celu zabezpieczenia transmisji przed interferencjami pomiędzy symbolami, zastosowano czas ochronny T_g, którego wartość również powinna być adaptowana do parametrów kanału propagacyjnego, gdyż również od tego parametru zależy uzyskiwana przepływność transmisji.

Kolejne założenie projektowe wynika z implementacji projektowanego systemu bezprzewodowej podwodnej łączności akustycznej. Mianowicie, zarówno w części nadawczej jak i odbiorczej wykorzystywane są moduły Arduino Giga R1 z wbudowanymi mikrokontrolerami STM32H747, odpowiedzialne za wstępne przetwarzanie sygnałów. Aby urządzenia te mogły realizować swoje zadania w czasie zbliżonym do rzeczywistego konieczne było ustawienie częstotliwości próbkowania sygnałów na 400 kHz oraz przyjęcie, że liczba składowych formujących widmo sygnału w odbiorniku wyniesie dokładnie 4096. Te dwa parametry determinują możliwe odstępy pomiędzy podnośnymi w sygnale MBFSK. W tab. 1 zamieszczono wybrane odstępy pomiędzy podnośnymi oraz liczbę bitów dostępną w danym symbolu.

Największy przyjęty odstęp pomiędzy podnośnymi wynika z maksymalnego przesunięcia Dopplera, wyznaczonego dla wzajemnej prędkości poruszania się obiektów 5 m/s, oraz spodziewanego rozproszenia Dopplera, które jest trudne do oceny, ponieważ zależy również od wielodrogowości obserwowanej w kanale propagacyjnym.

Tab. 1. Wykaz wybranych odstępów między podnośnymi

Odstęp [liczba składowych]	*Odstęp [Hz]*	*Liczba bitów na symbol*
4	97,7	480
8	195,3	240
12	293,0	160
16	390,6	120
20	488,3	96
24	585,9	80
32	781,3	60

Przyjęto, że czas trwania symbolu wynosi 40,96 ms. Stanowi to kompromis pomiędzy energią odbieranego sygnału, a uzyskiwaną przepływnością. Uwzględniając dodatkowo konieczność zastosowania czasu ochronnego, częstotliwość powtarzania symboli jest niewielka. Uznano więc, że każda ramka w projektowanym protokole powinna mieścić się w pojedynczym symbolu MBFSK. Stąd też należało wybrać kodowanie kanałowe, które nie będzie wymagało dużej nadmiarowości oraz długiego ciągu bitów, zważywszy na najmniejszą dostępną liczbę bitów w symbolu w Tab. 1. Kodowanie kanałowe, które wymagałoby odbioru większej liczby bitów prowadziłoby do dodatkowych opóźnień transmisji. Do tego celu wybrano więc kodowanie BCH (Bose–Chaudhuri–Hocquenghem). Kody te są systematyczne, dzięki czemu struktura ramki nie musi być ściśle związana z rozmiarem kodu. Należy wówczas przyjąć podczas kodowania i dekodowania, że dane, które nie zmieściły się w ramce wynoszą 1. Dodatkowo, ze względu na większe tłumienie w wodzie fal krótszych, nadmiarowość kodu BCH powinna być odwzorowywana na podnośnych o najmniejszych częstotliwościach.

Projektując protokół założono także, możliwość trzech trybów pracy systemu:

- bez potwierdzenia (NON ACK) – tryb przeznaczony do zdalnego sterowania pojazdami, gdzie dzięki zaniechaniu przesyłania potwierdzeń odbioru możliwe jest zmniejszenie opóźnienia podczas transmisji,
- zawsze potwierdzenie (ALL ACK) – po każdej odebranej ramce wymagane jest przesłanie potwierdzenia. Tryb przeznaczony do transmisji danych krytycznych lub w bardzo dynamicznie zmieniającym się kanale propagacyjnym. Charakteryzuje się najmniejszymi przepływnościami,

- potwierdzenie po bloku (BL ACK) – po kilku przesłanych ramkach (nazywanych dalej blokiem) następuje transmisja potwierdzenia odbioru. Liczba ramek w bloku jest ustalana w taki sposób by potwierdzenie nie było przesyłane rzadziej niż co 1 s.

Przyjęto również, że dane użytkownika muszą być uporządkowane bajtowo.

3. PROTOKÓŁ KOMUNIKACYJNY

Projektując protokół przy uwzględnieniu poczynionych założeń uznano, że wystarczą tylko cztery rodzaje ramek komunikacyjnych:

- Wywołanie (INIT) – ramka służąca do nawiązania połączenia,
- Potwierdzenie wywołania (ACK INIT) - potwierdzenie gotowości odbioru danych oraz wskazanie parametrów transmisji odpowiednich dla danego stanu kanału oraz maksymalizujących przepływność,
- Transmisja danych (DATA) – ramka zawierająca dane użytkownika,
- Potwierdzenie odbioru danych (ACK DATA) – potwierdzenie odbioru ramek.

Podczas wywołania przysyłany jest jedynie adres odbiorcy, adres nadawcy oraz tryb pracy (NON ACK, ALL ACK czy BL ACK). Zakłada się, że najbardziej optymalnym z punktu widzenia uzyskiwanych przepływności oraz adaptacyjności pracy jest tryb BL ACK. Ramka INIT kodowana jest kodem BCH(63,18), którego zdolność korekcyjna wynosi 10 bitów. Ponadto ramka ta jest transmitowana przy największym możliwym odstępie pomiędzy podnośnymi oraz największym czasie ochronnym, tak by odporność transmisji INIT na warunki propagacyjne była możliwie największa.

Odbiornik po otrzymaniu wiadomości INIT, może ocenić stan kanału propagacyjnego [1] i dobrać odpowiedni odstęp pomiędzy podnośnymi oraz czas ochronny T_g. Dlatego też ramka z potwierdzeniem odbioru wywołania, oprócz adresów nadawcy i odbiorcy, zawiera również informację o odstępie pomiędzy podnośnymi oraz o czasie T_g. Następna transmisja musi być zrealizowana z ustalonymi przez odbiornik parametrami. W tej ramce, ze względu na konieczność przesłania większej liczby danych niż w INIT, stosuje się kodowanie BCH(63,24). Nie jest używane dodatkowe kodowanie CRC (Cyclic Redundancy Check) w ramkach INIT i ACK INIT, ponieważ wybrane kody BCH umożliwiają również ocenę poprawności zrealizowanego dekodowania kanałowego.

Po pomyślnym odebraniu ramki ACK INIT strona nadawcza ma informację z jakimi parametrami będzie realizowana transmisja. Na podstawie podanego czasu ochronnego wyliczana jest liczba ramek w bloku po stronie nadawczej i odbiorczej. Transmisja danych użytkownika może zostać rozpoczęta w ramkach DATA. W ramce tej znajdują się standardowe pole jak adres nadawcy oraz odbiorcy. Ponadto, każda ramka ma swój przypisany numer w ramach bloku. Ponieważ na numer ramki w bloku można było przeznaczyć tylko 4 bity, dlatego też ramek w bloku nie może być więcej niż 15. Oznacza to, że gdy T_g zostanie zredukowane do 10 ms, czas trwania bloku będzie wynosił jedynie 750 ms. Numery ramek są niezbędne do ewentualnego wskazania ramek wymagających retransmisji. Ponadto w ramce DATA jeden bit przeznaczono na tzw. flagę końca transmisji (END). Gdy wartość tej flagi wynosi 0, wówczas transmisja trwa. Gdy flaga ta wynosi 1, oznacza to, że przesyłana jest właśnie ostatnia ramka danych i niezależnie od wyznaczonej liczby ramek w bloku, po tej ramce musi być odesłane potwierdzenie odbioru danych (ACK DATA). W ramce DATA jest również przewidziane 6-cio bitowe pole LB, które informuje o liczbie bajtów przesyłanych w ramce. Jeśli pole END wynosi 0, wówczas pole LB również musi wynosić 0. Te dwa pola mogą w takiej sytuacji również posłużyć do oceny poprawności dekodowania kanałowego. Jeśli natomiast pole END wynosi 1, to dzięki wartości z pola LB wiadomo, ile bajtów danych w ramce zawiera dane użytkownika. Wobec tego informacja o zakończeniu transmisji jest przesyłana łącznie na 7 bitach. Pozostałe bity służą jedynie do wypełnienie symbolu MBFSK. Po wymienionych polach ramki DATA znajduje się pole przeznaczone na dane użytkownika. Po nim przewidziano 5 bitów na sumę kontroli parzystości. Ze względu na różną liczbę bitów dostępnych w ramce DATA oraz ze względu na możliwe różne odstępy pomiędzy podnośnymi, zmianie ulega również używany kod BCH. W tab. 2 wypisano wszystkie stosowane kody BCH. W dwóch pierwszych wierszach zamieszczono kody używane w ramkach INIT i ACK INIT. Ramki z danymi są mniej zabezpieczone kodowo BCH niż ramki sygnalizacyjne. Wówczas kodowanie to jest już niewystarczające do ewentualnej detekcji błędnego dekodowania. Stąd konieczność użycia kodowania CRC.

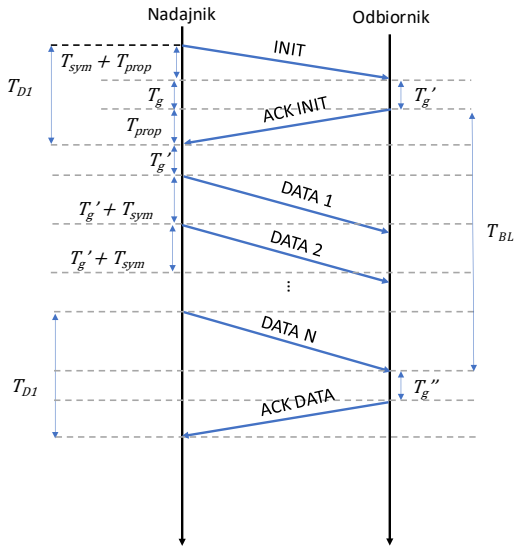
Tab. 2. Wykaz używanych kodów BCH

Długość słowa kodowego	*Liczba bitów danych*	*Liczba korygowanych błędów*	*Liczba bitów w symbolu MBFSK*
63	18	10	60
63	24	7	60
63	51	2	60
127	113	2	80
127	113	2	96
127	106	3	120
255	223	4	160
255	215	5	240
511	421	10	480

Po przesłaniu ustalonej liczby ramek z bloku, strona odbiorcza musi wysłać potwierdzenie odbioru danych ACK DATA. W tym momencie transmisji mogą ulec zmianie jej parametry tj. odstęp pomiędzy podnośnymi oraz czas T_g. Jeśli stracono zbyt dużo ramek lub w wyniku analizy sygnału odbieranego strona odbiorcza stwierdzi taką konieczność, w momencie wysyłania ACK DATA może narzucić zmiany wspomnianych parametrów. Oprócz pól adresów w ramce ACK DATA jest pole, które

określa odstęp pomiędzy podnośnymi podczas transmisji kolejnego bloku oraz pole, które definiuje czas T_g kolejnych transmisji. Następnie wyznaczono miejsce na piętnastobitowe pole, w którym wartościami jeden oznaczono ramki wymagające retransmisji. Oczywiście każda ramka oprócz elementów wymienionych powyżej zawiera również identyfikator ramki.

Na rys. 1 zaprezentowano schemat poprawnej komunikacji, gdy potwierdzenie odbioru przesyłane jest po bloku z uwzględnieniem odstępów czasowych pomiędzy kolejnymi aktywnościami urządzeń. T_{D1} jest to czas oczekiwania po wysłaniu INIT na odpowiedź. Po upływie tego czasu strona inicjująca może ponowić próbę transmisji INIT. Czas ten wynika z T_{sym} czyli czasu trwania symbolu MBFSK i jednocześnie pojedynczej ramki. Niezależnie od liczby bajtów dostępnych w ramce, każda trwa jednakowo długo. Kolejnym składnikiem T_{D1} jest omawiany już wcześniej czas ochronny T_g, który może przyjmować wartości od 10 ms do 320 ms. Czas oczekiwania na odpowiedź ACK INIT wynika również z czasu propagacji, który dla przyjętego zasięgu i prędkości propagacji fali może wynieść nawet 101 ms. Po odebraniu ACK INIT należy odczekać nowy czas ochronny T_g' przed rozpoczęciem wysyłania N ramek danych w bloku. Ponieważ może nastąpić sytuacja, w której odbiornik przestanie wykrywać transmisje lub odbierać je z błędami, co spowoduje, że nie będzie w stanie ocenić czy wszystkie ramki z bloku już nadeszły. Dlatego też zaproponowano odbiór w czasie T_{BL}. Po tym czasie wysyłane jest potwierdzenie ACK DATA z informacją, które ramki należy retransmitować niezależnie czy odebrano przewidzianą liczbę ramek z bloku. Przy wyznaczaniu T_{BL} po stronie odbiorczej przyjmuje się oczywiście największe możliwe opóźnienie propagacyjne. W związku z tym przy niezakłóconym odbiorze, rzeczywisty czas transmisji dla większości przypadków będzie krótszy niż T_{BL}.



Rys. 1. Schemat poprawnej komunikacji w trybie BL ACK

Przeprowadzono badania zaprojektowanego systemu transmisji z zaimplementowanym przedstawionym protokołem, jednakże bez włączonych mechanizmów adaptacji do kanału propagacyjnego. Badania te przeprowadzono w basenie Centrum Techniki Okrętowej w Gdańsku, którego długość wynosi 250 m, szerokość 12 m, a głębokość 6 m. Wstępne wyniki potwierdzają skuteczność zaproponowanego protokołu. Ponadto, w warunkach statycznych uzyskano 2 kb/s przepływności, natomiast podczas ruchu z prędkością 5 m/s 1,5 kb/s (liczba bitów w symbolu 480 i T_g, 60 ms). Różnica wynika z większej liczby wymaganych retransmisji ramek podczas ruchu.

4. PODSUMOWANIE

W referacie omówiono zaproponowany protokół komunikacji podwodnej dla wód płytkich i basenów portowych, w których panują bardzo trudne warunki propagacyjne. Szczegółowo omówiono strukturę ramkową, tryby pracy i ograniczenia wynikające z zastosowanej techniki modulacji. Przedstawiono także zależności czasowe uwzględnione w protokole. Zaprezentowano również schemat przykładowej poprawnej komunikacji w trybie transmisji blokowej. Wstępne wyniki przeprowadzonych badań zrealizowane w warunkach rzeczywistych potwierdzają słuszność przyjętych założeń i funkcjonalność zaprojektowanego protokołu.

LITERATURA

[1] Czapiewska A., Łuksza A., Studański R., Wojewódka Ł., Żak A. 2024. "Comparison of Doppler Effect Estimation Methods for MFSK Transmission in Multipath Hydroacoustic Channel". IEEE Access. vol. 12. doi: 10.1109/ACCESS.2024.3385441.

[2] Czapiewska A., Łuksza A., Studański R., Wojewódka Ł., Żak A. 2025. "Evaluation of Selected Modulation Techniques in Underwater Multipath Channel". IEEE Access. vol. 13. doi: 10.1109/ACCESS.2025.3530948.

[3] Huang J., Chi C., Wang W., Huang H. 2020. "A Sequence-Scheduled and Query-Based MAC Protocol for Underwater Acoustic Networks with a Mobile Node". Journal of Communications and Information Networks. vol. 5. no. 2. doi: 10.23919/JCIN.2020.9130431.

[4] Mizeraczyk J., Studanski R., Zak A., Czapiewska A. 2021. ‘‘A method for underwater wireless data transmission in a hydroacoustic channel under NLOS conditions’’. Sensors. vol. 21.

[5] NATO STANDARD ANEP-87, DIGITAL UNDERWATER SIGNALLING STANDARD FOR NETWORK NODE DISCOVERY & INTEROPERABILITY Edition A Version 2, 2024.

[6] Li Huizhe, Zhou Feng, Qiao Gang. 2024. "Hybrid Protocol Design of Underwater Monitoring Network Based on Cluster Topology". OES China Ocean Acoustics (COA). doi: 10.1109/COA58979.2024.10723473.

[7] Sun W., Gao X.. 2023. "Pre-allocated MAC Protocol Based on Underwater Acoustic Communication Network". 2023 8th International Conference on Signal and Image Processing (ICSIP), Wuxi, China, pp. 935-939, doi: 10.1109/ICSIP57908.2023.10271001.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

Tunelowanie ICMP jako medium szyfrowanej i niezawodnej komunikacji

ICMP tunneling as a medium for encrypted and reliable communication

Zuzanna Konopek[1], Jerzy Domżał[2], Robert Wójcik[3],

[1] Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, Kraków, konopek@student.agh.edu.pl
[2] Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, Kraków, jerzy.domzal@agh.edu.pl
[3] Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, Kraków, robert.wojcik@agh.edu.pl

DOI: 10.15199/59.2025.4.17

Streszczenie: Niniejszy artykuł omawia możliwości wykorzystania tunelowania ICMP jako medium do szyfrowanej i niezawodnej komunikacji. Przedstawiono mechanizm działania opracowanej biblioteki do tunelowania ICMP, a także omówiono wprowadzone środki zapewniające niezawodność, szyfrowanie oraz ukrycie komunikacji. Dodatkowo przeprowadzono porównanie tej biblioteki z innymi rozwiązaniami oraz dokonano analizy wyników badań nad jej funkcjonalnością. Przeprowadzone testy wykazały, że opracowana biblioteka umożliwia stabline tunelowanie ICMP z zachowaniem szyfrowania oraz mechanizmów zapewniających niezawodność transmisji. Analiza wyników potwierdziła, że zastosowanie zoptymalizowanego rozmiaru pakietów utrudnia wykrycie komunikacji przez systemy monitorujące ruch sieciowy.

Abstract: This article explores the potential of ICMP tunneling as a viable medium for encrypted and reliable communication. It outlines the mechanisms employed in the developed ICMP tunneling library, as well as the strategies implemented to ensure reliability, encryption, and covert communication. Furthermore, a comparative analysis of this library against other solutions is conducted, accompanied by an evaluation of the findings from a functional study. Tests proved that the developed library allows stabile ICMP tunneling while maintaining encryption and mechanisms to ensure transmission reliability. Analysis of the results confirmed that the use of an optimised packet size makes it more difficult for network traffic monitoring systems to detect the communication.

Słowa kluczowe: biblioteka, ICMP, niezawodność, tunelowanie, ukrywanie ruchu

Keywords: ICMP, library, reliability, traffic hiding, tunneling

1. WSTĘP

Współczesne sieci komputerowe stoją w obliczu dynamicznie zmieniających się wyzwań w zakresie bezpieczeństwa wysyłanych danych. Jednym z mniej oczywistych, a jednocześnie skutecznych sposobów komunikacji w środowiskach o ograniczonym dostępie jest wykorzystanie tunelowania ICMP (Internet Control Message Protocol). Chociaż protokół ICMP został pierwotnie zaprojektowany jako narzędzie diagnostyczne w sieciach komputerowych [1], jego łatwa dostępność, szerokie wdrożenie w infrastrukturach sieciowych i mała uwaga poświęcana analizie tego ruchu, czynią go atrakcyjnym medium komunikacyjnym, również w celach niezgodnych z jego pierwotnym przeznaczeniem.

Tunelowanie ICMP od dawna znajduje zastosowanie w działaniach złośliwego oprogramowania, w szczególności botnetów. Botnety, operujące jako zdecentralizowane struktury zarządzające grupami urządzeń zainfekowanych złośliwym kodem, coraz częściej wykorzystują techniki tunelowania, aby zapewnić sobie ukrytą, niezawodną komunikację z serwerami C&C (Command and Control). Jednak w wielu przypadkach taka komunikacja pozostaje niezaszyfrowana lub nieodpowiednio ukryta, co może ograniczać jej skuteczność w zaawansowanych scenariuszach operacyjnych. ICMP, dzięki swojej szerokiej dostępności i często pomijaniu go przez zapory ogniowe oraz systemy wykrywania włamań, staje się idealnym medium do przesyłania danych w sposób trudny do wykrycia. Dodatkowo ICMP jest często ignorowany przez mechanizmy takie jak np. Captive Portal, co sprawia, że odpowiednio zaprojektowana aplikacja może umożliwić dostęp do Internetu, enkapsulując zwykły ruch HTTP/HTTPS w komunikatach ICMP. Tego typu zastosowanie jeszcze bardziej podkreśla potencjał tego protokołu w zapewnianiu niekonwencjonalnych ścieżek komunikacji.

W niniejszym artykule przedstawiamy sposób implementacji biblioteki do tunelowania ICMP, która umożliwia szyfrowanie, zapewnia niezawodną komunikację oraz ukrywanie ruchu. Proponowane rozwiązanie zostało zaprojektowane z myślą o zastosowaniu w komunikacji botnetu w celach badawczych. Niemniej jednak biblioteka może być z powodzeniem wykorzystywana również w innych rozwiązaniach. Rozdział 2 omawia opis problemu, w rozdziale 3 przedstawiono mechanizm działania biblioteki oraz omówiono wprowadzone środki zapewniające niezawodność, szyfrowanie oraz ukrycie komunikacji, w rozdziale 4 wskazano użyte narzędzia, w rozdziale 5 przeprowadzono porównanie tej biblioteki z innymi rozwiązaniami, a w rozdziale 6 dokonano analizy wyników badań nad jej funkcjonalnością.

2. OPIS PROBLEMU

Protokół ICMP jest jednym z kluczowych elementów stosu protokołów sieciowych w modelu IP. Jego podstawowym zadaniem jest dostarczanie informacji diagnostycz-

nych i komunikacyjnych między urządzeniami w sieci, takich jak zgłaszanie błędów transmisji czy informowanie o niedostępności zasobów. Pakiety ICMP, mimo swojej użyteczności, charakteryzują się specyficzną konstrukcją i różnym zachowaniem w zależności od systemu operacyjnego, co ogranicza ich potencjalne zastosowanie jako środka komunikacji.

2.1 Struktura pakietu ICMP

Pakiet ICMP składa się z nagłówka (header) oraz opcjonalnej części danych, której zawartość i długość zależy od rodzaju wiadomości ICMP. Nagłówek ICMP ma stałą długość wynoszącą 8 bajtów i zawiera takie informacje jak typ wiadomości (1 bajt), kod (bardziej szczegółowy opis typu - 1 bajt), sumę kontrolną (2 bajty) oraz pola specyficzne dla danego typu pakietu. Bezpośrednio za nagłówkiem znajdują się dane, które mogą być dowolne i służą np. do przesyłania informacji pomocniczych w przypadku wiadomości *Echo Request* i *Echo Reply*.

Maksymalna wielkość całego pakietu ICMP, w tym danych, jest ograniczona przez maksymalną jednostkę transmisji (MTU) w danej sieci, najczęściej wynoszącą 1500 bajtów dla Ethernetu. Od tej wartości należy odjąć 20 bajtów dla nagłówka IP i 8 bajtów dla nagłówka ICMP, co daje maksymalną długość części danych wynoszącą 1472 bajty.

2.2 Brak niezawodności w ICMP

ICMP, jako protokół warstwy sieciowej, nie zapewnia niezawodności. Działa on w oparciu o protokół IP, który jest bezpołączeniowy i nie gwarantuje dostarczenia pakietów ani ich odpowiedniej kolejności. Dane przesyłane w części użytkowej pakietu ICMP są w pełni widoczne i nie są domyślnie szyfrowane, co oznacza, że mogą być łatwo przechwycone i przeanalizowane przez atakujących. Te cechy czynią ICMP podatnym na wykorzystanie w nieautoryzowanych działaniach, ale jednocześnie wskazują na jego ograniczenia jako medium komunikacyjnego wymagającego niezawodności.

2.3 Różnice w rozmiarach pakietu ICMP w systemach Windows i Linux

Wielkość pakietów ICMP, które można wysyłać lub odbierać, różni się między systemami operacyjnymi Windows i Linux, co wynika z różnych implementacji narzędzi, takich jak ping. Przykładowo, w przypadku użycia komendy ping:

- w systemie Windows część danych zawartych w pakiecie ICMP wynosi 32 bajty,
- w systemach operacyjnych opartych na jądrze Linux część danych w pakiecie ICMP wynosi 40 bajtów.

Wspomniane różnice w strukturze i właściwościach ICMP oraz brak niezawodności protokołu stawiają wyzwanie w zakresie jego adaptacji do niekonwencjonalnych zastosowań. Tunelowanie danych za pomocą ICMP wymaga uwzględnienia tych ograniczeń, a także potencjalnych rozbieżności w implementacjach systemowych, co czyni rozwój bardziej zaawansowanych mechanizmów komunikacyjnych na bazie ICMP szczególnie interesującym.

3. MECHANIZM DZIAŁANIA BIBLIOTEKI

Biblioteka opracowana w ramach niniejszej pracy umożliwia tunelowanie danych z wykorzystaniem protokołu ICMP. Mechanizm działania został zaprojektowany z uwzględnieniem kluczowych wyzwań charakterystycznych dla protokołu ICMP, takich jak brak wbudowanej niezawodności, jawność przesyłanych danych oraz konieczność ich ukrycia i zabezpieczenia przed nieautoryzowanym dostępem. W odróżnieniu od podejścia zaprezentowanego w pracy [2], w którym rozróżniano sposób transmisji plików i wiadomości tekstowych, niniejsza implementacja wykorzystuje jednolity mechanizm przesyłania danych niezależnie od ich formatu. Takie podejście umożliwia uzyskanie uniwersalnej biblioteki, zdolnej do obsługi różnych typów komunikacji w sposób spójny. W dalszej części przedstawiono szczegóły implementacyjne oraz zastosowane mechanizmy zwiększające niezawodność transmisji, zapewniające szyfrowanie i integralność danych, a także utrudniające wykrycie komunikacji — zgodnie z założeniami przedstawionymi w pracy [3].

3.1 Użycie biblioteki

Celem użycia biblioteki należy zainicjować klasę Icmp, podając adres IP hosta, z którym będziemy rozpoczynać komunikację. Należy zapewnić również hasło znane obydwu odbiorcom, przy użyciu którego będą szyfrowane wszystkie następne wiadomości. Opcjonalnie można również określić, w jakim trybie będzie następowała wymiana komunikacji. Biblioteka udostępnia 3 możliwe tryby szybkości, które wpływają na skrytość komunikacji:

- **default** - tryb domyślny, w którym wiadomości są wysyłane w odstępach czasowych zgodnych z wartością domyślną wysyłaną przez narzędzie ping; czyli 2 wiadomości na sekundę. Czas wysłania wiadomości jest losowany pomiędzy 0,4 a 0,5 sekundy. Na odpowiedź ACK biblioteka czeka 0,05 sekundy. Maksymalny rozmiar danych, jaki może być wysłany to 30 bajtów przed zaszyfrowaniem, po zaszyfrowaniu wiadomość będzie miała większy rozmiar.

- **fast** - tryb szybki, w którym wiadomości są wysyłane w krótkich odstępach. Czas wysłania wiadomości jest losowany pomiędzy 0,01 a 0,05 sekundy. Na odpowiedź biblioteka czeka 0,01 sekundy. Maksymalny rozmiar danych, jaki może być wysłany to 15 bajtów. Jest to maksymalny rozmiar danych, dla którego szyfrogram ma najmniejszą możliwą wartość. Dla wiadomości od 1 do 15 bajtów szyfrogram przyjmuje stałą wartość.

- **stealth** - tryb ukryty, w którym wiadomości są wysyłane w dłuższych odstępach. Czas wysłania wiadomości jest losowany pomiędzy 0,4 a 1 sekundą. Na odpowiedź biblioteka czeka 0,5 sekundy natomiast maksymalny rozmiar danych, jaki może być wysłany to 15 bajtów.

W tabeli 1 przedstawiono parametry w zależności od wybranego trybu. Kolumna *Rozmiar wiadomości* określa wielkość wiadomości przed zaszyfrowaniem.

Tabela 1: Parametry czasowe w zależności od wybranego trybu.

Tryb	Przedział [s]	Rozmiar wiadomości [B]
default	0,4 – 0,5	30
fast	0,01 – 0,05	15
stealth	0,4 – 1,0	15

3.2 Wysyłanie wiadomości

Po pomyślnym zainicjowaniu klasy można rozpocząć komunikację z hostem. Wysyłanie wiadomości odbywa się za pomocą metody `supervise_sending`, która jako argument przyjmuje wiadomość do wysłania. Następnie metoda dodaje na końcu wiadomości dwa znaki końca linii, które informują odbiorcę, że to koniec datagramu. Zmodyfikowana wiadomość jest dzielona na fragmenty o długości określonej przez wybrany tryb. Każdy fragment jest szyfrowany za pomocą algorytmu AES wariant Fernet, z kluczem podanym przy inicjalizacji klasy. Następnie tworzony jest pakiet ICMP z ustawioną wartością *Echo request* i jest wysyłany do odbiorcy. Metoda czeka na odpowiedź *Echo reply*, z takimi samymi danymi, z jakimi została nadana. Jeśli metoda nie otrzymała odpowiedzi w przewidzianym czasie, ponawia próbę wysłania określonego fragmentu wiadomości 10 razy, jeśli po 10 próbach nie otrzymała odpowiedzi, metoda zwraca błąd. W przeciwnym razie metoda kontynuuje wysyłanie fragmentów oryginalnej wiadomości.

3.3 Odbieranie wiadomości

Odbieranie wiadomości odbywa się za pomocą metody `supervise_receiving`, która nie przyjmuje argumentów. Metoda nasłuchuje na surowym gnieździe na pakiet ICMP z typem *ICMP Echo Request*. Po odebraniu takiego pakietu metoda pobiera dane zapisane w pakiecie (od 28. bajtu do końca), deszyfruje je i zapisuje do zmiennej `data` pod warunkiem, że posiada odpowiedni `sequence number`. Wiadomości są odbierane do momentu, w którym zmienna `data` zawiera dwa znaki końca linii, po czym zwraca otrzymaną wiadomość.

3.4 Niezawodność

Jednym z największych wyzwań podczas wykorzystania ICMP jako medium transmisji danych jest brak wbudowanego mechanizmu zapewniającego niezawodność. Aby temu sprostać, w bibliotece zastosowano metodę do sprawdzania potwierdzeń odbioru (ACK), która umożliwia kontrolowanie, czy wszystkie fragmenty danych zostały poprawnie dostarczone. Z każdym wysłanym pakietem *ICMP Echo Request* host, do którego było kierowane zapytanie, powinien odpowiedzieć *ICMP Echo Reply*. Host wysyłający nasłuchuje tej odpowiedzi w określonym czasie. Jeśli ją otrzyma, to oznacza, że pakiet został poprawnie dostarczony. Aby upewnić się, czy jest tak faktycznie, część danych powinna zawierać te same dane, które były wysłane. W przypadku braku odpowiedzi, pakiet jest ponownie wysyłany do 10 razy, jeśli po 10 próbach nie otrzymał odpowiedzi, metoda zwraca błąd. W przeciwnym wypadku metoda kontynuuje wysyłanie kolejnych fragmentów wiadomości wraz z ich potwierdzaniem. Potwierdzone wiadomości są trzymane w liście potwierdzeń, która jest zmienną klasy o nazwie `ack`.

3.5 Szyfrowanie

Bezpieczeństwo przesyłanych danych jest kluczowym aspektem w projektowaniu systemów opartych na tunelowaniu ICMP. Aby chronić dane przed dostępem osób nieuprawnionych, w bibliotece zastosowano algorytm szyfrowania AES-128 (Advanced Encryption Standard) w trybie CBC [4] nazwanej Fernet. Fernet wykorzystuje również SHA-256 do uwierzytelnienia wiadomości [5]. Natomiast IV jest generowany przy pomocy funkcji `os.urandom()`. Klucz szyfrowania jest generowany podczas inicjalizacji klasy i musi być znany obydwu stronom komunikacji. Wysyłane dane są szyfrowane przed wysłaniem, a odbierane dane są deszyfrowane po odebraniu. W ten sposób zapewniona jest poufność przesyłanych informacji. W implementacji tej biblioteki celowo wybrano słabszy algorytm szyfrowania, aby zwiększyć wydajność i zmniejszyć opóźnienia w komunikacji. Dodatkowo biblioteka ma służyć celom edukacyjnym, taki szyfr łatwiej będzie złamać, jeśli ktoś chciałby wykorzystać bibliotekę w nieetyczny sposób.

3.6 Ukrycie

Maskowanie komunikacji jest kluczowym aspektem tunelowania ICMP, zwłaszcza gdy celem jest uniknięcie wykrycia przez systemy monitorujące ruch sieciowy. W bibliotece zastosowano kilka środków mających na celu utrudnienie analizy ruchu, w tym: zmniejszenie części danych do jak najmniejszych wartości. Użyto również celowo najbardziej popularnego typu pakietów ICMP, aby zminimalizować podejrzenia. Dodatkowo zastosowano losowe opóźnienia w wysyłaniu pakietów, aby utrudnić analizę ruchu sieciowego. Wszystkie te środki mają na celu zwiększenie ukrycia tunelowanego ruchu ICMP.

4. UŻYTE NARZĘDZIA

Do implementacji biblioteki wykorzystano język Python w wersji 3.12.6. Wybór języka Python wynikał z jego popularności, wsparcia dla wielu bibliotek oraz przejrzystości tworzonego kodu. W ramach projektu wykorzystano następujące biblioteki:

- `os` – wykorzystywana do generowania odpowiedniego identyfikatora pakietu,
- `time` – używana jest do oczekiwania na odpowiedź przez określony czas,
- `struct` – używana jest do tworzenia nagłówka ICMP,
- `socket` – wykorzystywana jest w metodzie do tworzenia surowych gniazd oraz implementacji wysyłania i odbierania pakietów ICMP,
- `random` – używana jest do generowania losowych wartości czasu pomiędzy wysłaniami pakietów,
- `datetime` – używana jest do obliczania czasu trwania wysyłania pakietów, do celów testowych,
- `threading` – używana jest do kończenia funkcji oczekiwania na odpowiedź,
- `cryptography` – używana jest do importu funkcji szyfrującej.

Do tworzenia pakietów ICMP napisane zostały specjalne metody: do obliczania sumy kontrolnej, odbierania i nadawania pakietów ICMP i wyciągania z nich danych. Zostały również napisane metody szyfrujące i deszyfrujące odpowiednie wiadomości.

4.1 Narzędzia testowe

Testy biblioteki odbywały się na systemie z jądrem Linux. W celu przetestowania biblioteki wykorzystano narzędzia, takie jak `ping` oraz `wireshark`. Do prezentacji wyników wykorzystano `LibreOffice Calc`.

4.2 Środowisko testowe

Testy działania biblioteki przeprowadzono na interfejsie loopback. Podjęto również próbę wykonania analogicznych testów w środowisku, w którym laptop (z systemem Linux) oraz komputer stacjonarny (z systemem Windows), uruchamiający maszynę wirtualną z systemem Kali Linux, miały się ze sobą komunikować. Wszystkie testy zakończyły się powodzeniem i dały wyniki zgodne z tymi uzyskanymi dla interfejsu loopback.

5. PORÓWNANIE ROZWIĄZANIA

Porównując opracowane rozwiązanie z metodami opisanymi w pracy [6], można zaobserwować istotne różnice, które wskazują na poprawę w zakresie ukrywania i stabilności komunikacji. W pracy [6] autorzy skupili się na analizie zmiennych długości części użytkowej pakietów ICMP (payload), gdzie wartości te dochodziły nawet do 1000 bajtów. Tak duże rozbieżności mogą wskazywać na nietypowy ruch, co znacząco ułatwia jego wykrycie przez systemy monitorujące. W proponowanym rozwiązaniu długości payloadu zostały zoptymalizowane i utrzymane w stabilnym zakresie, co znacznie utrudnia identyfikację komunikacji na podstawie analizy tego parametru.

Kolejnym porównywanym aspektem była częstotliwość wysyłanych pakietów. W pracy [6] zauważono, że częstotliwość pakietów była bardzo rozrzucona, co mogło wzbudzać podejrzenia systemów IDS (Intrusion Detection Systems). W naszym rozwiązaniu częstotliwość jest stabilna i oparta na losowości w granicach naturalnych zachowań sieciowych (w trybie stealth), co umożliwia skuteczniejsze wtapianie się w typowy ruch ICMP. Dzięki temu ruch generowany przez opracowaną bibliotekę trudniej wyróżnić na tle innych danych diagnostycznych.

Jednym z kluczowych elementów analizy była również proporcja liczby wysyłanych request do reply. W pracy [6] przedstawiono wykresy, gdzie stosunek ten bywał bardzo niestabilny i w ekstremalnych przypadkach osiągał nawet 7 prób wysyłania reply na jedno request. Takie zachowanie jest łatwe do zidentyfikowania jako anomalne. W naszym rozwiązaniu stosunek ten niemal zawsze wynosi 1:1, ponieważ biblioteka czeka odpowiednią ilość czasu, zanim wyśle, co znacznie redukuje widoczność ruchu w sieci.

Praca [7] dostarcza dodatkowych wskaźników potencjalnie niebezpiecznego ruchu ICMP, takich jak obecność specyficznych ciągów znaków w danych użytkowych, np. `root`, `ls` czy `pwd`. W naszym rozwiązaniu problem ten został całkowicie wyeliminowany dzięki zastosowaniu szyfrowania. Szyfrowanie gwarantuje, że żadne czytelne ciągi znaków nie będą widoczne w pakietach ICMP, co uniemożliwia identyfikację przesyłanych informacji na podstawie ich zawartości. Dzięki temu biblioteka jest odporna na proste metody detekcji oparte na wyszukiwaniu charakterystycznych ciągów znaków.

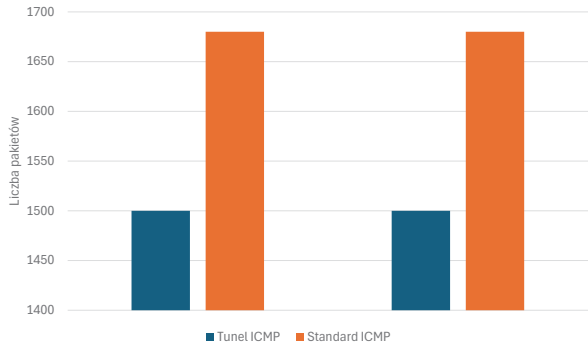
Dodatkowym aspektem, który warto poruszyć w kontekście porównania z istniejącymi rozwiązaniami, jest analiza przedstawiona w pracy [8]. Autorzy tej pracy skupili się na wykrywaniu anomalii w ruchu ICMP, zwracając uwagę na sytuacje, w których liczba odpowiedzi (reply) na jedno zapytanie (request) jest większa niż 1. Takie zachowanie jest nietypowe i często wskazuje na obecność niepożądanych mechanizmów w komunikacji ICMP. Algorytm zaproponowany w pracy [8] skutecznie identyfikuje tego rodzaju anomalie, które mogą wynikać z błędnych implementacji lub celowych prób zakłócenia ukrytej transmisji.

Prezentowane rozwiązanie całkowicie eliminuje ten problem dzięki zastosowaniu mechanizmu, który nie generuje dodatkowych odpowiedzi. Biblioteka działa w trybie pasywnym, co oznacza, że jedynie nasłuchuje odpowiedzi pochodzących z interfejsu sieciowego i nie wysyła ich ponownie. Retransmisji podlegają tylko pakiety *ICMP Echo Request*. Taka implementacja znacząco utrudnia wykrycie ukrytego tunelowania za pomocą algorytmu opisanego w pracy [8].

6. ANALIZA WYNIKÓW

6.1 Liczba zapytań do liczby odpowiedzi

Jak widać na Rysunku 1 stosunek liczby zapytań do liczby odpowiedzi wynosi 1:1. Zbędne wiadomości *ICMP Echo Reply* nie są generowane przez bibliotekę. Jeśli nastąpiłaby chwilowa niedostępność hosta, wtedy liczba zapytań mogłaby wzrosnąć, jednak nie udało się zestawić takiego scenariusza. Stosunek liczby zapytań do liczby odpowiedzi pozostaje stabilny na poziomie 1:1, co znacząco utrudnia wykrycie ukrytego tunelowania za pomocą algorytmu opisanego w pracy [6].

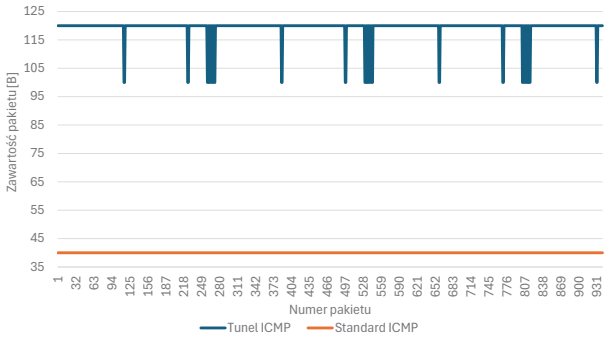


Rysunek 1: Liczba zapytań (po lewej) oraz liczba odpowiedzi (po prawej)

6.2 Rozmiar zawartości pakietu w proponowanej bibliotece

Jak przedstawia Rysunek 2 rozmiar zawartości pakietu (payload) w proponowanej bibliotece jest nieznacznie większy niż w przypadku zwykłego payloadu, w porównaniu do wartości rzędu 1000 B z pracy [6]. Wartość ta wynosi od 100 do 120 B, co jest optymalnym rozmiarem dla zapewnienia niezawodności i ukrycia komunikacji. Dzięki zastosowaniu szyfrowania żadne czytelne ciągi znaków nie są widoczne w pakietach ICMP, co uniemożliwia identyfikację przesyłanych informacji na podstawie ich zawartości, kosztem rozmiaru pakietu. Wahania widać tylko w momentach, w których wysyłane fragmenty danych były krótsze niż 15 bajtów. Przedstawiony wykres pokazuje

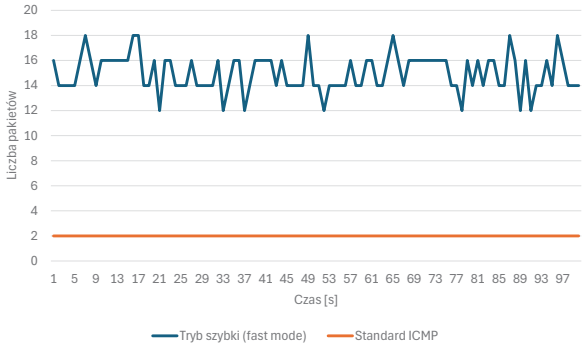
działanie biblioteki dla trybu default. W trybie fast i stealth rozmiar pakietu jest stały i wynosi 100 bajtów. Dalej jest to więcej niż domyślne 40 bajtów dla normalnego ruchu ICMP, ale znacząco mniej niż 1000 bajtów.



Rysunek 2: Rozmiar zawartości pakietu w proponowanej bibliotece w porównaniu do normalnego ruchu ICMP

6.3 Szybkość wysyłanych pakietów ICMP na sekundę

Rysunek 3 przedstawia szybkość wysyłanych pakietów ICMP na sekundę w zależności od wybranego trybu. W trybie fast, biblioteka wysyła od 12 do 18 pakietów na sekundę, co jest lekko większe niż wartość domyślna dla narzędzia ping. W trybie fast wiadomości są wysyłane w krótkich odstępach, co znacząco przyspiesza wykonywane testy. W trybie stealth biblioteka wysyła 1–2 pakiety na sekundę, co jest zgodne z wartościami domyślnymi wysyłanymi przez narzędzie ping. W trybie stealth wiadomości są wysyłane w dłuższych odstępach, co umożliwia ukrycie komunikacji w tle innych pakietów ICMP.



Rysunek 3: Liczba wysyłanych pakietów ICMP na sekundę — tryb fast mode

7. WNIOSKI

Porównania z wcześniejszymi rozwiązaniami dowodzą, że zastosowanie małych, stabilnych pakietów ICMP o stałej strukturze danych, połączonych z szyfrowaniem oraz pasywnym trybem działania, znacznie utrudnia detekcję tunelowania. Ponadto analiza wyników wykazała, że poprawa parametrów takich jak długość payloadu, identyfikator pakietu [9] czy proporcja liczby odpowiedzi sprawia, że proponowane rozwiązanie w pełni integruje się z typowym ruchem sieciowym. Opracowana biblioteka wykorzystuje surowe gniazda (raw sockets), co umożliwia manipulację pakietami ICMP bez użycia portów. Dzięki temu ruch generowany przez bibliotekę jest trudniejszy do wykrycia przy użyciu typowych mechanizmów monitorowania opartych na analizie portów.

W ramach dalszych prac planowana jest rozbudowa funkcjonalności biblioteki o możliwość równoczesnej obsługi wielu połączeń pochodzących od różnych klientów. Umożliwi to realizację bardziej zaawansowanych scenariuszy komunikacyjnych, takich jak komunikacja w architekturze typu klient-serwer. Rozwój ten stanowić będzie również podstawę dla przyszłej pracy magisterskiej, w której biblioteka zostanie zaadaptowana jako podstawowy komponent do implementacji botnetu wykorzystującego ruch ICMP jako kanał komunikacji serwera Command and Control.

LITERATURA

[1] J. Postel, “Internet Control Message Protocol, RFC 792,” Tech. Rep., Sep. 1981.

[2] Z. Trabelsi, W. El-Hajj, and S. Hamdy, “Implementation of an icmp-based covert channel for file and message transfer,” in *2008 15th IEEE International Conference on Electronics, Circuits and Systems*. IEEE, Aug. 2008, p. 894–897. [Online]. Available: http://dx.doi.org/10.1109/ICECS.2008.4674998

[3] B. Ray and S. Mishra, “A protocol for building secure and reliable covert channel,” in *2008 Sixth Annual Conference on Privacy, Security and Trust*. IEEE, Oct. 2008, p. 246–253. [Online]. Available: http://dx.doi.org/10.1109/PST.2008.26

[4] M. D. Hisbullah, C. A. Sari, E. H. Rachmawanto, L. B. Handoko, C. Umam, and F. Isinkaye, “Hybrid encryption based on fernet and rivest shamir adleman (rsa),” in *2023 International Seminar on Application for Technology of Information and Communication (iSemantic)*. IEEE, Sep. 2023, p. 352–357. [Online]. Available: http://dx.doi.org/10.1109/iSemantic59612.2023.10295309

[5] “Python Cryptography documentation, Fernet class.” [Online]. Available: https://cryptography.io/en/45.0.2/fernet

[6] H. Duan and B. Wang, “Detection of covert tunnel over internet control message protocol based on baseline features,” in *2023 IEEE 29th International Conference on Parallel and Distributed Systems (ICPADS)*, 2023, pp. 2210–2218.

[7] Z. Sui, H. Shu, F. Kang, Y. Huang, and G. Huo, “A comprehensive review of tunnel detection on multilayer protocols: From traditional to machine learning approaches,” *Applied Sciences*, vol. 13, no. 3, p. 1974, Feb. 2023. [Online]. Available: http://dx.doi.org/10.3390/app13031974

[8] S. Sayadi, T. Abbes, and A. Bouhoula, “Detection of covert channels over icmp protocol,” in *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*. IEEE, Oct. 2017, p. 1247–1252. [Online]. Available: http://dx.doi.org/10.1109/AICCSA.2017.60

[9] G. Kaddissi, M. W. Hatoum, K. Samrouth, and A. El Attar, “Securing networks against icmp covert channels: Detection and countermeasures,” in *2025 5th IEEE Middle East and North Africa Communications Conference (MENACOMM)*. IEEE, Feb. 2025, p. 1–8. [Online]. Available: http://dx.doi.org/10.1109/MENACOMM62946.2025.10911039

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

ANALIZA PORÓWNAWCZA WYMAGAŃ UNIJNEGO AKTU W SPRAWIE SZTUCZNEJ INTELIGENCJI ORAZ NORMY ISO/IEC 42001

COMPARISON BETWEEN EUROPEAN UNION'S AI ACT AND ISO/IEC 42001 REQUIREMENTS

Magdalena Ślusarczyk[1]; Dawid Juszka[2]

[1] Akademia Górniczo-Hutnicza, Instytut Telekomunikacji, Kraków, mslusarczyk@student.agh.edu.pl
[2] Akademia Górniczo-Hutnicza, Instytut Telekomunikacji, Kraków, dawid.juszka@agh.edu.pl

DOI: 10.15199/59.2025.4.18

Streszczenie: **Przedstawiono analizę spójności między unijnym rozporządzeniem tzw. aktem o sztucznej inteligencji (AI Act) a normą ISO 42001 dotyczącą systemów zarządzania sztuczną inteligencją (AI). Porównano wybrane kluczowe wymagania dla rozwiązań wysokiego ryzyka i oceniono, w jakim stopniu ISO 42001 może wspierać organizacje w osiągnięciu zgodności z AI Act. Wyniki umożliwiają organizacjom dążącym do wdrożenia frameworków zarządzania sztuczną inteligencją, zapewnienie zgodności z ciągle zmieniającymi się regulacjami w obszarze AI.**

Abstract: **This article analyzes the alignment between the European Union's Artificial Intelligence Act and the ISO 42001 standard for AI management systems. By comparing selected key requirements for high-risk systems, the study assesses the extent to which ISO 42001 can support organizations in achieving compliance with the AI Act. The findings highlight areas of the highest and lowest coverage, offering insights for organizations seeking to implement robust AI governance frameworks that will ensure compliance in everchanging AI regulatory landscape.**

Słowa kluczowe: **AI Act, ISO 42001, zarządzanie AI, zgodność, regulacje**

Keywords: **AI Act, ISO 42001, AI governance, compliance, regulations**

1. INTRODUCTION

W obliczu szybkiej transformacji wywołanej przez intensywny rozwój sztucznej inteligencji (SI), która stwarza nowe możliwości, a jednocześnie budzi obawy dotyczące potencjalnych zagrożeń i implikacji etycznych, Unia Europejska postanowiła opracować Akt w sprawie sztucznej inteligencji (AI Act, AIA) [1]. Rozporządzenie to jest odpowiedzią na potrzebę proaktywnego zarządzania SI i ma na celu ustanowienie zharmonizowanych ram prawnych dla sztucznej inteligencji, zapewniając jej wiarygodność oraz tworzenie procedur jej rozwoju i wdrażania z uwzględnieniem wszystkich potencjalnych skutków wykorzystania takich systemów.

AIA przyjmuje podejście oparte na ryzyku, kategoryzując systemy SI w zależności od ich potencjalnej szkodliwości i nakładając surowsze wymogi na rozwiązania wysokiego ryzyka, szczególnie w tak wrażliwych sektorach jak infrastruktura krytyczna, opieka zdrowotna, transport i egzekwowanie prawa. Usługi telekomunikacyjne są wykorzystywane we wszystkich tych sektorach, a jeśli zawierają określone funkcje SI, mogą podlegać kategorii wysokiego ryzyka. Przykładami takich systemów mogą być narzędzia do zarządzania ruchem sieciowym oparte na SI, rozwiązania do przechwytywania i nadzoru sieciowego wykorzystujące SI lub narzędzia cyberbezpieczeństwa analizujące ruch sieciowy i przewidujące ataki. Wymogi dla systemów wysokiego ryzyka obejmują kluczowe aspekty wytwarzania, wdrażania i dystrybucji systemu SI, takie jak zarządzanie ryzykiem, zarządzanie danymi, nadzór człowieka, czy solidność. AIA nakłada również pewne obowiązki na państwa członkowskie UE, związane z nadzorem nad firmami opracowującymi rozwiązania SI w ich kraju oraz z wyznaczaniem organów notyfikujących.

Równolegle Międzynarodowa Organizacja Normalizacyjna (ISO) wprowadziła normę ISO 42001 [4], standard dla systemów zarządzania SI. Norma ta proponuje działania, jakie organizacje mogą podjąć w celu ustanowienia, wdrożenia, utrzymania i ciągłego doskonalenia systemów i procedur, aby skutecznie zarządzać ryzykiem i możliwościami związanymi z SI. Ma ona na celu pomóc organizacjom w odpowiedzialnym pełnieniu ich roli w odniesieniu do systemów SI (np. w użytkowaniu, rozwijaniu, monitorowaniu lub dostarczaniu produktów lub usług wykorzystujących SI).

Niniejszy artykuł przedstawia analizę porównawczą AIA i normy ISO 42001, badając zbieżność ich kluczowych wymagań. Oceniamy, w jakim stopniu norma ISO 42001 obejmuje obowiązki dla systemów SI oraz ich dostawców i podmiotów wdrażających, określone w AIA. Nasza analiza pokazuje, które obszary unijnego rozporządzenia są już objęte normą ISO. Ostatecznie dążymy do oceny potencjału normy ISO 42001 jako podstawy do budowania zgodności z znacznie szerszymi i bardziej szczegółowymi regulacjami, takimi jak AIA.

2. POWIĄZANE PRACE

Naukowcy z Uniwersytetu Oksfordzkiego stworzyli narzędzie o nazwie capAI [2], procedurę służącą do przeprowadzania oceny zgodności systemów sztucznej inteligencji zgodnie z AIA. CapAI dostarcza organizacjom praktycznych wskazówek, jak przełożyć ogólne zasady etyczne na weryfikowalne kryteria, które pomagają kształtować projektowanie, rozwój, wdrażanie i

użytkowanie etycznej sztucznej inteligencji. Narzędzie to może być wykorzystane nie tylko do wykazania, że rozwój i działanie systemu SI są godne zaufania, ale także do lepszego zrozumienia treści AIA.

W pracy [3] D. Golpayegani, H. J. Pandit i D. Lewis przedstawili wstępne porównanie normy ISO/IEC 42001 z unijną listą oceny godnej zaufania sztucznej inteligencji (ALTAI) oraz proponowanym AIA, wykorzystując ontologię wyższego poziomu do zapewnienia interoperacyjności semantycznej między dokumentami dotyczącymi godnej zaufania SI. Przedstawili oni to porównanie w formie grafu zasobów RDF (Resource Description Framework), aby umożliwić jego dalszy rozwój. Badanie to uwypukliło luki i zbieżności w badanych dokumentach, dostarczając spostrzeżeń organom normalizacyjnym, ustawodawcom i twórcom SI w celu zwiększenia harmonizacji między wymogami prawnymi a normatywnymi.

Innym opracowaniem porównującym AIA z innymi standardami związanymi z SI i bezpieczeństwem jest praca [5] autorstwa T. R. McIntosha, T. Susnjaka, T. Liu, P. Wattersa, D. Xu, D. Liu, R. Nowrozy'ego i M. N. Halgamuge. W badaniu tym przeanalizowano gotowość do integracji czterech standardów w obszarze cyberbezpieczeństwa – NIST CSF 2.0, COBIT 2019, ISO 27001:2022 oraz ISO 42001:2023 – pod kątem możliwości, adresacji ryzyk i zgodności z przepisami przy wdrażaniu dużych modeli językowych (LLM). Ich analiza ujawniła niedociągnięcia w nadzorze nad ryzykiem związanym z LLM w tych ramach. Porównawcza analiza luk wykazała, że nowa norma ISO 42001:2023, zaprojektowana specjalnie dla systemów zarządzania SI, zapewnia najbardziej kompleksowe pokrycie dla systemów LLM, podczas gdy COBIT 2019 jest najbardziej zgodny z AIA.

3. ZAKRES

Ponieważ zakres AIA jest znacznie szerszy niż zakres normy ISO 42001, obejmując nie tylko wymagania dla dostawców i podmiotów wdrażających systemy SI, ale także wymagania dotyczące importerów, dystrybutorów, organów notyfikujących i jednostek notyfikowanych UE i jej państw członkowskich, a także związane z normami, ocenami zgodności, certyfikacją i rejestracją systemów SI, niniejszy artykuł skupia się wyłącznie na obszarach związanych z samym systemem SI oraz jego dostawcami i podmiotami wdrażającymi. Uzasadnieniem tego jest fakt, że ISO 42001 to norma określająca wymagania dotyczące ustanowienia, wdrożenia, utrzymania i ciągłego doskonalenia systemu zarządzania SI w kontekście organizacji, niezwiązanej terytorialnie z UE. Biorąc to pod uwagę, znaczna część wymagań AIA domyślnie nie będzie objęta normą ISO.

Niniejszy artykuł przedstawia wyniki wyłącznie dla systemów wysokiego ryzyka, ponieważ jest to najważniejsza i najbardziej rygorystyczna część AIA. Obowiązki dotyczące systemów SI ogólnego przeznaczenia i innych, które nie należą do kategorii wysokiego ryzyka, wykraczają poza zakres tego opracowania.

Podczas analizy nie uwzględniono następujących artykułów i sekcji AIA, dotyczących systemów SI wysokiego ryzyka: Art. 21 Współpraca z właściwymi organami, Art. 22. Upoważnieni przedstawiciele dostawców systemów AI wysokiego ryzyka, Art. 23 Obowiązki importerów, Art. 24 Obowiązki dystrybutorów, Art. 25 Odpowiedzialność w całym łańcuchu wartości AI, Rozdział 4, Sekcja 4: Organy notyfikujące i jednostki notyfikowane, Rozdział 4, Sekcja 5: Normy, ocena zgodności, certyfikaty, rejestracja, Rozdział VI Środki wspierające innowacyjność.

4. PODEJŚCIE DO ZAGADNIENIA

Na początku analizy porównawczej dokładnie zbadano zakres AIA w celu zidentyfikowania konkretnych, istotnych wymagań, zgodnie z opisem w sekcji **Error! Reference source not found.**. Każde z wybranych wymagań AIA zostało następnie porównane z klauzulami normy ISO 42001, a także z Załącznikiem A tej normy, który zawiera zestaw referencyjnych kontroli, które organizacje mogą wykorzystać do zarządzania ryzykiem związanym z SI i osiągania celów biznesowych. Wynikają one bezpośrednio z klauzul normy i dostarczają dodatkowego opisu do niektórych jej wymagań. Proces mapowania pomiędzy dwoma dokumentami obejmował wyszukiwanie w normie ISO 42001 identycznych terminów lub słów kluczowych użytych w wymaganiach AIA lub ich synonimów, a także analizę wymagań ISO w celu znalezienia tych o podobnym znaczeniu lub celach, nawet jeśli nie używają one jawnie tej samej terminologii co AIA.

Po zidentyfikowaniu potencjalnych dopasowań oceniono poziom zgodności między każdym wymaganiem AIA a zmapowaną(-ymi) klauzulą(-ami) ISO 42001. Przypisano trzy poziomy pokrycia:

- Pełne pokrycie: Wymaganie ISO w pełni odnosi się do intencji i zakresu wymagania AIA – ma to samo znaczenie lub jest szerzej zdefiniowane, a zatem obejmuje wymaganie AIA.
- Częściowe pokrycie: Wymaganie ISO częściowo odnosi się do wymagania AIA, ale z pewnymi lukami, ograniczeniami lub nie jest tak szczegółowe.
- Brak pokrycia: Norma ISO nie odnosi się do wymagania AIA – nie znaleziono odpowiednich wymagań.

To ustrukturyzowane podejście pozwoliło na kompleksową i obiektywną ocenę, w jakim stopniu organizacja wdrażająca normę ISO 42001 spełnia również wymagania AIA, jednocześnie podkreślając obszary, w których konieczne mogą być dalsze działania w celu zapewnienia zgodności.

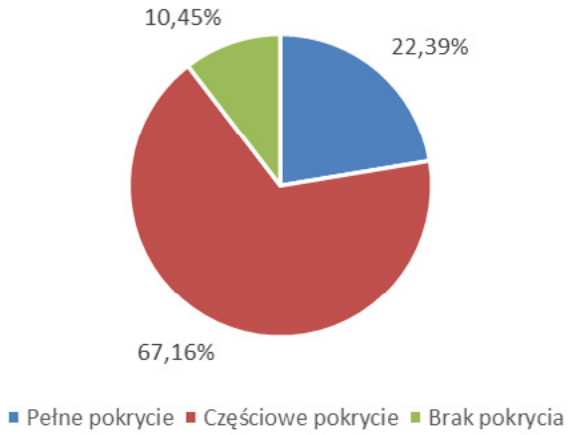
5. WYNIKI

Analiza wykazała, że znaczna większość wymagań AIA nie mogła być w pełni pokryta przez normę ISO 42001. Mimo tego, że ISO 42001 dostarcza podstawę do

ustanowienia solidnego systemu zarządzania SI, skupia się ona głównie na ustanawianiu polityk, standardów i procedur związanych z rozwojem i użytkowaniem systemów SI, a nie na szczegółowych wymaganiach technicznych dla systemów SI wysokiego ryzyka, tak jak zostało określone w AIA.

Zostało to zilustrowane na Rys. 1, który przedstawia wizualną reprezentację wyników. Stwierdzono, że zaledwie 22,39% analizowanych wymagań AIA zostało w pełni uwzględnionych przez odpowiednie podpunkty normy ISO 42001. Chociaż większość (67,16%) wymagań miała pewien stopień zbieżności z ISO, odpowiednie wytyczne w tymże standardzie często nie miały takiej specyficzności i szczegółowości, jak w AIA. Co więcej, dla 10,45% wymagań nie było możliwe zmapowanie żadnego wymagania z normy ISO, co pokazuje dość znaczącą lukę w pokryciu.

Wynik ten jest konsekwencją natury i celu każdego z tych instrumentów. AIA, jako prawnie wiążący dokument, określa jednoznaczne i szczegółowe wymagania, skoncentrowane na ograniczaniu konkretnego ryzyka związanego z tworzeniem i użytkowaniem systemów SI wysokiego ryzyka. Z drugiej strony, norma ISO 42001 służy jako bardziej ogólne narzędzie do ustanowienia kompleksowego systemu zarządzania SI.



Rys. 1. Pokrycie wymagań AIA dla systemów wysokiego ryzyka przez ISO 42001.

Gdy przyjrzymy się bardziej szczegółowemu podziałowi wymagań, biorąc pod uwagę obszar, którego dotyczą, możemy jeszcze wyraźniej dostrzec naturę normy ISO 42001. Tab. 1 pokazuje pokrycie przez ISO w każdej kategorii wymagań AIA. Obszary w tym przypadku to tytuły artykułów AIA, które składają się z wielu ustępów, z których każdy stanowi wymaganie.

Zgodnie z oczekiwaniami, najlepsze pokrycie osiągnięto w kategoriach związanych z ustanawianiem systemów zarządzania, polityk i procedur oraz dokumentowaniem systemu, takich jak:

- System zarządzania ryzykiem – wymagania dotyczące systemu zintegrowanego jako ciągły proces w całym cyklu życia SI. ISO w pełni obejmuje ogólne wymagania tego artykułu, ale nie odnosi się do niektórych konkretnych kwestii, takich jak ocena ryzyka związanego ze zdrowiem czy określanie sposobu testowania systemu SI;
- Dokumentacja techniczna – przepisy dotyczące przygotowywania i aktualizowania dokumentacji technicznej systemu SI – ISO również wymaga sporządzenia szczegółowej dokumentacji technicznej, ale nie jest obowiązkowe, aby była ona zgodna z jednym znormalizowanym formatem;
- Przejrzystość i udostępnianie informacji podmiotom stosującym – artykuł dotyczący tworzenia systemów z zachowaniem przejrzystości i dostarczania odpowiedniej dokumentacji - przejrzystość nie jest bezpośrednio wspomniana w ISO 42001, ale istnieją obowiązki dotyczące dostarczenia wszystkich niezbędnych instrukcji, aby umożliwić prawidłowe wdrożenie systemu;
- System zarządzania jakością – postanowienia dotyczące dokumentacji i wdrożenia odpowiednich strategii zapewniających odpowiednią jakość systemu SI podczas całego jego cyklu życia - w pełni objęte przez ISO 42001;
- Ocena wpływu systemów SI wysokiego ryzyka na prawa podstawowe – wymóg przeprowadzenia oceny opisującej, jak i kiedy system będzie używany, na kogo może wpłynąć i jakie ryzyka może stwarzać – ISO nie opisuje oceny wpływu w tak dużym szczególe jak AIA, ale wspomina o konieczności uwzględnienia w ocenie wszystkich obowiązujących przepisów prawa i regulacji, co sprawia, że ten artykuł jest w pełni objęty.

Stwierdzono, że inne kategorie są omówione w ISO 42001 w znacznie mniejszym stopniu niż w AIA ze względu na mniejszą dokładność ISO 42001. Zaskakująco, artykułem, który w ogóle nie został objęty, było Przechowywanie dokumentacji – ISO nie zawiera żadnych konkretnych wymagań dotyczących tego, jak długo należy przechowywać dokumenty związane z systemem SI.

Tab. 1. Pokrycie wymagań AIA przez ISO 42001 z podziałem na obszary tematyczne.

	% wszystkich wymagań w obszarze		
Obszar	Pełne pokrycie	Częściowe pokrycie	Brak pokrycia
System zarządzania ryzykiem	40,00	50,00	10,00
Dane i zarządzanie danymi	0,00	100,00	0,00

Dokumentacja techniczna	50,00	50,00	0,00
Rejestrowanie zdarzeń	0,00	100,00	0,00
Przejrzystość i udostępnianie informacji podmiotom stosującym	50,00	50,00	0,00
Nadzór ze strony człowieka	0,00	100,00	0,00
Dokładność, solidność i cyberbezpieczeństwo	25,00	75,00	0,00
Obowiązki dostawców systemów AI wysokiego ryzyka	14,29	42,86	42,86
System zarządzania jakością	100,00	0,00	0,00
Prowadzenie dokumentacji	0,00	0,00	100,00
Automatycznie generowane rejestry zdarzeń	0,00	100,00	0,00
Działania naprawcze i obowiązek informacyjny	0,00	100,00	0,00
Obowiązki podmiotów stosujących systemy AI wysokiego ryzyka	22,22	55,56	22,22
Ocena skutków systemów AI wysokiego ryzyka dla praw podstawowych	100,00	0,00	0,00
Prowadzone przez dostawców monitorowanie po wprowadzeniu do obrotu i plan monitorowania systemów AI wysokiego ryzyka po ich wprowadzeniu do obrotu	33,33	66,67	0,00
Zgłaszanie poważnych incydentów	0,00	100,00	0,00

6. WNIOSKI I MOŻLIWOŚCI ROZWOJU

Wyniki wskazują, że norma ISO 42001 skupia się głównie na procesach organizacyjnych, zarządzaniu ryzykiem i kwestiach etycznych, podczas gdy AIA zagłębia się w szczegółowe specyfikacje, procedury oceny zgodności oraz obowiązki prawne dla dostawców i podmiotów wdrażających. Sprawia to, ze konieczne jest, aby organizacje wyszły poza ramy normy ISO 42001, szczegółowo przeanalizowały każde wymaganie i wdrożyły dodatkowe środki w celu zapewnienia pełnej zgodności z AIA.

Jednakże, pomimo braku pełnego pokrycia AIA, norma ISO 42001 pozostaje cennym narzędziem dla organizacji dążących do odpowiedzialnego rozwijania i wdrażania SI. Przyjmując tę normę, organizacje mogą ustanowić solidny system zarządzania SI, wzmocnić wewnętrzne procesy i procedury oraz stworzyć podstawę do wdrożenia dalszych kontroli, co ostatecznie pozwoli im na zgodność z różnymi regulacjami, w tym z AIA.

Niniejsze badanie stanowiwstępną eksplorację zgodności między AIA a normą ISO 42001. Przyszłe badania mogłyby zagłębić się w konkretne aspekty Aktu i normy, dostarczając bardziej szczegółowych wytycznych dla organizacji. Potencjalne kierunku rozwoju w tej dziedzinie obejmują:

- analizę wymagań dla innych systemów SI nienależących do kategorii wysokiego ryzyka lub systemów SI ogólnego przeznaczenia;
- opracowanie wytycznych specyficznych dla danego sektora;
- zbadanie, w jaki sposób proces certyfikacji ISO 42001 może być wykorzystany w procedurze oceny zgodności w ramach Aktu w sprawie SI;
- analizę rzeczywistych przykładów organizacji wdrażających zarówno wymagania ISO 42001, jak i AIA.

Dalsze badania mogą przyczynić się do rozwoju kompleksowych wytycznych dotyczących odpowiedzialnego użytkowania i rozwoju sztucznej inteligencji.

REFERENCES

[1] European Union. 2024. “Regulation - EU - 2024/1689 - EN - EUR-Lex.”. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

[2] Floridi, L., Holweg, M., Taddeo, M., Amaya, J., Mökander, J., Wen, Y. 2022. “capAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act”. https://ssrn.com/abstract=4064091

[3] Golpayegani, D., Pandit, H.J., Lewis, D. 2023. “Comparison and Analysis of 3 Key AI Documents: EU’s Proposed AI Act, Assessment List for Trustworthy AI (ALTAI), and ISO/IEC 42001 AI Management System.” Artificial Intelligence and Cognitive Science. AICS 2022. Communications in Computer and Information Science, vol 1662.

[4] ISO. 2023. “ISO/IEC 42001:2023,” https://www.iso.org/standard/81230.html.

[5] Timothy R. McIntosh, Teo Susnjak, Tong Liu, Paul Watters, Dan Xu, Dongwei Liu, Raza Nowrozy, Malka N. Halgamuge, 2024. “From COBIT to ISO 42001: Evaluating cybersecurity frameworks for opportunities, risks, and regulatory compliance in commercializing large language models”. Computers & Security, vol 144.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

METROLOGIA PRAWNA W ASPEKCIE USTAW i ROZPORZĄDZEŃ KRAJOWYCH DOYCZĄCYCH PÓL ELEKTROMAGNETYCZNYCH

LEGAL METROLOGY IN THE ASPECT OF NATIONAL ACTS AND REGULATIONS CONCERNING ELECTROMAGNETIC FIELDS

Waldemar Karcz

Emitel S.A. Szczecin, waldemar.karcz@emitel.pl

DOI: 10.15199/59.2025.4.19

Streszczenie**: W referacie zostały przedstawione rozbieżności pomiędzy metrologią prawną i metrologią naukową w zakresie pomiarów EM i w ocenie zgodności. Polski ustawodawca w zakresie metrologii prawnej zatrzymał się na roku 2005 i mocy EIRP≥ 15 W i w ten sposób blokuje możliwości stosowania uproszczonych procedur zgodności urządzeń instalowanych w stacjach bazowych. Według polskiego ustawodawcy metrologii prawnej argument siły jest ważniejszy od siły argumentów metrologii naukowej.**

Abstract**: The paper presents the discrepancies between legal metrology and scientific metrology in the field of EM measurements and in the assessment of conformity. The Polish legislator in the field of legal metrology stopped at 2005 and on the basis of EIRP≥ 15 W and thus blocks the possibility of applying simplified procedures for the conformity of devices installed in base stations. According to the Polish legislator of legal metrology, the argument of force is more important than the arguments of scientific metrology**

Słowa kluczowe**: gęstość mocy, metrologia naukowa, metrologia prawna, techniki pomiarowe**

Keywords**: legal metrology, measurement techniques, power density, scientific metrology**

1. WSTĘP

20 maja obchodzony był Światowy Dzień Metrologii, upamiętniający rocznicę podpisania Konwencji Metrycznej w Paryżu w 1875 r. [7] Traktat ten stanowi podstawę ogólnoświatowego, spójnego systemu pomiarowego, który stanowi podstawę odkryć naukowych i innowacji, produkcji przemysłowej i handlu międzynarodowego, a także poprawy jakości życia i ochrony globalnego środowiska. Tematem Światowego Dnia Metrologii 2025 są Pomiary na wszystkie czasy, dla wszystkich ludzi , wybrane, aby podkreślić znaczenie pomiarów w kształtowaniu naszej przeszłości, teraźniejszości i przyszłości. Temat ten nabrał znaczenia w 2025 r., który oznacza 150 rocznicę Konwencji Metrycznej. Na całym świecie krajowe laboratoria metrologiczne nieustannie rozwijają naukę pomiaru, opracowując i weryfikując nowe techniki pomiarowe na niezbędnym poziomie zaawansowania. Krajowe instytuty metrologiczne uczestniczą w porównaniach pomiarów koordynowanych przez Międzynarodowe Biuro Miar i Wag lub przez Regionalne Organizacje Metrologiczne, aby zapewnić niezawodność wyników pomiarów na całym świecie. Międzynarodowa Organizacja Metrologii Prawnej (OIML) opracowuje Międzynarodowe Rekomendacje, których celem jest dostosowanie i harmonizacja wymagań na całym świecie w wielu dziedzinach. OIML prowadzi również System Certyfikacji OIML (OIML-CS), który ułatwia międzynarodową akceptację i globalny handel przyrządami pomiarowymi.

2. METROLOGIA

2.1. Metrologia prawna

Metrologia jest nauką zajmująca się sposobami dokonywania pomiarów oraz zasadami interpretacji uzyskanych wyników. Metrologia prawna (ang. legal metrology) jest to dział metrologii (nauki o pomiarach) odnoszący się do działań, które wynikają z wymagań ustawowych i dotyczą pomiarów, jednostek miar, przyrządów pomiarowych i metod pomiarowych, które przeprowadzane są przez organy rządowe i parlamentarne. W odniesieniu do ustaw i rozporządzeń dotyczących ochrony przed wpływem pól elektromagnetycznych i metod pomiarowych oraz prezentacji wyników w Polsce są to Ustawa Prawo ochrony środowiska i Rozporządzenie Ministra Klimatu z dnia 17 lutego 2020 r.

2.2. Metrologia naukowa

Metrologia naukowa zajmuje się tworzeniem, utrzymywaniem i ulepszaniem systemów odniesienia pomiarów. Obejmuje ona ustanawianie i utrzymywanie standardów pomiarowych, które są akceptowalne na arenie międzynarodowej i które są powiązane z podstawowym układem SI.W ten sposób metrologia naukowa przyczynia się do tworzenia i wdrażania standardów pomiarowych w badaniach, eksperymentach i innowacjach. W konsekwencji zapewnia porównywalną i niezawodną jakość pomiarów w różnych dziedzinach nauki na całym świecie. W odniesieniu do aktów zaliczanych do metrologii naukowej w zakresie pól elektromagnetycznych są to normy PN-EN IEC 62232:23-07[6], PN-EN 50383:2011/AC oraz szereg rekomendacji ITU K.100 [5], ETSI TR 102 457 v2.1.0.

3. PRAWO UE ZWIĄZANE Z PEM

3.1.Kraje europejskie bez Polski [4]

Dopuszczalne poziomy pól elektromagnetycznych, gęstości mocy oraz SAR opublikowane zostały w marcu 2020 r. przez Międzynarodową Komisję ds. Ochrony Przed Promieniowaniem Niejonizującym w dokumencie ICNIRP 2020 [2], zawierającym wytyczne dotyczące ochrony ludzi narażonych na działanie pola elektromagnetycznego o częstotliwościach radiowych w zakresie od

100 kHz do 300 GHz. Wytyczne ICNIRP 2020 zastępują wcześniejsze wytyczne z 1998 r. (ICNIRP1998) oraz uzupełniają wytyczne z 2010 r. (ICNIRP 2010) dotyczące pól niskich częstotliwości w zakresie częstotliwości od 100 kHz do 10 MHz. Zalecenie 1999/519/EC, jako zbiór wytycznych, stało się podstawą do określenia dopuszczalnych wartości poziomów PEM w środowisku w przepisach krajowych państw członkowskich Unii Europejskiej. W krajach: Czechy, Cypr, Estonia, Finlandia, Francja, Hiszpania, Irlandia, Malta, Portugalia, Rumunia i Węgry zalecenie 1999/519/EC zostało transponowane do obowiązującego ustawodawstwa krajowego. W tych krajach nie ma obowiązków wykonywania pomiarów PEM. W Czechach należy posiadać wyniki obliczeń (symulacji) potwierdzające zgodność z dopuszczalnymi poziomami PEM. W wielu z tych państw stosuje się stacje monitorujące poziomy PEM.

3.2. Stan prawny w Polsce – metrologia prawna według polskiego ustawodawcy.

Zmiany dopuszczalnych poziomów pól elektromagnetycznych w środowisku zostały wprowadzone Rozporządzeniem Ministra Zdrowia z dnia 17 grudnia 2019 r. Zmiany te weszły w życie 1 stycznia 2020 r. [13] Jednocześnie 1 stycznia 2020 r. przestało obowiązywać Rozporządzenie Ministra Środowiska z dnia 30.10.2003 r. Rozporządzenie to zostało zastąpione Rozporządzeniem Ministra Klimatu z dnia 17 lutego 2020 r. [11] w sprawie sposobów dotrzymywania dopuszczalnych poziomów pól elektromagnetycznych w środowisku. Pozytywna zmiana dopuszczalnych poziomów pól w środowisku wprowadziła nowe limity PEM dla poszczególnych zakresów częstotliwości zgodne z wytycznymi ICNIRP [2]. Niestety pozostałe rozporządzenia nie zostały zaktualizowane i dodatkowo w ramach tak zwanego „sprawnego i szybkiego procesu legislacyjnego" utrzymano w POŚ [14] artykuł 122a oraz utrzymano w mocy obowiązek zgłaszania instalacji radiokomunikacyjnych, których anteny emitują moc EIRP ≥15W. Pomimo zwiększenia, dla zakresu częstotliwości powyżej 2 GHz, dopuszczalnego poziomu gęstości mocy z 0,1 W/m² do 10 W/m² oraz z 01 W/m^2 do 2 W/m^2 dla 400 MHz ustawodawca w ustawie [14] POŚ w § 338 narzuca obowiązek wykonania pomiarów PEM w środowisku przez akredytowana laboratoria bez względu na fakt, że pozom PEM pochodzący od radiolinii jest kilkaset razy mniejszy od czułości sond pomiarowych wykorzystywanych do pomiarów przez te laboratoria i podobnie dla stacji bazowych z zakresu 380 – 400 MHz. Art. 338a „Kto, będąc obowiązany do wykonywania pomiarów pól elektromagnetycznych w środowisku, na podstawie art. 122a, nie wykonuje tych pomiarów, podlega karze aresztu albo ograniczenia wolności albo karze grzywny". To jest metrologia prawna obowiązująca w Polsce od 2005 roku. Art. 122a wymusza wykonywanie pomiarów PEM dla instalacji, której anteny emitują moc EIRP ≥ 15 W. Przed przystąpieniem do pomiarów zgodnie z Art. 121 i 122 [14] w połączeniu z ust. 5 pkt załącznika do rozporządzenia [12] należy wykonać obliczenia w celu wyszukania miejsc, w których występują poziomy PEM zbliżone do zbliżonych do poziomów dopuszczalnych. Wyniki obliczeń dla radiolinii zostały zestawione w tabeli nr 1. Parametry LR: 23 GHz z P=22 dBm, antena 03 m, G 35,6 dBi, EIRP w osi głównej 575,4 W.

Tab.1. Rozkład gęstości mocy w osi gł. i na wys. 1m n.p.t.

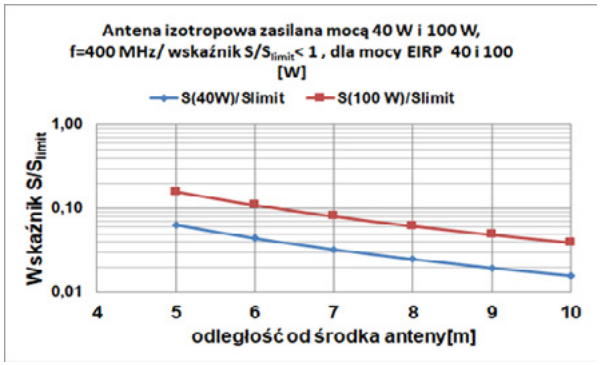
Pion [m]	*$S_{oś\ gł.}[Wm^{-2}]$*	*$S_{h1[m]}[Wm^{-2}]$*	*$S_{czułość\ sondy[Wm-2]}$*
10	0,458	5,79E-08	1,7E-03
20	0,114	8,75E-07	1,7E-03
30	0,051	1,75E-06	1,7E-03
40	0,029	1,62E-06	1,7E-03
50	0,018	1,86E-06	1,7E-03
60	0,013	1,92E-06	1,7E-03

Poziomy gęstości mocy w pionach na wys. 1 m n.p.t. są o więcej niż trzy rzędy wielkości mniejsze od czułości sond pomiarowych (tabela 1).

4. KRYTERIUM MOCY EIRP ≥15W

4.1. Geneza określenia instalacji mogących negatywnie oddziaływać na środowisko ze względu na moc EIRP ≥15 W

Określenie „ moc EIRP niemniejsza niż 15 W" pojawia się po raz pierwszy w Ustawie Prawo ochrony środowiska [14] jako kryterium kwalifikujące instalację do obowiązku uzyskania pozwolenia na emisję. Art. 234 miał następujące brzmienie: „Pozwolenie na emitowanie pól elektromagnetycznych jest wymagane dla instalacji radiokomunikacyjnych, radionawigacyjnych i radiolokacyjnych, których równoważna moc promieniowana izotropowo jest równa 15 W lub wyższa, emitujących pola elektromagnetyczne o częstotliwościach od 0,03 MHz do 300 000 MHz" . Następnie w 2005 roku wprowadzone zostały zmiany w POŚ [14] polegające na rezygnacji z obowiązku uzyskiwania pozwolenia na emisję. Został usunięty artykuł 234, a do rozdziału 6 został dodany kontrowersyjny art. 122a zobowiązujący prowadzących instalację oraz użytkowników urządzeń emitujących pola elektromagnetyczne o częstotliwościach od 30 kHz do 300 GHz, których równoważna moc promieniowana izotropowo wynosi nie mniej niż 15 W, do wykonania pomiarów poziomów pól elektromagnetycznych w środowisku. O ile ustawodawca do 2005 r. mógł wymagać pozwolenia na emisje na podstawie mocy EIRP ≥ 15 W bez merytorycznego uzasadnienia, to wprowadzenie art.122a powinno być uzasadnione w oparciu o dopuszczalne poziomy PEM w miejscach dostępnych dla ogółu ludności.



Rys. 1. Wskaźnik S/Slimi$_t$, moc EIERP 40W i 100W

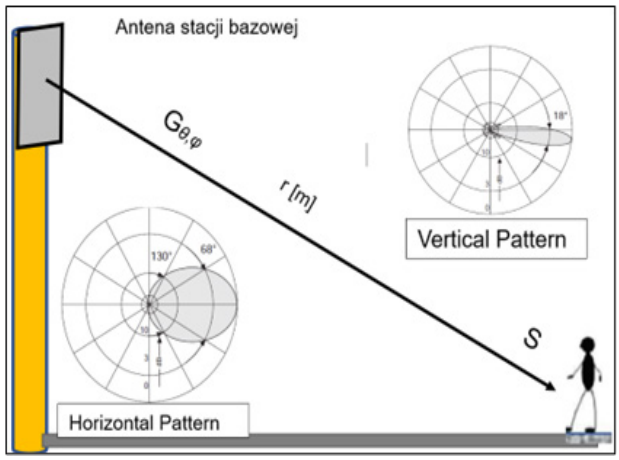
Wskaźnik S/Slimit przy mocy EIRP 40 W i 100 W i częstotliwości 400 MHz, w odległości od 5 do 10 metrów od anteny przedstawiony jest na rys. 1. Wskaźnik S/Slimit

jest znacznie niższy od 1 (Slimit = 2 W/m^2). Moc EIRP występuje tylko w kierunku wiązki głównej, a w kierunkach do miejsc dostępnych dla ogółu ludności moc jest mniejsza o współczynnik zależny od charakterystyki anteny przy zmianie kąta odchylenia od kierunku maksymalnego promieniowania w płaszczyźnie E I H. Analizując stan faktyczny należy zauważyć, że parametrem określającym narażenie ludzi na promieniowanie jest wartość skuteczna natężenia pola elektrycznego lub wartość średnia gęstości mocy, a nie tylko i wyłącznie moc EIRP. Gęstość mocy i natężenie pola wyznacza się z następujących zależności:

$$S = \frac{P \cdot G \cdot f(\theta,\varphi)}{4\pi r^2} \qquad (1)$$

gdzie:
S – gęstość mocy [W/m^2]
G – zysk anteny [W/W]
P – moc na wejściu anteny [W]
r – odległość od środka elektrycznego anteny do punktu, w którym będzie wyznaczana gęstość mocy w środowisku dostępnym dla ludności.
$f(\theta,\varphi)$ – charakterystyka promienia anteny, θ, φ kąt elewacji i azymutu.

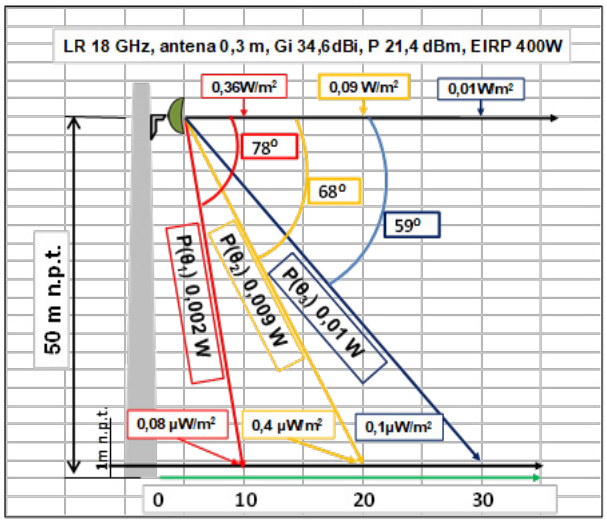
Na Rys. 2. przedstawiony został typowy przypadek oceny narażenia na częstotliwości radiowe. Przypadek ten został przedstawiony w normie [6]. Jako pierwsze oszacowanie można zastosować przestrzennie uśrednioną i przestrzennie szczytową równoważną gęstość mocy korzystając z pomocy prostych wzorów sferycznych opisanych w równaniu (2).



Rys. 2. Typowy przypadek oceny narażenia na RF

$$S = \frac{P \cdot G_{\theta,\varphi}}{4\pi r^2} = \frac{EIRP_{\theta,\varphi}}{4\pi r^2} \qquad (2)$$

W przeciwieństwie do art. 122a i art.338a, 338b [14] sposób wyznaczania gęstości mocy w kierunku od środka anteny do punktu obserwacji, w miejscach dostępnych dla ogółu ludności, opisany w normie [6] jest logiczny i wynika z elementarnych podstaw teorii pola elektromagnetycznego. Jeszcze bardziej nielogiczne jest stosowanie art. 122a, 338a i 338b do instalacji radiolinii. Przykładowo, przęsło łącza radioliniowego o długości 12 km, z antenami 0,3m, f =18GHz, EIRP 400W, a odpowiednio do kierunków P(θ_1) 0,002 W, P(θ_2) 0,009 W, P(θ3) 0,01 W.



Rys. 3. Rozkład gęstości mocy na wys. 1m n.p.t.

Wartości gęstości mocy na poziomie 1 m n.p.t. są niemierzalne miernikami z sondami pomiarowymi, którymi pomiary wykonują akredytowane laboratoria. Postępując zgodnie z normą PN-EN IEC 62232 [6] została wyznaczona ocena zgodności dla anteny z Rys. 3. Jest to obszar ograniczony powierzchnią cylindra o średnicy anteny D=0,3m i długości walca CDlos=0,68 m.

4.2. Instalacje mogące negatywnie oddziaływać na środowisko a moc EIRP

Obecnie kryterium mocy EIRP ≥ 15 W występuje w rozporządzeniach wykonawczych jako swoisty dogmat nie podlegający sprawdzeniu. Analizując stan faktyczny należy zauważyć, że parametrem określającym narażenie ludzi na promieniowanie jest wartość skuteczna natężenia pola elektrycznego lub wartość średnia gęstości mocy. Gęstość mocy w dowolnym punkcie obserwacji wyznacza się z następujących zależności:

$$S = \frac{P \cdot G \cdot F_{f(\varphi,\theta)}}{4\pi \cdot d^2}; E = \sqrt{\frac{30 \cdot P \cdot G \cdot F_{f(\varphi,\theta)}}{d^2}} \qquad (1)$$

gdzie:
S – gęstość mocy [W/m^2]
E – natężenie pola [V/m]
P – moc na wejściu anteny [W]
G – zysk anteny w odniesieniu do anteny izotropowej
$F_{f(\varphi,\theta)}$ – współczynnik zależny od kąta φ i θ
d – odległość od środka elektrycznego anteny do punktu, w którym wyznaczana jest gęstość mocy [m]

4.3. Ochrona przed polami EM-metrologia naukowa

Ochrona przed polami elektromagnetycznymi polega na zapewnieniu jak najlepszego stanu środowiska poprzez utrzymanie poziomów pól elektromagnetycznych poniżej dopuszczalnych wartości określonych w Rozporządzeniu Ministra Zdrowia[13]. Ten cel osiąga się poprzez:

a) instalowanie anten na wysokich podporach
b) dobór anten o kształtowanej odpowiednio charakterystyce w płaszczyźnie pionowej i poziomej
c) dobór optymalnej mocy EIRP
d) zachowanie odpowiedniej odległości od miejsc dostępnych dla ludności

Parametrami charakteryzującymi oddziaływanie na środowisko są gęstość mocy lub natężenie pola. Znajomość wartość mocy EIRP dotyczącej danej instalacji jest warunkiem koniecznym, ale niewystarczającym aby przesądzać o wartości gęstości mocy lub o wartości natężenia pola w miejscach dostępnych dla ludności. Miejsca dostępne dla ludności znajdują się w obszarach pola dalekiego, gdzie prawdziwa jest następująca zależność:

$$S = |E| \cdot |H| = \frac{E^2}{120 \cdot \pi} = 120 \cdot \pi \cdot H^2 \qquad (4)$$

gdzie:

E – wartość skuteczna natężenia pola elektrycznego
H – wartość skuteczna natężenia pola magnetycznego

SEMINARIUM OIML 2002

Seminarium OIML, Wrzesień 2002 Tematem przewodnim było „Jaka będzie metrologia prawna w roku 2020 ?" Ogólne wprowadzenie do rozpoczęcia Seminarium przedstawił Gerard Faber, Prezes CIML

Rola państwa w metrologii prawnej będzie w 2020 r. inna niż ta z 2002. Państwo będzie miało cztery obowiązki w zakresie metrologii prawnej: a) tworzenie i utrzymywanie krajowego systemu metrologii; b) opracowywanie ustawodawstwa i zapewnianie jego wdrażania; c) definiowanie ogólnej polityki metrologii i akredytacji; d) współpraca globalna i regionalna. Wspomniałem o słowach „krajowy system metrologii". Postulaty zebrane w trakcie tego seminarium w 2002 r zostały spełnione. Zostało to sprawdzone 2020 r. W końcu powstały krajowe systemy pomiarowe, regionalne systemy pomiarowe i globalny system pomiarowy, wszystko do siebie pasujące.

PODSUMOWANIE

Jednak w Polsce, w zakresie metrologii prawnej dotyczącej pomiarów PEM ustawodawca zatrzymał się na 2005 roku z mocą EIRP ≥ 15 W. Wtedy zgodnie z Rozporządzeniem Ministra Środowiska z 30.10.2003 r dopuszczane natężenie pola wynosiło E ≤ 7 V/m i dopuszczalna gęstość mocy S ≤ 01 W/m². Nie nastąpiła zmiana nawet po 17.02.2020 r, kiedy zmieniły się limity natężenia pola i limity gęstości mocy [11] i [13].

W zakresie metrologii naukowej Polska należy do IEC (Międzynarodowej Komisji Elektrotechnicznej). Normy IEC umożliwiają wszystkie rodzaje oceny zgodności. Są one napisane w taki sposób, aby zgodność z nimi mogła ocenić każda zainteresowana strona. Polska metrologia naukowa ma dostęp nie tylko do norm IEC, ale również wnosi duży wkład w rozwój radiokomunikacji i teleinformatyki.

Norma PN-EN IEC 62232 2023-07E jest wzbogacona o nowe osiągnięcia i nowe opracowania z okresu od 2017 do 2022 i zawiera zaktualizowane opracowania z norm np. PN-EN 62232:2018 i PN-EN50383:2011. Norma ta nie tylko otwiera możliwości wyznaczania gęstości mocy i natężenia pola wokół instalacji klasycznych stacji bazowych, ale rozwiązuje również problemy zgodności nowych instalacji z antenami Massive MIMO i z SMART.

Od 2018 mamy polską normę PN-EN 62232 i należy korzystać z niej w celach tworzenia dokumentacji zgodności i obliczeń (symulacji) zamiast wykonywać pomiary PEM z powodu mocy EIRP ≥ 15 W.

Korzystając z normy PN-EN IEC 62232 i z rekomendacji ETSI TR 102 457 v2.1.0). W prosty sposób można wyznaczyć obszar zgodności dla anten parabolicznych radiolinii, zamiast stosować się do artykułu 122a, [14]

WHO i Komitet Naukowy ds. Zagrożeń dla Zdrowia stwierdziły, że narażenia z sieciami i ich użytkowaniem nie prowadzi do negatywnych skutków dla zdrowia, jeżeli jest poniżej limitów zalecanych przez ICNIRP. Na całym świecie krajowe agencje ds. zdrowia, rządowe organy regulacyjne, laboratoria badawcze, operatorzy komórkowi i producenci przeprowadzili szeroko zakrojone testy w sieciach komercyjnych i testowych w celu określenia poziomów narażenia na promieniowanie elektromagnetyczne 5G. Poziomy PEM 5G są podobne do innych technologii bezprzewodowych i nie przekraczają 1% gęstości mocy dopuszczalnych limitów określonych przez ICNIRP [2].

LITERATURA

[1] Faber Gerard *General introduction to the Seminar September 2002*
[2] ICNIRP *Guidelines for limiting exposure to electromagnetic fields (100 kHz-300 GHz)*, Health Phys, March 2020
[3] IEC 62232:2022 *Determination of RF field strength, power density and SAR in the vicinity of base stations for the purpose of evaluating human exposure*
[4] Instytut Łączności PIB *Analiza sposobów monitorowania PEM w wybranych krajach europejskich*
[5] ITU K.100 2021, *Measurement of radio frequency electromagnetic fields to determine compliance with human exposure limits when a base station is put into service*
[6] PN-EN IEC 62232:2023-07E Wyznaczanie natężenia pola RF, gęstości mocy i SAR w otoczeniu stacji bazowych dla oceny poziomu ekspozycji człowieka
[7] Message from Ms Audrey Azoulay, Director-General of UNESCO, World Metrology Day 20 May 2024
[8] Rozporządzeniu Rady Ministrów z dnia 10 września 2019 r. w sprawie przedsięwzięć mogących znacząco oddziaływać na środowisko(Dz.U. z 2019 r. poz. 1839)
[9] Rozporządzenie Rady Ministrów z 5 maja 2022 r. zmieniające rozporządzenie w sprawie przedsięwzięć mogących znacząco oddziaływać na środowisko (Dz.U. z 2022 r. poz. 1071)
[10] Rozporządzenie Ministra Środowiska z dnia 2 lipca 2010 r. w sprawie zgłoszenia instalacji wytwarzających pola elektromagnetyczne Dz.U. 2010 nr 130 poz. 879

[11] Rozporządzenie Ministra Klimatu z dnia 17 lutego 2020 *r w sprawie sposobów sprawdzania dotrzymania dopuszczalnych poziomów pól elektromagnetycznych w środowisku* (Dz. U. poz. 258)
[12] Rozporządzenie Ministra Klimatu z dnia 6 maja 2022 r. *zmieniające rozporządzenie w sprawie sposobów sprawdzania dotrzymania dopuszczalnych poziomów pól elektromagnetycznych w środowisku.* Dz. U dnia 26 maja 2022 r. Poz. 1121
[13] Rozporządzenie Ministra Zdrowia z dnia 17 grudnia 2019 r., *w sprawie dopuszczalnych poziomów pól elektromagnetycznych w środowisku,* Dz. U. 2019 poz. 2448
[14] Ustawa z dnia 27 kwietnia 2001 r., P*rawo ochrony środowiska*. Dz. U. 2022 poz. 2556
[15] Zalewski Paweł *Metrologia; Definicje i pojęcia podstawowe*

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

ANALIZA WPŁYWU SYGNAŁÓW INTERFERUJĄCYCH NA PRACĘ INTERFEJSU RADIOWEGO 5G-NR

IMPACT OF INTERFERING SIGNALS ON THE OPERATION OF THE 5G-NR RADIO INTERFACE

Piotr Rajchowski; Krzysztof Cwalina; Olga Błaszkiewicz; Alicja Olejniczak; Jarosław Sadowski;

Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk,
{piorajch, kkcwalina, olga.blaszkiewicz, alicja.olejniczak, jarsadow1}@pg.edu.pl

DOI: 10.15199/59.2025.4.20

Streszczenie: W artykule przedstawiono wyniki badań związanych z analizą wpływu rodzaju i parametrów sygnałów interferujących na pracę interfejsu radiowego 5G-NR. W pierwszej kolejności przedstawiono przykładowe sygnały interferujące i ich wpływ na sygnały warstwy fizycznej interfejsu radiowego 5G-NR. Następnie zweryfikowano stopień wpływu sygnałów interferujących na jakość procesu synchronizacji częstotliwości i czasu oraz dekodowania wiadomości rozgłoszeniowych w programowym odbiorniku sygnałów 5G-NR.

Abstract: In the paper the results of research related to the analysis of the impact of the type and parameters of interfering signals on the operation of the 5G-NR radio interface are presented. Firstly, authors performed an analysis of sample interfering signals and their influence on the 5G-NR radio interface physical layer. During the numerical studies impact of interfering signals on the quality of the frequency and time synchronization process and decoding of broadcast messages was verified.

Słowa kluczowe: 5G-NR, Interferencje, OFDM, Zagłuszanie

Keywords: 5G-NR, Interferences, Jamming, OFDM

1. WSTĘP

Sieci komórkowe czwartej (4G) oraz piątej (5G) generacji na przestrzeni ostatnich lat stały się powszechnym i podstawowym elementem masowej łączności radiowej. Wzrost wymaganej pojemności podyktowany rosnącą liczbą terminali użytkownika (User Equipment - UE) obsługiwanych przez człowieka bądź traktowanych np. jako urządzenia Internetu Rzeczy (Internet of Things - IoT) bezpośrednio przełożył się na znaczące zwiększenie liczby stacji bazowych eNodeB (Evolved Node B) bądź gNodeB (Next Generation Node B). Zjawisko to jest szczególnie widoczne w środowisku miejskim.

Wysokie zagęszczenie stacji gNodeB sprzyja powstawaniu interferencji własnych, co może być szczególnie widoczne w przypadku pracy w paśmie n77, gdzie w rzeczywistych implementacjach sieci komórkowych stacje gNodeB generują np. trzy niezależne wiązki w ramach jednego sektora [1, 2]. Emisja wielu wiązek z sygnałami rozgłoszeniowymi, tj. z blokiem synchronizacyjnym (Synchronization Signal Block - SSB) w środowisku miejskim prowadzi do odbioru przez terminal licznych odbić sygnału, a dynamiczna regulacja mocy może skutkować chwilowym pogorszeniem jakości odbioru w sąsiednich sektorach.

Osobny problem w systemach komórkowych stanowi celowe zakłócanie interfejsu radiowego, a emisja sygnałów niepożądanych w łączu w dół może być znacznie trudniej wykrywalna przez stację gNodeB. W literaturze przedmiotu można znaleźć wyniki badań eksperymentalnych zarówno w zakresie próby ślepego, bądź siłowego zakłócania interfejsów radiowych, jak i tzw. inteligentnych sygnałów interferujących [3]. Aktualna dostępność narzędzi programistycznych ułatwia przeprowadzanie eksperymentów naukowych, lecz także generowanie sygnałów nieautoryzowanych [4].

W niniejszym artykule przedstawiono wyniki badań numerycznych, których celem było zweryfikowanie stopnia negatywnego wpływu sygnałów interferujących na poprawność działania wybranych elementów toru odbiorczego interfejsu radiowego 5G - New Radio (5G-NR). Skupiono się na numerycznym ujęciu błędów procesu synchronizacji częstotliwości i czasu oraz określeniu bitowej stopy błędów (Bit Error Rate - BER) wyznaczanej podczas dekodowania wiadomości rozgłoszeniowej Master Information Block (MIB). Badania przeprowadzono dla łącza w dół interfejsu radiowego 5G-NR, choć warto zauważyć, iż rodzaje interferencji wskazywane w artykule są tożsame dla interfejsów radiowych Long Term Evolution (LTE) oraz NarrowBand IoT (NB-IoT), po zmianie ich parametrów.

2. RODZAJE ANALIZOWANYCH SYGNAŁÓW INTERFERUJĄCYCH

W niniejszym rozdziale omówiono rodzaje i parametry interferencji wskazywane w literaturze przedmiotu. Przyjęto, że rozważania prowadzone będą w odniesieniu do interfejsu radiowego 5G-NR i jego warstwy fizycznej.

2.1 Rodzaje interferencji własnych

Poziom interferencji własnych jest jednym z głównych parametrów podlegających optymalizacji jednoczęstotliwościowej sieci takich jak 5G. Współdzielenie fragmentu przydzielonego pasma częstotliwości wymaga poprawnego planowania radiowego, a zmienność środowiska

pracy stacji gNodeB wymaga realizacji np. cyklicznych pomiarów parametrów pracy interfejsu radiowego. W interfejsie radiowym 5G-NR zoptymalizowano np. sposób przesyłania symboli referencyjnych analizowanych w procesie estymacji parametrów kanału radiowego, lecz dalej możliwe jest wskazanie przypadków charakterystycznych. Przykładem jest emisja bloku SSB, który to przesyłany jest synchronicznie i periodycznie przez każdą ze stacji gNodeB.

Zakładana synchronizacja stacji gNodeB [5], a zwłaszcza sąsiadujących tzw. głowic inteligentnych, często skutkuje odbieraniem w obszarze miejskim nakładających się z niewielkim przesunięciem czasowym Δt składowych propagacji wielodrogowej. W związku z powyższym dochodzi do interferencji pomiędzy symbolami referencyjnymi Demodulation Reference Signal (DMRS), bądź symbolami przenoszącymi dane kanału rozgłoszeniowego Physical Broadcast Channel (PBCH) z pobliskiego sektora. Jeden z zabiegów minimalizujący niniejsze zjawisko został wskazany w [1], tzn. w sąsiednich sektorach przypisywany jest inny identyfikator komórki co ma na celu minimalizowanie interferencji pomiędzy symbolami DMRS poprzez utworzenie siatki zasobów z odmiennymi punktami alokacji.

Niemniej jednak wspomniana w Rozdziale 1 emisja wielu wiązek w sektorze przez tzw. głowice inteligentne może prowadzić do powstania innego rodzaju interferencji pomiędzy sąsiednimi podsektorami bądź sektorami. Jedną z nich jest możliwość częściowego nakładania się bloków SSB przy nieidealnej synchronizacji czasowej pomiędzy głowicami radiowymi i seryjnej emisji tychże bloków. Zjawisko to będzie prowadziło do pogorszenia jakości detekcji sygnału synchronizacyjnego Primary Synchronization Signal (PSS), jako że jest on przesyłany jako pierwszy z elementów bloku SSB [1].

Powiązanym przykładem interferencji własnych będzie wzajemna, pełna interferencja bloków SSB pomiędzy sąsiednimi sektorami. Pomimo, że każdy z sektorów może mieć przypisany inny identyfikator komórki N_{ID}^{cell} to sygnał PSS może mieć taką samą postać, zgodnie z zasadą jego wytwarzania [1].

2.2 Rodzaje interferencji celowych

Drugą grupą interferencji wskazywanych w publikacjach naukowych są tzw. interferencje celowe. Jednym z przykładów stanowi emitowanie w paśmie pracy interfejsu radiowego szerokopasmowego, filtrowanego szumu gaussowskiego [3]. Niniejszy sposób zakłócania pracy interfejsu może być dostosowany do jego struktury, tj. emisja może zostać ograniczona do czasu trwania bloku SSB, elementów zasobów kanału współdzielonego Physical Data Shared Channel (PDSCH) bądź do podnośnych przypisanych symbolom referencyjnym.

Odmiennym sposobem celowego zakłócania może być emitowanie sygnałów radiowych o parametrach tożsamych z konfiguracją warstwy fizycznej interfejsu radiowego, np. emisja zmodyfikowanych sygnałów synchronizacyjnych. W tym miejscu warto zauważyć, iż realizacja synchronicznego ataku na interfejs 5G-NR, oraz LTE i NB-IoT, jest możliwa bez konieczności odbioru sygnałów tegoż interfejsu ponieważ stacje gNodeB (bądź eNodeB) są synchronizowane czasowo sygnałem Pulse Per Second (PPS) systemu Global Positioning System (GPS) [6].

3. REALIZACJA BADAŃ NUMERYCZNYCH

W niniejszym rozdziale przedstawiono przyjętą metodykę badań numerycznych w zakresie programowego środowiska testowego oraz metodyki badań.

3.1 Środowisko testowe

W celu przeprowadzenia badań numerycznych w środowisku Matlab zaimplementowano fragment toru odbiorczego UE oraz generator syntetycznego sygnału interfejsu radiowego 5G-NR. Przyjęto koncepcję odbiornika definiowalnego programowo, gdzie wejście stanowią próbki sygnału radiowego zaś poszczególne funkcje realizowane są przez moduły programowe. Przyjęto, iż parametry sygnału interfejsu radiowego będą odpowiadać konfiguracji stosowanej przez operatorów komercyjnych w Polsce [2]. Podstawowe parametry zawarto w Tab. 1.

Tabela 1. Podstawowe parametry wytworzonego sygnału interfejsu 5G-NR podczas badań numerycznych

Parametr	*Wartość*
Odstęp pomiędzy podnośnymi	30 kHz
Tryb pracy	TDD
Sposób transmisji bloku SSB	burst
Liczba bloków SSB w ramce radiowej	4

3.2 Metodyka badań i scenariusze badawcze

Mając na względzie przedstawione w rozdziale 2 sygnały interferujące oraz ich parametry przyjęto, iż nacisk zostanie położony na określenie stopnia wpływu interferencji na wybrane elementy kluczowe tj. sygnał synchronizacyjny PSS oraz symbole referencyjne DMRS bloku SSB. Występowanie sygnału interferującego w czasie trwania emisji wspomnianych elementów warstwy fizycznej bezpośrednio oddziałuje na poprawność realizowanego procesu synchronizacji częstotliwości i czasu, a także estymacji i korekcji kanału radiowego podczas próby dekodowania wiadomości rozgłoszeniowej MIB. Z punktu widzenia użytkowego, bezpośrednim skutkiem obserwowanym może być wydłużony czas nawiązywania połączenia UE ze stacją gNodeB bądź całkowita niezdolność wykrycia i podłączenia się do komórki.

W pierwszym scenariuszu badawczym (S1) dokonano zakłócenia sygnału PSS dodając do niego sygnał interferujący w postaci fragmentu czwartego symbolu OFDM (Orthogonal Frequency Division Multiflexing) bloku SSB. Scenariusz ten odpowiada opisanemu w rozdziale 2.1 częściowemu nakładaniu się bloków SSB, tj w zakresie do 500 ns. Tak zmodyfikowany sygnał był następnie przetwarzany przez opracowane moduły programowe realizujące proces synchronizacji częstotliwości i czasu. Każdą z prób przeprowadzono dla pięciu przyjętych wartości stosunku sygnału użytecznego do sygnału interferującego SJR (Signal to Jamming signal Ratio), gdzie stosunek do szumu (Signal to Noise Ratio - SNR) sygnału użytecznego w każdej z prób wynosił 10 dB.

Drugi scenariusz (S2) zakładał podobny sposób zakłócania sygnału PSS, tj. dodanie fragmentu sygnału interferującego, lecz tym razem przyjęto iż interferencja ta

będzie miała charakter celowy. Stąd, do początku sygnału PSS dodawano inny fragment sygnału PSS, również o maksymalnym czasie trwania 500 ns.

Należy zauważyć, iż przyjęty w scenariuszu S1 i S2 czas trwania interferencji sygnału PSS dobrano na podstawie możliwego maksymalnego stopnia niezsynchronizowania głowic radiowych w trybie tożsamym z konfiguracją sieci komercyjnej tj. 260 ns [2], [5]. Jako metrykę oceny poprawności synchronizacji częstotliwości i czasu przyjęto odpowiednio estymatę odchyłki częstotliwości (Carrier Frequency Offset - CFO) oraz błąd czasu detekcji sygnału PSS.

W drugim etapie badań numerycznych podjęto próbę zakłócenia procesu estymacji i korekcji kanału radiowego oraz dekodowania wiadomości rozgłoszeniowej MIB. W przyjętym scenariuszu trzecim (S3) do sygnału łącza w dół, wyłącznie w czasie trwania wybranych elementów zasobów kanału PBCH, tj. symboli DMRS, dodawano szum gaussowski o zmiennej mocy SJR. Stosunek mocy dodawanego szumu zawierał się w zakresie od -15 dB do 5 dB względem mocy symboli DMRS sygnału użytecznego. Ponadto eksperyment powtórzono dla różnych wartości SNR sygnału użytecznego, tj. od 0 dB do 10 dB. Niniejszy scenariusz odpowiada przypadkowi występowania dopasowanej interferencji celowej, bądź interferowania bloków SSB nadawanych z różnych stacji gNodeB, tj. o różnej wartości $mod_4(N_{ID}^{cell})$.

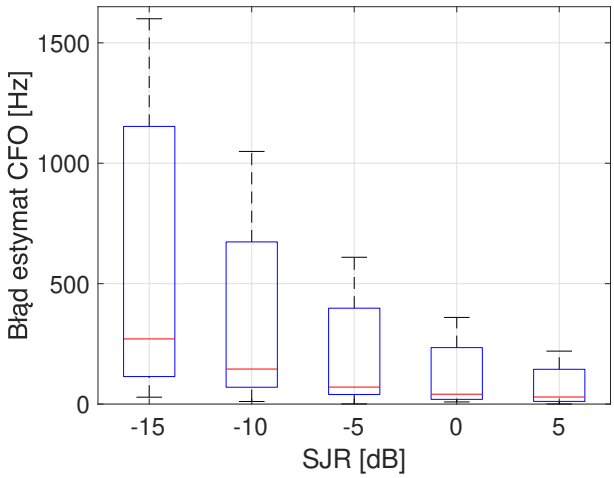
Kolejny scenariusz badawczy (S4) zakładał wprowadzenie do sygnału interferencji w postaci zmiany fazy symboli referencyjnych DMRS w sygnale o wybranych wartościach SNR, tj. od 0 dB do 10 dB. Przypadek ten odpowiada powstaniu interferencji pomiędzy symbolami DMRS kanału PBCH gdy sygnał jest emitowany przez stacje gNodeB o tej samej wartości $mod_4(N_{ID}^{cell})$.

W scenariuszach S3 i S4 sygnał poddany interferencjom doprowadzono do modułu programowego estymacji i korekcji kanału radiowego (oraz modułów realizujących pozostałe operacje przetwarzania zgodnie ze standardem), a następnie do modułu dekodowania polarnego i dekodowania wiadomości MIB. Analizie poddano bity na wejściu (WE) i wyjściu (WY) dekodera polarnego, gdzie jako metrykę oceny działania fragmentu toru odbiorczego przyjęto bitową stopę błędów (BER). Warto również nadmienić, iż w scenariuszach S3 i S4 założono idealną synchronizację częstotliwościową i czasową toru odbiorczego.

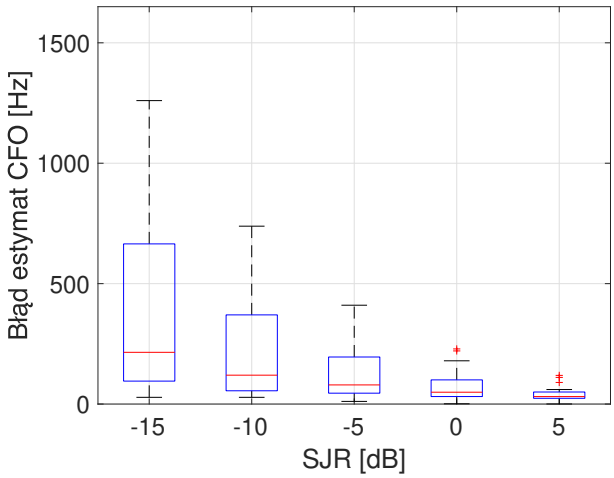
3.3 Analiza uzyskanych wyników

Na Rys. 1 przedstawiono pierwszy zbiór wyników numerycznych dotyczący scenariusza S1. Przedstawione wykresy pudełkowe reprezentują uzyskiwane błędy estymat CFO dla różnego stosunku SJR, w tym wartości minimalne, maksymalne, dla 25/75% przypadków oraz medianę (czerwona linia). Warto zauważyć, iż jedna reprezentacja statystyczna na Rys. 1 dla danego SJR odpowiada analizowanemu zakresowi czasu trwania interferencji tj. do 500 ns. Przyjmując maksymalną dozwoloną odchyłkę częstotliwości równą 360 Hz [7] przy częstotliwości środkowej przydzielonego w Polsce pasma n77 [2] stwierdzić można, iż jedynie dla sygnałów z SJR wyższym od 0 dB możliwe jest spełnienie wymaganego kryterium jakościowego dla 75% przypadków i czasie trwania sygnału interferującego do 500 ns. Dla tychże dwóch przypadków maksymalny błąd czasu detekcji sygnału PSS nie przekroczył 1,37% czasu trwania przedrostka cyklicznego tj. 2.34 μs.

Analiza wyników dla scenariusza S2, Rys. 2, wykazała, że sygnał interferujący w postaci fragmentu sygnału PSS mniej skuteczniej zaburza proces estymacji częstotliwości i czasu względem przypadku S1. Widoczne jest, że dla sygnałów z SJR wyższym od 0 dB możliwe jest spełnienie wymaganego kryterium jakościowego w 100% przypadków, a wartości maksymalne oraz dla 75% są niższe względem scenariusza S1. Dla tychże dwóch przypadków scenariusza S2 również nie przekroczono maksymalnego błędu czasu detekcji sygnału PSS.



Rysunek 1. Błędy estymat CFO w funkcji czasu nałożenia sygnału interferującego na sygnał użyteczny, scenariusz S1.

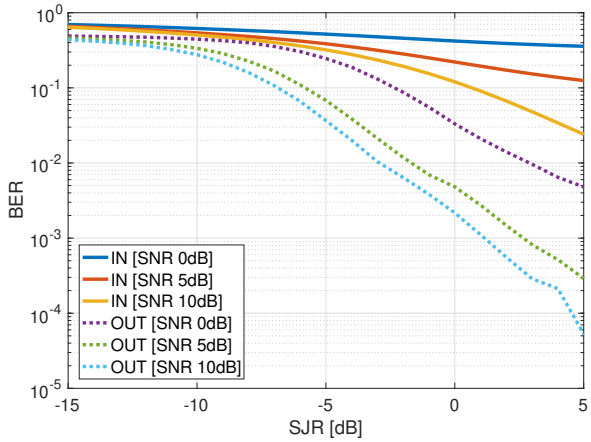


Rysunek 2. Błędy estymat CFO w funkcji czasu nałożenia sygnału interferującego na sygnał użyteczny, scenariusz S2.

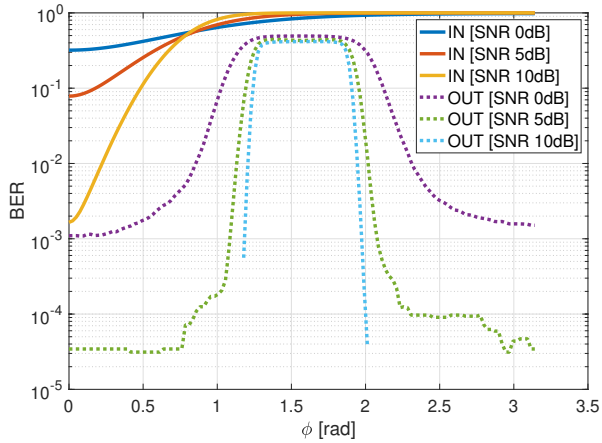
Na Rys. 3 przedstawiono wyniki uzyskane dla scenariusza S3. Zgodnie z oczekiwaniami zastosowane w interfejsie 5G-NR [8] kodowanie polarne znacząco redukuje obserwowaną bitową stopę błędów w trudnych warunkach odbioru, zarówno na wejściu (WE) jak i wyjściu (WY) dekodera. Przekłada się to na wysoką odporność interfejsu na analizowane w S3 interferencje o charakterze szumu.

Dla przypadku S4, Rys. 4, widoczne jest, że sygnał interferujący wywołujący zmianę fazy symboli DMRS znacząco zwiększa stopę błędów, zarówno dokonując ob-

serwacji na wyjściu modułu estymacji i korekcji kanału radiowego (wejście dekodera polarnego - WE) jak i na wyjściu (WY) dekodera. Stan ten jest szczególnie widoczny jeśli zmiana fazy ϕ zawiera się w zakresie od $\frac{\pi}{4}$ do $\frac{3}{4}\pi$, co jest zbieżne ze sposobem wytwarzania symboli DMRS, które to są symbolami modulacji Quadrature Phase Shift Keying (QPSK). Warto zauważyć, że zgodnie z [7] wymagane prawdopodobieństwo niewykrycia bądź błędnej detekcji bloku MIB nie może przekroczyć 1 % przy SNR w zakresie od -10 dB do -5 dB (zależnie od konfiguracji interfejsu), co nie jest możliwe do osiągnięcia przy występowaniu interferencji o wskazanych parametrach.



Rysunek 3. BER w kanale PBCH dla sygnału z wpływem interferencji o zmiennym SJR, scenariusz S3.



Rysunek 4. BER w kanale PBCH przy interferencji wpływającej na fazę symboli DMRS, scenariusz S4.

4. PODSUMOWANIE

Na podstawie przeprowadzonych badań można stwierdzić, iż występowanie w sygnale interfejsu radiowego 5G-NR sygnałów interferujących w sposób znaczący zaburza bądź całkowicie uniemożliwia realizację procesu podłączania się terminala do stacji gNodeB. Występowanie rzeczonych interferencji z punktu widzenia użytkownika objawiać się będzie ponownym podłączaniem się do stacji gNodeB bądź niemożliwością użycia wybranego pasma częstotliwości, co sumarycznie może mieć odzwierciedlenie w obniżeniu jakości świadczonych usług lub braku dostępu do nich. Względem wcześniej przeprowadzonych badań, [3], przedstawione metody przejawiają wyższą skuteczność zwłaszcza mając na uwadze wydajność energetyczną. Na aktualnym etapie prac badawczych autorzy nie mogą numerycznie porównać wpływu zakłócania celowanego na wybrane elementy siatki zasobów względem szerokopasmowego zagłuszania zaporowego [3].

Rzeczone porównania stanowi plan dalszych prac badawczych, gdzie również zweryfikowany zostanie wpływ szerszego zbioru sygnałów interferujących. Ponadto opracowane zostaną metody detekcji interferencji dla łącza w dół m.in. interfejsu radiowego 5G-NR. Zakłada się, że estymowane parametry sygnałów niepożądanych będą przekazywane do programowych modułów przetwarzania sygnału, min. estymacji i korekcji parametrów kanału radiowego, w celu kompensowania ich negatywnego wpływu.

5. FINANSOWANIE

Przedstawione prace badawcze zrealizowano w ramach projektu „Definiowalny programowo układ do wykrywania i kompensowania wpływu zakłóceń wewnątrz i zewnątrz systemowych oraz transmisji niepożądanych w interfejsie radiowym 4G-LTE, 5G-NR" finansowanego ze środków krajowych przez Narodowe Centrum Badań i Rozwoju w ramach programu Lider XIV, umowa nr LIDER14/0116/2023 z dnia 13.02.2024.

LITERATURA

[1] 3GPP - 3rd Generation Partnership Project, 3GPP, *5G-NR, Physical channels and modulation*, Report TS 38.211, V18.2.0., 2024.

[2] Rajchowski, P., „Examination of 5G NR, LTE, and NB-IoT Radio Interfaces and Their Vulnerabilities to Interference", Journal of Telecommunications and Information Technology, 2024, str. 93–100. doi: https://doi.org/10.26636/jtit.2024.4.1960.

[3] Skokowski P. i inni, „Jamming and jamming mitigation for selected 5G military scenarios", Proc. Comput. Sci., vol. 205, str. 258–267, Jan. 2022. doi: 10.1016/j.procs.2022.09.027.

[4] Wu J., i inni, „UE MMSE-IRC Receiver for Suppressing Inter-cell and Inter-user Interference in 5G NR Standardization", IEEE 8th International Conference on Computer and Communications (ICCC), 2022, doi: 10.1109/ICCC56324.2022.10065799.

[5] 3GPP - 3rd Generation Partnership Project, 3GPP, *Base Station (BS) radio transmission and reception*, Report TS 38.104, V18.9.0., 2025.

[6] Magiera J., Rajchowski P., „Celowe zanieczyszczanie pilotów w łączu w górę w interfejsie 5G NR ", Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne, vol. 4, 2024 doi: https://doi.org/10.15199/59.2024.4.42.

[7] 3GPP - 3rd Generation Partnership Project, 3GPP, *User Equipment (UE) radio transmission and reception*, Report TS 38.101, V18.9.0., 2025.

[8] 3GPP - 3rd Generation Partnership Project, 3GPP, *5G-NR, Multiplexing and channel coding*, Report TS 38.212, V17.1.0., 2022.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

ANALIZA WPŁYWU ŚRODOWISKA TESTOWEGO NA SKUTECZNOŚĆ DETEKCJI INTERFERENCJI W INTERFEJSIE RADIOWYM 5G-NR

INFLUENCE OF THE TEST ENVIRONMENT ON THE INTERFERENCE DETECTION PERFORMANCE IN A 5G-NR RADIO INTERFACE

Krzysztof Malon[1]; Paweł Skokowski[1];
Olga Błaszkiewicz[2]; Krzysztof Cwalina[2]; Alicja Olejniczak[2]; Piotr Rajchowski[2]

[1] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa,
{krzysztof.malon, pawel.skokowski}@wat.edu.pl
[2] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk,
{piorajch, kkcwalina, olga.blaszkiewicz, alicja.olejniczak}@pg.edu.pl

DOI: 10.15199/59.2025.4.21

Streszczenie: W artykule przedstawiono wyniki badań symulacyjnych w zakresie detekcji interferencji w interfejsie radiowym 5G-NR w środowisku miejskim. Na drodze badań symulacyjnych analizie poddano parametr SINR (Signal to Noise and Interference Ratio) dla łącza w dół, co umożliwiło wykrycie źródeł sygnału zakłócającego. Implementacja oraz modyfikacja metody detekcji interferencji wykazała, iż wymagane jest np. adaptacyjne wyznaczanie parametrów metody bazując m.in. na modelach środowiska propagacyjnego.

Abstract: In the paper authors present the results of simulation studies on interference detection in the 5G-NR radio interface in an urban environment. The Signal to Noise and Interference Ratio (SINR) parameter determined in downlink transmission was analyzed to detect a source of the interfering signal. The implementation and introduced modification showed that i.e. adaptive selection of methods' parameters is required to adjust its operation according to propagation environment.

Słowa kluczowe: 5G-NR, detekcja, interferencje, jamming

Keywords: 5G-NR, detection, interferences, jamming

1. WSTĘP

Współczesne systemy radiokomunikacyjne o strukturze komórkowej takie jak 4G i 5G projektowane są w celu obsługi dużej liczby użytkowników, którym to świadczona jest przede wszystkim usługa transmisji danych. Ustandaryzowane przez 3rd Generation Partnership Project (3GPP) interfejsy radiowe takie jak 4G - Long Term Evolution (4G-LTE) [1], 5G - New Radio (5G-NR) [2], czy też Narrowband Internet of Things (NB-IoT) [1] stały się powszechne zarówno w sieciach komercyjnych jak i coraz popularniejszych sieciach prywatnych. Rozwój technologii wraz z rosnącym zapotrzebowaniem rynku zaowocował upowszechnieniem się interfejsu radiowego 5G-NR, zwłaszcza w ośrodkach miejskich. Widoczne jest to w liczbie zmodernizowanych stacji bazowych, łączących funkcjonalność eNodeB (eNB) jak i gNodeB (gNB). Ponadto zmiany prawne otworzyły możliwość pracy interfejsu radiowego 5G-NR w nowych pasmach częstotliwości, zwłaszcza części pasma n77 [3], a także widoczna jest zróżnicowana, lecz powtarzalna konfiguracja interfejsów u różnych operatorów sieci komercyjnych [4].

Jednym z obserwowanych negatywnych skutków rzeczonego rozwoju sieci 4G i 5G jest wzrost poziomu interferencji własnych oraz interferencji zewnętrznych. Konieczność współpracy wielu stacji gNodeB, podyktowana zapotrzebowaniem na pojemność sieci, prowadzi do występowania interferencji własnych, co również jest pochodną jednoczęstotliwościowego charakteru sieci 5G. Zauważalna jest zatem potrzeba realizacji prac badawczych w zakresie opracowywania metod do wykrywania obecności, estymowania rodzaju oraz wpływu występujących interferencji.

W literaturze przedmiotu wskazywane jest kilka kluczowych i charakterystycznych rodzajów sygnałów interferujących tj. filtrowany szum typu Additive White Gaussian Noise (AWGN), sygnał zmodulowany np. QPSK czy też sygnał danego bądź innego interfejsu sieci 4G/5G [5, 6, 7, 8, 9, 10, 11]. Dodatkowo przyjmuje się, iż mogą mieć one charakter intencjonalny bądź losowy.

Ponadto, widoczne są prawidłowości w zakresie metod stosowanych podczas wykrywania czy w sygnale analizowanego interfejsu radiowego występuje sygnał interferujący. W [12] wskazano, iż detekcja może odbywać na podstawie analizy skuteczności detekcji wiadomości rozgłoszeniowych (np. Master Information Block - MIB), bądź analizie symboli referencyjnych np. Demodulation Reference Symbols (DMRS) w interfejsie 5G-NR [13]. Osobnym przypadkiem [14] może być podejście bazujące na analizie procesu synchronizacji częstotliwości i czasu tj. detekcji pierwszego i drugiego sygnału synchronizacyjnego (Primary Synchronization Signal - PSS, Secondary Synchronization Signal - SSS). W literaturze wskazywane są także metody bazujące na analizie sygnałów radiowych odebranych przez układy antenowe typu Multiple Input, Multiple Output (MIMO) [15]. Alternatywnym podejściem wydaje się być analiza parametru niezależnego od interfejsu radiowego i estymowanego przez terminal użytkownika (User Equipment - UE), tj. Signal-to-Interference-plus-Noise Ratio (SINR) [16].

W niniejszym artykule autorzy skupili się na analizie skuteczności metody detekcji interferencji wraz z możliwością lokalizowania ich źródła [16], zakładając użycie interfejsu radiowego 5G-NR oraz uwzględniając w badaniach numerycznych profil kanału w środowisku miejskim.

2. METODA WYKRYWANIA INTERFERENCJI

W zaproponowanej metodzie przedstawiono podejście do wykrywania zakłócania dla interfejsu radiowego 5G-NR, które wykorzystuje UE jako rozproszone czujniki do zbierania i analizy otrzymanych raportów. Metoda detekcji zakłócania oparta jest na metryce SINR mierzonej w UE. Do detekcji sygnału zakłócającego wykorzystano znormalizowane raporty pomiarowe z każdego UE dotyczące estymowanej wartości metryki SINR. Rozwiązanie opiera się na wcześniej istniejących standardowych procedurach pomiarowych UE i nie wymaga żadnych modyfikacji oprogramowania. W ramach opisywanej metody, obszar analizy dzielony jest na mniejsze elementy nazywane przedziałami. Oznaczmy $s_j^{(i)}$, gdzie $j \in \{1, \cdots, N_i\}$, jako zebrane pomiary SINR w przedziale i, gdzie liczba pomiarów N_i zależy od konfiguracji raportowania oraz określonego okresu zbierania T. Dla każdego przedziału obliczana jest metryka S_i:

$$S_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbb{I}\left(s_j^{(i)} < \lambda_i\right), \tag{1}$$

gdzie $\mathbb{I}(\cdot)$ oznacza funkcję wskaźnikową, a λ_i to próg SINR dla przedziałów. W rezultacie uzyskiwana jest macierz detekcji $\mathbf{S}_d$, a następnie po zastosowaniu globalnego progu detekcji μ wyznaczane są elementy macierzy detekcji $\mathbf{D}_d$:

$$D_i = \begin{cases} 1 & \text{if } S_i > \mu, \\ 0 & \text{if } S_i \leq \mu. \end{cases} \tag{2}$$

Oprócz detekcji interferencji metoda ta umożliwia także estymowanie położenia urządzenia interferującego. W tym celu obliczana jest macierz lokalizacji $\mathbf{S}_l$ jako rezultat dwuwymiarowej korelacji macierzy $\mathbf{S}_d$ z macierzą kwadratową $\mathbf{K}$. Po zastosowaniu progu decyzyjnego ω generowana jest końcowa macierz $\mathbf{D}_l$ zawierająca wartości wskazujące na obecność lub brak urządzenia interferującego w poszczególnych przedziałach.

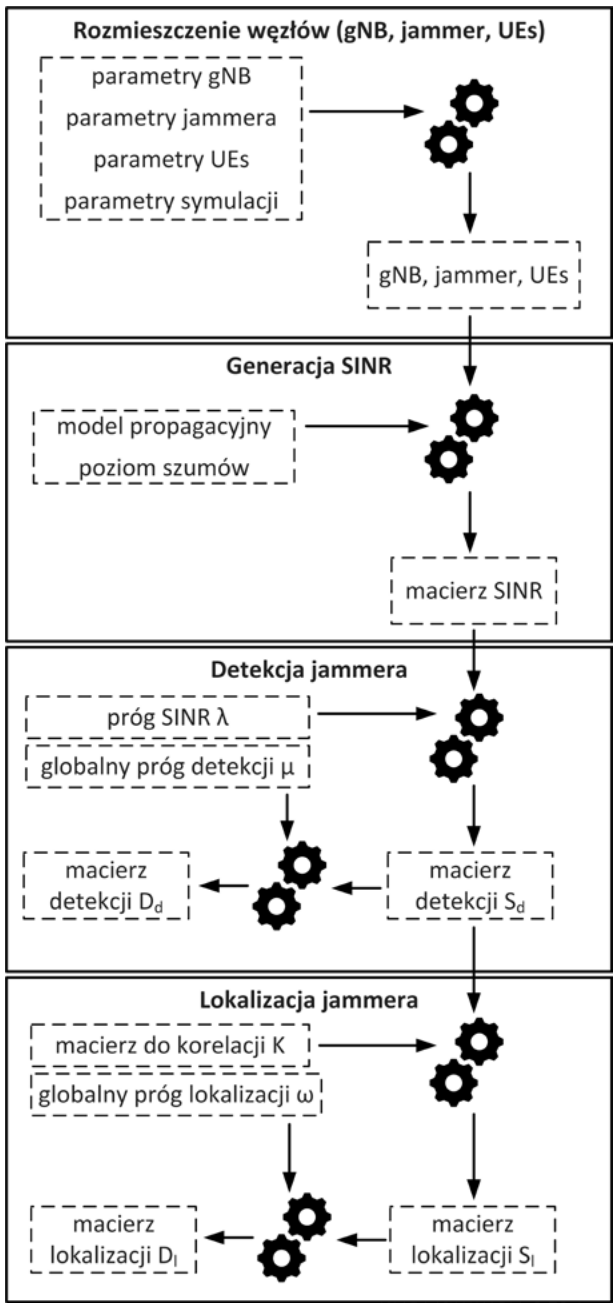
3. IMPLEMENTACJA METODY

Implementacja metody wykrywania interferencji w kanale radiowym wykonana w środowisku MATLAB została podzielona na cztery zasadnicze części związane kolejno z:

- rozmieszczeniem węzłów na wybranym obszarze,
- generacją wartości SINR dla każdego terminala użytkownika,
- wykonaniem algorytmu metody detekcji,
- lokalizacją urządzenia interferującego (ang. jammer).

Każda z wymienionych sekcji posiada możliwość zdefiniowania wartości parametrów oraz może być uruchamiana niezależnie od pozostałych pod warunkiem zachowania w przestrzeni roboczej oprogramowania wymaganych zmiennych. Na Rys.1 zaznaczono dane, które wykorzystywane są w kolejnych częściach oprogramowania.

W ramach pierwszego etapu należy określić obszar symulacji w obrębie dwóch zaimplementowanych lokalizacji, dla których zaimportowano warstwę budynków wykorzystywaną podczas obliczeń propagacyjnych. Kolejnym istotnym elementem jest zdefiniowanie liczby przedziałów oraz terminali użytkowników znajdujących się w każdym z nich. Dla gNB, urządzenia interferującego oraz urządzeń końcowych należy także określić częstotliwość pracy, wysokość zawieszenia i rodzaj anteny, a także moc nadajnika oraz lokalizację. W symulacji zaimplementowano losowe rozmieszczenie UE wewnątrz każdego przedziału z możliwością wskazania czy urządzenia mogą znajdować się na budynkach.



Rysunek 1. Schemat implementacji metody wykrywania interferencji

W kolejnym kroku, zdefiniowane parametry oraz wybrany model propagacyjny i poziom szumów (zależny od pasma sygnału i współczynnika szumów) wykorzystywane są do przeprowadzenia symulacji, której efektem jest

macierz SINR. Dwa ostatnie moduły zawierają implementację metody detekcji i lokalizacji urządzenia interferującego zgodnie z opisem zamieszczonym w poprzedniej sekcji. Stosując wartości progowe λ, μ, ω oraz macierz do korelacji wyznaczane są macierze detekcji i lokalizacji $\mathbf{D}_d$ i $\mathbf{D}_l$.

4. BADANIA SYMULACYJNE

W celu zweryfikowania działania i analizy wybranej metody wykrywania interferencji przeprowadzono badania symulacyjne uwzględniające określony obszar oraz parametry.

4.1 Scenariusz badawczy

Na potrzeby realizacji badań symulacyjnych wybrano obszar kampusu Politechniki Gdańskiej (600 m x 600 m), na którym w centralnym punkcie umiejscowiono stację gNB (Rys.2). Urządzenie interferujące umieszczono na budynku (żółty okrąg na Rys.2), natomiast terminale użytkowników rozmieszczono w sposób losowy. Parametry symulacji zawarto w tabelach (Tabela 1 i Tabela 2). Przyjęte w wariancie 1 wartości parametrów metody zostały dobrane eksperymentalnie na podstawie wyznaczonych wartości SINR (zgodnie z modelem Free space) tak aby urządzenie interferujące lokalizować w testowych, korzystnych warunkach. Następnie w wariancie 2 zachowano wartości parametrów do weryfikacji skuteczności metody w przypadku wykorzystania modelu propagacyjnego Longley-Rice. Konfiguracja wariantu 3 stanowi natomiast próbę wyznaczenia parametrów, które w jak najlepszym stopniu zapewnią właściwe wskazanie lokalizacji źródła emisji sygnału interferującego.

Tabela 1. Podstawowe parametry symulacji

Parametr	*Wartość*
Liczba przedziałów	100
Liczba UE w przedziale	6
UE na budynkach	Nie
Wysokość zawieszenia anteny gNB	15 m
Moc nadajnika gNB	20 W
Wysokość zawieszenia anteny jammera	5 m
Moc nadajnika jammera	1 W
Wysokość zawieszenia anteny UE	1,5 m
Częstotliwość pracy	3,5 GHz
Pasmo sygnału	50 MHz
Współczynnik szumów	7 dB
Macierz do korelacji	[-1 -1 -1; -1 8 -1; -1 -1 -1]

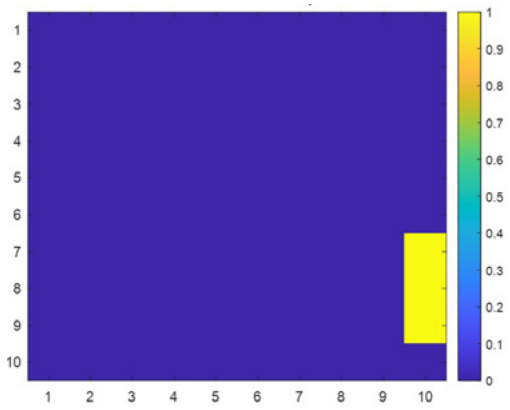
Rysunek 2. Rozmieszczenie UEs, gNB oraz jammera na mapie dla badanej lokalizacji

Tabela 2. Parametry wariantów symulacji

Wariant	*Model propagacyjny*	λ	μ	ω
1	Free space	0	0,3	1
2	Longley-Rice	0	0,3	1
3	Longley-Rice	-50	0,3	4

4.2 Rezultaty i analiza wyników

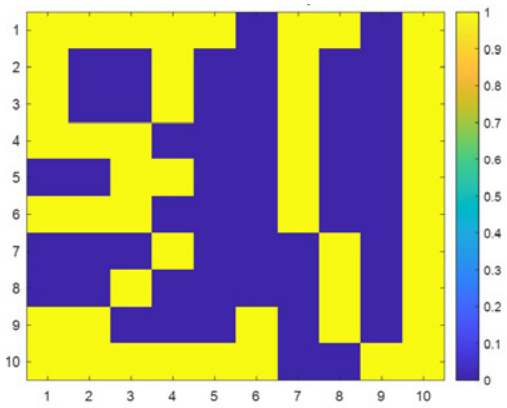
Rezultaty przeprowadzonych badań zobrazowano w postaci graficznej reprezentującej macierze lokalizacji $\mathbf{D}_l$ zawierające wartości wskazujące na estymowaną obecność (wartość 1, kolor żółty) lub brak (wartość 0, kolor niebieski) urządzenia interferującego w poszczególnych przedziałach (Rys.3 - Rys.5).



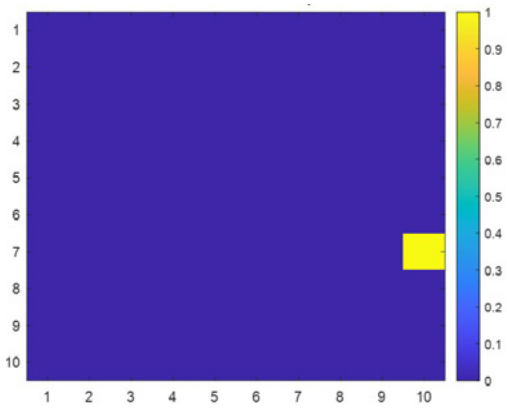
Rysunek 3. Macierz lokalizacji $\mathbf{D}_l$ dla wariantu 1

Przedstawione wyniki potwierdzają istotny wpływ doboru właściwych wartości parametrów dla metody detekcji i lokalizacji urządzenia interferującego. W przypadku wariantu 2 zastosowanie modelu propagacyjnego uwzględniającego budynki (Longley-Rice) prowadzi do powstania wielu błędnych decyzji, które można sklasyfikować jako fałszywe alarmy. Algorytm zwraca decyzję o występowaniu jammera w miejscach (przedziałach), gdzie

następuje spadek wartości poziomu sygnału skutkujący pogorszeniem metryki SINR. Po zmianie wartości parametrów w wariancie 3 uzyskano wyniki, które w dużo większym stopniu reprezentują symulowaną sytuację.



Rysunek 4. Macierz lokalizacji $\mathbf{D}_l$ *dla wariantu 2*



Rysunek 5. Macierz lokalizacji $\mathbf{D}_l$ *dla wariantu 3*

Prowadzi to do wniosku, że jednym z kierunków dalszych prac nad tą metodą mógłby być automatyczny dobór wartości parametrów związanych z progami detekcji. Należy jednak zwrócić uwagę na fakt, iż zaprezentowana zmiana wartości tych parametrów powoduje również zmniejszenie czułości algorytmu na wykrycie urządzeń interferujących. Kolejną propozycją modyfikacji przedstawionej metody jest jej uzupełnienie o wykonanie pomiarów Signal-to-Noise Ratio (SNR) lub wstępne symulacje z estymacją tej metryki na danym obszarze, w celu wyznaczenia wartości referencyjnych w określonych lokalizacjach w sytuacji braku sygnału interferującego. W efekcie można uzyskać mapę referencyjną uwzględnianą w kolejnych etapach algorytmu co powinno znacząco zmniejszyć prawdopodobieństwo występowania fałszywych alarmów spowodowanych spadkami poziomu sygnału, bez pogarszania czułości metody.

5. PODSUMOWANIE

Podsumowując można stwierdzić, iż zaprezentowana metoda wykrywania i lokalizowania źródła interferencji charakteryzuje się następującymi właściwościami:

- elastyczność i uniwersalność ze względu na możliwość zastosowania w wielu systemach radiokomunikacyjnych, w których estymowany jest SINR,
- możliwość sterowania parametrami algorytmu,
- niska złożoność obliczeniowa,
- zadowalające wyniki w sytuacji otwartej przestrzeni, gdzie nie występują przeszkody na ścieżkach propagacji sygnału,
- wrażliwość na rzeczywiste zjawiska propagacji fali radiowej m.in. tłumienie fal radiowych, szczególnie w środowiskach zurbanizowanych gdzie spadek wartości poziomu sygnału użytecznego powoduje zmniejszenie SINR co jest błędnie interpretowane przez algorytm jako potencjalne wystąpienie źródła interferencji,
- możliwości rozszerzenia i modyfikacji metody, np. automatyczny dobór parametrów lub generacja referencyjnej mapy SNR.

W ramach prac opracowano narzędzie aplikacyjne do badań metody wykrywania interferencji. Pozwala ono na symulację i sprawdzenie wpływu wielu parametrów oddziałujących na skuteczność metody, w tym modeli propagacyjnych czy elementów fizycznych (jak rodzaj anteny oraz jej charakterystyka, moc nadajnika, wysokość rozmieszczenia elementów) mających wpływ na otrzymywane rezultaty. Implementacja metody detekcji i lokalizacji może zostać również wykorzystania do analizy danych z zarejestrowanych rzeczywistych wartości metryki SINR na przykład z wykonanych kampanii pomiarowych. Należy zwrócić uwagę na fakt, że badania metody i implementacja zostały rozszerzone o model 3D w porównaniu z wcześniejszymi pracami. Niezaprzeczalną zaletą stworzonego oprogramowania jest wykorzystanie modelu deterministycznego (Longley-Rice'a) w porównaniu do modelu statystycznego przyjętego w artykule bazowym.

Przedstawione w artykule rezultaty stanowią wstępną weryfikację prezentowanej metody. W ramach kierunków dalszych prac można wskazać opracowanie algorytmów automatycznego doboru parametrów modelu oraz zastosowanie referencyjnej mapy wartości SNR. Niewątpliwie, planowane przeprowadzenie rzeczywistych pomiarów wartości SINR, stanowiących dane wejściowe do prezentowanej metody, umożliwi finalną weryfikację wyników symulacyjnych. Podczas kolejnych etapów prac przewidywane są również badania porównawcze z innymi metodami wykrywania interferencji w interfejsie radiowym.

6. FINANSOWANIE

Przedstawione prace badawcze zrealizowano w ramach projektu „Definiowalny programowo układ do wykrywania i kompensowania wpływu zakłóceń wewnątrz i zewnątrz systemowych oraz transmisji niepożądanych w interfejsie radiowym 4G-LTE, 5G-NR" finansowanego ze środków krajowych przez Narodowe Centrum Badań i Rozwoju w ramach programu Lider XIV, umowa nr LIDER14/0116/2023 z dnia 13.02.2024.

LITERATURA

[1] 3GPP - 3rd Generation Partnership Project, 3GPP, "LTE, Evolved Universal Terrestrial Radio Access (E-UTRA, Physical channels and modulation", TS 36.211, V13.4.0., 2017.

[2] 3GPP - 3rd Generation Partnership Project, 3GPP, *5G-NR, Physical channels and modulation*, Report TS 38.211, V18.2.0., 2024.

[3] SI2PEM, Mapa pola elektromagnetycznego, Online: https://si2pem.gov.pl, Data dostępu: 20.04.2025.

[4] Rajchowski, P., „Examination of 5G NR, LTE, and NB-IoT Radio Interfaces and Their Vulnerabilities to Interference", Journal of Telecommunications and Information Technology, 2024, str. 93–100. doi: https://doi.org/10.26636/jtit.2024.4.1960.

[5] Philippe G, Montaigne F., Schiel J.C., Georgeaux E., Gruet C., Roy P.Y., „LTE resistance to jamming capability", 2013 Military Communications and Information Systems Conference, Saint-Malo, France, 2014.

[6] Xu S., Xin J., Xiong S., Sun Z., „Performance Analysis of CRS Interference Mitigation Algorithm in LTE and NR Coexistence Scenario", 2nd Information Communication Technologies Conference (ICTC), str. 184–188, Nanjing, China, 2021, doi: https://doi.org/10.1109/ICTC51749.2021.9441653.

[7] Krenz R., i Brahma S., „Jamming LTE Signals", IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom), str. 72–76, Constanta, Romania, 2015, doi: https://doi.org/10.1109/BlackSeaCom.2015.7185089.

[8] Skokowski P. et al.„Jamming and jamming mitigation for selected 5G military scenarios", Proc. Comput. Sci., vol. 205, str. 258–267, Jan. 2022. doi: 10.1016/j.procs.2022.09.027.

[9] Mazuro M., Skokowski P., Kelner J.K. „Comparison of Selected Estimation Methods of Interference Level in Radio Channel", 2024 International Radar Symposium (IRS), Wroclaw, Poland, 2024, str. 150-152.

[10] Skokowski P. et al., „Practical Trial for Low-Energy Effective Jamming on Private Networks With 5G-NR and NB-IoT Radio Interfaces", in IEEE Access, vol. 12, str. 51523-51535, 2024. doi: 10.1109/ACCESS.2024.3385630.

[11] Skokowski P., Dułowicz J. „Autonomous Mobile System for Detecting and Jamming Cellular Network Signals Using a Software Defined Radio Integrated into a UAV Platform", 2023 Signal Processing Symposium (SPSympo), Karpacz, Poland, 2023, str. 149-151. doi: 10.23919/SPSympo57300.2023.10302702.

[12] Morillo G., Roedig U., Pesch D., „Detecting Targeted Interference in NB-IoT". 19th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT), Pafos, Cyprus, 2023, doi: https://doi.org/10.1109/DCOSS-IoT58021.2023.00080.

[13] Sun Y., Jiaqi K., Wenshu S., Shunqing Z., Xiaojing C., Nan D., „Semi-Blind Multi-Cell Interference Detection and Cancellation in 5G Uplink OFDM Systems", International Wireless Communications and Mobile Computing (IWCMC), str. 167–171, Dubrovnik, Croatia, 2022, doi: https://doi.org/10.1109/IWCMC55113.2022.9824834.

[14] Cai B., Weiliang X., Han G., „Analysis and Field Trial on Interference Coexistence of 5G NR and 4G LTE Dynamic Spectrum Sharing", International Wireless Communications and Mobile Computing (IWCMC), str. 1281–1285. Harbin City, China, 2021, doi: https://doi.org/10.1109/IWCMC51323.2021.9498772.

[15] Wesołowski K., "A Simple Algorithm for Jamming Detection in OFDM Systems", 2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring), 2023, doi: https://doi.org/10.1109/VTC2023-Spring57618.2023.10200416.

[16] Forssell H., Mungara R. K., Ferrante G. C., Tullberg H., „UE-Assisted Jamming Detection in 5G NR", IEEE International Conference on Communications (ICC 2023), 2023, str. 5216–5220. doi: 10.1109/ICC45041.2023.10279513.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

Budowa bliźniaka cyfrowego środowiska radiowego z użyciem metody śledzenia promieni

Building a Digital Twin of a radio environment using a Ray Tracing method

Marcin Pakuła[1]; Marcin Hoffmann[1]; Paweł Kryszkiewicz[1]

[1] Instytut Radiokomunikacji, Politechnika Poznańska, Poznań,
marcin.pakula@student.put.poznan.pl,marcin.hoffmann@put.poznan.pl
pawel.kryszkiewicz@put.poznan.pl

DOI: 10.15199/59.2025.4.22

Streszczenie: W pracy przedstawiono budowę cyfrowego bliźniaka środowiska radiowego przy użyciu metody śledzenia promieni (ang. ray tracing). Zaprezentowano jak uwzględnić rzeczywisty układ budynków i materiałów z którego są zbudowane. Zaprezentowano wyniki badań dla różnych konfiguracji obliczeń propagacyjnych. Przeanalizowano ich wpływ na wybrane metryki radiowe, a także miary właściwe dla systemów wieloantenowych. Otrzymane wyniki pozwalają ocenić, jak bardzo typowe modele odbiegają od zbliżonych do rzeczywistych warunków propagacyjnych, a także które zjawiska propagacyjne mają największy wpływ na złożoność obliczeniową i jakość kanału radiowego.

Abstract: The paper presents the construction of a digital twin of the radio environment using the ray tracing method. It shows how to take into account the real layout of buildings and the materials they are made of. It presents the results of research for different configurations of the propagation model. Their impact on selected radio metrics, as well as measures specific to multi-antenna systems, is analyzed. The obtained results allow us to assess how much typical models deviate from propagation conditions close to reality, and which propagation phenomena have the greatest impact on computational complexity and radio channel quality. [1]

Słowa kluczowe: Cyfrowy Bliźniak, Technika śledzenia promieni, Systemy wieloantenowe

Keywords: Digital Twin, MIMO, Ray-Tracing

1. WSTĘP

Wraz z rozwojem kolejnych generacji systemów radiokomunikacyjnych rośnie zapotrzebowanie na ilość danych przesyłanych drogą radiową. Jest to naturalny trend, który jednak stawia poważne wyzwanie przed operatorami, w zakresie odpowiedniego planowania oraz konfiguracji sieci, pozwalającego na znaczną poprawę jakości usług użytkowników końcowych. Dla dokładnego planowania sieci potrzebny jest odpowiedni model systemu radiokomunikacyjnego, który pozwoli na przetestowanie wielu możliwych konfiguracji sieci i optymalizację jej parametrów pod kątem różnych funkcji celu. Model spełniający te założenia możemy nazwać Cyfrowym Bliźniakiem (ang. *Digital Twin*). Nazwa ta nawiązuje do przeniesienia w świat cyfrowy danego sytemu lub obiektu - w tym przypadku systemu radiokomunikacyjnego - aby analizować jego zachowanie w określonych warunkach. Dotychczas w literaturze [1] spotykano głównie modele, które skupiają się na odwzorowaniu statystycznym spadku mocy pomiędzy stacją bazową, a użytkownikiem. Pomimo, że jest to ważny parametr opisujący kanał, w nowoczesnych systemach, w których używa się wielu anten, konieczne jest uwzględnienie innych zjawisk w kanale takich jak wielodrogowość czy korelacja przestrzenna między antenami [11].

Wraz ze wzrostem mocy obliczeniowej komputerów coraz większe znaczenie zyskuje technika śledzenia promieni (ang. ray tracing), która śledzi poszczególne ścieżki propagacyjne fal elektromagnetycznych pomiędzy anteną nadawczą a odbiorczą [2, 3]. Plusem takiego podejścia jest możliwość zdefiniowania środowiska o odpowiednim dla naszego przypadku poziomem szczegółowości którego chcemy analizować. Dodatkowo parametry symulacji również mogą być modyfikowane co wpływa bezpośrednio na sposób propagacji promieni pomiędzy nadajnikiem, a odbiornikiem. Do kluczowych ustawień należą: liczba odbić, liczba dyfrakcji oraz uwzględnienie bądź nie rozpraszania dyfuzyjnego (ang. diffuse scattering).

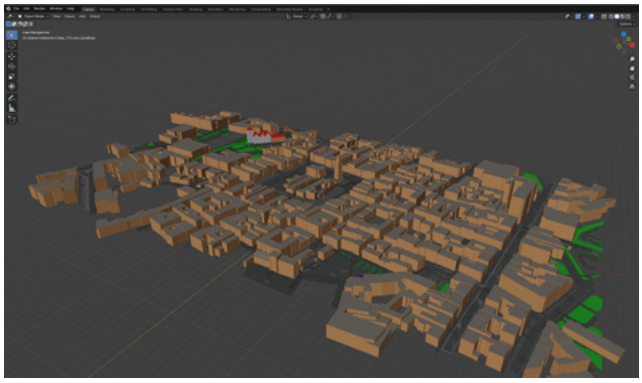
W pracy zaprezentowano budowę cyfrowego bliźniaka z użyciem oprogramowania firmy Remcom - Wireless InSite oraz przedstawiono, które parametry powinny zostać użyte w symulacjach. W wynikach przedstawiono moc sygnału odbieranego, współczynnik K-Rice'a oraz rangę kanału - metrykę używaną dla systemów wieloantenowych. W rozdziale II przedstawiono budowę cyfrowego bliźniaka. W rozdziale III wytłumaczono poszczególne metryki opisujące kanał radiowy. W rozdziale IV podsumowano wyniki dla różnych parametrów cyfrowego bliźniaka. Podsumowanie umieszczono w rozdziale 5.

2. BUDOWA BLIŹNIAKA CYFROWEGO

W celu wiernego odwzorowania środowiska rzeczywistego wykorzystano dane przestrzenne pochodzące z serwisu OpenStreetMap [9], obejmujące wybrany przez

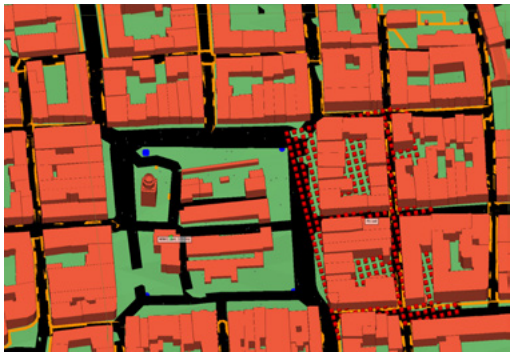
[1]Praca powstała w ramach projektu OPUS finansowanego przez Narodowe Centrum Nauki nr 2021/41/B/ST7/00136.

użytkownika obszar. Następnie dane te zostały przekształcone w formacie kompatybilnym z oprogramowaniem Blender [8], a konkretnie zapisane jako pliki z rozszerzeniem „*.obj". Rysunek 1 przedstawia zrzut ekranu z aplikacji Blender z zaimportowaną mapą pochodzącą z OpenStreetMap. Model zawiera trójwymiarowe reprezentacje budynków, ulic, ścieżek rowerowych, torów tramwajowych, zbiorników wodnych oraz terenów zielonych. Wszystkie obiekty zostały osadzone na warstwie topograficznej odwzorowującej ukształtowanie terenu.



Rysunek 1: Mapa centrum Poznania w oprogramowaniu Blender

Głównym komponentem Cyfrowego Bliźniaka jest program Wireless InSite, do którego zaimportowano mapę przedstawioną na Rys. 2. Wszystkie elementy, takie jak budynki, ulice, drogi i tory tramwajowe, zostały dopasowane do warstwy terenu odwzorowującej rzeczywistą topografię, zachowując ich rzeczywiste wzajemne położenie.



Rysunek 2: Mapa centrum Poznania w oprogramowaniu Wireless InSite

Do każdego materiału użytego w scenariuszu przypisywane są parametry elektryczne na podstawie rekomendacji ITU [4]. W tak przygotowanym środowisku zdefiniowano parametry symulacyjne. Wyznaczono obszar symulacyjny, czyli region, w którym program oblicza interakcje promieni z otoczeniem. Do obliczeń wykorzystano model propagacyjny X3D Ray Model, który jest przystosowany do symulacji z użyciem anten MIMO.

Istotnym aspektem przy tworzeniu Bliźniaka Cyfrowego jest kompromis pomiędzy dokładnością symulacji, a jej złożonością obliczeniową. Kluczowy wpływ mają na to parametry takie jak maksymalna liczba odbić oraz dyfrakcji, które mogą wystąpić pomiędzy nadajnikiem, a odbiornikiem. Wraz ze wzrostem tych wartości, obliczanych jest więcej ścieżek propagacyjnych, co z jednej strony zwiększa realizm modelu, ale z drugiej – znacząco wydłuża czas symulacji. Może się jednak okazać, że moc dodatkowych ścieżek będzie miała pomijalny wpływ na jakość modelowania kanału.

Program Wireless InSite umożliwia również uwzględnienie rozpraszania dyfuzyjnego [7] dla wybranych materiałów. Zjawisko to polega na oddziaływaniu fali z materią, która powoduje zmianę kierunku rozchodzenia się fali. Na to zjawisko ma wpływ nieregularność powierzchni odbijającej falę takich jak tynk czy cegła. Uwzględnienie tego zjawiska pozwala uzyskać bardziej realistyczne wyniki, jednak wiąże się z dodatkowym obciążeniem obliczeniowym [12], wynikającym z większej liczby analizowanych promieni.

3. MIARY JAKOŚCI WYGENEROWANEGO KANAŁU RADIOWEGO

W symulacjach opartych na śledzeniu promieni uzyskuje się informacje o promieniach docierających z każdej anteny nadawczej do odbiorczej, różniących się mocą, opóźnieniem i fazą — zależnie od liczby oraz rodzaju odbić i dyfrakcji. Zwiększenie ich maksymalnej liczby prowadzi do bogatszego kanału wielodrogowego (ang. *richmultipath*).

W systemach MIMO kluczowe jest, czy możliwe jest przesyłanie danych wieloma niezależnymi ścieżkami. Dzięki temu wielodrogowość (np. odbicia od budynków) może być wykorzystana korzystnie, zwiększając dywersyfikację przestrzenną przez prekodowanie. Zjawisko to, znane jako channel hardening w Massive MIMO [10], pozwala zwiększyć przepustowość i ograniczyć wpływ zaników sygnału.

Jedną z metryk opisującą "różnorodność" kanału jest współczynnik K dla kanału Rice [5], obliczany jako stosunek mocy najsilniejszej ścieżki dochodzącej do danego użytkownika do sumy mocy pozostałych ścieżek

$$K = \frac{\max\limits_{i} P_i}{\sum\limits_{i=1}^{N} P_i - \max\limits_{i} P_i}, \tag{1}$$

gdzie P_i to moc i-tej scieżki z N docierających do odbiornika. Większy współczynnik K odpowiada większej mocy w ścieżce dominującej - co odpowiada rzadszemu kanałowi, natomiast im mniejszy współczynnik K tym więcej mocy jest przekazywane przez ścieżki odbite. W idealnym kanale widoczności między antenami(ang. $LineOfSight$) wartość K powinna wynosić nieskończoność ze względu na tylko jedną ścieżkę, natomiast w idealnym kanale Rayleigha ze względu na nieskończoną liczbę ścieżek o równych mocach średnich powinna dążyć do zera.

Ranga kanału (ang. *ChannelRank*) jest metryką określającą, ile niezależnych strumieni transmisyjnych można przesłać przez dany kanał. Z matematycznego punktu widzenia odpowiada liczbie istotnych wartości własnych macierzy korelacji kanału. W literaturze [6, 5] często stosuje się podejście, w którym wszystkie wartości własne porównywane są do największej z nich, a za istotne uznaje się te, które różnią się od niej o mniej niż 20 dB. Pozostałe wartości są pomijane. Im większa wartość rangi kanału, tym większą ilością kanałów niezależnych możemy przesyłać sygnał. Gdy pomiędzy anteną nadawczą a odbiorczą nie występują przeszkody (kanał LOS), ranga

kanału wynosi 1. W przypadku idealnego kanału Rayleigha ranga może osiągnąć wartość równą maksimum z liczby anten nadawczych i odbiorczych.

Użyty Ray Tracer zwraca dla k-tej anteny (z K dotępnych) zbiór N_k ścieżek docierających, charakteryzowanych dla i-tej ścieżki mocą $P_{i,k}$, przesunięciem w fazie $\phi_{i,k}$, a także opóźnieniem propagacyjnym $\tau_{i,k}$. Można zatem wyznaczyć współczynnik kanału na k-tej antenie dla n-tej podnośnej jako

$$h_{n,k} = \sum_{i=1}^{N} \sqrt{P_{i,k}} \cdot e^{j\phi_{i,k}} \cdot e^{-j2\pi f_n \cdot \tau_{i,k}} \qquad (2)$$

Moc ścieżki i czas przyjścia zostały wcześniej znormalizowane odpowiednio względem mocy anteny nadawczej oraz pierwszej ścieżki, docierającej do odbiornika. W taki sposób zebrano współczynniki kanału dla wszystkich odbiorników. Można wyznaczyć uśrednioną po N_{RB} częstotliwościach (podnośnych/blokach zasobów) macierz autokorelacji jako

$$R_{k1,k2} = \frac{1}{N_{\mathrm{RB}}} \sum_{n=1}^{N_{\mathrm{RB}}} h_{n,k1} h^*_{n,k2}, \qquad (3)$$

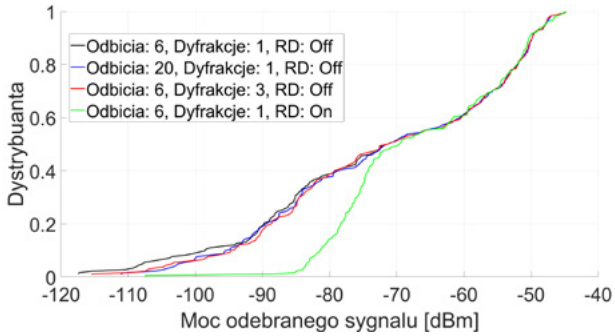
gdzie * oznacza sprzężenie liczby zespolonej. Z uśrednionej macierzy korelacji przestrzennych w danym odbiorniku, obliczono następnie wartości własne macierzy. Te wartości służą do wyznaczenia rangi kanału jak napisano wyżej. W ten sposób w każdej lokalizacji odbiornika uzyskujemy informację o ilości niezależnych kanałów, które mogę nam posłużyć do równoległego przesyłu sygnału.

4. WYNIKI

W celu uzyskania wyników symulacyjnych na mapie rozlokowano nadajnik oraz zestaw odbiorników. Nadajnik wyposażono w antenę MIMO 8×4, zbudowaną z dipoli półfalowych o poziomej polaryzacji, emitującą sygnał o łącznej mocy 30 dBm na nośnej 3,5 GHz. Szyk antenowy został umieszczony na jednym z budynków, na wysokości 3 metrów ponad jego dachem, co odpowiada łącznej wysokości 20 metrów nad poziomem gruntu.

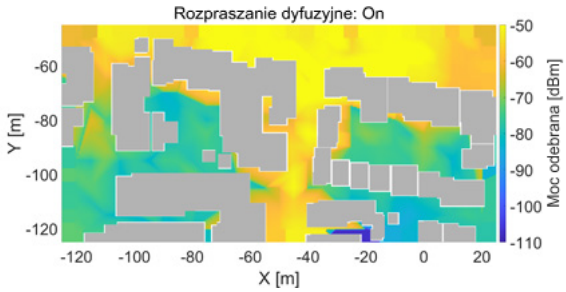
Odbiorniki rozmieszczono w równomiernej siatce kartezjańskiej, z krokiem co 5 metrów. Zaznaczono je czerwonymi kropkami na Rys. 2. Przyjęto założenie, że pozycje odbiorników znajdują się wyłącznie poza budynkami — w przypadku, gdy punkt siatki wypada wewnątrz budynku, zostaje pominięty. Każdy z 182 odbiorników wyposażono w antenę dookólną, która jest ustawiona na wysokości 0.6 m nad powierzchnią gruntu.

Na rysunku 3 przedstawiono dystrybuantę mocy sygnału odbieranego w wyznaczonych punktach odbiorczych. Moc sygnału dla około połowy odbiorników o najlepszych warunkach propagacyjnych nie zmienia się istotnie przy zmianie parametrów metody śledzenia promieni. Jest to spowodowane prawdopodobnie bezpośrednią widocznością pomiędzy anteną nadawczą, a odbiorczą, co można wywnioskować z wyższych wartości mocy odbieranej. Dla wartości poniżej -70 dBm widzimy znaczącą różnicę między symulacją z włączonym rozpraszaniem dyfuzyjnym (RD: ON), a pozostałymi symulacjami z wyłączonym rozpraszaniem. Dodanie 20 odbić oraz dodanie 3 dyfrakcji zwiększają moc odebraną o kilka decybeli w kilku punktach w stosunku do symulacji z 6 odbciami i 1 dyfrakcją. Dodanie rozpraszania dyfuzyjnego poprawia jakość kanału nawet o 20 dB.

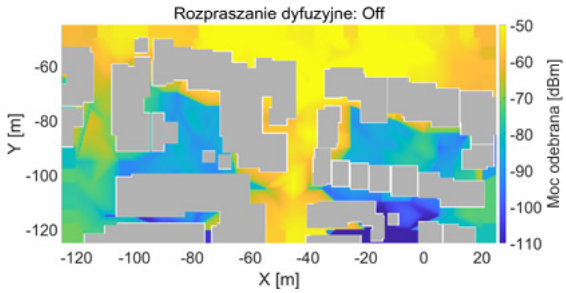


Rysunek 3: Dystrybuanta mocy sygnału odebranego

Na rysunkach 4 i 5 pokazano mapę zsumowanej mocy odebranego sygnału dla wszystkich anten nadawczych. Dla przejrzystości wyników W miejscach gdzie nie ma odbiorników zastosowano interpolacje liniową.



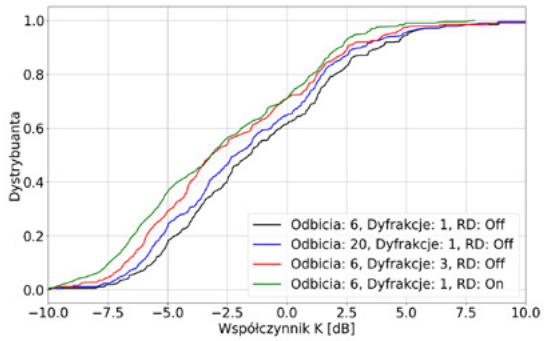
Rysunek 4: Mapa mocy odebranego sygnału dla symulacji bez rozpraszania dyfuzyjnego



Rysunek 5: Mapa mocy odebranego sygnału dla symulacji z rozpraszaniem dyfuzyjnym

Na rysunku 4 widzimy symulację uwzględniającą 6 odbić oraz 1 dyfrakcję, bez rozpraszania dyfuzyjnego, a na rysunku 5 z uwzględnieniem tego rozpraszania. W oprogramowaniu Wireless InSite istnieje możliwość określenia, które materiały mają uwzględniać rozpraszanie dyfuzyjne, a także całkowitego wyłączenia tego zjawiska. W przeprowadzonych symulacjach rozpraszanie dyfuzyjne uwzględniono jedynie dla powierzchni budynków. W obszarach odsłoniętych nie widać znaczącej różnicy mocy docierającej. Jednak dla miejsc między budynkami możemy zauważyć wzmocnienie sygnału w przypadku uwzględnienia rozproszenia dyfuzyjnego, nawet do kilkunastu dB. Oznacza to, że więcej promieni dotarło do tych punktów oraz że ich moc była na tyle duża, że realnie wpływa na moc odebraną. Do analizy niezależności kanałów posłużyliśmy się współczynnikiem K-Rice'a, którego rozkład dla różnej

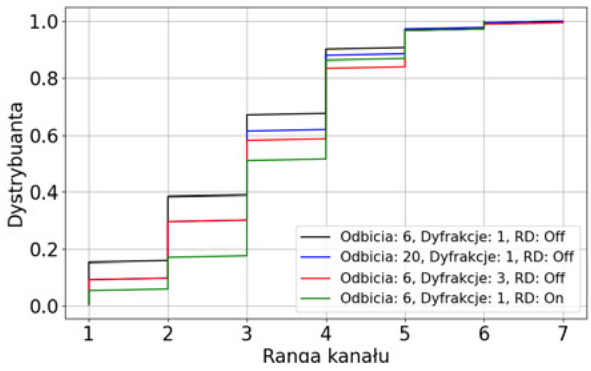
liczby odbić, dyfrakcji i rozproszenia dyfuzyjnego porównano na Rys. 6.



Rysunek 6: Dystrybuanta współczynnika K-Rice'a dla różnych parametrów symulacji

Dodanie większej liczby odbić w symulacji zwiększyło liczbę promieni dochodzących do odbiornika, a co za tym idzie przesunęło całą krzywą w lewą stronę - czyli wartość współczynnika K się zmniejszyła dla wszystkich odbiorników. Warto zauważyć, że zwiększenie z 1 do 3 dyfrakcji ma większy wpływ na współczynnik K niż zwiększenie z 6 do 20 odbić, co pokazuje większy wpływ dyfrakcji na wyniki. Względem symulacji wyjściowej - z 1 dyfrakcją i 6 odbiciami - największa zmiana zaszła po włączeniu rozpraszania dyfuzyjnego. Dla większości punktów, różnica ta nie przekracza jednak 2 dB.

Ze względu na zastosowanie anteny MIMO, istotną miarą jakości kanału jest jego ranga, która odzwierciedla liczbę niezależnych strumieni danych, jakie można jednocześnie przesłać do konkretnego użytkownika. Na rysunku 7 przedstawiono wykres dystrybuanty rangi kanału wyznaczonej w punktach odbiorczych.



Rysunek 7: Dystrybuanta rangi kanału w punktach odbiorczych

Zgodnie z oczekiwaniami, wzrost liczby odbić oraz uwzględnienie dyfrakcji skutkują zwiększeniem rangi kanału — im więcej zjawisk propagacyjnych, tym większa liczba niezależnych ścieżek transmisji. Szczególnie widoczny jest wpływ rozpraszania dyfuzyjnego, które prowadzi do najwyższych wartości rangi w większości punktów pomiarowych. Warto zauważyć, że te wartości są znacząco niższe niż teoretyczna maksymalna ranga, jaka mogłaby zostać osiągnięta w idealnym kanale Rayleigha z 32 antenami nadawczymi, gdzie ranga kanału wynosiłaby 32.

Wyniki wskazują, że rozpraszanie dyfuzyjne ma większy wpływ na wyniki niż zwiększenie maksymalnych interakcji promieni z otoczeniem. Warto zauważyć, że włączenie rozpraszania dyfuzyjnego zwiększa liczbę ścieżek docierających do odbiornika, co pozytywnie wpływa na realizm modelowanego kanału. Jednocześnie jednak znacząco wydłuża czas trwania symulacji — nawet kilkunastokrotnie.

5. PODSUMOWANIE

W pracy przedstawiono budowę cyfrowego bliźniaka dla środowiska radiowego z użyciem metody śledzenia promieni zaimplementowanej w oprogramowaniu Wireless InSite. Następnie, porównano wyniki otrzymane dla różnych parametrów symulacyjnych w środowisku miejskim bazującym na rzeczywistej topografii pobranej z serwisu Open Street Maps. Przedstawione wyniki pokazują większy wpływ rozpraszania dyfuzyjnego niż zwiększenie liczby odbić i dyfrakcji. W przypadku uwzględnienia rozproszenia dyfuzyjnego moc odebrana zwiększa się nawet o 15 dB.

LITERATURA

[1] Al-Samman, A. M. et al. 2016. "Statistical Modelling of Experimental mm-Wave Indoor Channels." *PLoS ONE* 11(9): e0163034.

[2] Ai, B. et al. 2019. "High-Performance Ray-Tracing Platform for 5G and Beyond." *IEEE Commun. Surveys Tuts.* 21(1): 10–27.

[3] Grupe, N. et al. 2024. "3-D Ray Tracing Channel Predictor for Each Effect." *IEEE Open J. Antennas Propag.* 5(2).

[4] ITU-R. 2012. *Recommendation ITU-R P.1238-7: Indoor Propagation Prediction.* Geneva.

[5] Sloane, W. et al. 2023. "Analysis of mmWave Channel Sparsity." *IEEE Antennas Wirel. Propag. Lett.* 22(4).

[6] Schützenhöfer, D. et al. 2020. "Massive MIMO O2I Channel Rank Analysis." In *Proc. IEEE SPAWC*, Atlanta, USA.

[7] Degli-Esposti, V. et al. 2007. "Scattering From Buildings: Measurement and Modelling." *IEEE Trans. Antennas Propag.* 55(1): 143–153.

[8] Blender Online Community. 2025. "Blender – 3D Modelling Package." https://www.blender.org, accessed Apr. 25, 2025.

[9] OpenStreetMap contributors. 2025. "OpenStreetMap." https://www.openstreetmap.org, accessed Apr. 25, 2025.

[10] Björnson, E., J. Hoydis, and L. Sanguinetti. 2017. *Massive MIMO Networks.* Found. Trends Signal Process. 11(3–4).

[11] Chen, X. 2015. "Antenna Correlation in Multi-Antenna Systems." *PIER B* 62: 241–253.

[12] Vitucci, E. M. et al. 2023. "Heuristic Model for Diffuse Scattering." *IEEE Trans. Antennas Propag.* 71(7): 5564–5575.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

Detekcja Preambuł PRACH jako Problem Klasyfikacji Wieloklasowej: Podejście Uczenia Maszynowego z wykorzystaniem SVM

PRACH Preamble Detection as a Multi-Class Classification Problem: A Machine Learning Approach Using SVM

Filip Ferenc (filip.ferenc@nokia.com)[1,2]; Michal Szczachor (michal.szczachor@nokia.com)[2]

[1]Wrocław University of Science and Technology,
27 Wybrzeże S. Wyspiańskiego, 50-370 Wrocław
[2]Nokia, Szybowcowa 2, 54-130 Wrocław

DOI: 10.15199/59.2025.4.23

Streszczenie: Prezentowane badania poruszają problem detekcji preambuł w procedurze Random Access (RA) w sieciach LTE/5G, reformułując go jako zadanie klasyfikacji wieloklasowej. Autorzy proponują i oceniają zastosowanie algorytmu Support Vector Machine (SVM) w porównaniu z szeroko przyjętym analitycznymi metodami detekcji. Proponowane rozwiązanie ma na celu zwiększenie efektywność detekcji (MDR) i zmniejsza wskaźnik pominiętych detekcji (FAR). Podejście to może znacząco poprawić dokładność detekcji, dlatego wyniki podkreślają potencjał technik uczenia maszynowego w optymalizacji procedur RA oraz poprawie ogólnej dostępności sieci.

Abstract: This study addresses the problem of preamble detection in the Random Access (RA) procedure of LTE/5G networks by reformulating it as a multi-class classification task. The authors propose and evaluate the use of a Support Vector Machine (SVM) algorithm in comparison with conventional detection methods. The proposed solution aims to improve the assignment of preamble indices, thereby enhancing detection efficiency and reducing the rate of missed detections. This approach can be used to significantly improve detection accuracy, therefore the findings highlight the potential of machine learning techniques in optimizing Random Access procedures and improving overall network accessibility.

Słowa kluczowe: PRACH, 5G, LTE, SVM, Uczenie Maszynowe
Keywords: PRACH, 5G, LTE, SVM, Machine Learning

1. WSTĘP

Autorzy proponują wykorzystanie uczenia maszynowego do wykrywania preambuł PRACH w sieciach LTE/5G, co jest kluczowym procesem w procedurze dostępu losowego określonej w specyfikacjach 3GPP [1, 2]. Poprzednie podejścia badały sieci neuronowe i metody grupowania w celu realizacji tego zadania. [3, 4]. To badanie koncentruje się na SVM (Support Vector Machine) ze względu na jego równowagę pomiędzy wysoką dokładnością a dobrą zdolnością generalizacji. Proponowane podejście oparte na uczeniu maszynowym jest porównywane z konwencjonalną, analityczną metodą detekcji PRACH, która zazwyczaj wykorzystuje obliczoną korelację odebranego sygnału do zestawu znanych sekwencji Zadoff-Chu w celu zidentyfikowania przesyłanej preambuły. Chociaż podejście oparte na korelacji jest optymalne w idealnych warunkach kanałowych, jego wydajność może się pogarszać w bardziej realistycznych środowiskach z wielościeżkowym propagacją, zakłóceniami lub przesunięciami częstotliwości sygnału. Modele uczenia maszynowego natomiast mają potencjał adaptacji do takich zakłóceń sygnału, co umożliwia bardziej niezawodną detekcję w nieidealnych scenariuszach.

2. METRYKI WYKORZYSTANE DO OCENY ALGORYTMU

Ocena wydajności odbiornika PRACH [5] opiera się na kilku podstawowych wskaźnikach, przede wszystkim wskaźniku False Alarm Rate (FAR) i wskaźniku Miss Detection Rate (MDR).

Dodatkowo, skuteczność algorytmu SVM została oceniona na podstawie dokładności zbioru walidacyjnego oraz macierzy pomyłek.

Metryki FAR, MDR oraz macierz pomyłek są ze sobą powiązane. W ocenach opartych na algorytmach uczenia maszynowego istotne jest przeanalizowanie całej macierzy pomyłek, aby uzyskać kompleksową ocenę wydajności modelu.

3. BAZA DANYCH

Zestaw danych użyty do oceny algorytmu analitycznego oraz do szkolenia i testowania modelu opartego na SVM został wygenerowany przy użyciu symulatora warstwy fizycznej LTE/5G opracowanego przez Nokia Solutions & Networks. To konfigurowalne narzędzie umożliwia kompleksowe symulacje sieci LTE/5G, wspiera rejestrowanie wybranych zmiennych oraz dostarcza statystycznych wskaźników, takich jak FAR i MDR.

Zestaw danych został wygenerowany w czterech scenariuszach kanałów propagacyjnych (AWGN, EPA, EVA, ETU) dla 64 preambuł w zakresie 41 wartości SNR. Zakres punktów SNR uwzględnionych w symulacji przyj-

mował wartości całkowite od -20 do 20 dB. Każda próbka została zebrana podczas jednosekundowej symulacji, z oknem transmisji dostępnym co 10 ms [6, 7], co dało w sumie 26 240 próbek sygnału. Ponadto, dane zostały zarejestrowane dla dwóch anten odbiorczych, dlatego każda preambuła przy danym SNR była powiązana z 2×100 próbkami sygnału. Po ostatecznym przetworzeniu kompletny zestaw danych składał się z około 2 milionów próbek sygnałów.

4. PRZETWARZANIE WSTĘPNE

Zespolone sygnały Zadoff-Chu [8] o długości 839 próbek, zostały rozdzielone na ich rzeczywiste i urojone składniki. Składniki te zostały następnie połączone sekwencyjnie, aby utworzyć reprezentację wektorową zawierającą 1678 próbek.

Wektory wynikowe zostały zebrane w macierz, gdzie każdy wiersz odpowiadał pojedynczemu sygnałowi, a każda kolumna reprezentowała konkretna próbkę. W tej konfiguracji każda próbka była traktowana jako niezależna cecha, co pozwoliło na uchwycenie odmiennych charakterystyk sygnału w kolumnach macierzy. Ta ustrukturyzowana reprezentacja umożliwiła kompleksową analizę za pomocą Principal Component Analysis (PCA) oraz ułatwiła klasyfikację sygnałów przy użyciu algorytmu SVM.

4.1 Analiza głównych składowych (PCA)

Analiza głównych składowych ang. Principal Component Analysis (PCA) [9] to technika statystyczna, która wykorzystuje dekompozycję macierzy danych [10] w celu określenia kierunków, wzdłuż których maksymalizowana jest wariancja danych. Ta metoda przekształca układ współrzędnych w taki sposób, że pierwszy główny komponent jest wyrównany z kierunkiem największej wariancji. Kolejne składowe główne są wzajemnie ortogonalne i uporządkowane zgodnie z ich odpowiednim wkładem w całkowitą wariancję, w porządku malejącym.

Wyniki analizy PCA przedstawiono poniżej. Rysunki 1 oraz 2 ilustrują skumulowaną wariancję w funkcji liczby komponentów, a także wizualizację danych rzuconych na pierwsze dwa główne składniki.

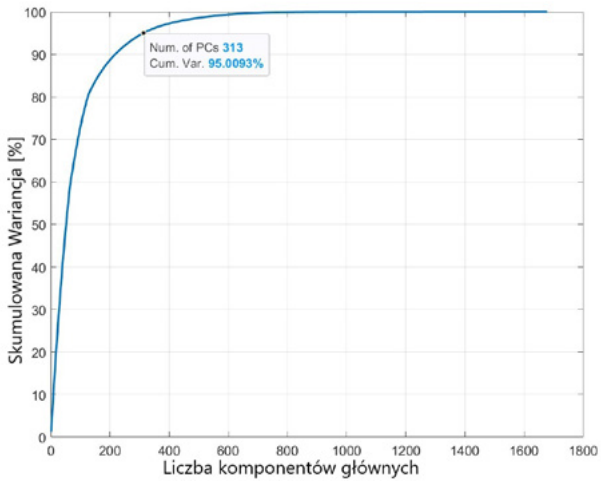


Figure 1: Wykres skumulowanej wariancja w zależności od liczby głównych komponentów.

Na podstawie analizy wykresów przedstawiających zależność między wariancją skumulowaną a liczbą składowych głównych (rys. 1) można stwierdzić, że przy wykorzystaniu 314 składowych głównych osiągnięto wariancję skumulowaną na poziomie 95%, zmniejszając tym samym liczbę parametrów z 1678 cech.

Wizualizacja danych rzutowanych na pierwsze dwa główne komponenty sugeruje wysoki stopień separowalności klas. Rysunek 2 pokazuje, że separacja klas poprawia się w miarę zwiększania się odległości od środka układu współrzędnych.

Algorytm PCA może być skutecznie zaimplementowany poprzez dekompozycję wartości osobliwych (SVD) [11].

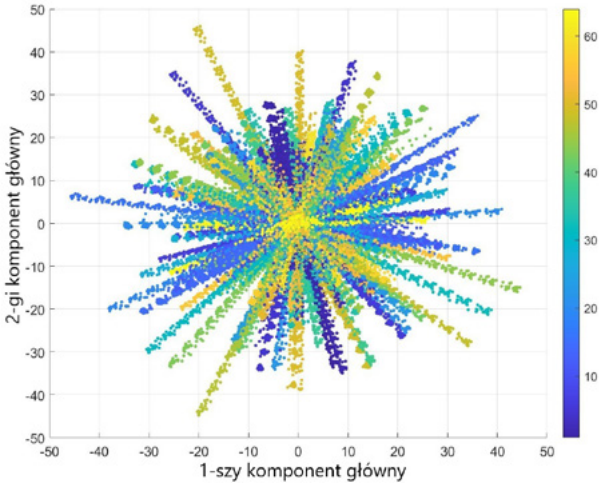


Figure 2: Wizualizacja pierwszych dwóch komponentów głównych.

5. MASZYNA WEKTORÓW NOŚNYCH (SVM)

Maszyna wektorów nośnych ang. Support Vector Machine (SVM) [12] jest szeroko stosowanym i wysoce skutecznym algorytmem uczenia nadzorowanego w uczeniu maszynowym. Jego głównym celem jest wyznaczenie optymalnej hiperpłaszczyzny w przestrzeni cech, która maksymalnie oddziela punkty danych należące do różnych klas. Hiperpłaszczyzna jest zdefiniowana w taki sposób, że margines między najbliższymi punktami danych z przeciwnych klas jest maksymalizowany. Te kluczowe punkty danych, określane jako wektory nośne, odgrywają kluczową rolę w kształtowaniu granic klasyfikacji.

5.1 Optymalizacja parametrów modelu

Podczas projektowania modelu zastosowano pięciokrotną walidację krzyżową. Procedura ta polegała na podzieleniu danych na pięć grup, z jedną grupą pełniącą rolę zbioru testowego, a pozostałymi jako zbiorami treningowymi w każdej iteracji, uśredniając wyniki i ograniczając wpływ losowego podziału danych na końcową ocenę modelu.

Aby zmaksymalizować dokładność modelu, użyto optymalizacji hiperparametrów [13, 14], która automatycznie przeszukiwała przestrzeń możliwych konfiguracji parametrów, wybierając te, które zapewniły najlepszą dokładność przewidywania modelu.

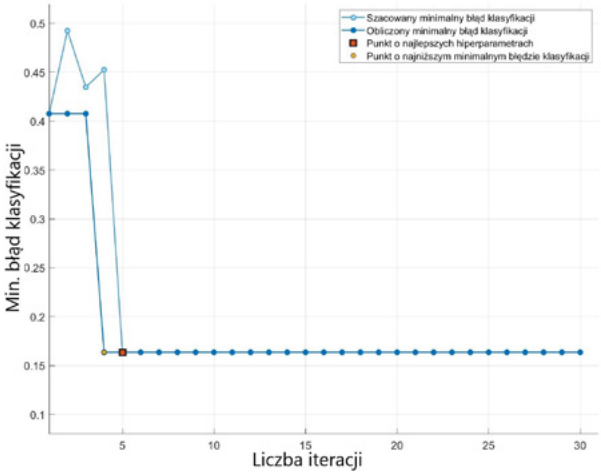


Figure 3: Wykres minimalnego błędu klasyfikacji w zależności od liczby iteracji.

Optymalizacja parametrów modelu SVM została przeprowadzona w 30 iteracjach przy użyciu algorytmu optymalizacji bayesowskiej. Algorytm ten iteracyjnie aktualizuje założenia dotyczące przestrzeni hiperparametrów, aby znaleźć optymalne wartości minimalizujące funkcję błędu.

Rysunek 3 pokazuje związek między minimalnym błędem klasyfikacji a liczbą iteracji, a także wyniki optymalizacji parametrów modelu.

Parametry uzyskane po optymalizacji SVM przedstawiono w Tabeli 1.

Table 1: Parametry modelu SVM

Parametr	**Wartość**
Kernel	Quadratic
Box Constraint (C parameter)	0.09
Kernel Scale	1
Multiclass Coding	One-vs-All
Data Standardization	No

5.2 Wyniki treningu modelu

Dla wytrenowanego modelu została wygenerowana macierz pomyłek. Obliczono również dokładność modelu na zbiorze walidacyjnym, która wyniosła 94,8%.

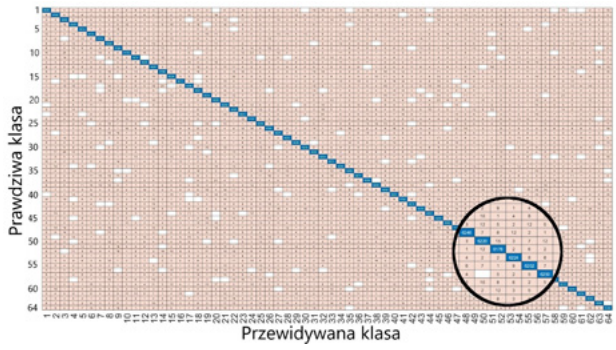


Figure 4: Macierz pomyłek modelu SVM.

Dokładność osiągnięta w 64 klasach stanowi satysfakcjonujący wynik, potwierdzający skuteczność procesu optymalizacji parametrów.

6. TESTOWANIE MODELU

Aby zweryfikować skuteczność opracowanego detektora PRACH i porównać jego działanie z algorytmem analitycznym, przeprowadzono symulacje przy użyciu symulatora LTE/5G, który wcześniej służył do generowania bazy danych. Symulacje te skoncentrowały się na scenariuszu procedury wczesnego dostępu i zostały przeprowadzone zarówno dla opracowanego detektora, jak i algorytmu analitycznego. Eksperymenty obejmowały różne warunki propagacyjne, co pozwoliło na kompleksową ocenę obu rozwiązań w różnych scenariuszach.

6.1 Kanały propagacyjne

Testy przeprowadzono na następujących modelach kanałów propagacyjnych [15, 5]: Additive White Gaussian Noise (AWGN), Extended Pedestrian A (EPA), Extended Vehicular A (EVA), Extended Typical Urban (ETU), TDLC300 i TDLD30.

Analizując działanie detektora preambuły PRACH w tych warunkach kanałowych, oceniono jego odporność i zdolność do adaptacji do różnych scenariuszy występujących w rzeczywistych warunkach propagacyjnych.

7. WYNIKI

Zgodnie ze standardami 3GPP [15] prawdopodobieństwo wystąpienia fałszywego alarmu nie powinno przekraczać 0,1%, natomiast prawdopodobieństwo wykrycia preambuły powinno wynosić co najmniej 99%. Zapewnienie tak wysokiej precyzji jest kluczowe dla prawidłowego funkcjonowania sieci LTE i 5G.

Rysunki 5 i 6 ilustrują porównanie wyników FAR i MDR między detektorem SVM a algorytmem analitycznym. Wyniki obejmują wszystkie wymienione kanały propagacji. Wartości MDR i FAR na wykresach są identyczne, ponieważ każdy przypadek błędnej klasyfikacji preambuły był jednocześnie interpretowany przez detektor SVM jako przypisanie preambuły innej niż ta faktycznie wysłana (FA) oraz jako niemożność poprawnego wykrycia preambuły, która rzeczywiście została przesłana (MD).

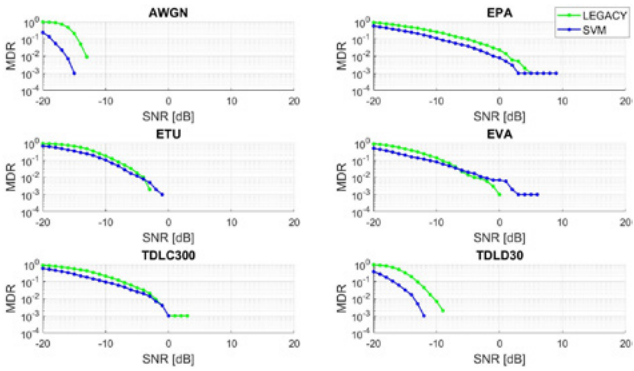


Figure 5: Porównanie wyników MDR dla algorytmu analitycznego i detektora SVM.

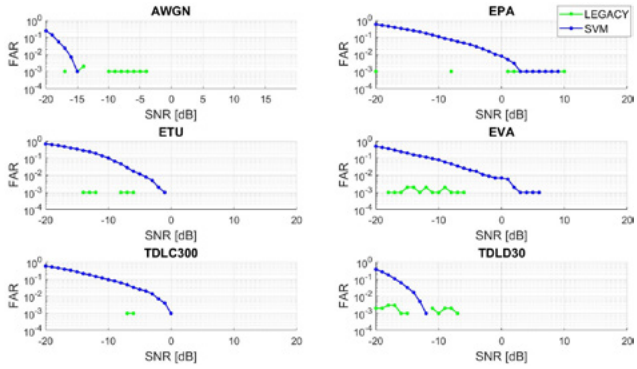


Figure 6: Porównanie wyników FAR dla algorytmu analitycznego i detektora SVM.

Algorytm analityczny osiągnął poziom odniesienia MDR na poziomie -13 dB dla kanału AWGN, 2 dB dla kanału EPA, -3 dB dla kanału EVA, -4 dB dla ETU, -2 dB dla TDLC300 i -11 dB dla TDLD30. Fałszywe alarmy występowały tylko w wybranych punktach SNR; Ich liczba nie przekraczała jednak 0,3% we wszystkich analizowanych kanałach propagacji. Najgorszy wynik FAR uzyskano dla kanału TDLD30. Wyniki FAR dla algorytmu analitycznego są nierówne.

Detektor oparty na SVM osiągnął poziom odniesienia wskaźnika MDR dla poziomu SNR: -16 dB w kanale AWGN, -1 dB dla EPA, -3 dB dla EVA, -5 dB dla ETU, -3 dB dla TDLC300 oraz -14 dB w kanale TDLD30. Dla wskaźnika FAR detektor osiągnął wartość 0,1% dla poziomów SNR: -15 dB w kanale AWGN, 3 dB w kanale EPA, 3 dB w kanale EVA oraz -1 dB w kanale ETU.

8. WNIOSKI

Porównanie detektora opartego na SVM z algorytmem analitycznym ujawnia, że, w przypadku kanału AWGN, opracowany detektor osiągnął referencyjną wartość MDR przy poziomie SNR o 3 dB niższym. Podobne poprawę o 3 dB zaobserwowano dla kanału EPA, 1 dB dla kanałów ETU i TDLC300 oraz 3 dB dla kanału TDLD30. Jedynym wyjątkiem był kanał EVA, gdzie algorytm analityczny przewyższał detektor oparty na SVM o 1 dB. Takie różnice w poziomach SNR są bardzo istotne w systemach telekomunikacyjnych i mogą znacząco wpłynąć na ogólną efektywność operacyjną.

Jak opisano w poprzedniej sekcji, założenie o nieignorowaniu klasy szumów skutkuje tym, że algorytm zawsze generuje detekcję, co przyczynia się do wysokiego wskaźnika fałszywych alarmów (FA). W tym kontekście algorytm SVM ma potencjał, aby służyć jako komplementarne podejście do klasycznego rozwiązania, wspomagając redukcję wskaźnika nieudanych detekcji (MD).

Uzyskane wyniki są obiecujące, pokazując odpowiedni dobór parametrów w procesie optymalizacji i podkreślając potencjał do dalszych ulepszeń opracowanego rozwiązania.

REFERENCES

[1] 3rd Generation Partnership Project (3GPP), "Study on New Radio (NR) Access Technology; Physical Layer Aspects," 3GPP Technical Report TR 38.802, vol. V14.2.0, Mar. 2017.

[2] 3rd Generation Partnership Project (3GPP), "Study on New Radio Access Technology; Radio Access Architecture and Interfaces," 3GPP Technical Report TR 38.801, vol. V14.0.0, Mar. 2017.

[3] N. Modina, R. Ferrari, and M. Magarini, "A Machine Learning-Based Design of PRACH Receiver in 5G," in Proceedings of the 2nd International Workshop on Recent Advances in Cellular Technologies and 5G for IoT Environments (RACT-5G-IoT), Leuven, Belgium, Apr. 2019, pp. 3. [Online]. Available: https://re.public.polimi.it/retrieve/handle/11311/1124713/477363/RACT-5G-IoT_3_6791_&_6834.pdf

[4] R. Singh, A. K. Yerrapragada, and R. K. Ganti, "A Machine Learning based Hybrid Receiver for 5G NR PRACH," arXiv preprint, 2024. [Online]. Available: https://arxiv.org/abs/2411.08919.

[5] 3rd Generation Partnership Project (3GPP), "NR; User Equipment (UE) radio transmission and reception; Part 4: Performance requirements (Release 18)," 3GPP Technical Specification TS 38.101-4, vol. V18.5.0, Sep. 2024.

[6] Keysight Technologies, "Understanding the 5G NR Physical Layer." [Online]. Available: https://www.keysight.com/us/en/assets/9921-03326/training-materials/Understanding-the-5G-NR-Physical-Layer.pdf

[7] X. Lin, J. Li, R. Baldemair, et al., "5G New Radio: Unveiling the Essentials of the Next Generation Wireless Access Technology," arXiv, 2018. [Online]. Available: https://arxiv.org/abs/1806.06898

[8] J. G. Andrews, "A Primer on Zadoff-Chu Sequences," Nov. 2022. [Online]. Available: https://ar5iv.labs.arxiv.org/html/2211.05702

[9] I. T. Jolliffe and J. Cadima, "Principal Component Analysis: A Review and Recent Developments," Philosophical Transactions of the Royal Society A, vol. 374:20150202, 2016. doi: 10.1098/rsta.2015.0202.

[10] T. Eriksen and N.u. Rehman, "Data-driven nonstationary signal decomposition approaches: a comparative analysis," Scientific Reports, vol. 13, 2023. doi: 10.1038/s41598-023-28390-w.

[11] S. L. Brunton and J. N. Kutz, "Singular Value Decomposition (SVD)," in Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control, Cambridge University Press, 2019, pp. 3-46.

[12] C. M. Bishop, "Pattern Recognition and Machine Learning," Springer, New York, 2006. Chapter 7: Sparse Kernel Machines.

[13] MathWorks, "Classification Learner App." [Online]. Available: https://www.mathworks.com/help/stats/classification-learner-app.html

[14] MathWorks, "Statistics and Machine Learning Toolbox." [Online]. Available: https://www.mathworks.com/help/stats/index.html?s_tid=CRUX_lftnav

[15] 3rd Generation Partnership Project (3GPP), "Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Base Station (BS) radio transmission and reception, Release 18," 3GPP Technical Specification TS 36.104, vol. V18.5.0, Mar. 2024. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2412

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

FWA PERFORMANCE IMPROVEMENT IN 5G NETWORKS THROUGH MIMO PRECODING OPTIMIZATION

Jakub Borkowski[1]; Jacek Wszołek[2]; Michał Wągrowski[3]; Jukka Lempiainen[4];

[1] AGH University in Cracow, Kraków: jakub.borkowski@agh.edu.pl
[2] AGH University in Cracow, Kraków: jacek.wszolek@agh.edu.pl
[3] AGH University in Cracow, Kraków: michal.wagrowski@agh.edu.pl
[4] Tampere University, Faculty of Information Technology and Communication Sciences: jukka.lempiainen@tuni.fi

...

DOI: 10.15199/59.2025.4.24

Abstract: FWA is widely recognized as an efficient, scalable, and cost-effective solution for delivering high-speed broadband services, especially where fiber rollout is challenging. Despite 5G's support for high data rates using MIMO, real-world FWA often suffers from dominant LOS propagation and static terminals, which limit spatial diversity and MIMO rank. This study evaluates 3GPP closed-loop precoding in such conditions and introduces an adaptive phase-shift optimization approach to enhance spatial decorrelation. Analyses performed in a simulated MIMO 4×4 environment show up to 30-50% capacity gain with the proposed method in static LOS-dominant scenarios. Conducted measurements further confirms the limited use of higher-rank MIMO in practice, which highlights the potential of the proposed method to improve the performance of FWA under constrained spatial conditions.

Keywords: FWA, MIMO, LTE, 5G, precoding, PMI

1. INTRODUCTION

Fixed Wireless Access (FWA) presents a compelling business case for operators looking to expand coverage while optimizing capital expenditures due to lower deployment costs and a faster roll-out process compared to traditional wired infrastructure, [3], [10]. It serves as a strong complement to fiber optic networks, filling in last-mile connectivity gaps and offering a viable alternative in hard-to-reach areas. FWA plays a pivotal role in bridging the digital divide, particularly in undeserved rural regions, by providing affordable Internet access where wired solutions are economically unfeasible [7]. According to the latest GSA report, FWA over 5G has already been launched by nearly 500 operators in about 180 countries [6].

Multiple Input Multiple Output (MIMO) is a fundamental technology in current radio systems that drives high data rates without the use of additional spectrum. For most applications, the advantages of MIMO are realized in a rich scattering environment with multipath propagation. However, considering strong LOS presence in FWA links and a fact that CPEs remain static, it is becoming challenging to achieve optimal transmission characteristics. As a result, MIMO may often under-perform in such LOS-dominant scenarios, leading to poor channel capacity and suboptimal utilization of available radio resources. This limitation highlights the need for advanced techniques or alternative solutions to enhance MIMO performance in static or low mobility scenarios such as FWA.

In this paper, FWA performance is analyzed in indoor and macrocellular environment with varying LOS conditions and three different antenna configurations. Moreover, an iterative phase shift optimization approach is proposed to boost a 3GPP-standardized closed-loop MIMO precoding to increase decorrelation between MIMO channels and hence to improve the FWA performance. The performance of the precoding algorithm is simulated in order to understand a potential Single User (SU)-MIMO capacity gain that could be achieved by a codebook-unconstrained optimization of the phase shifts.

1.1 Higher-rank MIMO availability

Classical cross-polarized antennas are generally sufficient to enforce a satisfactory level of decorrelation between parallel MIMO channels, enabling MIMO to operate in 2x2 spatial multiplexing mode even in strong LOS environments. However, the challenge lies in utilizing MIMO spatial multiplexing functionality with ranks higher than 2, due to the small distance between the antenna elements in base station antennas (and even more so in terminals). Challenges with FWA under-performing in LOS-dominant environments were identified in [9], [4]. In [11], researchers proposed the use of the Zero-Forcing (ZF) precoding scheme at the base station to improve MIMO channel performance. Moreover, in [8] an innovative approach was studied and presented to increase the capacity of LOS dominant cells through polarization adaptation.

1.2 3GPP closed-loop MIMO functionality

In a closed-loop MIMO, the codebook-based precoding mechanism is a fundamental aspect of MIMO operation, enabling efficient spatial multiplexing and beamforming. The codebook, as defined by 3GPP standards, is a predefined set of precoding matrices that specify how transmitted symbols are mapped onto the antenna elements, including the phase shifts applied to each symbol

[1], [2]. For a given antenna setup (e.g., 2x2, 4x4, or 8x8 MIMO), the codebook contains a collection of precoding matrices, each corresponding to a specific PMI (Precoding Matrix Indicator). The PMI indicates the optimal precoding matrix based on the observed channel conditions by the UE.

The codebook is designed to balance performance and complexity, ensuring that the UE can efficiently compute and feedback the PMI while maintaining high spectral efficiency. For example, in a 2x2 MIMO system, the codebook might include matrices as following with each column representing a different precoding vector, and the phase shifts are applied to the symbols transmitted from each antenna element.

$$W_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, W_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & j \\ 1 & -j \end{bmatrix} \quad (1)$$

The phase shifts applied to the symbols are constrained by the 3GPP-standardized codebook [1], [2] and are designed to exploit the spatial characteristics of the channel. The allowed phase shifts are typically quantized to simplify implementation and reduce feedback overhead. Common phase shifts include

- - 0°, 90°, 180°, and 270°: These correspond to multiplicative factors of 1, j, -1, and -j, respectively.
- - 45° increments: In some codebooks, finer granularity is supported, allowing phase shifts in 45° increments (e.g., $e^{\pi/4}$, $e^{\pi/2}$, $e^{3\pi/4}$, $e^{j\pi}$, etc.) ([1]-from Rel. 10, [2] Type 1 – 45°, Type II – even more accurate phase shifts are supported).

2. MIMO RANK MEASUREMENT RESULTS

SU-MIMO performance was analyzed in a typical macrocellular urban environment in an LTE (a) and 5G (b) commercial network in Poland , as well as in the 5G standalone (SA) laboratory network (c).

(a) The analyzed network cluster consisted of four LTE base stations in a city of Cracow that were arranged in three-sectored deployment with antenna height exceeding average rooftop level by about 20% of the building height. Users were located in mixed NLOS and LOS positions at street level. The sites were configured with passive MIMO 2x2 in the low band: $800MHz$ layer (Id: 1, in Tab. 1), as well as with MIMO 4x4 in the FDD midbands: $1800MHz$ (Id: 2), $2100MHz$ (Id: 3) and $2600MHz$ (Id: 4) layers. Table 1 illustrates a summary of 24-hour average of reported RI (Rank Indicators) by the UE in the analyzed network cluster. Observations indicate that very rarely higher ranks (>2) were reported even if the terminals are not static and the channel is not LOS dominant.

Table 1: Statistic of reported RI (Rank Indicator) – average of 24h observation of commercial network cluster consisting of 4 three-sectored sites in urban macro-cellular environment of Cracow.

Id	*RI=1*	*RI=2*	*RI=3*	*RI=4*
1	81.1%	18.9%	0	0
2	50.8%	45.0%	3.2%	1%
3	52.4%	42.5%	3.1%	2%
4	45.2%	48.5%	4.3%	2%

(b) The outdoor to indoor macro-cell urban measurements scenario was also realized in the city center of Cracow. The serving NR cell was equipeed in 64T64R Massive (M)-MIMO antenna operating on $100MHz$ bandwidth. The purpose was to perform measurements under good radio channel conditions in order to enable connections to switch into a higher-rank transmission mode. The measured RSRP (Reference Signal Received Power) of the serving cell was in the range of $RSRP_{serving} = \langle -99; -100 \rangle [dBm]$. At the same time, the RSRQ (Reference Signals Received Quality) of the serving cell $RSRQ_{serving} = \langle -10; -10.5 \rangle [dB]$. The terminal, during the entire connection, used $MCS = 18$, which leads to 64-QAM modulation and an IP layer downlink speed of around 450 Mbit/s. However, $RI = 3$ and $RI = 4$ were never observed.

(c) Indoor measurements were performed in the 5G Standalone network at $3.9MHz$ carrier frequency. The CPE was 8 meters from the base station, with a single glass window as the only obstruction between them. The following were noted: RI=4 was never observed while $RI = 3$ was an observer for 5% of the measurement time. For 95% of the measurement time $RI = 2$ was recorded.

3. MIMO PRECODING OPTIMIZATION FOR FWA

In FWA scenarios, where the UE remains static, the radio channel exhibits significantly higher temporal stability compared to dynamic mobile environments. This stability allows for an extended link adaptation (LA) procedure, enabling more accurate optimization of the precoding matrix selection to maximize MIMO performance. In contrast, for dynamic UEs with rapidly varying channels (conventional mobile phone use case), the LA process must operate under strict time constraints, trading off optimal precoder selection for low latency. Consequently, FWA systems could leverage prolonged channel coherence times to achieve higher spectral efficiency through refined MIMO configurations. The codebook-constrained 3GPP standardized precoding offers low-feedback overhead and efficiency, although in LOS-dominant environments or in case of highly correlated channels, optimization of precoding seems applicable. This could be more relevant in static or low-mobility scenarios, such as FWA, IoT, etc. where channel remains relatively constant over time.

In our study, we aim to investigate the trade-offs between standardization-compliant and adaptive MIMO precoding strategies, focusing on their suitability for static FWA scenarios, where channel coherence time critically influences the feasibility of iterative optimization. Namely, we compare two distinct MIMO precoding strategies under a simulated 4×4 MIMO transmission system: a 3GPP-approach (Algorithm 1) and proposed Adaptive Phase-shift Optimization method (Algorithm 2).

Algorithm 1 is based on the 3GPP standardized codebook-based precoding, as defined in the 3GPP LTE [1] and NR [2] specifications. In this approach, the UE estimates the downlink channel, determines the optimal RI and PMI and reports them to the base station (BS). The BS, in turn, applies the precoding matrix from a standardized Discrete Fourier Transform (DFT)-based codebook corresponding to the reported RI and PMI. In order to emulate this process, the simulator performs an exhaustive search over the 3GPP-style DFT-based codebook across all allowable RIs from 1 up to 4. These

DFT-based matrices simulate the 3GPP Type-I codebook structure and are normalized to ensure fair power distribution across the antennas. The selected precoder is then applied to both the data symbols and the random Gaussian test symbols for channel evaluation, capacity computation, and bit error rate (BER) analysis using minimum mean square error (MMSE) equalization at the receiver. The algorithm adheres to the feedback-driven operation of practical systems and conforms to 3GPP-defined procedures with assumptions of perfect feedback and noiseless reporting. Therefore, it effectively provides an upper bound on the performance achievable with standardized 3GPP codebooks, assuming ideal channel knowledge and codebook quantization on the UE side. This makes the Algorithm 1 a representative benchmark for the best-case performance within the limitations of real-world 3GPP-compliant systems.

In contrast, Algorithm 2 (Adaptive Phase-shift Optimization) adopts a more flexible and codebook-free adaptive optimization strategy, which is well-suited for static or semi-static UEs, such as FWA scenarios. In this approach, the BS iteratively applies a series of phase shift combinations at the transmit antennas and evaluates the resulting MIMO channel capacity. The iterative phase shift optimization starts from applying a precoding matrix as selected by the 3GPP standardized closed-loop PMI-based selection (Algorithm 1), then it evaluates various phase shift combinations while monitoring the resulting MIMO channel performance (based on the UE feedback) and converges with the optimum phase shifts matrix. This adaptive approach ensures better alignment with the channel's spatial characteristics, even in strong LOS conditions, improving both throughput and reliability compared to a standardized 3GPP precoding. Based on repeated UE channel feedback (up to 100 iterations), the BS selects the phase configuration that yields the highest spectral efficiency. This method allows for a precoder selection without the limitations of a predefined codebook, enabling finer granularity in spatial processing. As such, Algorithm 2 reflects a non-quantized idealized adaptation process, making it suitable for theoretical exploration of capacity-optimal strategies in static settings. Clearly, the brute-force phase search method used can be computationally demanding. As part of ongoing research, we investigate the use of gradient-based, metaheuristic, and machine learning-based techniques to reduce convergence time and computational complexity while preserving near-optimal capacity performance.

4. SIMULATION ENVIRONMENT AND RESULTS

The functionality of Algorithm 1 and Algorithm 2 was implemented in MATLAB using the 5G Toolbox to assess and compare their performance. The simulator models a 4×4 MIMO system with cross-polarized antenna arrays, where elements are spaced at half-wavelength intervals ($\approx 4.17[cm]$ for the $3.6[GHz]$ operating frequency) to balance spatial correlation and compact array design. The wireless channel is modeled as a composite of deterministic LOS and stochastic NLOS components, reflecting realistic propagation conditions in urban microcellular scenarios. The LOS component is constructed using array response vectors at both the transmitter and the receiver, representing a Rank-1 deterministic channel with a single dominant path. In turn, the NLOS component is generated as a Rayleigh fading matrix with independent and identically distributed complex Gaussian entries, incorporating cross-polarization effects through co- and cross-polarized sub-path scaling based on a defined cross-polarization power ratio. Namely, the combination of LOS and NLOS components in the MIMO channel matrix H is modelled by the following Eq. 2:

$$H = \sqrt{\beta}\left(\sqrt{\rho}H_{LOS} + \sqrt{1-\rho}H_{NLOS}\right) \tag{2}$$

where, ρ is a LOS dominance factor, β is a large-scale path loss gain (computed from 3GPP UMi path loss model) H_{LOS} is a rank-1 deterministic channel using transmit and receive steering vectors and H_{NLOS} is a Rayleigh fading matrix with cross-polarized substructure.

The relative strength of LOS and NLOS components is controlled by a LOS dominance factor, enabling simulation of diverse propagation environments. Large-scale fading is introduced via path loss model, which is according to the 3GPP Urban Micro (UMi) channel model [5], accounting for a distance-dependent attenuation (0–1000 m transmitter-receiver separation). To ensure a statistical robustness, 1000 terminal locations are randomly distributed within the coverage area, each simulated over a 100 MHz channel bandwidth. The channel matrix H is normalized by its Frobenius norm to isolate spatial characteristics from absolute gain effects. The simulation uses link adaptation limited to QPSK and 16-QAM modulation and demodulation is done after the MMSE equalization. The channel capacity $[bits/s/Hz]$ is then determined under a given decoder based on the following Eg. 3:

$$C = log_2 det\left(I_{N_r} + \frac{\gamma}{N_t}\cdot HWW^H H^H\right) \tag{3}$$

where, I_{N_r} is identity matrix of size $N_r \times N_t$, representing the noise covariance matrix, γ is an SNR, H is the channel matrix of size $N_r \times N_t$, (N_t = 4,number of transmit antenna, N_r = 4,number of receive antennas) representing the wireless channel as provided in Eq.‘2 and W is the precoding matrix of size $N_r \times d$ (where d is the number of data streams or rank).

The conducted simulations demonstrate that the Adaptive Phase-shift Optimization (Algorithm 2) consistently outperforms the 3GPP codebook-based precoding (Algorithm 1), even when the latter represents an idealized upper bound of a standardized precoding performance. Maximum number of iterations for precoding optimization was limited to 100 in the algorithm implementation, although the simulation results indicated that this does not constitute a bottleneck in finding the optimum combination of the phase shifts, as on average within 60-80 iterations, the algorithm was reaching the optimization convergence. As illustrated in Fig. 1, the Algorithm 2 achieves a higher spectral efficiency across all tested SNR conditions under various simulated LOS-dominance scenarios: (a) - 20%, (b)- 60% and (c)- 80%. Furthermore, Fig. 1 reveals that the Algorithm 2 sustains higher average RI values, indicating its robustness in maintaining high-rank MIMO transmission. Clearly for highly dynamic environments (regular mobile phone use case) there is no room to compromise duration of the link adaption convergence, although in our opinion in more static environments (FWA, cameras, IoT, etc.) this could be studied further to understand if such direction could be practically applicable. Correspondingly, future work will explore reduced-complexity implementations of the

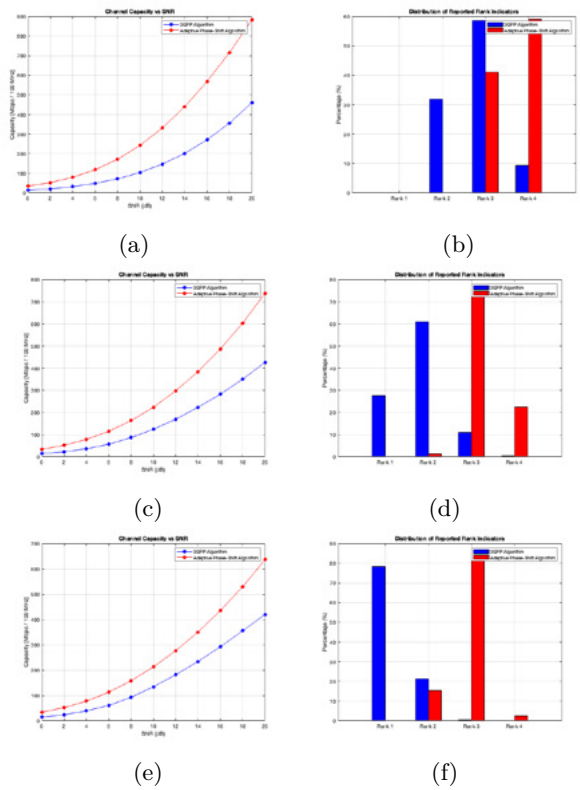


Figure 1: Simulation results illustrating comparison of SU-MIMO channel capacity ((a), (c) and (e)) and reported Rank Indicators ((b), (d) and (f)) in three propagation environments differentiated by LOS component dominance: (a)&(b) 20%, (c)&(d) 60% and (e)&(f) 80%.

Adaptive Phase-shift Optimization mechanism to bridge the gap between theoretical gains and practical deployment constraints.

5. CONCLUSIONS AND FUTURE WORK

The widespread deployment of FWA systems has revealed a significant technical challenge: conventional MIMO techniques, while effective in mobile scenarios with rich scattering, often underperform in LOS-dominant environments where user equipment remains stationary. This limitation rises from the inherent correlation between antenna elements in static configurations, which restricts higher-rank transmission and reduces spectral efficiency. Our research investigates this problem by exploring the possibility of improving MIMO performance through an adaptive precoding system tailored to the stable channel condition characteristics of FWA deployments. Although the work continues, the initial simulations demonstrate that iterative phase shift optimization can achieve capacity improvements of $30\% - 50\%$ compared to standardized 3GPP codebook-based approaches under varying LOS / NLOS conditions. Crucially, while such iterative optimization would be impractical for mobile terminals, where rapid channel variations impose strict latency constraints on link adaptation, the quasistatic nature of FWA links allows extended convergence times without compromising reliability. This distinction positions our method as particularly relevant for fixed or low-mobility use cases, including not only FWA but also IoT gateways and industrial sensors.

Finally, in order to ensure that the performance gains observed in simulations translate to more realistic deployment scenarios, future work must incorporate channel models through ray tracing simulations. Ray tracing allows for a precise modeling of real-world signal propagation, accounting for site-specific reflections, diffractions, and material interactions.

REFERENCES

[1] 3GPP. *TS 36.213: E-UTRA; Physical layer procedures for control*. standard Version 18.3.0. 3GPP, 2025.

[2] 3GPP. *TS 38.214: 5G NR; Physical layer procedures for data*. standard Version 18.6.0. 3GPP, 2025.

[3] Isiaka Alimi. *5G Fixed Wireless Access*. Springer Cham, 2025.

[4] Frode Bohagen, Pal Orten, and Geir E. Oien. "Design of Optimal High-Rank Line-of-Sight MIMO Channels". In: *IEEE Transactions on Wireless Communications* 6.4 (2007), pp. 1420–1425. DOI: 10.1109/TWC.2007.348338.

[5] ETSI. *TS 36.814: E-UTRA Further advancements for E-UTRA physical layer aspects*. standard Version 9.2.0. ETSI, 2017.

[6] GSA. *GSA Report, FWA Market*. 2024. URL: https://gsacom.com/paper/fwa-market-november-2024/ (visited on 05/10/2025).

[7] S. M. Mohd Hassan et al. "Bridging the Digital Divide in Malaysia using Fixed Wireless Access". In: *2021 26th IEEE Asia-Pacific Conference on Communications (APCC)*. 2021, pp. 74–78. DOI: 10.1109/APCC49754.2021.9609825.

[8] Jukka Lempiäinen et al. "Polarization adaptation to improve cell border area bitrates and system capacity in small cells". In: *Wireless Networks* 30 (Dec. 2023), pp. 1–11. DOI: 10.1007/s11276-023-03613-0.

[9] Gregory Morozov, Alexei Davydov, and Dmitry Dikarev. "Line-of-Sight MIMO for 5G Millimeter Wave Fixed Wireless Access Applications". In: *GLOBECOM 2024 - 2024 IEEE Global Communications Conference*. 2024, pp. 5441–5446. DOI: 10.1109/GLOBECOM52923.2024.10901740.

[10] Kudret Topyan and Mehmet Ulema. "Architectural and Financial Considerations for Deploying 5G Based Fixed Wireless Access". In: *2020 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom)*. 2020, pp. 1–6. DOI: 10.1109/BlackSeaCom48709.2020.9235003.

[11] Ami Wiesel, Yonina C. Eldar, and Shlomo Shamai. "Zero-Forcing Precoding and Generalized Inverses". In: *IEEE Transactions on Signal Processing* 56.9 (2008), pp. 4409–4418. DOI: 10.1109/TSP.2008.924638.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

Implementacja mechanizmów oszczędzania energii w nadajnikach 5G – studium przypadku

IMPLEMENTATION OF ENERGY SAVING MECHANISMS IN 5G TRANSMITTERS – CASE STUDY

Dawid Rosołowski[1]; Daniel Gryglewski[1]; Wojciech Wojtasiak[1]; Wojciech Kazubski[1]; Michał Kajczuk[2]

[1] Instytut Radioelektroniki i Technik Multimedialnych, Politechnika Warszawska, dawid.rosolowski@pw.edu.pl
[2] IT Partners Telco Sp. z o.o.

DOI: 10.15199/59.2025.4.25

Streszczenie: **W referacie przedstawiono dwa mechanizmy sterowania wzmacniaczami nadajnika TDD w celu ograniczenia zużycia energii m.in. w typowym dobowym rytmie pracy gNB (stacji bazowej 5G). Pierwszy polega na wystarczająco szybkim wyłączaniu wzmacniaczy TX w szczelinach nietransmisyjnych. Drugi dostosowuje maksymalny poziom mocy wyjściowej nadajnika do natężenia ruchu sieciowego w określonych przedziałach godzinnych doby. Obie procedury zweryfikowano eksperymentalnie w pasmie n78 (TD 3700) (3,3 GHz ÷ 4,2 GHz) na przykładzie pięciu wzmacniaczy mocy Dohert'yego sterowanych sygnałem NR 50 MHz.**

Abstract: **The paper describes two control mechanisms for TDD transmitter amplifiers to reduce power consumption in a typical daily rhythm of a gNB (5G base station) operation. The first mechanism is to switch off amplifiers in non-transmission slots quickly enough. The second adjusts a maximum transmitter output power level to the network traffic volume in defined hourly intervals of the day. Both methods were verified experimentally in the n78 (TD 3700) band (3,3 GHz ÷ 4,2 GHz) using five Doherty power amplifiers controlled by a NR 50 MHz signal as an example.**

Słowa kluczowe: **Oszczędzanie energii w sieci 5G, Aktywny tryb niskiego zużycia energii, ALP, Tryb uśpienia, TRX_Control, ASM, Rekonfiguracja wzmacniaczy TX**

Keywords: **Network Energy Savings 5G, NES, Active Low Power Mode (ALP), Sleep modes, TRX_Control, ASM, Reconfiguration of TX amplifiers**

1. WSTĘP

Wdrażanie zaleceń dyrektywy unijnej **RED** (**R**adio **E**quipment **D**irective) [3] w kontekście harmonijnego rozwoju i normalizacji systemów radiowych działających na terenie Unii Europejskiej, wiąże się z coraz większym zainteresowaniem technikami redukcji zużycia energii na poziomie infrastruktury radiowej. W systemach LTE (4G) obecnie powszechnie stosowaną praktyką jest użycie wysokosprawnych wzmacniaczy mocy (np. w konfiguracji Dohert'yego) [7] w połączeniu z redukcją szczytów (Crest Factor Reduction) oraz linearyzacją cyfrową (Digital Predistortion) [8].

Postulowana dalsza redukcja zużycia energii w sieciach 5G / Open RAN [2] wymaga szczegółowej wiedzy o sposobie transmisji, konfiguracji zasobów, zaimplementowanych procedurach rekonfiguracji i diagnostyki z uwzględnieniem adaptacji na podstawie bieżącej analizy ruchu sieciowego i monitorowania widma w otoczeniu gNB. Cechy współczesnych systemów radiowych i potencjalne mechanizmy do wykorzystania przy optymalizacji poboru energii to: nieciągła praca nadajnika TDD (tylko w przydzielonych szczelinach czasowych), agregacja komponentów widmowych dla osiągniecia pojemności sieci generowanej na potrzeby użytkowników, dostępność statystyk obciążenia sieci, informacje o interferencjach międzysystemowych, czy realizacja przetwarzania sygnałów radiowych w większości w domenie cyfrowej (technologia Software Defined Radio) z dużą rekonfiguracją parametrów w locie.

Dla systemów 5G nie wypracowano jeszcze konkretnych, wiążących i jednoznacznych mechanizmów osiągania wyższej sprawności energetycznej. Sugerowane procedury oszczędzania energii NES (Network Energy Savings) dla systemów Open RAN TDD są opisane w raporcie O-RAN.WG7.NES.0-v02.00 [6]. Oprócz scenariuszy związanych z zależną od obciążenia sieci dynamiczną alokacją zasobów radiowych takich jak: komponenty, podpasma i wiązki w systemach MIMO przez włączenie/wyłącznie odpowiednich elementów w warstwie sprzętowej, zdefiniowane są dwa tryby oszczędzania energii przez kontrolowane przechodzenie urządzenia radiowego w tryb ze zmniejszonym zużyciem energii: **aktywny tryb niskiego zużycia mocy** (**A**ctive **L**ow **P**ower – ALP) i **tryb uśpienia**: **TRX_CONTROL i Advanced Sleep Mode** (ASM). ALP redukuje pobór mocy w aktywnym trybie pracy urządzenia radiowego RU (Radio Unit) z zachowaniem ogólnej wydajności sieci np. przez zmniejszenie mocy nadajnika z dopuszczalnym pogorszeniem ACPR. Efektywność wykorzystania trybu ALP zależy od tego, czy warunki sieciowe utrzymują się w dłuższym przedziale czasu, np. kilka godzin i wiąże się z niewielką dobową liczbą rekonfiguracji odpowiednio dobraną do potrzeb. Uruchomienie uśpienia może być wykonane przy pomocy poleceń: TRX_CONTROL i ASM. Polecenie TRX _CONTROL umożliwia jednostce DU (Distributed Unit) wyłączenie pewnej części układu antenowego jednostki radiowej (RU) na pewien czas. ASM pozwala jednostce DU wyłączyć część przetwarzania jednostki RU. ASM może wyłączyć wybraną nośną, jeden z szyków TX lub RX, albo wszystkie szyki. Tryb ASM przewiduje m.in. szybkie usypianie i wybudzanie komponentów toru radiowego w czasie proporcjonalnym do

trwania slotów lub nawet symboli, czyli rzędu mikrosekund.

Oszczędzanie energii w sieci realizowane jest z zachowaniem pewnych reguł:

- koordynacja użycia trybów oszczędnościowych inicjowana jest z poziomu nadrzędnego kontrolera umieszczonego w węźle centralnym,
- stacje bazowe deklarują swoje możliwości rekonfiguracji, które kontroler zna i decyduje o ich wykorzystaniu bazując na bieżącej analizie ruchu oraz predykcji,
- redukcja energii następuje poprzez wyłączanie lub przejście w tryb uśpienia nieużywanych sprzętowych lub algorytmicznych komponentów lub bloków funkcjonalnych – redukcja komponentów, wyłączenie pasm w danej lokalizacji, uśpienie nieużywanego transceivera, przy czym postulowane czasy wybudzenia powinny się mieścić pomiędzy 10 ms a 100 ms,
- możliwe jest intencjonalne pogorszenie jakości transmisji (pogorszenie liniowości nadajnika, zmniejszenie czułości / odporności odbiornika) o ile nie następuje interferencja z innymi systemami i zachowana jest dopuszczalna jakość transmisji.

W celu realizacji postulatów oszczędzania energii konieczne jest inteligentne zarządzanie wykorzystaniem zasobów bardziej zatomizowanych niż dotychczas. Część z mechanizmów wykonawczych, jak na przykład wyłączanie całych torów, monitorowanie na bieżąco zużycia energii w odniesieniu do istotnych elementów funkcjonalnych jest już implementowana na poziomie wydzielonych systemów zasilania.

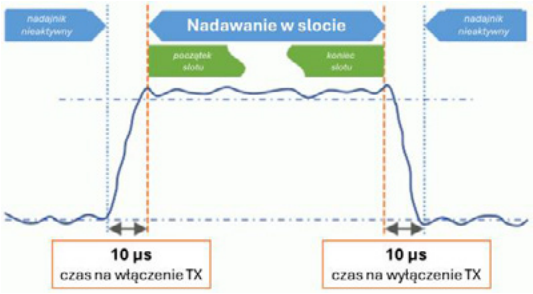
W artykule zaproponowano realizację mechanizmu rekonfiguracji wzmacniaczy mocy torów nadawczych dla potrzeb NES. Sprawdzono także zasadność użycia tego mechanizmu dla energetycznego bilansu łącza. W następnym rozdziale zaprezentowano metody rekonfiguracji wzmacniaczy mocy RF. Rozdział 3 zawiera wyniki eksperymentów przeprowadzonych w rzeczywistym systemie 5G z wykorzystaniem 5 wzmacniaczy w pasmie n78 (3,3 GHz ÷ 4,2 GHz). Referat kończy podsumowanie, w którym oceniono potencjalne oszczędności energetyczne oraz wpływ metod na jakość transmitowanego sygnału.

2. Mechanizmy rekonfiguracji wzmacniaczy TX

W przypadku realizacji postulowanych trybów oszczędnościowych w stacji bazowej 5G, zakłada się możliwość implementacji szybkiego wygaszania nadajnika w obrębie ramki TDD (ASM) lub/i dostosowania poziomu nadawanej mocy do potrzeb wynikających z dobowej aktywności użytkowników (ALP). W kolejnych podpunktach opisano pokrótce ww. mechanizmy.

2.1. Mechanizm szybkiego wygaszania

Nadajnik realizujący transmisję z czasowym dostępem do medium radiowego może być dezaktywowany poza szczelinami czasowymi wyznaczonymi do jego pracy. Przykładowy fragment ramki z wyróżnionym slotem faktycznej aktywności nadajnika pokazano na rys. 1.



Rys. 1. Slot nadawczy – fragment ramki TDD nadawanej przez stację bazową 5G.

W czasie trwania 10 ms ramki systemu TDD może dochodzić do wielokrotnych przejść pomiędzy stanami nadawanie/odbiór zarówno w stacji bazowej jak i w urządzeniach abonenckich. W takim czasie nie możliwe jest wyłączanie i włączenie całego toru nadawczego w RU. Można natomiast wyłączyć wybrane podzespoły toru. Ze względu na dominujący udział w bilansie energetycznym, szczególnie cenne jest kontrolowane włączanie i wyłączanie wzmacniacza mocy. Czas przejścia ze stanu wyłączenia do włączenia i vice versa, zgodnie ze specyfikacją w systemach 5G, musi trwać nie dłużej niż 10 µs. Kluczowanie stanem danego wzmacniacza jest tym bardziej uzasadnione energetycznie, im w mniejszym stopniu wypełnia on alokowany uprzednio w systemie przedział czasu na nadawanie.

Fizyczna implementacja mechanizmu wybudzania wzmacniacza sprowadza się do włączenia zasilania tuż przed slotem emisji i wyłączenia tuż po zakończeniu nadawania w czasie nie przekraczającym kilku mikrosekund. W celu sterowania kluczowaniem, na podstawie konfiguracji przydziału ramki, w torze nadawczym musi zostać wygenerowany sygnał sterujący z odpowiednim wyprzedzeniem obejmującym czas potrzebny do przejścia wzmacniacza w stan stabilnej pracy.

W ofercie rynkowej pojawiły się ostatnio scalone wzmacniacze Doherty'ego o mocy średniej do 10 W [5] [1]. Niektóre posiadają dedykowane wyprowadzenie (ENAble) pozwalające na włączanie wzmacniacza. Określenie faktycznego czasu przełączenia wymaga eksperymentalnej weryfikacji. Brak takiej funkcji wymusza zaprojektowanie odpowiedniego układu sterowania dla skokowej zmiany napięć na bramkach tranzystorów – od zatkania do napięcia zdefiniowanego dla normalnej pracy.

Wyniki pomiaru czasów włączania/wyłączania pięciu zintegrowanych wzmacniaczy Doherty'ego przedstawiono w p.3.1.

2.2. Mechanizm poziomowania mocy wyjściowej

W systemach 5G TDD istnieje widoczna zależność poziomu średniej mocy nadajnika od natężenia ruchu w sieci. Na podstawie analizy dobowego rytmu obciążenia sieci w danej lokalizacji w dłuższym okresie czasu da się przewidzieć poziom średniej mocy nadawania w określonych porach dnia. Dla przykładu w nocy z dużym prawdopodobieństwem można stwierdzić, że moc średnia na-

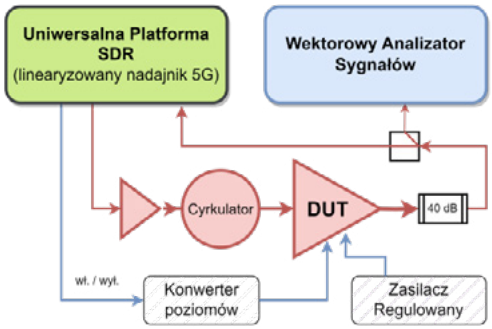
dawania jest najniższa w cyklu dobowym. W konsekwencji sprawność wzmacniacza mocy jest również najniższa. Największą sprawność uzyskuje się w otoczeniu punktu 1dB kompresji, czyli w momentach szczytowego obciążenia sieci. Stąd pomysł na oszczędność energii poprzez regulowanie maksymalnego poziomu mocy wyjściowej nadajnika w zadanych przedziałach godzinowych doby, w których przewidywana średnia nadawana moc nie przekracza zadanych wartości.
W tym celu zbadano możliwości regulacji maksymalnej mocy na wyjściu nadajnika przy pomocy napięcia zasilania wzmacniaczy oraz ocenie jakości transmitowanego widma w takich warunkach.

3. Weryfikacja mechanizmów oszczędzania energii we wzmacniaczach mocy

W eksperymentach wykorzystano pięć płytek testowych z monolitycznymi wzmacniaczami Doherty'ego produkcji NXP i Ampleon polecanych na pasma 3,3 GHz ÷ 3,8 GHz lub 3,8 GHz ÷ 4,2 GHz. Wyniki zawarte w tabelach zmierzono dla częstotliwości środkowej podanych powyżej pasm. W celu uzyskania możliwości szybkiego przełączania wzmacniaczy, zmodyfikowano obwody polaryzacji poszczególnych stopni mocy.

3.1. Skokowa regulacja mocy wyjściowej

Badania przeprowadzono w układzie pomiarowym, którego schemat blokowy pokazano na rys.2. *Uniwersalna Platforma Radiowa* [4] pełni funkcję linearyzowanego przez DPD nadajnika reprezentatywnego sygnału 5G o szerokości 50 MHz.



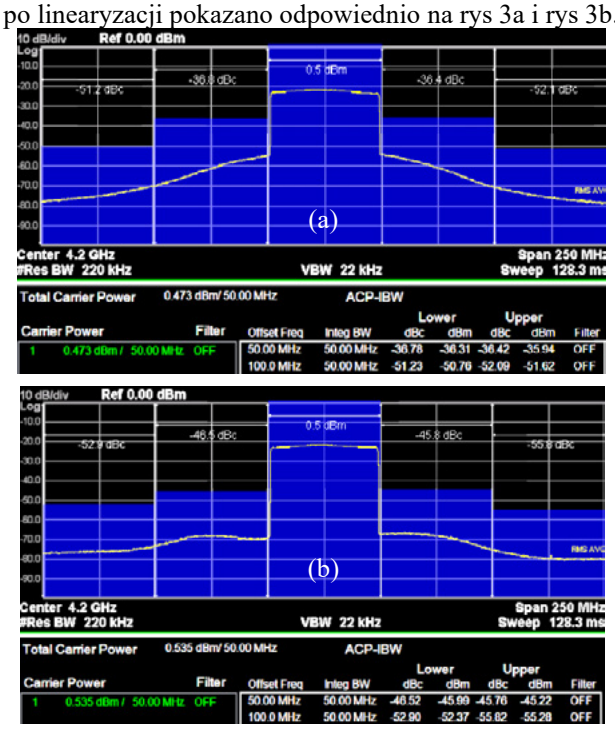
Rys. 2. Stanowisko do pomiaru liniowości wzmacniaczy w trybie redukcji mocy wyjściowej.

Podczas testów poszukiwano nastaw napięć zasilania, które pozwolą uzyskać trzy poziomy mocy tj. moc maksymalną ($\boldsymbol{P_{MAX}}$), połowę ($\boldsymbol{P_{MAX-3dB}}$) i czwartą część mocy maksymalnej ($\boldsymbol{P_{MAX-6dB}}$) przy zachowaniu maski widma z ACPR poniżej -44 dBc w kanałach sąsiednich. W tab.1 zebrano uzyskane wyniki. Jaśniejszą czcionką oznaczono sprawność dodaną (PAE) bez redukcji napięcia zasilania, a pogrubioną po celowym zmniejszeniu mocy. Zmiana warunków zasilania badanych stopni mocy umożliwiła poprawę sprawności i jednocześnie spełnić wymagania dotyczące maski widma. Wzrost sprawności wyniósł ok. 3 punktów procentowych tj. ok. 20%.

Tab. 1. Nastawy U_{DD} vs moc wyjściowa, ACPR i sprawność dodana wzmacniaczy Doherty'ego.

Wzmacniacz	U_{DD} [V] / ***PAE*** [%]		
	P_{MAX}	$P_{MAX-3dB}$	$P_{MAX-6dB}$
B11G33742N81D P_{MAX}=38,5dBm@4GHz	28/**22**	23/14,8→**18**	19,5/9,8→**14**
B11G3338N81D P_{MAX}=37,3dBm@3,6GHz	28/**20,5**	24/15,4 → **18**	20/10 →**14**
BLM9D3842-16AM P_{MAX}=27,2dBm@3,6GHz	28/**21**	24/12,9→**15**	21/8,2→**11**
BLM9D3438-16AM P_{MAX}=26,4 dBm@4GHz	28/**15**	24/10,3→**12**	20/5,7→**8**
RAPIDRF-36SL039 P_{MAX}=34,2dBm@3,6GHz	28/**24**	23,5/16→**19**	21/15,7→**21**

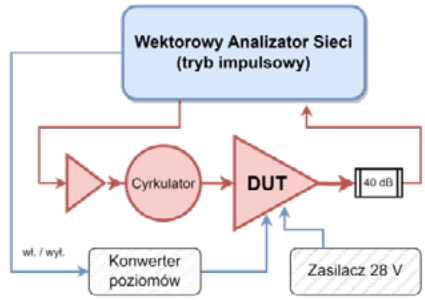
Przykładowe widmo sygnału na wyjściu wzmacniacza B11G3742N81D dla mocy $\boldsymbol{P_{MAX-6dB}}$ = 30,7 dBm bez i po linearyzacji pokazano odpowiednio na rys 3a i rys 3b.



Rys. 3. Widmo sygnału na wyjściu wzmacniacza B11G3742N81D: bez linearyzacji (a) i po zastosowaniu DPD (b).

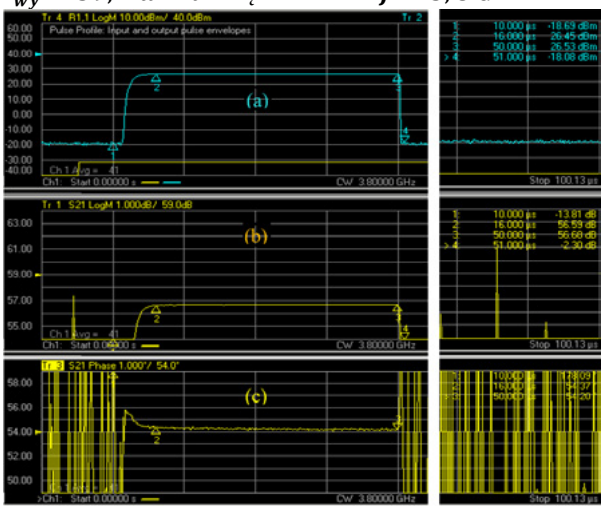
3.1. Szybkie przełączanie

Czasy włączania i wyłączania wzmacniaczy zmierzono na stanowisku pomiarowym pokazanym na rys.3. W tym celu dokonano pomiaru transmitancji przy pomocy analizatora sieci (Keysight N5242A) w trybie impulsowym dla maksymalnej i średniej mocy.



Rys. 4. Stanowisko do pomiaru czasów włączania i wyłączania wzmacniaczy.

Przykład zmierzonych charakterystyk mocy wyjściowej (a), wzmocnienia (b) i fazy transmitancji (c) wraz z zaznaczonymi obszarami włączania (markery: 1→2) i wyłączenia (markery: 3→4) układu B11G3742N81D pokazano na rys.5. Pomiary wykonano dla mocy wyjściowej $\boldsymbol{P_{wy} = 39,2\ dBm}$ i częstotliwości $\boldsymbol{f = 3,8\ GHz}$.



Rys. 5. Czasy włączania i wyłączania wzmacniacza B11G3742N81D.

Czas włączania $\boldsymbol{t_{on}}$ wyznacza położenie markera 2, czyli moment ustalenia się poziomu wzmocnienia $\boldsymbol{|s_{21}|}$ i fazy $\boldsymbol{\Delta arg(s_{21})}$. Zauważalny dryft fazy od chwili włączenia może być skutkiem efektu samonagrzewania się struktury półprzewodnikowej. W związku z tym przeanalizowano również zachowanie się fazy w dłuższej perspektywie czasu. Wyniki pomiaru czasów przełączania ($\boldsymbol{t_{on}, t_{off}}$), zmian wzmocnienia ($\Delta\boldsymbol{|s_{21}|}$) i fazy po upływie 10 ms od włączenia zebrano w tab.2.

Tab. 2. Czasy włączenia i wyłączenia oraz zmiany transmitancji.

Typ wzmacniacza	t_{on} *[us]*	t_{off} *[us]*	$\Delta\lvert s_{21}\rvert$* *[dB]*	$\lvert\Delta arg S_{21}\rvert$* *[°]*
B11G33742N81D	1,5	1	< 0,3	< 3
B11G3338N81D	1,5	1	< 0,3	< 2,2
BLM9D3842-16AM	0,5	0,5	< 0,3	< 2
BLM9D3438-16AM	0,5	0,5	< 0,2	< 1,9
RAPIDRF-36SL039	0,5	0,5	< 0,5	< 1,9

* po 10 ms od włączenia dla $\boldsymbol{P_{MAX}}$

Uzyskanie czasów przełączania rzędu 0,5 ÷1,5 µs, przy jednoczesnym zapewnieniu stabilności wzmacniaczy było możliwe dzięki przeprojektowaniu obwodów zasilania. Ponadto osiągnięto brak wpływu poziomu mocy sygnału na czas przełączania wzmacniacza.
Badania (rys.2) pokazały również, że uzyskane widma po linearyzacji są zbliżone do tych bez kluczowania wzmacniacza, a zysk energii zasilania wzmacniacza jest wprost proporcjonalny do stosunku czasu aktywności wzmacniacza do czasu przydzielonego na nadawanie. Dla typowego scenariusza w gNB z 30 % wypełnieniem slotu czasowego przydzielonego na nadawanie można oszczędzić do 70 % zużywanej energii.

4. Podsumowanie

W referacie przestawiono dwie propozycje redukcji mocy pobieranej przez najbardziej energochłonny składnik systemu 5G tj. przez wzmacniacze mocy. Propozycje te mogą być podstawą wprowadzenia oszczędnościowych trybów zużycia energii w nadajnikach gNB zgodnych z postulowanymi w normie O-RAN procedurami ALP i ASM. Przeprowadzone badania wzmacniaczy mocy typu Doherty'ego wykazały możliwość poprawy sprawności o około 3 punkty procentowe w wyniku regulacji poziomu maksymalnej mocy nadajnika w odpowiedzi na obciążenie sieci, przy czym ta korekta mocy może być dokonywana kilka razy na dobę. Szybkie włączenie wzmacniacza na czas bieżącego slotu nadawania w trakcie wykrycia mniejszego ruchu w sieci jest możliwe – zmierzone czasy przełączania wyynosiły max. 1,5 µs. Ten mechanizm zaoszczędzi energię proporcjonalnie do ilorazu czasu trwania slotu do maksymalnego czasu trwania slotu skonfigurowanego dla maksymalnego obciążenia sieci. Należy zwrócić uwagę, że opisane w normach inne niewykorzystywane do tej pory na szeroką skalę procedury polegające na wyłączaniu bloków sprzętowych po stronie cyfrowej mogą znacząco zmniejszyć pobór energii.

PODZIĘKOWANIA

Referat powstał w ramach projektu PAPR:" Innowacyjna jednostka radiowa S-RU sieci 5G Open RAN z możliwością programowej zmiany funkcjonalności przez podmioty trzecie" nr FENG.01.01-IP.02-0280/23

LITERATURA

[1] **Ampleon.** *Product Catalog: RF Power Solutions for Wireless Infrastructure.* April 2025. Accessed May 24, 2025. https://www.ampleon.com/documents/brochure-leaflet/product-catalog.pdf.

[2] **ETSI.** *ETSI TS 138 104: Technical Specification for the 3GPP System Architecture Evolution (SAE) for Public Land Mobile Networks (PLMN) — Security Aspects.* ETSI, 2019.

[3] **European Union.** *Directive 2014/53/EU of the European Parliament and of the Council of 16 April 2014 on the harmonisation of the laws of the Member States relating to the making available on the market of radio equipment and repealing Directive 1999/5/EC. Official Journal of the European Union*, L 153 (May 22, 2014): 62–106.

[4] **Korpas, Przemysław**, Dawid Rosołowski, Michał Kajczuk, Daniel Gryglewski, and Wojciech Wojtasiak. 2023. "A Universal Hardware Platform for an LTE-Advanced Base Station Prototyping." *Electronics (Switzerland)* 12: 1–18.

[5] **NXP Semiconductors.** *RapidRF Smart LDMOS Front-End Designs User Guide.* January 2023. Accessed May 24, 2025. www.nxp.com/docs/en/user-guide/RapidRFSLUG.pdf.

[6] **O-RAN Alliance.** *O-RAN Network Energy Savings Procedures and Performance Metrics 2.0.* O-RAN Technical Report R003, wersja 02.00, 2024.

[7] **Wang, Zhancang.** *High-Efficiency Load Modulation Power Amplifiers for Wireless Communications.* Amsterdam: Academic Press, 2016.

[8] **Wood, John**, and David R. Bogdanović. *Behavioral Modeling and Linearization of RF Power Amplifiers.* Norwood, MA: Artech House, 2005.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

IMPLEMENTACJA SYNCHRONICZNEJ GENERACJI ZAKŁÓCEŃ CELOWYCH SYSTEMÓW LTE I 5G W ŚRODOWISKU GNU RADIO Z WYKORZYSTANIEM PLATFORMY USRP X310

IMPLEMENTATION OF SYNCHRONOUS GENERATION OF INTENTIONAL INTERFERENCES AGAINST LTE AND 5G SYSTEMS IN GNU RADIO USING USRP X310 PLATFORM

Konrad BOŻEK[1,2]; Paweł SKOKOWSKI[3]; Jan M. KELNER[4]

[1] Wojskowa Akademia Techniczna, Szkoła Doktorska, Warszawa, konrad.bozek@wat.edu.pl
[2] Sieć Badawcza Łukasiewicz – Przemysłowy Instytut Automatyki i Pomiarów, Warszawa, konrad.bozek@piap.lukasiewicz.gov.pl
[3] Wojskowa Akademia Techniczna, Wydział Elektroniki, Instytut Systemów Łączności, Warszawa, pawel.skokowski@wat.edu.pl
[4] Wojskowa Akademia Techniczna, Wydział Elektroniki, Instytut Systemów Łączności, Warszawa, jan.kelner@wat.edu.pl

DOI: 10.15199/59.2025.4.26

Streszczenie: **W niniejszej pracy opisano koncepcję selektywnego zagłuszania systemów LTE i 5G, implementację w środowisku GNU Radio algorytmu do synchronizacji sygnałów LTE oraz synchronicznej generacji zakłóceń celowych na platformie USRP X310 z wykorzystaniem znaczników czasowych. Zaprezentowano stanowisko testowe, w skład którego wchodzi m.in. prywatna stacja bazowa LTE oraz przedstawiono wyniki przeprowadzonych na nim testów.**

Abstract: **The paper describes the concept of selective jamming against LTE and 5G systems, the algorithm implementation in GNU Radio to the LTE signals' synchronization, and the synchronous generation of the intentional interferences on USRP X310 platform using timing tags. The testbed with private LTE base station and results of performed test were presented.**

Słowa kluczowe: **5G, GNU Radio, LTE, USRP, zagłuszanie**

Keywords: **5G, GNU Radio, interference, jamming, LTE, USRP**

1. WSTĘP

W systematyce ataków na komórkowe systemy telekomunikacyjne zagłuszanie można sklasyfikować jako atak na warstwę fizyczną powodujący blokadę usług DoS (ang. *Denial of Service*). Celem tego typu ataków jest spowodowanie jak największej stopy błędów w strumieniu danych na wyjściu odbiornika poprzez degradację odbieranego sygnału radiowego. Ponadto od współczesnych urządzeń zagłuszających oczekuje się, aby efekt ten uzyskany był przy wypromieniowaniu jak najmniejszej mocy.

Jak wynika z opisanych w literaturze badań zarówno w przypadku systemów LTE (ang. *Long Term Evolution*) [3], jak i 5G NR (ang. *New Radio*) [4] poszczególne kanały i sygnały warstwy fizycznej cechuje różna podatność na zakłócenia. Zróżnicowanie to wynika przede wszystkim z zastosowania różnych modulacji i kodowania korekcyjnego oraz liczby elementów zasobów RE (ang. *Resource Elements*) wykorzystywanych przez dany sygnał lub kanał w stosunku do całkowitej ich liczby w ramce.

Z przeprowadzonej w [3] oceny podatności na zagłuszanie kanałów i sygnałów warstwy fizycznej systemów LTE wynika, że najbardziej efektywne są ataki na kanały: wskaźnika formatu kanałów sterujących PCFICH (ang. *Physical Control Format Indicator Channel*), rozgłoszeniowy PBCH (ang. *Physical Broadcast Channel*) oraz podstawowy sygnał synchronizacji PSS (ang. *Primary Synchronization Signal*). Wśród ataków na warstwę fizyczną systemów 5G jako najskuteczniejsze zostały w [4] wskazane ataki na: referencyjny sygnał demodulacji dla kanału rozgłoszeniowego PBCH DM-RS (ang. *PBCH DeModulation Reference Signal*), podstawowy sygnał synchronizacji PSS (ang. *Primary Synchronization Signal*), drugorzędny sygnał synchronizacji SSS (ang. *Secondary Synchronization Signal*) oraz kanał rozgłoszeniowy PBCH [5].

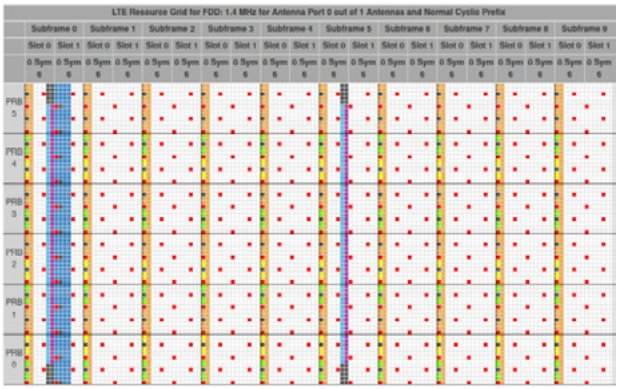
2. ZAGŁUSZANIE SELEKTYWNE

2.1. Koncepcja

W systemach LTE i 5G wykorzystywana jest technika zwielokrotnienia OFDM (ang. *Orthogonal Frequency-Division Multiplexing*), w której podnośne (w dziedzinie częstotliwości) i symbole modulacji (w dziedzinie czasu) tworzą siatkę tzw. elementów zasobów RE, przy czym jeden element zasobów odpowiada jednemu symbolowi modulacji na pojedynczej podnośnej. Dwanaście elementów zasobów na dwunastu kolejnych podnośnych w ramach jednego symbolu modulacji tworzy blok zasobów RB (ang. *Resource Block*). Poszczególne sygnały i kanały

warstwy fizycznej mapowane są w obrębie ramki na poszczególne elementy zasobów, co sytuuje je w określonych obszarach przestrzeni czasowo-częstotliwościowej.

Poszczególne kanały i sygnały warstwy fizycznej systemów LTE i 5G, ich mapowanie na elementy zasobów, stosowane modulacje i schematy kodowania zostały zdefiniowane w specyfikacjach [1] i [2], choć istnieje wiele innych źródeł, w których informacje zawarte w oficjalnych dokumentach zostały przedstawione w formie o wiele bardziej przystępnej. Wśród nich na uwagę zasługują bardzo dobre kompendium technologii LTE [6] i 5G [7] oraz interaktywna wizualizacja struktury czasowo częstotliwościowej ramek LTE [8] (rys.1).



Rys. 1 Przykładowa wizualizacja struktury ramki LTE wygenerowana w [8].

Realizacja selektywnego zagłuszania sygnałów i kanałów warstwy fizycznej systemów LTE i 5G w pierwszej kolejności wymaga ustalenia ich usytuowania w dziedzinie czasowo-częstotliwościowej. O ile sygnały synchronizacyjne i kanał rozgłoszeniowy zajmują stałe miejsce w ramce, to alokacja pozostałych sygnałów i kanałów warstwy fizycznej zależy od szerokości pasma kanału radiowego LTE/5G oraz innych parametrów systemowych zakodowanych w sygnałach synchronizacyjnych (fizyczny identyfikator komórki) i przesyłanych w kanałach rozgłoszeniowych i sterujących. Na podstawie tych informacji możliwe jest wygenerowanie sygnału zagłuszającego o dopasowanych do sygnału zagłuszanego widmie i strukturze czasowej.

2.2. Stanowisko badawcze

Na rys. 2 przedstawiono schemat blokowy stanowiska badawczego do selektywnego zagłuszania systemów LTE.

Rys. 2 Schemat blokowy stanowiska badawczego do selektywnego zagłuszania systemów LTE

Stanowisko składa się z dwóch platform radia programowego SDR (ang. *Software Defined Radio*): USRP X310 i BladeRF, dwóch komputerów PC z systemem operacyjnym Ubuntu 22.04, analizatora sygnałów N9020A oraz terminala abonenckiego. Platforma BladeRF połączona za pomocą interfejsu USB 3.0 z komputerem z zainstalowanym oprogramowaniem srsRAN 4G pełni funkcję prywatnej stacji bazowej LTE, do której łączy się terminal abonencki. Rolę zagłuszacza pełni platforma USRP X310 połączona za pomocą światłowodowego interfejsu Ethernet 10Gb/s z komputerem z zainstalowanym środowiskiem GNU Radio, w którym testowane są opracowywane algorytmy. Sygnały wygenerowane przez prywatną stację bazową i zagłuszacz obserwowane są za pomocą analizatora widma.

3. IMPLEMENTACJA ALGORYTMÓW

3.1. Synchronizacja

W systemach LTE i 5G synchronizacja czasowa odbywa się na podstawie sygnałów PSS i SSS, które bez względu na szerokość pasma kanału radiowego transmitowane są na 62 (LTE) i 127 (5G) podnośnych.

W systemach LTE w trybie dupleksu częstotliwościowego FDD (ang. *Frequency Division Duplex*) sygnał PSS transmitowany jest na ostatnim (#6) symbolu OFDM, w szczelinach #0 i #10. Ponieważ sygnał ten transmitowany jest dwukrotnie w czasie trwania ramki w odstępach 5ms, nie jest możliwe wyłącznie na jego podstawie jednoznaczne określenie jej początku. Do tego celu niezbędne jest wykorzystanie drugorzędnego sygnału synchronizacji SSS, który co prawda również transmitowany jest w dwóch szczelinach #0 i #10 na przedostatnich (#5) symbolach OFDM, jednakże w odróżnieniu od PSS, sygnał SSS w tych szczelinach generowany jest na podstawie dwóch różnych ciągów, co umożliwia jednoznaczne określenie początku ramki.

W zaimplementowanym w środowisku GNU Radio algorytmie odbierany przez odbiornik platformy USRP X310 sygnał radiowy zostaje przetworzony do postaci ciągu próbek IQ, który z wyjścia bloku „UHD: USRP Source" podawany jest do wejścia układu synchronizacji. Funkcję korelatorów pełnią w nim filtry o skończonej odpowiedzi impulsowej (SOI) dopasowane do danego sygnału synchronizacji. Rząd filtrów wynosi 256, co wynika z przyjętej częstotliwości próbkowania, przy której pracują korelatory, tj. 3,84 MHz. Należy zaznaczyć, iż ze względu na liczbę podnośnych równą 62 wykorzystywanych w systemach LTE do transmisji sygnałów synchronizacji PSS i SSS, istnieje możliwość decymacji ciągu próbek z wyjścia bloku „UHD: USRP Source" do częstotliwości próbkowania 960 kHz (po uprzednim poddaniu go filtracji dolnoprzepustowej), dzięki czemu rząd dopasowanego filtru korelacyjnego może zostać ograniczony do 64. Podobną operację można zastosować w odniesieniu do sygnału 5G, i wówczas, przy założeniu 15 kHz rastra podnośnych, częstotliwość próbkowania sygnału podawanego na filtry korelacyjne może zostać obniżona do 1,92 MHz, co wynika z liczby podnośnych równej 127

wykorzystywanych w systemach 5G do transmisji sygnałów synchronizacji PSS i SSS. Wówczas rząd filtru korelacyjnego wynosiłby 128.

W prezentowanym przykładzie źródłem synchronizacji jest sygnał SSS w szczelinie #10. Przy tak dobranych wartościach współczynników dopasowanego filtru korelacyjnego kształt przebiegu na jego wyjściu będzie miał charakterystyczne maksima wskazujące na ostatnią próbką symbolu OFDM z sygnałem SSS (symbol #5) w szczelinie #10. Należy pamiętać, że w przypadku zastosowania przed korelatorem filtru dolnoprzepustowego, sygnał na wyjściu korelatora będzie dodatkowo opóźniony względem sygnału na wyjściu bloku „UHD: USRP Source" o wartość opóźnienia grupowego tego filtru. Ponieważ amplituda przebiegu na wyjściu filtru korelacyjnego zależy od mocy odbieranego sygnału, w celu jej uniezależnienia w algorytmie zastosowano normowanie względem wartości RMS z ostatnich próbek sygnału w liczbie równej rzędowi filtru korelacyjnego.

3.2. Generowanie sygnału zagłuszającego

Jak już wspomniano, selektywne zagłuszanie polega na dopasowaniu widma impulsów sygnału zagłuszającego i czasu, w którym odbywa się ich emisja, do położenia elementów zasobów, na które zmapowany został zagłuszany sygnał lub kanał warstwy fizycznej. Biorąc pod uwagę uwarunkowania sprzętowe nadajników urządzeń zagłuszających, w szczególności stopni mocy wzmacniaczy w.cz., bardzo ważne jest również to, aby obwiednia impulsów zagłuszających miała możliwie stałą wartość, czyli jak najmniejszy stosunek mocy szczytowej do mocy średniej PARP (ang. *Peak to Average Power Ratio*).

W dalszej części pracy skupimy się na zagadnieniu synchronizacji czasowej pomijając nieco kwestię odpowiedniego dopasowania widma sygnału zagłuszającego do sygnału zagłuszanego. Dla uproszczenia rozważań oraz wyraźniejszego zobrazowania na spektrogramie za sygnał zagłuszający przyjęto jednotonowy sygnał o częstotliwości równej $f_n + 384$ kHz, gdzie f_n jest częstotliwością środkową kanału radiowego łącza „w dół", zaś 384 kHz stanowi 1/10 częstotliwości próbkowania w badanym algorytmie. Czas trwania impulsu zagłuszającego ustalono na $71\frac{17}{48}$ µs, co odpowiada czasowi trwania symbolu OFDM łącznie z prefiksem cyklicznym (symbol #0 ma dłuższy niż pozostałe prefiks cykliczny, przez co ich łączny czas trwania wynosi $71\frac{7}{8}$ µs).

W opisywanym przykładzie sygnał zagłuszający generowany jest z częstotliwością próbkowania 3,84 MHz z ciągu N_j próbek w postaci:

$$j[n] = \cos\left(2\pi \cdot \frac{n}{10}\right) + j \cdot \sin\left(2\pi \cdot \frac{n}{10}\right) \qquad (1)$$

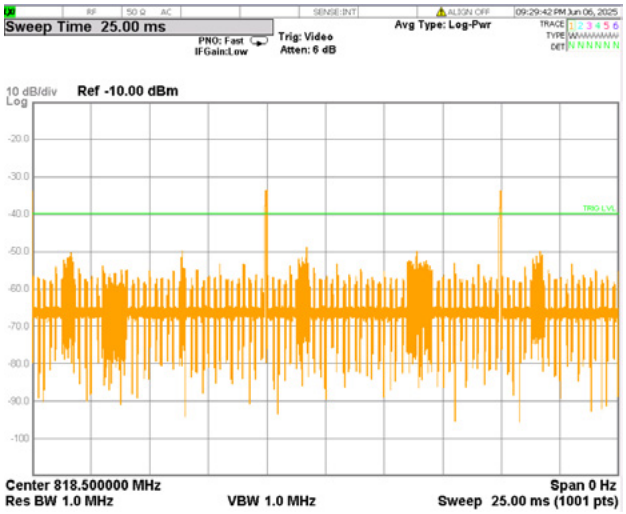
$$n \in [0, N_j - 1] \qquad N_j = 274$$

Emisja sygnału zagłuszającego z nadajnika platformy USRP X310 rozpoczyna się z ustalonym opóźnieniem po wykryciu sygnału synchronizacji. W środowisku GNU Radio interfejsem do nadajnika jest blok „UHD: USRP Sink" skonfigurowany do pracy w trybie pakietowym [9].

Po każdym wykryciu sygnału synchronizacji do bloku „UHD: USRP Sink" wysyłany jest ciąg próbek sygnału zagłuszającego zgodnie z (1), w którym pierwsza z nich opatrzona jest w znaczniki (ang. *tags*) o kluczach „tx_time" oraz „tx_pkt_len". Wartość skojarzona z kluczem „tx_time" określa chwilę czasu, w której ma nastąpić rozpoczęcie emisji sygnału zagłuszającego, zaś wartość skojarzona z kluczem „tx_pkt_len" określa czas emisji i jest równa liczbie próbek w wysłanym ciągu. Wartość skojarzona z kluczem tx_time jest wyliczana dla każdej kolejnej emisji sygnału zagłuszającego jako suma wartości stempla czasowego dla impulsu z synchronizatora (wartość skojarzona z kluczem „rx_time") oraz zadanego opóźnienia.

4. WYNIKI BADAŃ

Pracę układu synchronicznej generacji zakłóceń celowych zweryfikowano w układzie pokazanym na rys. 2. Po uruchomieniu prywatnej stacji LTE i zweryfikowaniu jej prawidłowego działania poprzez upewnienie się, że terminal abonencki zalogował się do niej, uruchomiono zagłuszacz z opóźnieniem startu emisji ustalonym na 3,5 ms. Sygnały LTE i zagłuszający były obserwowane na pracującym w trybie Span Zero analizatorze widma nastrojonym na środkową częstotliwość kanału „w dół", co przedstawiono na poniższym rys. 3.

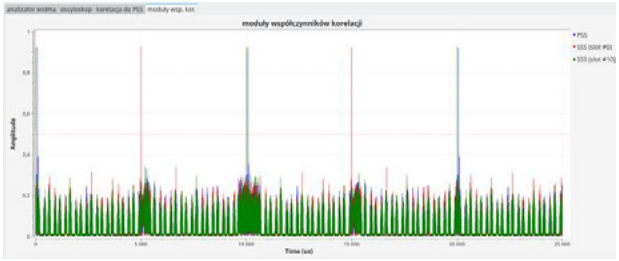


Rys. 3 Sygnał zagłuszający na tle sygnału LTE

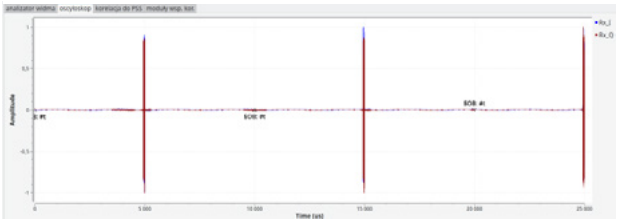
Stworzona w środowisku GNU Radio aplikacja realizująca selektywne zagłuszanie sygnału LTE umożliwia także obserwację widma i przebiegu czasowego sygnału z wyjścia odbiornika platformy USRP X310 oraz unormowanych modułów przebiegów czasowych z wyjść filtrów korelacyjnych (modułów współczynników korelacji). W opisywanym przykładzie odbiornik oprócz sygnału LTE odbiera także wyemitowany przez nadajnik sygnał zagłuszający, co

pozwala na jego obserwację w oknie aplikacji na tle sygnału LTE.

Rysunki 4 i 5 przedstawiają przebiegi modułów współczynników korelacji odbieranego sygnału LTE do wzorca sygnału PSS (kolor niebieski) oraz do wzorców sygnału SSS dla szczelin #0 (kolor czerwony) i #10 (kolor zielony) oraz oscylogramy sygnałów IQ z wyjścia odbiornika z widocznymi znacznikami pochodzącymi z synchronizatora przy ustalonej na 5 ms wartości opóźnienia startu emisji sygnału zagłuszającego.



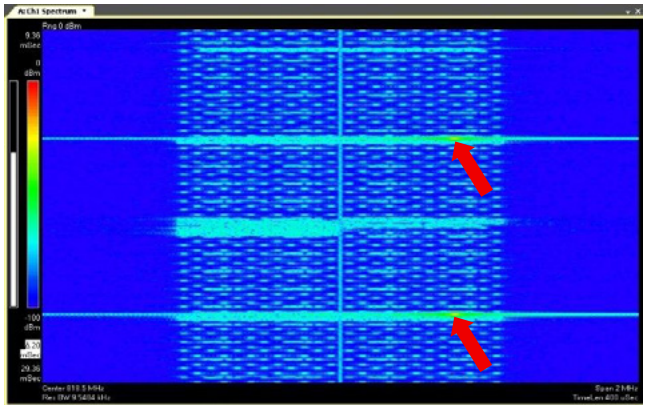
Rys. 4 Moduły współczynników korelacji przy 5 ms opóźnieniu startu emisji sygnału zagłuszającego



Rys. 5 Przebieg sygnału z wyjścia odbiornika przy 5 ms opóźnieniu startu emisji sygnału zagłuszającego

Ustawienie opóźnienia startu emisji sygnału zagłuszającego na 5 ms spowodowało, że rozpoczyna się ona wraz z początkiem ostatniego (#6) symbolu pierwszej (#0) szczeliny w ramce, w którym przesyłany jest sygnał PSS, co spowodowało jego zagłuszenie widoczne na rys. 4 jako wygaszenie pików na przebiegu modułu współczynnika korelacji do PSS (kolor niebieski) w okolicach 5 ms i 15 ms na skali czasu.

Na znajdującym się na rys. 6 spektrogramie zarejestrowanym za pomocą analizatora N9020A sygnał zagłuszający można rozpoznać jako żółte punkty na tle sygnału LTE.



Rys. 6 Spektrogram z widocznym sygnałem zagłuszającym na tle sygnału LTE

5. PODSUMOWANIE

W pracy przedstawiono wykorzystanie zaimplementowanego na platformach radia programowego USRP mechanizmu synchronizacji emisji sygnału z wykorzystaniem znaczników w kontekście realizacji selektywnego zagłuszania sygnałów i kanałów warstwy fizycznej systemów LTE i 5G. Zaprezentowane wyniki testów potwierdzają, iż mimo że sygnał służący do synchronizacji nadajnika przetwarzany jest w systemie niebędącym systemem czasu rzeczywistego i wprowadzającym niedające się przewidzieć opóźnienia, możliwe jest wygenerowanie sygnału zagłuszającego w reżimie czasowym umożliwiającym w ocenie autorów realizację selektywnego zagłuszania systemów LTE i 5G.

6. PODZIĘKOWANIA

Prace realizowane są w ramach VI edycji programu „Doktorat Wdrożeniowy” na podstawie umowy nr DWD/6/0465/2. Publikacja została sfinansowana przez Wojskową Akademię Techniczną w ramach grantu nr UGB/22-059/2025/WAT.

LITERATURA

[1] ETSI. 2024. „LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation (3GPP TS 36.211 version 18.0.1 Release 18)”
[2] ETSI. 2024. „5G; NR; Physical channels and modulation (3GPP TS 38.211 version 18.2.0 Release 18)”
[3] M. Lichtman, R. P. Jover, M. Labib, R. Rao, V. Marojevic, and J. H. Ree. 2016. „LTE/LTE-A Jamming, Spoofing, and Sniffing: Threat Assessment and Mitigation”. IEEE Communications Magazine vol. 54, no.4:54-61
[4] M. Lichtman, R. Rao, V. Marojevic, J. Reed and R. P. Jover. 2018. „5G NR Jamming, Spoofing, and Sniffing: Threat Assessment and Mitigation”. IEEE International Conference on Communications Workshops (ICC Workshops):1-6
[5] P. Skokowski, J.M. Kelner, K. Malon, K. Maślanka, A. Birutis, M.A. Vazquez, S. Saha, W. Low, A. Czapiewska, P. Rajchowski, J. Magiera, and S.J. Ambroziak, 2022. “Jamming and jamming mitigation for selected 5G military scenarios”, Procedia Computer Science, vol. 205: 258–267.
[6] https://www.sharetechnote.com/html/Handbook_LTE.html (dostęp 30 czerwca 2025)
[7] https://www.sharetechnote.com/html/5G/Handbook_5G_Index.html (dostęp 30 czerwca 2025)
[8] https://dhagle.in/LTE.php (dostęp 30 czerwca 2025)
[9] https://wiki.gnuradio.org/index.php/USRP_Sink (dostęp 30 czerwca 2025)

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

MECHANIZM PRZECIWDZIAŁANIA ZAKŁÓCENIOM W SIECI 5G

MECHANISM FOR MITIGATING INTERFERENCE IN 5G NETWORKS

Bartłomiej Bryś; Piotr Pyda

Wojskowy Instytut Łączności - Państwowy Instytut Badawczy, Zegrze Płd., {b.brys; p.pyda}@wil.waw.pl

DOI: 10.15199/59.2025.4.27

Streszczenie: **Artykuł przedstawia implementację nowatorskiego mechanizmu przeciwdziałania zakłóceniom w sieciach komórkowych (w tym 5G) – mechanizm Anti-interference system (AIM). Zakłócenia mogą znacząco wpływać na jakość połączeń oraz wydajność komunikacji. Zaimplementowane oprogramowanie umożliwia monitorowanie istotnych parametrów wskazujących na zakłócenia tj. stosunek sygnał-szum SNR (*ang. Signal-to-Noise Ratio*) oraz energia na element EPRE (*ang. Energy per Resource Element*).**

Abstract: **The article presents the implementation of an innovative interference mitigation mechanism for cellular networks (including 5G) – the Anti-Interference Mechanism (AIM). Interference can significantly impact connection quality and communication performance. The implemented software enables monitoring of key parameters indicative of interferences, such as SNR and EPRE.**

Słowa kluczowe: **bezpieczeństwo cyberfizyczne, detekcja zagłuszania radiowego, odporność komunikacji bezprzewodowej, przeciwdziałanie zakłóceniom w sieci 5G.**

Keywords: **5G interference mitigation, cyber-physical security, radio jamming detection, wireless communication resilience.**

1. WSTĘP

Wiadomości przesyłane z wykorzystaniem sieci radiowych są narażone na pojawiające się w tej sieci różnego rodzaju zakłócenia, takie jak interferencje elektromagnetyczne, szumy atmosferyczne, czy kolizje sygnałów wynikające z przeciążenia pasma. Te czynniki mogą prowadzić do utraty danych, opóźnień w transmisji lub obniżenia jakości połączenia. Aby zminimalizować te problemy, stosuje się techniki takie jak korekcja błędów, modulacja adaptacyjna czy wybór mniej zatłoczonych kanałów. W przypadku krytycznych zastosowań, takich jak komunikacja w systemach bezpieczeństwa, dodatkowe zabezpieczenia, np. szyfrowanie i redundancja danych, są niezbędne do zapewnienia niezawodności. Wyróżnia się następujące typy zakłóceń w zależności od przyczyn ich powstawania, są to: zakłócenia spowodowane naturalnymi zjawiskami, zakłócenia generowane przez człowieka oraz zakłócenia związane z propagacją fal radiowych. Do zakłóceń generowanych przez człowieka możemy zaliczyć te powstałe w wyniku działania: innych urządzeń radiowych, urządzeń przemysłowych oraz celowego wprowadzania zakłóceń (np. w wojsku wprowadzane przez przeciwnika). Aby zapewnić poprawne przesyłanie wiadomości należy przeciwdziałać pojawiającym się zakłóceniom. W literaturze spotyka się różnego rodzaju metody, z których najbardziej popularne to:

- stosowanie odporniejszych metod modulacji, w tym z rozproszonym widmem,
- wykorzystanie kodów korekcyjnych,
- stosowanie anten MIMO/Massive MIMO (multiple-input-multiple-output), które w 5G umożliwiają ukierunkowaną transmisję sygnału tylko do wybranego użytkownika,
- dynamiczne zarządzanie zasobami radiowymi,
- separację logiczną i izolację sieci (tzw. network slicing),
- optymalizację sieci (np. SON – Self-Organizing Networks). Sieć 5G automatycznie wykrywa źródła zakłóceń i optymalizuje swoje ustawienia (np. moc nadawania, wybór częstotliwości). Zmniejsza to konieczność interwencji człowieka w zarządzaniu zakłóceniami.

Przykładami wykorzystania powyższych metod niech będą poniższe opracowania i wdrożenia. Grupa WB dla Wojska Polskiego opracowała taktyczną radiostacje MANET, która charakteryzuje się niską gęstością widmową promieniowanego sygnału, hoppingiem oraz wbudowanymi algorytmami szyfrującymi. Zastosowany jest również algorytm przeciwdziałania zakłóceniom oraz samoorganizujące się sieci MANET [3]. W [7] opisano zaawansowane algorytmy analizy pasma radiowego (RF) w celu automatycznego przydzielania kanałów i zarządzania mocą transmisji punktów dostępowych, co pomaga w redukcji zakłóceń. W artykule [2] na temat problemów z działaniem sieci bezprzewodowych w bliskim sąsiedztwie opisano metodę Adaptive Frequency Hopping (AFH), która pozwala na skuteczne eliminowanie interferencji poprzez dynamiczne dostosowywanie wykorzystywanych kanałów częstotliwości. W [8] autorzy analizują transmisję z krótkookresowymi zakłóceniami i rozważają wstępnie zakodowaną transmisję oraz nowe detektory co powoduje odporność na ten typ zakłóceń. W [4] autorzy w pracy zawarli kompleksowy przegląd technologii znanej jako Massive MIMO, jej wyzwań, dostępnych rozwiązań i nowych kierunków badań. W [6] przy zastosowaniu Massive MIMO w technologii 5G oparte na architekturze wielowarstwowej wykorzystującej kilka stacji bazowych o niskiej mocy wewnątrz małej komórki proponują zarządzanie zaso-

bami dzięki czemu można zminimalizować zakłócenia między kilkoma stacjami bazowymi i różnymi urządzeniami.

2. SYSTEM PRZECIWDZIAŁANIA ZAKŁÓCENIOM

Artykuł dotyczy zagadnień realizowanych w projekcie 5gStar[1], który był realizowany w ramach konsorcjum naukowego utworzonego przez WIŁ-PIB, Politechnikę Poznańską oraz Grandmetrics Sp. z.o.o. Głównym celem projektu było opracowanie metod i technik identyfikacji oraz przeciwdziałania nowym, zaawansowanym atakom na infrastrukturę dostępową oraz aplikacje 5G. W ramach projektu opracowano elementy detekcji (sonda, honeypot) oraz przeciwdziałania atakom (firewall aplikacyjny, firewall, polityka bezpieczeństwa), a także aplikacja monitorowania bezpieczeństwa pozwalająca na ocenę ryzyka i raportowanie. Opracowany system umożliwia integrację z systemami klienckimi oferując ochronę infrastruktury teleinformatycznej, wykrywanie zagrożeń, ochronę prywatności, poufności, integralności i dostępności. Wśród zagadnień rozpatrywanych w projekcie znajduje się opracowanie metod identyfikacji ataków na aplikacje sieciowe oraz mechanizm przeciwdziałania zakłóceniom w sieciach komórkowych, w tym 5G. Zakłócenia te występują w czasie działania systemu pomiędzy stacją bazową i użytkownikiem. Zadaniem oprogramowania jest wykrywanie i neutralizacja występujących tego typu zakłóceń. Mechanizm opiera się na danych, które są zbierane w stacji bazowej tj. stosunek sygnał-szum (SNR) oraz energia na jednostkę zasobu (EPRE). Dzięki zebraniu danych, zaprogramowany algorytm jest w stanie przeliczać prawdopodobieństwo wystąpienia zakłóceń i podjąć działanie. Proces detekcji zakłóceń obejmuje kilka kluczowych etapów. Zaczynając od monitorowania parametrów SNR i EPRE, detektor regularnie zbiera dane na temat stosunku sygnał-szum (SNR) oraz energii na element (EPRE). Te parametry są kluczowe dla oceny jakości połączenia i identyfikacji potencjalnych zakłóceń. Następnym procesem jest analiza danych w czasie rzeczywistym, która jest możliwa do odczytania przez użytkownika. Zebrane dane są natychmiast analizowane w celu wykrycia nieprawidłowości. Algorytmy detekcji analizują zmiany w SNR i EPRE, aby wykryć potencjalne źródła zakłóceń. Na podstawie analizy, system podejmuje odpowiednie działania w celu przeciwdziałania zakłóceniom. Wysyła do stacji bazowej kwerendę o zmniejszenie własnej mocy nadawczej, Powoduje to wzrost mocy u klientów, którzy będą w stanie „przebić" się przez występujące zakłócenia.

3. WYZWANIA ZWIĄZANE Z DETEKCJĄ ZAKŁÓCEŃ W SIECIACH 5G

Jednym z kluczowych problemów, które należało rozwiązać podczas projektowania detektora zakłóceń w sieciach 5G, było zapewnienie możliwie najkrótszego czasu reakcji na wykryte zakłócenia oraz skutecznego przeciwdziałania ich wpływowi na jakość usług. W sieciach 5G, gdzie niezawodność i jakość sygnału mają kluczowe znaczenie, nawet krótkotrwałe zakłócenia mogą prowadzić do znaczących problemów, takich jak spadek przepustowości, zwiększone opóźnienia, a w skrajnych przypadkach – całkowite rozłączenie połączenia. W przypadku sieci 5G, monitorowanie jakości sygnału może być realizowane za pomocą różnorodnych sensorów i urządzeń pomiarowych, takich jak analizatory sygnału, skanery widma oraz zaawansowane stacje bazowe wyposażone w funkcje monitorowania parametrów radiowych. Jednakże, w rzeczywistych scenariuszach, dostępność takich urządzeń pomiarowych może być ograniczona, co utrudnia ciągłe monitorowanie i natychmiastową reakcję na zmiany w jakości sygnału.

Jeśli stacje bazowe oraz urządzenia użytkowników końcowych nie są wyposażone w odpowiednie mechanizmy monitorujące, mogą nie wykrywać występujących zakłóceń, co prowadzi do pogorszenia jakości usług. W takiej sytuacji, system detekcji zakłóceń staje się kluczowym elementem, umożliwiającym skuteczne monitorowanie i zarządzanie jakością sygnału. Aby zminimalizować wpływ zakłóceń, kluczowe jest zapewnienie możliwie najkrótszego czasu przekazywania informacji o wykrytych zakłóceniach do odpowiednich modułów systemu zarządzania siecią. System detekcji zakłóceń musi być zdolny do szybkiego zbierania, analizowania i przekazywania danych dotyczących jakości sygnału, takich jak stosunek sygnał-szum (SNR) oraz energia na element (EPRE). W ten sposób możliwe jest natychmiastowe wykrycie zakłóceń i podjęcie odpowiednich działań korekcyjnych.

4. KONCEPCJA DETEKTORA

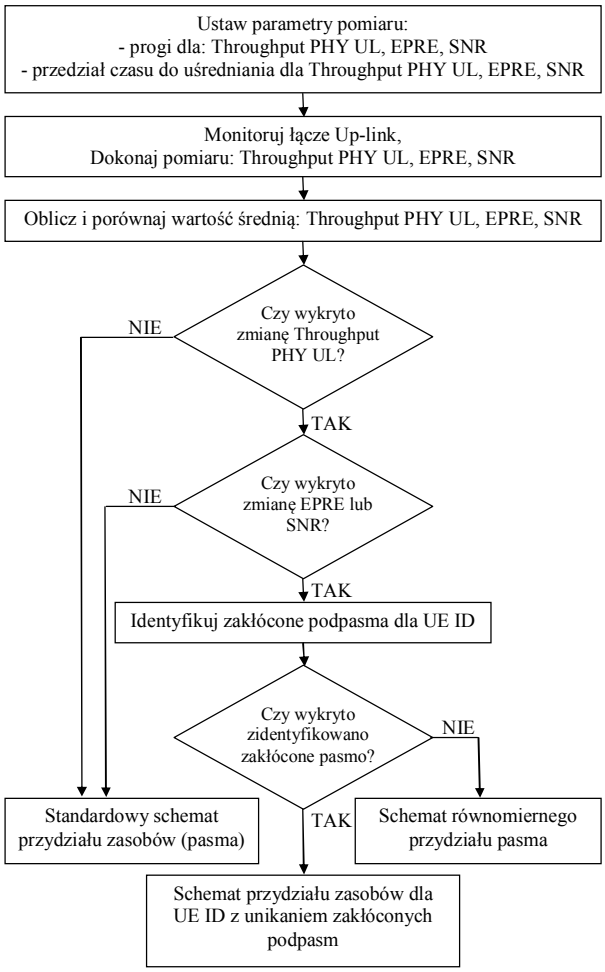
Detektor zakłóceń opracowany w ramach projektu jest nowatorskim rozwiązaniem, które łączy zaawansowaną analizę danych z możliwościami monitorowania sieci 5G w czasie rzeczywistym. Mechanizm jest oparty na monitorowaniu kluczowych parametrów sygnału, takich jak stosunek sygnał-szum (SNR) oraz energia na element (EPRE), co pozwala na dokładną ocenę jakości połączenia i identyfikację potencjalnych zakłóceń. Dzięki zastosowaniu algorytmów detekcji, jest w stanie szybko identyfikować zakłócenia i podejmować skuteczne działania w celu ich neutralizacji.

Mechanizm działa zgodnie z algorytmem przedstawionym na rysunku 1. Rozpoczyna się od importowania niezbędnych modułów, takich jak sys, time, json oraz create_connection z modułu websocket.

- sys – udostępnia funkcje i zmienne związane z interpreterem Pythona i systemem operacyjnym.
- time - obsługuje operacje związane z czasem, np. pomiar czasu, opóźnienia, formatowanie dat,
- json - służy do pracy z danymi w formacie JSON (serializacja i deserializacja),

[1] Zaawansowane metody i techniki wykrywania i przeciwdziałania atakom na infrastrukturę dostępową i aplikacje 5G. Praca finansowana przez Narodowe Centrum Badań i Rozwoju, projekt 5gStar, nr umowy: CYBERSECIDENT/487845/IV/NCBR/2021

- create_connection - importuje funkcję create_connection z modułu websocket, która umożliwia tworzenie połączeń WebSocket.



Rysunek 1 Propozycja algorytmu unikania zakłóceń [5]

Następnie zdefiniowane są dwie globalne zmienne last_good_snr i previous_change_power_value. Są zainicjalizowane na None. Będą przechowywać poprzednie wartości SNR i zmiany mocy. Funkcja change_rb otwiera połączenie websocket do adresu ws://127.0.0.1:9001 i wysyła wiadomość zmieniającą wzmocnienie komórki cell_id o identyfikatorze 2 na -30 [1]. Funkcja change_rb_reverse również otwiera połączenie websocket do tego samego adresu, ale resetuje wzmocnienie do 0. Funkcja calculate_average oblicza średnią z listy wartości i zwraca ją z zaokrągleniem do dwóch miejsc po przecinku lub 0.0 jeśli lista jest pusta. Funkcja measurement inicjuje połączenie websocket do adresu ws://127.0.0.1:9001, a następnie w pętli wysyła wiadomości ue_get z żądaniem statystyk. Zapisuje wartości pusch_snr oraz epre do odpowiednich list pusch_snr_values oraz EPRE_values W pętli wewnętrznej wykonuje 10 pomiarów w odstępach 0.5 sekundy, aż do zebrania wystarczającej liczby danych Po zebraniu danych, funkcja oblicza średnie wartości pusch_snr i epre, zamyka połączenie websocket i zwraca obie średnie. Funkcja main rozpoczyna działanie programu, wywołując change_rb, czekając 10 sekund, wywołując change_rb_reverse i ponownie czekając 10 sekund. Globalne zmienne last_good_snr oraz previous_change_power_value są ustawiane na None. Następnie w nieskończonej pętli while funkcja wywołuje measurement i przechowuje wynik w zmiennej pusch_snr_average. Jeżeli zmienna last_good_snr jest pusta oraz aktualna wartość pusch_snr_average jest o 5 mniejsza niż poprzednia wartość previous_pusch_snr_average, to zmienna last_good_snr zostaje ustawiona na poprzednią wartość pusch_snr_average. Wywoływana jest funkcja change_rb, po czym program czeka 60 sekund, następnie wywołuje change_rb_reverse i czeka 5 sekund. Ponownie wykonuje measurement i jeżeli aktualna wartość pusch_snr_average jest mniejsza o 2.5 od last_good_snr, pętla wykonuje dodatkowe zmiany mocy i pomiary aż do ustabilizowania wartości pusch_snr_average. Ostatecznie zmienna previous_pusch_snr_average jest aktualizowana na wartość pusch_snr_average.

4.1. Kluczowe cechy i funkcje detektora zakłóceń

Do kluczowych cech i funkcji detektora zakłóceń należą:

1) Monitorowanie parametrów sygnału w czasie rzeczywistym.
   System regularnie zbiera dane na temat SNR i EPRE z urządzeń sieciowych, co pozwala na bieżąco oceniać jakość połączenia. Monitorowanie w czasie rzeczywistym umożliwia natychmiastową reakcję na wszelkie zmiany w jakości sygnału.
2) Analiza danych.
   Zebrane dane są analizowane za pomocą algorytmów detekcji, które identyfikują zmiany w SNR i EPRE wskazujące na potencjalne zakłócenia. Algorytmy te są w stanie rozróżniać pomiędzy naturalnymi fluktuacjami sygnału a rzeczywistymi zakłóceniami, dzięki zastosowaniu stosunkowo prostego mechanizmu detekcji niestabilności łącza, rozumianego jako krótkotrwałe, losowe zmiany parametrów transmisyjnych, takich jak opóźnienie, przepustowość czy poziom mocy sygnału.
3) Automatyczne działania korygujące.
   Na podstawie analizy danych, detektor podejmuje automatyczne działania w celu przeciwdziałania zakłóceniom. Przykładem jest wysłanie komendy do stacji bazowej o zmniejszenie własnej mocy nadawczej, co powoduje zwiększenie mocy sygnału u użytkowników końcowych, umożliwiając im "przebicie" się przez zakłócenia. Pozwala to na kilka kluczowych korzyści, przykładowo - redukcja interferencji międzykomórkowej. Są to zakłócenia sygnału, które mogą występować między sąsiednimi komórkami w sieci komórkowej. Dzięki temu poprawia się jakość połączeń w całej sieci, a użytkownicy w różnych komórkach mogą korzystać z lepszego sygnału. Kolejnym przykładem jest elastyczność w zarządzaniu zasobami sieciowymi, gdzie operatorzy będą w stanie dynamicznie dostosowywać moc nadawczą stacji bazowych i urządzeń końcowych w zależności od bieżących warunków sieciowych i wymagań użytkowników. Redukcja mocy nadawczej stacji bazowej prowadzi do zmniejszenia jej zużycia energii. Zwiększona moc sygnału użytkowników, choć wymaga większego zużycia energii po

stronie urządzeń końcowych, jest zazwyczaj mniej znacząca w skali całego systemu.

5. MOŻLIWOŚĆ DALSZEGO ROZWOJU

Mechanizm został zaprojektowany z myślą o przyszłym rozwoju i łatwej integracji nowych funkcji. Możliwości dalszego rozwoju obejmują:

- Dodanie nowych algorytmów detekcji - możliwość implementacji bardziej zaawansowanych algorytmów uczenia maszynowego i sztucznej inteligencji, które będą w stanie lepiej identyfikować wzorce zakłóceń i przewidywać ich występowanie.
- Integracja z systemami analizy dużych zbiorów danych (*ang. Big Data*) - umożliwienie analizy większych zbiorów danych z wielu źródeł, co pozwoli na bardziej kompleksowe i dokładne monitorowanie sieci.
- Współpraca z innymi systemami zarządzania siecią: Integracja z istniejącymi systemami zarządzania siecią NMS (*ang. Network Management Systems*) w celu wymiany informacji.
- Adaptacyjna konfiguracja parametrów - wprowadzenie mechanizmów automatycznej adaptacji parametrów monitorowania i detekcji w zależności od warunków sieciowych i rodzaju zakłóceń. Operator może sam ustalić zakres wahań, które mogą zostać pominięte oraz ustawić zakres zbierania danych według własnych preferencji.

6. ADAPTACJA MECHANIZMU

Kluczową cechą zaprojektowanego mechanizmu przeciwdziałania zakłóceniom w sieci 5G jest jego elastyczność i możliwość adaptacji do zmieniających się warunków oraz specyficznych potrzeb operatorów sieci. Oprogramowanie umożliwia wprowadzanie własnych danych i dostosowywanie parametrów monitorowania oraz detekcji zakłóceń w celu zoptymalizowania jego działania. Przez możliwość dokonywania zmian w kodzie można określić:

1) Ilość pobieranych elementów EPRE i SNR. Użytkownicy mogą określić, ile próbek danych dotyczących energii na element (EPRE) oraz stosunku sygnał-szum (SNR) ma zostać pobranych w każdym cyklu pomiarowym. Dzięki temu możliwe jest dostosowanie ilości zbieranych danych do specyfiki sieci oraz wymagań dotyczących dokładności monitorowania.

2) Czas zbierania danych. Możliwość ustawienia czasu, przez jaki dane mają być zbierane.

3) Mechanizm umożliwia ustawienie progów, które określają, jak duże zmiany w wartościach EPRE oraz SNR są jeszcze uznawane za normalne, a nie za oznakę zakłóceń. Dzięki temu można odfiltrować naturalne wahania poziomu sygnału, które są nieszkodliwe i przez to uniknąć fałszywych alarmów. Pozwala to na dokładniejsze wykrywanie rzeczywistych zakłóceń w sieci.

7. PODSUMOWANIE

Przedstawiono implementację mechanizmu przeciwdziałania zakłóceniom w sieciach 5G, który jest częścią projektu 5gSTAR realizowanego przez opisane konsorcjum. Celem opracowanego mechanizmu było wykrywanie i neutralizowanie zakłóceń w komunikacji radiowej pomiędzy stacją bazową i użytkownikami. Mechanizm ten opiera się na monitorowaniu kluczowych parametrów, takich jak stosunek sygnał-szum (SNR) oraz energia na element (EPRE), które pozwalają na wykrywanie zakłóceń w czasie rzeczywistym. W artykule omówiono także wyzwania związane z detekcją zakłóceń w sieciach 5G, w tym konieczność szybkiej reakcji na zmiany w jakości sygnału, co jest kluczowe dla utrzymania wysokiej niezawodności i jakości usług. Zaprojektowany detektor zakłóceń łączy analizę danych w czasie rzeczywistym z automatycznym podejmowaniem działań korygujących, takich jak zmiana mocy nadawczej stacji bazowej, co pozwala na zmniejszenie zakłóceń międzykomórkowych i poprawę jakości połączeń w sieci. Wykazana jest również możliwość rozwoju mechanizmu, w tym implementację bardziej zaawansowanych algorytmów detekcji oraz integrację z systemami analizy dużych zbiorów danych i innymi systemami zarządzania siecią. Natomiast nie zaprezentowano metryk skuteczności, ani wyników numerycznych, które pozwoliłyby ocenić działanie mechanizmu, np. w porównaniu z innymi rozwiązaniami spotykanymi w literaturze.

LITERATURA

[1] Amarisoft. 2025. "Remote API - Python", https://tech-academy.amarisoft.com/RemoteAPI_Python.html.

[2] Dębowski Jacek. 2007. „Problemy z działaniem sieci bezprzewodowych w bliskim sąsiedztwie", ElektronikaB2B, https://elektronikab2b.pl/technika/1462-problemy-z-dzialaniem-sieci-bezprzewodowych-w-bliskim-sasiedztwie?

[3] Grupa WB. 2025. „Taktyczna radiostacja MANET PERAD 6010", online: https://www.wbgroup.pl/produkt/taktyczna-radiostacja-wojskowa-perad/.

[4] Pereira de Figueiredo Felipe Augusto. 2022."An Overview of Massive MIMO for 5G and 6G", *IEEE Latin America Transactions*, vol. 20, no. 6, 931-940.

[5] Romanik Janusz. 2023. "Algorytm unikania zakłóceń 5G", raport wewnętrzny WIŁ-PIB.

[6] Venkatesan Mithra, Kulkarni Anju; Menon Radhika, Prasad Shashikant. 2022. "Interference Mitigation Approach using Massive MIMO towards 5G networks", *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, Ravet, India, pp. 1-5.

[7] WiFi. 2020. „Architektura sieci Wi-Fi – wprowadzenie", online: https://www.wifi.com.pl/wifi/technologia-wifi/architektura-sieci-wifi-wprowadzenie.

[8] Zhou Qi, Ma Xiaoli. 2014."Shot Interference Detection and Mitigation for Heterogeneous Networks", *IEEE Transactions on Vehicular Technology*, vol. 63, no. 1, 225-236.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

Metody redukcji strat propagacyjnych w sieciach wspomaganych technologią IPR

Methods for Reducing Path Loss in RIS-Aided Networks

Adam Samorzewski[1]; Adrian Kliks[2,3];

[1] Politechnika Poznańska, Poznań, Polska, adam.samorzewski@doctorate.put.poznan.pl
[2] Politechnika Poznańska, Poznań, Polska, adrian.kliks@put.poznan.pl
[3] Uniwersytet Techniczny w Luleå, Luleå, Szwecja, adrian.kliks@ltu.se

DOI: 10.15199/59.2025.4.28

Streszczenie: W artykule przedstawiono analizę propagacji sygnału radiowego w Poznaniu w kontekście rozmieszczenia Inteligentnych Powierzchni Reconfigurowalnych w sieciach 5G/6G. Badanie obejmuje osiem stacji bazowych z antenami SISO lub MIMO, a symulacje uwzględniają warunki LOS i NLOS w złożonym środowisku miejskim. Wyniki pokazują, że odpowiednie rozmieszczenie inteligentnych matryc znacząco redukuje tłumienie propagacyjne i poprawia zasięg sygnału, zwłaszcza tam, gdzie standardowe stacje bazowe są niewystarczające. Technologia powierzchni rekonfigurowalnych okazuje się również opłacalnym sposobem na poprawę jakości usług bez konieczności gęstego rozmieszczania stacji bazowych.

Abstract: This paper presents an analysis of radio signal propagation in the city of Poznan, focusing on the deployment of Reconfigurable Intelligent Surfaces in 5G/6G networks. The study involves eight base stations equipped with SISO or MIMO antennas. Simulations consider both LOS and NLOS conditions in a complex urban environment. The results show that proper placement of intelligent arrays significantly reduces path loss and improves signal coverage, especially in areas where conventional base stations are insufficient. Reconfigurable surface technology also proves to be a cost-effective way to enhance service quality without the need for dense deployment of base stations.

Słowa kluczowe: Inteligentne Powierzchnie Reconfigurowalne, Komunikacja Bezprzewodowa, Optymalizacja Strat Propagacyjnych, Propagacja Sygnału Radiowego, Sieci Mobilne.

Keywords: Mobile Networks, Path Loss Optimization, Radio Signal Propagation, Reconfigurable Intelligent Surfaces, Wireless Communications.

1. WSTĘP

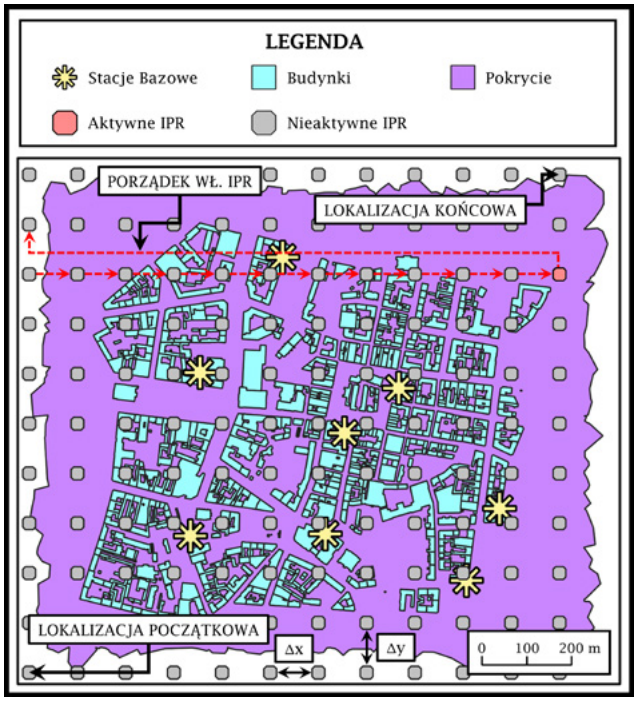
Inteligentne Powierzchnie Reconfigurowalne (IPR) to nowatorska technologia, która umożliwia dynamiczną modyfikację środowiska propagacyjnego w systemach komunikacji bezprzewodowej. W przeciwieństwie do klasycznych metod formowania wiązki po stronie nadajnika, IPR wykorzystują matryce pasywnych lub aktywnych elementów do sterowania odbiciem fal elektromagnetycznych w czasie rzeczywistym. Dzięki temu możliwe jest kierowanie wiązką, ograniczanie zakłóceń i minimalizacja strat propagacyjnych (*ang. Path Loss* – PL), co przekłada się na lepszy zasięg, szczególnie w trudnych warunkach miejskich i rozległych obszarach wiejskich [1]–[6].

W artykule przeanalizowano wpływ rozmieszczenia matryc IPR na tłumienie sygnału radiowego w miejskim środowisku Poznania. Uwzględniono różne lokalizacje IPR, oceniając ich skuteczność w poprawie zasięgu.

Struktura artykułu: rozdział 2 przedstawia scenariusz badawczy, rozdział 3 opisuje konfigurację symulacyjną, rozdział 4 prezentuje wyniki, a rozdział 5 zawiera podsumowanie i wnioski.

2. ROZWAŻANY SCENARIUSZ

Scenariusz badawczy przedstawiony na Rys. 1 został osadzony w Poznaniu (Stary Rynek i okolice), z wykorzystaniem rzeczywistych danych przestrzennych [7].



Rysunek 1: Schemat metodyki badania

Sieć mobilna obejmuje osiem Stacji Bazowych (SB), każda obsługująca trzy komórki pracujące na 800, 2100

i 3500 MHz. Lokalizacje SB ustalono na podstawie danych operatora komórkowego [8]. Anteny SISO (*ang. Single Input Single Output*) zastosowano na 800 i 2100 MHz, natomiast MIMO (*ang. Multiple Input Multiple Output*) z 64 elementami na 3500 MHz [6], [9].

Zidentyfikowano 132 potencjalne lokalizacje dla matryc IPR, rozmieszczone według siatki generowanej automatycznie przez oprogramowanie symulacyjne.

Dla każdej lokalizacji IPR przeprowadzono pomiary strat propagacyjnych, tworząc cztery przypadki symulacyjne: z wyłączonymi i włączonymi IPR (minimalne tłumienie), średnie tłumienie przy aktywnych IPR oraz minimalne tłumienie przy wzmocnieniu sygnałów w relacji IPR–SB. Wyniki posłużyły do dalszej analizy, w tym generowania funkcji CDF (*ang. Cumulative Distribution Function*) oraz tabel oceny wydajności systemu.

3. KONFIGURACJA SYMULACYJNA

Badanie zostało przeprowadzone w formie symulacji z wykorzystaniem oprogramowania o nazwie GRAND (*ang. Green Radio Access Network Design*) [9, 6]. Narzędzie GRAND to autorskie oprogramowanie zaimplementowane w języku Java służące do planowania sieci mobilnych. GRAND umożliwia projektowanie sieci na podstawie danych wejściowych dotyczących zasięgu, zabudowy, SB oraz technologii, a także pozwala na integrację elementów wspomagających, takich jak IPR. Konfigurację symulacji oparto na rzeczywistych danych pobranych z baz danych Urzędu Miasta Poznania [7] i jednego z polskich operatorów komórkowych [8].

Głównym celem badania była analiza rozkładu strat propagacyjnych sygnału radiowego na wybranym obszarze miejskim — zarówno bez zastosowania, jak i z zastosowaniem matryc IPR w różnych lokalizacjach — w celu identyfikacji ich najbardziej optymalnego rozmieszczenia. Do oceny tłumienia sygnału radiowego na całym obszarze zastosowano następujące modele propagacyjne:

- 3GPP TR 38.901 UMa (*ang. Urban Macro*) – propagacja sygnału w scenariuszach LOS (*ang. Line Of Sight*) i NLOS (*ang. Non-Line Of Sight*), gdzie fale radiowe są transmitowane *bezpośrednio* z SB [10],
- RIS-FFBC (*ang. Reconfigurable Intelligent Surface Far Field Beamforming Case*) – propagacja sygnału z SB przez odbicie fal radiowych od matryc IPR [3].

W oparciu o wspomniane modele, przeprowadzono symulacje strat propagacyjnych analizując cztery scenariusze: bez IPR (BS – *ang. Base Stations*), z aktywowanymi IPR wybierając minimalną wartość PL (RIS – *ang. Reconfigurable Intelligent Surfaces*), z IPR wzmacniającym sygnał z SB (RIS,BS) oraz uśredniającym sygnały z obu źródeł (AVG – *ang. Average*).

Założono możliwość obracania IPR w celu optymalizacji strat propagacyjnych, uzyskując stałe wartości kątów padającej i odbitej wiązki sygnału radiowego: $\theta_t = \frac{\pi}{4}$, $\varphi_t = \pi$, $\theta_r = \frac{\pi}{4}$, $\varphi_r = 0$, zgodnie z wynikami z [3]. Ponadto przyjęto, że matryce IPR spełniają warunek pola dalekiego względem SB i punktów pomiarowych [3].

Konfiguracja symulacji obejmuje 132 powtórzenia, z których każde koncentruje się na oszacowaniu rozkładu strat propagacyjnych w badanym obszarze — zarówno z matrycą IPR umieszczoną w z góry określonej lokalizacji, jak i bez niej. Parametry wykorzystane w symulacji przedstawiono w Tab. 1 oraz 2.

Tabela 1: Konfiguracja parametrów (SB) [6]–[11]

	Parametr	Znak	Jednostka	Wartość		
				Komórka Sieciowa		
				1.	2.	3.
Gen.	Liczba	K_{BS}	–	8		
	Technologia	–	–	5G		
Pasmo	Częstotliwość	f	[MHz]	800	2100	3500
	Szerokość pasma kanału	B_w	[MHz]	80	120	120
	L. użytych podnośne	$N_{SC,u}$	–	320		
	L. wszyst. podnośnych	$N_{SC,t}$	–	512		
	Współ. próbkowania	SF	–	1.536		
	Współ. ponow. użycia pilota	RF	–	1		
	Czas koherencji	t_c	[ms]	50		
	Szer. pasma koherencji	B_c	[MHz]	1		
	Prze. cykl pracy	S	[%]	0	0	25
Transceivery	Wysokość zawie. anteny	h_{BS}	[m]	(27, 47)		
	L. elemen. antenowych	M_{BS}	–	1	1	64
	Zysk antenowy	G_a	[dBi]	16	18	24
	Tł. kabla antenowego	L_f	[dBi]	2	2	3
	Maks. moc nadaw.	P_{TX}	[dBm]	43		
	Współ. szumów	NF	[dB]	8	8	7
Propagacja	Model tł. ścieżki	–	–	3GPP TR 38.901 UMa [10]		
	Mar. interferencji	IM	[dB]	2		
	Mar. Dopplera	DM	[dB]	3		
	Mar. zaników	FM	[dB]	10		
	Mar. shadowingu	SM	[dB]	12.8	15.2	10
	Implementation Loss	IL	[dB]	0	0	3

Tabela 2: Konfiguracja parametrów (IPR) [3]

Parametr	Znak	Jednostka	Wartość
Liczba lokalizacji	K_{RIS}	–	132
Liczba ele. kolumny IPR	M_{RIS}	–	102
Liczba ele. wiersza IPR	N_{RIS}	–	100
Szerokość kolumny IPR	d_m	[m]	0.01
Szerokość wiersza IPR	d_n	[m]	0.01
Współ. amp. syg. odbitego	A	–	0.9
Wysokość zawieszenia	h_{RIS}	[m]	40
Model tł. ścieżki	–	–	RIS-FFBC [3]

4. WYNIKI

W następnej sekcji przedstawiono wyniki uzyskane dla badanego scenariusza. Na początku omówione zostają parametry ilustrujące pozytywny wpływ wdrożenia technologii IPR na rozkład strat propagacyjnych sygnału radiowego. Następnie zaprezentowano wyniki w formie tabel (Tab. 3 i 4) oraz wykresów (Rys. 2), po czym następuje ich analiza i omówienie.

4.1 Miary wydajności

Do oceny poprawy strat propagacyjnych sygnału radiowego w scenariuszu z zastosowaniem urządzeń IPR, w porównaniu do konfiguracji bez jakichkolwiek elementów wspomagających (właśnie takich jak matryce IPR), wykorzystano parametry zdefiniowane przez Rów. (1)–(12).

Równania (1) oraz (2) definiują wzory użyte do oszacowania minimalnych wartości PL w określonym punkcie mapy dla scenariusza z aktywnymi IPR $\left(\dot{\mathrm{P}}\mathrm{L}_{\mathrm{RIS}}\right)$ oraz scenariusza referencyjnego bez IPR $\left(\dot{\mathrm{P}}\mathrm{L}_{\mathrm{BS}}\right)$. Dodatkowo, równanie (3) oblicza średnią wartość strat propagacyjnych, uwzględniając zarówno sygnał odbierany *bezpośrednio* (w warunkach LOS lub NLOS) z SB, jak i sygnał odbity od najkorzystniej umieszczonej matrycy IPR dla danej lokalizacji (punktu pomiarowego).

$$\dot{\mathrm{PL}}_{\mathrm{RIS}}(x,y) = \min_{i,j}\Big\{\mathrm{PL}_{\mathrm{LOS},i}(x,y)\,;\mathrm{PL}_{\mathrm{NLOS},i}(x,y)\,;\mathrm{PL}_{\mathrm{RIS},i,j}(x,y)\Big\}, \quad (1)$$

$$\dot{\mathrm{PL}}_{\mathrm{BS}}(x,y) = \min_{i}\Big\{\mathrm{PL}_{\mathrm{LOS},i}(x,y)\,;\mathrm{PL}_{\mathrm{NLOS},i}(x,y)\Big\}, \quad (2)$$

$$\dot{\mathrm{PL}}_{\mathrm{AVG}}(x,y) = \frac{\dot{\mathrm{PL}}_{\mathrm{RIS}}(x,y) + \dot{\mathrm{PL}}_{\mathrm{BS}}(x,y)}{2}, \quad (3)$$

gdzie $x = 0,\ldots,X-1$ oraz $y = 0,\ldots,Y-1$ oznaczają współrzędne konkretnego punktu pomiarowego strat propagacyjnych, przy czym X i Y odpowiadają całkowitej liczbie rozpatrywanych współrzędnych w odpowiednich kierunkach. Podobnie, $i = 0,\ldots,K_{\mathrm{BS}}-1$ oraz $j = 0,\ldots,K_{\mathrm{RIS}}-1$ pełnią funkcję identyfikatorów komórek SB oraz urządzeń IPR (gdzie j zależy od bieżącej lokalizacji matrycy), przy czym K_{BS} i K_{RIS} oznaczają łączną liczbę, odpowiednio, rozlokowanych komórek SB i lokalizacji IPR. Wartości $\mathrm{PL}_{\mathrm{LOS}}$, $\mathrm{PL}_{\mathrm{NLOS}}$ oraz $\mathrm{PL}_{\mathrm{RIS}}$ odpowiadają stratom propagacyjnym doświadczanym przez terminale mobilne przy odbiorze sygnału odpowiednio: bezpośrednio (LOS), poprzez zakłócające przeszkody (NLOS) oraz przez odbicia od powierzchni IPR.

Przed wprowadzeniem ostatniego podstawowego parametru strat propagacyjnych, istotne jest wyjaśnienie dodatkowych czynników uwzględnionych w równaniach (4) oraz (5). Pierwsze z nich definiuje średnią moc nadawczą $\left(\overline{P}_{\mathrm{TX}}\right)$ dla pojedynczej komórki sieciowej w badanym obszarze. Wzór ten wyprowadzono w następujący sposób:

$$\overline{P}_{\mathrm{TX}} = \frac{1}{K_{\mathrm{BS}}}\sum_{i=0}^{K_{\mathrm{BS}}-1} P_{\mathrm{TX},i}, \quad (4)$$

gdzie $P_{\mathrm{TX},i}$ oznacza moc nadawczą i-tej komórki sieciowej ($i = 0,\ldots,K_{\mathrm{BS}}-1$). Parametr ten, reprezentujący średnią moc nadawczą, jest następnie wykorzystywany w kolejnym równaniu do obliczenia mocy sygnału odbieranego w określonej lokalizacji na badanym obszarze (P_{RX}). Obliczenie mocy odbieranej odbywa się według wzoru:

$$P_{\mathrm{RX}}(x,y) = \overline{P}_{\mathrm{TX}} - \dot{\mathrm{PL}}(x,y). \quad (5)$$

Po wprowadzeniu równań (4) oraz (5), możemy teraz zdefiniować ostatni kluczowy parametr strat propagacyjnych, który jest istotny dla oceny korzyści wynikających z wdrożenia technologii IPR w sieci dostępu radiowego (*ang. Radio Access Network* – RAN). Poniższe równanie przedstawia minimalne tłumienie propagacyjne sygnału radiowego w scenariuszu, w którym w danym punkcie pomiarowym dwa sygnały — jeden nadany w warunkach LOS lub NLOS z SB, a drugi odbity przez IPR — konstruktywnie interferują $\left(\dot{\mathrm{PL}}_{\mathrm{RIS,BS}}\right)$:

$$\dot{\mathrm{PL}}_{\mathrm{RIS,BS}}(x,y) = \overline{P}_{\mathrm{TX}} - 10\cdot\log_{10}\Big(P_{\mathrm{RX,RIS}}(x,y) + P_{\mathrm{RX,BS}}(x,y)\Big), \quad (6)$$

gdzie $P_{\mathrm{RX,RIS}}$ oraz $P_{\mathrm{RX,BS}}$ oznaczają moce sygnałów odbieranych w określonym punkcie pomiarowym w badanym obszarze, pochodzących odpowiednio z komórki sieciowej poprzez powierzchnię IPR oraz bezpośrednio w warunkach LOS lub NLOS. Ponieważ każda komórka sieciowa nadaje sygnał radiowy z jednakową mocą, można przyjąć założenie, że $\overline{P}_{\mathrm{TX}} = P_{\mathrm{TX}}$.

Poniższe równania opisują metodę zastosowaną do wyznaczenia minimalnych $\left(\widetilde{\mathrm{PL}}\right)$, maksymalnych $\left(\widehat{\mathrm{PL}}\right)$ oraz średnich $\left(\overline{\mathrm{PL}}\right)$ wartości strat propagacyjnych. Wartości te zostały obliczone na podstawie oszacowanego rozkładu strat sygnału radiowego w analizowanym obszarze.

$$\widetilde{\mathrm{PL}} = \min_{x,y}\left\{\dot{\mathrm{PL}}(x,y)\right\}, \quad (7)$$

$$\widehat{\mathrm{PL}} = \max_{x,y}\left\{\dot{\mathrm{PL}}(x,y)\right\}, \quad (8)$$

$$\overline{\mathrm{PL}} = \frac{1}{XY}\sum_{y=0}^{Y-1}\sum_{x=0}^{X-1}\dot{\mathrm{PL}}(x,y). \quad (9)$$

Na końcu, opisane powyżej parametry zostały wykorzystane do procentowej oceny wzrostu mocy sygnału radiowego w analizowanym obszarze, wynikającej z rozmieszczenia urządzeń IPR w różnych lokalizacjach. Wskaźniki zysku wyrażono w procentach w równaniach (10)–(12), gdzie wartością referencyjną są tłumienie propagacyjne zmierzone w scenariuszu bazowym bez zastosowania IPR (BS). Kluczowe wskaźniki obejmują minimalne tłumienie propagacyjne zarejestrowane w danym punkcie przy aktywnym IPR (RIS) oraz średnią moc sygnału odbieranego w określonej lokalizacji, uzyskaną przez połączenie sygnału bezpośredniego (LOS lub NLOS) z najkorzystniejszym sygnałem odbitym od IPR (RIS,BS).

$$G_{\mathrm{PL}}^{\min} = \left[1 - \frac{\left\{\widetilde{\mathrm{PL}}_{\mathrm{RIS}} \vee \widetilde{\mathrm{PL}}_{\mathrm{RIS,BS}}\right\}}{\widetilde{\mathrm{PL}}_{\mathrm{BS}}}\right]\cdot 100\ \%, \quad (10)$$

$$G_{\mathrm{PL}}^{\max} = \left[1 - \frac{\left\{\widehat{\mathrm{PL}}_{\mathrm{RIS}} \vee \widehat{\mathrm{PL}}_{\mathrm{RIS,BS}}\right\}}{\widehat{\mathrm{PL}}_{\mathrm{BS}}}\right]\cdot 100\ \%, \quad (11)$$

$$G_{\mathrm{PL}}^{\mathrm{avg}} = \left[1 - \frac{\left\{\overline{\mathrm{PL}}_{\mathrm{RIS}} \vee \overline{\mathrm{PL}}_{\mathrm{RIS,BS}}\right\}}{\overline{\mathrm{PL}}_{\mathrm{BS}}}\right]\cdot 100\ \%. \quad (12)$$

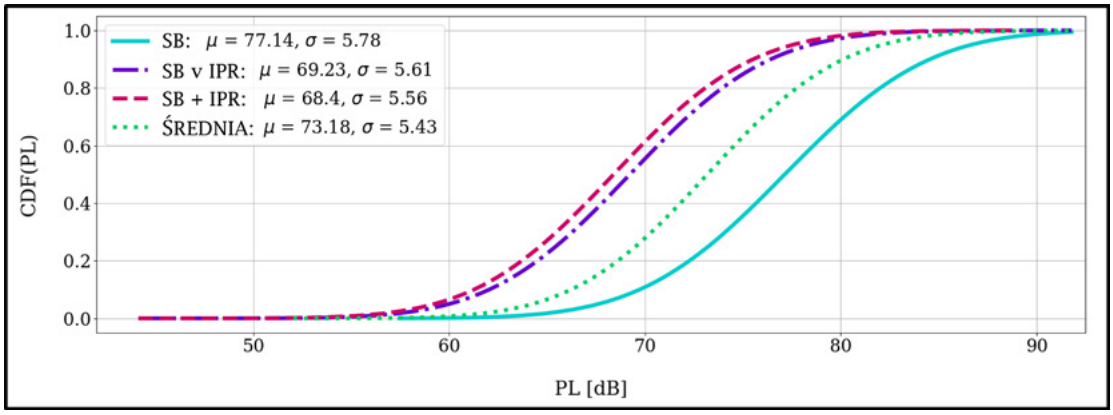
4.2 Analiza rezultatów

Tab. 3 przedstawia minimalne, maksymalne i średnie wartości PL dla czterech scenariuszy: bez IPR (BS), z IPR (RIS), z superpozycją sygnałów (RIS,BS) oraz z uśrednianiem (AVG). Uwzględniono sygnały nadawane *bezpośrednio* (LOS/NLOS) oraz odbite przez IPR.

Wartości z Tab. 3 posłużyły także do obliczenia zysków, przedstawionych w Tab. 4. Najlepsze rezultaty osiągnięto w podejściu RIS,BS, gdzie zredukowano tłumienie do: min. 44,13 dB, maks. 88,77 dB, średnio 68,4 dB, co oznacza poprawę odpowiednio o ok. 23,2 %, 3,28 % i 11,33 % względem przypadku BS.

Podejście z IPR dało zyski: 23,03 % (min.), 0 % (maks.), 10,25 % (średnia), z wartościami: 44,23 dB, 91,78 dB i 69,23 dB. Najmniejsze korzyści uzyskano dla AVG: zysk 9,42 % (min.), 0 % (maks.), 5,12 % (średnia), przy wartościach 52,05 dB, 91,78 dB i 73,18 dB.

Skuteczność poszczególnych podejść dodatkowo potwierdzają wykresy CDF (Rys. 2), ukazujące rozkład strat i udział punktów pomiarowych w danym przedziale tłumienia. Podkreślają one efektywność technologii IPR, zwłaszcza w gęsto zabudowanych środowiskach miejskich.



Rysunek 2: Wykresy CDF rozkładów PL dla wszystkich scenariuszy (BS – niebieski, RIS – fioletowy, RIS,BS – czerwony, AVG – zielony)

Tabela 3: Wpływ IPR na rozkład PL

Parametr	Wartość			
	SB (BS)	*SB* ∨ *IPR* (RIS)	*SB* + *IPR* (RIS,BS)	*ŚREDNIA* (AVG)
PL	57.46 dB	44.23 dB	44.13 dB	52.05 dB
$\widetilde{\text{PL}}$	91.78 dB	91.78 dB	88.77 dB	91.78 dB
$\overline{\text{PL}}$	77.14 dB	69.23 dB	68.4 dB	73.18 dB

Tabela 4: Zysk w redukcji PL przy użyciu IPR

Parameter	Wartość		
	SB ∨ *IPR* (RIS)	*SB* + *IPR* (RIS,BS)	*ŚREDNIA* (AVG)
$G_{\text{PL}}^{\text{min}}$	23.03 %	23.2 %	9.42 %
$G_{\text{PL}}^{\text{max}}$	0 %	3.28 %	0 %
$G_{\text{PL}}^{\text{avg}}$	10.25 %	11.33 %	5.12 %

5. WNIOSKI

Badanie wykazało, że zastosowanie IPR skutecznie redukuje tłumienie propagacyjne w sieciach mobilnych, szczególnie w podejściu ze wzmocnieniem sygnału (RIS,BS). Scenariusze z aktywnymi IPR (RIS) i uśrednianiem (AVG) również przyniosły poprawę, choć w mniejszym stopniu. Wnioski te potwierdzają wykresy CDF, wskazując na potencjał IPR w środowiskach miejskich.

PODZIĘKOWANIA

Praca została zrealizowana w ramach projektu nr 2021/43/B/ST7/01365 ufundowanego przez Narodowe Centrum Nauki (NCN) w Polsce.

LITERATURA

[1] C. Huang, A. Zappone, G.C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable Intelligent Surfaces for Energy Efficiency in Wireless Communication," *IEEE Transactions on Wireless Communications*, vol. 18, no. 8, pp. 4157—4170, 2019. DOI: 10.1109/TWC.2019.2922609.

[2] M. Di Renzo et al., "Smart Radio Environments Empowered by Reconfigurable Intelligent Surfaces: How It Works, State of Research, and The Road Ahead," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 11, pp. 2450–2525, 2020. DOI: 10.1109/JSAC.2020.3007211.

[3] W. Tang et al., "Wireless Communications With Reconfigurable Intelligent Surface: Path Loss Modeling and Experimental Measurement," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 421–439, 2021. DOI: 10.1109/TWC.2020.3024887.

[4] Y. Liu et al., "Reconfigurable Intelligent Surfaces: Principles and Opportunities," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1546–1577, 2021. DOI: 10.1109/COMST.2021.3077737.

[5] Y. Yuan, R. He, B. Ai, Z. Zhang, and Y. Jin, "RIS-Assisted Mobile Channels With Directional Transmission: Modeling and Characteristic Analysis," *IEEE Transactions on Wireless Communications*, vol. 23, no. 11, pp. 17021–17036, 2024. DOI: 10.1109/TWC.2024.3449991.

[6] A. Samorzewski, M. Deruyck, and A. Kliks, "Energy Consumption in RES-Aware 5G Networks," in *GLOBECOM 2023 – 2023 IEEE Global Communications Conference*, Kuala Lumpur, Malaysia, 2023, pp. 1024–1029. DOI: 10.1109/GLOBECOM54140.2023.10437451.

[7] *Poznan – Model 3D*. SIP (ang. Spatial Information System). [Online]. Available: http://sip.poznan.pl/model3d/#/legend.

[8] *Database and map of BTS station locations / UKE permits*. BTSearch. [Online]. Available: http://beta.btsearch.pl.

[9] G. Castellanos, S. De Gheselle, L. Martens, N. Kuster, W. Joseph, M. Deruyck, and S. Kuehn, "Multi-objective optimization of human exposure for various 5G network topologies in Switzerland," *Computer Networks*, vol. 2016, 2022. DOI: 10.1016/j.comnet.2022.109255.

[10] 3GPP, "Technical Specification Group Radio Access Network; Study on channel model for frequencies from 0.5 to 100 GHz (Release 18)," TR 38.901 v18.0.0, 2024.

[11] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency," *Foundations and Trends® in Signal Processing*, vol. 11, no. 3–4, pp. 154–655, 2017. DOI: 10.1561/2000000093.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

Ocena dokładności estymacji charakterystyki odbiciowej matrycy RMA na podstawie modelu Tanga

Evaluation of the accuracy of the reflectivity characteristics of the RIS matrix based on the Tang model

Paweł Hatka[1]; Karolina Lenarska[2]; Adrian Kliks[3];

[1] Politechnika Poznańska, Poznań, pawel.hatka@student.put.poznan.pl
[2] Politechnika Poznańska, Poznań, karolina.lenarska@put.poznan.pl
[3] Politechnika Poznańska, Poznań, adrian.kliks@put.poznan.pl

DOI: 10.15199/59.2025.4.29

Streszczenie: Jednym z intensywnie rozwijanych obszarów telekomunikacji bezprzewodowej ostatnich lat jest praktyczne zastosowanie rekonfigurowalnych macierzy antenowych (RMA). W pracy przedstawiono wyniki prac eksperymentalnych, których celem była weryfikacja dokładności teoretycznego modelu Tanga opisującego zachowanie sygnału w momencie odbicia od powierzchni RMA. Badania przeprowadzono w laboratorium będącym przykładem typowego pomieszczenia biurowego.

Abstract: One of the intensively developed areas of wireless telecommunications in recent years is the practical application of reconfigurable intelligent surfaces (RIS). The paper presents the results of experimental work aimed at verifying the accuracy of the theoretical Tang model describing the behavior of the signal at the moment of reflection from the RIS. The research was carried in a laboratory being an example of a typical office space.

Słowa kluczowe: Rekonfigurowalne Matryce Antenowe, model analityczny, weryfikacja eksperymentalna

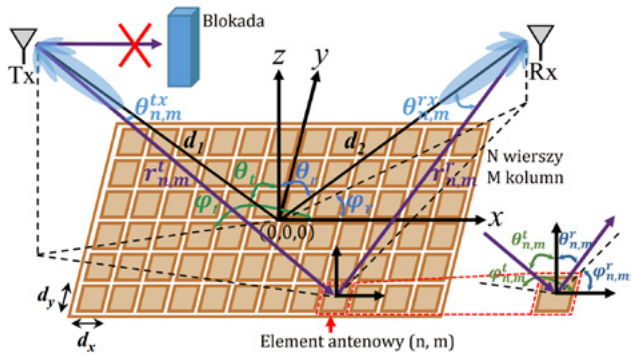
Keywords: Reconfigurable Intelligent Surfaces, analytic model, experimental verification

1. WSTĘP

Rekonfigurowalne Macierze Antenowe (RMA), czyli Reconfigurable Intelligent Surfaces (RIS), stanowią jeden z niezwykle intensywnie rozwijanych obszarów telekomunikacji bezprzewodowej. Popularność ta wynika z innowacyjnego podejścia do realizacji systemów bezprzewodowych, w którym zakłada się możliwość dynamicznego wpływania na środowisko propogacyjne poprzez zmianę sposobu odbicia sygnału od matryc RMA. W swojej podstawowej formie matryce RMA są urządzeniami pasywnymi, składajacymi się z licznych elementów antenowych, indywidualnie sterowanych w celu osiągnięcia określonej charakterystyki odbicia. Uzyskuje się to poprzez precyzyjną modyfikację amplitudy i fazę odbijanych fal elektromagnetycznych. W literaturze odnaleźć można różnorodne przykłady realizacji macierzy RMA o różnej rozdzielczości sterowania [1, 3]. Dokładne modelowanie działania macierzy RMA to jedno z kluczowych wyzwań. Na szczególną uwagę zasługują prace [4, 5], które stosunkowo precyzyjnie opisują charakterystykę odbiciową RMA. Zaproponowany model, opiera się na pewnych uproszczeniach — między innymi zakłada stałą w czasie i identyczną dla wszystkich elementów amplitudę odbicia. Weryfikacja tego modelu była prowadzona w warunkach idealnych, tj. w komorze bezechowej. Z praktycznego punktu widzenia istotne jest jednak zbadanie, jak model ten sprawdza się w realistycznych środowiskach. W pracy przedstawiono ocenę dokładności modelu Tanga w estymacji mocy sygnału odbieranego w przypadku zastosowania macierzy RMA dla środowisk biurowych. Celem pracy jest próba odpowiedzi na pytanie, jak dokładnie zaproponowany model odzwierciedla rzeczywistą propagację sygnału oraz czy istnieje potrzeba jego dalszego dostosowania w celu zwiększenia dokładności predykcji. Wstępne wyniki stanowią punkt wyjścia dla dalszych prac, których celem będzie rozwinięcie modelu Tanga z uwzględnieniem specyfiki bardziej złożonych środowisk rzeczywistych.

2. MODEL TEORETYCZNY TANGA

W niniejszej pracy estymacja mocy odebranej została przeprowadzona w oparciu o model teoretyczny zaproponowany przez Tanga w pracy [4]. W rozważanym scenariuszu komunikacja bezprzewodowa między nadajnikiem a odbiornikiem jest wspomagana przez matrycę RMA, ze względu na brak bezpośredniej ścieżki sygnału między nimi, co zilustrowano na Rysunku 1.



Rysunek 1: Komunikacja wspomagana przez RMA bez bezpośredniej ścieżki między nadajnikiem a odbiornikiem [4]

Wzór na moc sygnału odebranego, zastosowany w

niniejszym artykule, dotyczy scenariusza, w którym odbiornik znajduje się w polu bliskim macierzy RMA. W takim przypadku, dzięki odpowiedniej konfiguracji elementów antenowych matrycy RMA, możliwe jest ukierunkowanie odbitego sygnału w stronę odbiornika, co pozwala na maksymalizację mocy sygnału odbieranego. Wzór przedstawia się następująco:

$$P_r = P_t \frac{G_t G_r G d_x d_y \lambda^2 A^2}{64\pi^3} \cdot \cdot \left| \sum_{m=1-\frac{M}{2}}^{\frac{M}{2}} \sum_{n=1-\frac{N}{2}}^{\frac{N}{2}} \sqrt{F^{\mathrm{C}}_{n,m}} \cdot \frac{e^{\frac{-j(2\pi(r^t_{n,m}+r^r_{n,m})-\lambda\phi_{n,m})}{\lambda}}}{r^t_{n,m} r^r_{n,m}} \right|^2 \quad (1)$$

gdzie P_t oznacza moc sygnału transmitowanego przez antenę nadawczą, G_t i G_r to odpowiednio zysk anteny nadawczej oraz zysk anteny odbiorczej, natomiast G reprezentuje zysk pojedynczego elementu antenowego matrycy RMA. Symbol λ odnosi się do długości fali, natomiast A oznacza amplitudę sygnału dla każdego elementu antenowego (model zakłada jednakową amplitudę dla wszystkich elementów). Parametr $\phi_{n,m}$ oznacza fazę przypisaną elementowi o indeksie (n, m), natomiast $F^{\mathrm{C}}n, m$ to zbiorcza charakterystyka promieniowania uwzględniająca znormalizowaną charakterystykę promieniowania nadajnika (F^{tx}), odbiornika (F^{rx}) oraz elementu antenowego (F) zgodnie ze wzorem:

$$F^{\mathrm{C}}_{n,m} = F^{\mathrm{tx}}\left(\theta^{\mathrm{tx}}_{n,m}, \varphi^{\mathrm{tx}}_{n,m}\right) \cdot F\left(\theta^{t}_{n,m}, \varphi^{t}_{n,m}\right) \cdot F\left(\theta^{r}_{n,m}, \varphi^{r}_{n,m}\right) \cdot F^{\mathrm{rx}}\left(\theta^{\mathrm{rx}}_{n,m}, \varphi^{\mathrm{rx}}_{n,m}\right) \quad (2)$$

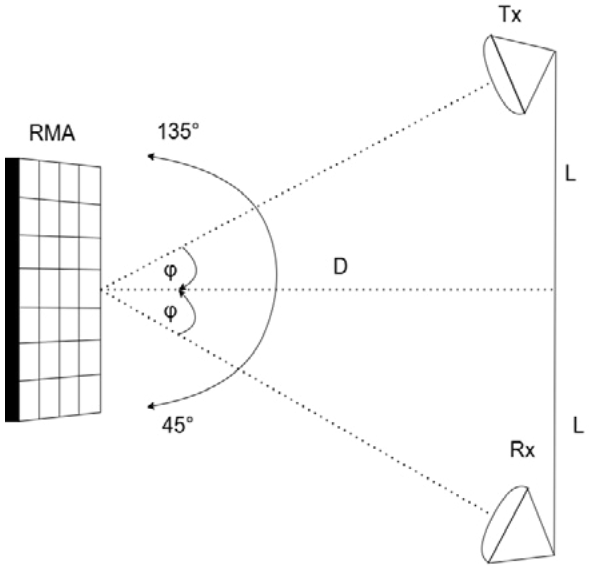
3. ŚRODOWISKA POMIAROWE

Pomiary przeprowadzono dla trzech różnych konfiguracji, różniących się zarówno środowiskiem propagacyjnym, jak i zastosowanymi antenami. Pierwszy zestaw pomiarów wykonano w środowisku nieidealnym, wykorzystując anteny kierunkowe firmy Ainfo (parametry anten zamieszczono w Tabeli 1).

Tabela 1: Parametry wykorzystywanych anten

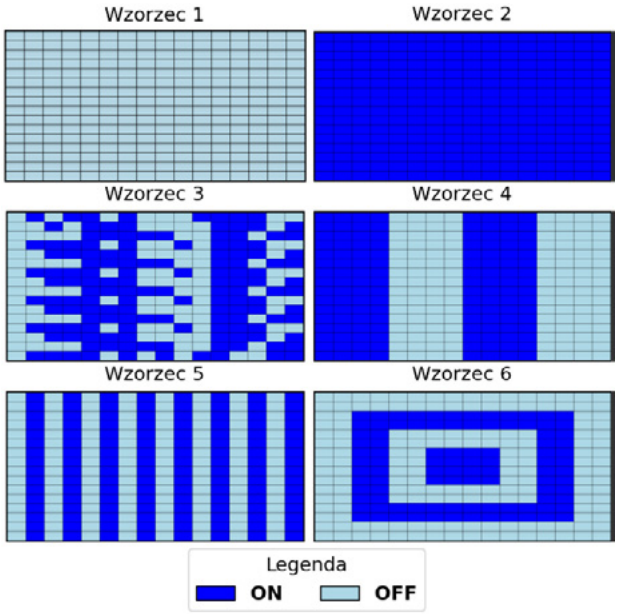
Parametr	**METIS**	**AINFO**
Model	14HV	LB-187-20-C-SF
Częst. [GHz]	5.1 – 5.9	3.95 – 5.85
Zysk [dBi]	14	20
Kąt 3dB – E [°]	32	16
Kąt 3dB – H [°]	35	18

Drugi scenariusz realizowano również w środowisku nieidealnym, tym razem z wykorzystaniem anten mikropaskowych firmy Metis. Scenariusz ten został szczegółowo opisany w artykule [2]. Trzeci zestaw pomiarów zrealizowano w warunkach idealnych – w komorze bezechowej – również przy użyciu anten firmy Metis. Zachowanie tej samej konfiguracji antenowej pozwoliło na pełną porównywalność wyników i umożliwiło izolowaną analizę wpływu środowiska propagacyjnego na efektywność RMA. Eliminacja odbić i zakłóceń charakterystycznych dla środowisk rzeczywistych pozwoliła skupić się na dokładnej ocenie elektromagnetycznych właściwości powierzchni rekonfigurowalnej. W każdym z opisanych środowisk pomiarowych mierzono odbiciowe właściwości matrycy RMA, przy ustalonej konfiguracji toru nadawczo-odbiorczego, przedstawionego na Rysunku 2. Wszystkie pomiary wykonano dla trzech różnych odległości pomiędzy powierzchnią RMA a środkiem anten nadawczej i odbiorczej: 1 m, 1,5 m oraz 2 m.



Rysunek 2: Widok z góry systemu pomiarowego

W celu analizy charakterystyki kierunkowej odbicia, powierzchnia RMA była obracana wyłącznie w płaszczyźnie azymutu w zakresie kątów od 45° do 135°, z rozdzielczością kątową wynoszącą 1,8°. Pomiary przeprowadzono dla 27 wzorców fazowych RMA; w niniejszym artykule zaprezentowano wyniki dla 6 konfiguracji, zilustrowanych na Rysunku 3. Wybrane wzorce to najbardziej reprezentatywne przypadki, jednocześnie możliwie najbardziej różniące się od siebie, aby uchwycić jak najszerszy zakres charakterystyk odbicia.



Rysunek 3: Wizualizacja wybranych wzorców

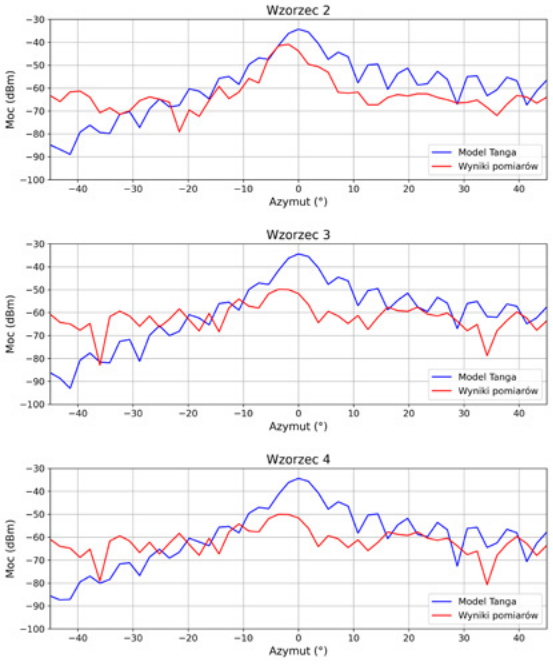
4. WYNIKI

Porównanie uzyskanych wyników z danymi pochodzącymi z modelu teoretycznego umożliwia kompleksową analizę wpływu zarówno środowiska pomiarowego, jak i parametrów antenowych na całkowitą charakterystykę układu Rekonfigurowalnej Macierzy Antenowej. Dzięki temu możliwa jest walidacja modelu oraz identyfikacja ewentualnych odchyleń wynikających z czynników praktycznych, takich jak interferencje, niedoskonałości elementów RMA czy nieliniowości układu. W Tabeli 2 zawarto wartości kluczowych parametrów podstawionych do wzoru Tanga.

Tabela 2: Kluczowe wartości parametrów zastosowanych we wzorze Tanga

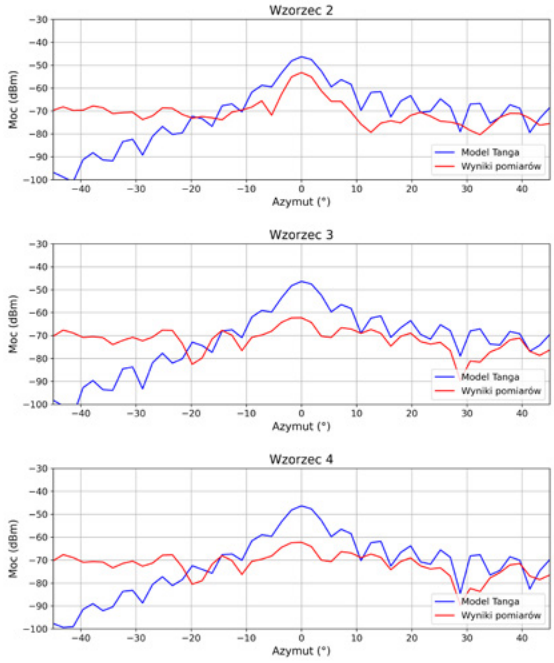
Parametr	Wartość
Liczba elementów w matrycy	16×16
P_t [dBm]	-10
A	0.8
dx [mm]	20
dy [mm]	13
λ [cm]	5.45
G [dBi]	1

Na Rysunkach 4 oraz 5 przedstawiono porównanie mocy odebranej obliczonej za pomocą modelu teoretycznego Tanga (linia niebieska) z rzeczywistymi wynikami pomiarów (linia czerwona) dla pierwszego i trzeciego środowiska pomiarowego, trzech wybranych wzorców i odległości $D = 2m$.



Rysunek 4: Porównanie wyników uzyskanych dla dla Środowiska #1 dla 3 wybranych wzorców dla D = 2m

W scenariuszu pierwszym (Rysunek 4) obserwujemy dobre dopasowanie modelu Tanga do wyników pomiarowych. Szczególnie widoczne jest to w zakresie centralnym (azymut od około -10° do 10°), gdzie krzywe modelu i pomiarów są zbliżone. Dla wzorca 2 zgodność jest bardzo dobra, natomiast dla wzorców 3 i 4 zauważalne są niewielkie różnice w kształcie charakterystyk. Wysoki poziom mocy w tym scenariuszu wynika bezpośrednio z charakterystyki zastosowanych anten, co przekłada się na wyższy poziom mocy odebranej.



Rysunek 5: Porównanie wyników uzyskanych dla dla Środowiska #3 dla 3 wybranych wzorców

W scenariuszu trzecim (Rysunek 5) zgodność pomiędzy modelem a wynikami pomiarowymi jest największa. Jest to potwierdzenie wysokiej dokładności modelu Tanga w warunkach idealnych i pozwala na skuteczną jego walidację przy braku wpływu otoczenia.

W celu precyzyjnej oceny dokładności modelu teoretycznego Tanga dla wszystkich przeprowadzonych pomiarów obliczono wartości następujących metryk: pierwiastek z uśrednionego błędu kwadratowego (ang. *Root Mean Square Error*, RMSE) (3), średni błąd bezwzględny (ang. *Mean Absolute Error*, MAE) (4) oraz suma kwadratów różnic (ang. *Sum of Squared Differences*, SSD) (5).

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{n\in N}\big(\mathbf{x}[n]-\mathbf{y}[n]\big)^2} \tag{3}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{n\in N}\big|\mathbf{x}[n]-\mathbf{y}[n]\big| \tag{4}$$

$$\mathrm{SSD} = \sum_{n\in N}\big(\mathbf{x}[n]-\mathbf{y}[n]\big)^2 \tag{5}$$

gdzie $\mathbf{x}, \mathbf{y}$ to wektory danych o długości N.

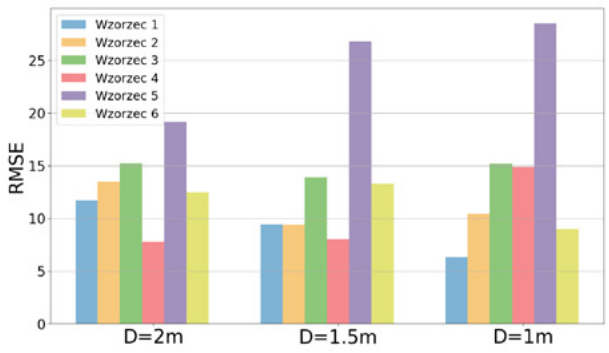
Wartości metryk RMSE, MAE oraz SSD, dla wszystkich badanych scenariuszy, zaprezentowano w Tabeli 3. Na wykresach słupkowych (Rysunek 6, 7 oraz 8) zobrazowano porównanie wartości trzech omawianych metryk dla sześciu analizowanych wzorców oraz trzech odległości między matrycą RMA a odbiornikiem, w ramach pierwszego środowiska pomiarowego.

Na podstawie danych z Tabeli 3 oraz przedstawionych wykresów (Rysunek 6, 7 oraz 8) można zauważyć,

że wartości metryk RMSE, MAE oraz SSD systematycznie maleją wraz ze wzrostem odległości między nadajnikiem a matrycą RMA w każdym z trzech analizowanych środowisk pomiarowych, co świadczy o poprawiającej się zgodności estymacji mocy odebranej (obliczonej na podstawie wzoru Tanga) z danymi pomiarowymi. Największe błędy występują w drugim środowisku pomiarowym dla najmniejszej odległości (1 m), co może wynikać ze specyfiki anten METIS użytych podczas pomiarów. W pierwszym środowisku testowym, mimo że również wykonywano pomiary w warunkach nieidealnych, zastosowano nowsze anteny AINFO, co mogło pozytywnie wpłynąć na dokładność estymacji. Najlepszą zgodność osiągnięto w trzecim środowisku pomiarowym, gdzie pomiary realizowano w warunkach kontrolowanych – w komorze bezechowej – co znacząco ograniczyło wpływ odbić i zakłóceń. Wszystkie trzy metryki wykazują spójne trendy, przy czym RMSE – jako bardziej wrażliwa na duże odchylenia – przyjmuje wyższe wartości niż MAE, natomiast SSD, jako metryka sumaryczna, osiąga największe liczby, ale jej zmienność odzwierciedla tę samą charakterystykę błędów. Wnioski te wskazują, że wykorzystany model Tanga lepiej odwzorowuje rzeczywistość w środowiskach o mniejszym stopniu złożoności i przy większych odległościach od źródła sygnału.

Tabela 3: Metryki RMSE, MAE oraz SSD

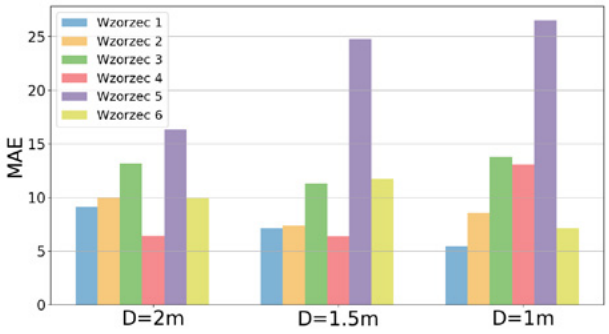
Środ. Pom.	*D [m]*	*RMSE*	*MAE*	*SSD*
#1	1	14.03	12.39	6458.89
	1.5	13.47	11.42	5759.77
	2	13.30	10.81	4910.53
#2	1	16.33	14.18	7722.02
	1.5	11.84	9.63	3913.04
	2	11.98	9.47	3862.77
#3	1	13.88	11.94	5947.28
	1.5	11.99	9.83	4106.07
	2	11.44	9.53	3619.57



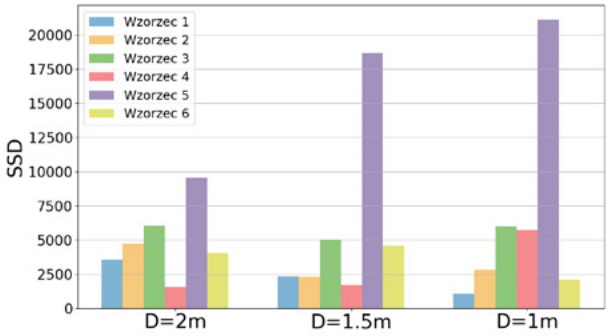
Rysunek 6: RMSE w środowisku pomiarowym #1

5. WNIOSKI

Przeprowadzone badania wskazały relatywnie duże podobieństwo pomiędzy charakterystykami uzyskanymi z modelu teoretycznego i wynikami pomiarów. Podobieństwo to jest jednak widoczne jedynie przy ogólnym, zgrubnym porównaniu charakterystyk odbiciowych. Głębsza analiza wykazuje znaczne różnice w wartościach (sięgające nawet ponad 10 dB) dla poszczególnych kątów azymutu, co wskazuje na konieczność dostosowania modelu Tanga do warunków panujących wewnątrz pomieszczeń.



Rysunek 7: MAE w środowisku pomiarowym #1



Rysunek 8: SSD w środowisku pomiarowym #1

LITERATURA

[1] Araghi, A., i in. 2022. "Reconfigurable Intelligent Surface (RIS) in the Sub-6 GHz Band: Design, Implementation, and Real-World Demonstration." *IEEE Access* 10: 2646–2655. https://doi.org/10.1109/ACCESS.2022.3140278.

[2] Hatka, P., Garczyk, M., Płaczkiewicz, P., Brząkała, D., Cichoń, K., i Kliks, A. 2024. "RIS-Measurements for Codebook Design." *Proc. 2024 20th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*: 437–442. https://doi.org/10.1109/WiMob61911.2024.10770406.

[3] Heinrichs, M., Sezgin, A., i Kronberger, R. 2023. "Open Source Reconfigurable Intelligent Surface for the Frequency Range of 5 GHz WiFi." *2023 IEEE International Symposium on Antennas and Propagation (ISAP)*: 1–2. Kuala Lumpur, Malezja. https://doi.org/10.1109/ISAP57493.2023.10389095.

[4] Tang, W., i in. 2021. "Wireless Communications With Reconfigurable Intelligent Surface: Path Loss Modeling and Experimental Measurement." *IEEE Transactions on Wireless Communications* 20(1): 421–439. https://doi.org/10.1109/TWC.2020.3024887.

[5] Tang, W., i in. 2022. "Path Loss Modeling and Measurements for Reconfigurable Intelligent Surfaces in the Millimeter-Wave Frequency Band." *IEEE Transactions on Communications* 70(9): 6259–6276.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

SYSTEM STEROWANIA REKONFIGUROWALNYMI MACIERZAMI ANTENOWYMI: DETEKCJA ZAJĘTOŚCI WIDMA

CONTROL SYSTEM FOR RECONFIGURABLE INTELLIGENT SURFACES: SPECTRUM SENSING

Marta Sieradzka[1]; Łukasz Kułacz[2]; Adrian Kliks[2]

[1] Politechnika Poznańska, Poznań, marta.sieradzka.1@student.put.poznan.pl
[2] Politechnika Poznańska, Poznań, imie.nazwisko@put.poznan.pl

DOI: 10.15199/59.2025.4.30

Streszczenie: W ciągu ostatnich kilku lat rekonfigurowalne macierze antenowe (RMA) stały się intensywnie rozwijającym się obszarem badań, oferującym możliwość wpływania na proces propagacji sygnału w czasie rzeczywistym. W niniejszej pracy przedstawiono modułowy system sterowania RMA oparty na programowalnej logice, który umożliwia bieżące dostosowywanie charakterystyki odbiciowej macierzy. Skuteczność zaprojektowanego systemu została potwierdzona w scenariuszu lokalizowania użytkowników, wykorzystującym różnice w mocy sygnału odbieranego dla różnych konfiguracji macierzy.

Abstract: Over the past few years, Reconfigurable Intelligent Surfaces (RISs) have emerged as a rapidly evolving research area, enabling dynamic control of signal propagation in real time. This work presents a modular RIS control system based on programmable logic, allowing for real-time adjustment of the proposed system is demonstrated in a user localization scenario, which leverages variation in received signal power across different RIS configurations.

Słowa kluczowe: Książka Kodowa, Rekonfigurowalne Macierze Antenowe, System Sterownia

Keywords: Codebook, Reconfigurable Intelligent Antennas, Control System

1. WSTĘP

Rekonfigurowalne Macierze Antenowe (RMA, ang. Reconfigurable Intelligent Surfaces) od kilku lat są intensywnie badane ze względu na zdolność do wpływania na środowisko propagacyjne [1]. W najprostszej formie RMA to (pół)pasywne układy złożone z wielu małych elementów antenowych, każdy indywidualnie sterowany, by uzyskać pożądane charakterystyki odbicia. Opisywane są wówczas jako refleksyjne, ponieważ wpływają na odbicie sygnału od powierzchni. Zainteresowanie technologią wynika z potencjału do zrewolucjonizowania sieci bezprzewodowych nowej generacji, oferując wysoki poziom kontroli nad medium transmisyjnym przy niskim zużyciu energii i opłacalności wdrożenia.

Mimo postępów w projektowaniu i optymalizacji RMA, istnieją wyzwania praktyczne, takie jak adaptacja w czasie rzeczywistym i efektywne sterowanie. Przykłady implementacji to m.in. refleksyjne RMA działająca na 3,5 GHz z 2430 elementami na powierzchni 1,2 × 1,2 m^2 [2] oraz system z 256 elementami sterowanymi z 2-bitową rozdzielczością [3]. Inicjatywa OpenSourceRIS [4] opracowała binarne RMA 16×16 elementów z uproszczonym sterowaniem, przydatnym w eksperymentach.

Istnieją także inne typy RMA, takie jak refrakcyjne, transmisyjne czy jednocześnie transmisyjno-refleksyjne (STAR-RIS) [5], które szczegółowo omówiono w [6]. Literatura podkreśla różnorodne zastosowania RMA, m.in. zwiększanie mocy odbieranego sygnału, rozszerzanie zasięgu, minimalizowanie zakłóceń czy pochłanianie niepożądanego promieniowania [7].

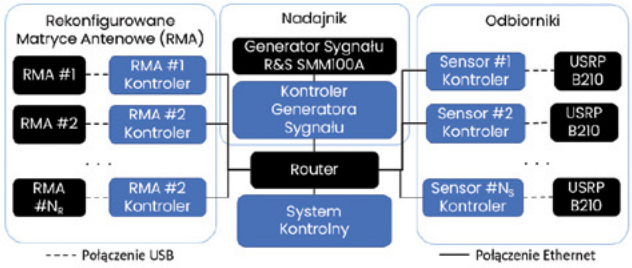
RMA poprzez konroler może zmieniać wzorzec odbicia. W przyszłych systemach telekomunikacyjnych kluczowym wyzwaniem jest zapewnienie odpowiedniej komunikacji między poszczególnymi elementami systemu, w tym matrycami RMA. Jako, że pełne wykorzystanie ich potencjału zależy od skuteczności systemu sterowania istotna jest minimalizacja opóźnień i zapewnienie szybkich reakcji na zmiany otoczenia. Różnorodność RMA wymaga modułowego, opartego na oprogramowaniu systemu, umożliwiającego efektywne zarządzanie konfiguracją i szybki wybór wzorca odbicia z tzw. książki kodowej RMA [8, 9, 10].

Celem tej pracy jest zaproponowanie skalowalnego, responsywnego systemu sterowania RMA, umożliwiającego integrację z przyszłymi systemami komunikacji. Skoncentrowano się na projektowaniu i ocenie zdalnego systemu sterowania RMA, co jest nowością artykułu. Zaprezentowane zostały wyniki eksperymentów ze zdalnym, indywidualnym sterowaniem fizycznie wytworzonych płyt OpenSourceRIS [4]. System został przetestowany podczas eksperymentów sprzętowych w Politechnice Poznańskiej. Wybrany został przypadek poprawy skuteczności wykrywania zajętości widma przez dynamiczny wybór wzorca odbicia RMA. Korzystając ze scentralizowanego kontrolera, system skutecznie dobierał wzorce RMA, poprawiając efektywność wykrywania widma. Automatyczne pętle sprzężenia zwrotnego pozwalały na ciągłą optymalizację przy minimalnej ingerencji użytkownika.

2. ARCHITEKTURA SYSTEMU

Głównym celem pracy było opracowanie programowalnego systemu sterowania RMA, umożliwiającego skuteczną komunikację każdego z elementów systemu. Sys-

tem składa się z czterech komponentów: kontrolera, nadajnika, RMA oraz czujników (rys. 1).



Rysunek 1: Schemat architektury systemu.

Komponenty łączą się przez sieć Ethernet, komunikując się protokołem ZeroMQ, który zapewnia niskie opóźnienia i wysoką stabilność w porównaniu z innymi przetestowanymi protokołami. Kontroler indywidualnie dostosowuje parametry RMA dla osiągnięcia zadanego celu. Cel optymalizacji został przygotowany w taki sposób, aby poprzez łatwą modyfikację system pozostał elastyczny dla różnych zastosowań. Czujniki okresowo raportują poziomy mocy odbieranej do kontrolera, co pozwala na optymalizację w czasie rzeczywistym. Funkcje poszczególnych modułów:

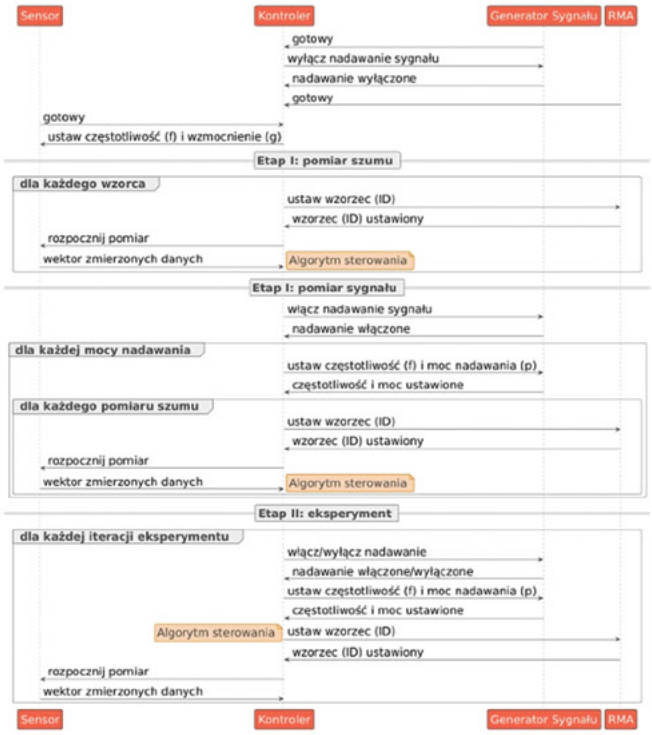
- *Kontroler systemu* — zarządza systemem, wysyłając polecenia do komponentów zapewniając synchronizację działań. Odbiera dane z czujników i wybiera ustawienia RMA według aktualnie stosowanego algorytmu.
- *Generator sygnału* — nadaje sygnał po komendzie kontrolera systemu, który ustala częstotliwość, moc i włączenie/wyłączenie transmisji. Kontroler generatora tłumaczy komendy systemowe na polecenia sprzętowe, umożliwiając pracę z różnymi generatorami.
- *Czujniki* — odbierają próbki sygnału, mierząc średnią moc w wyznaczonym paśmie. Kontroler czujnika steruje parametrami pomiaru, a warstwa abstrakcji umożliwia obsługę różnych urządzeń (np. USRP B210).
- *RMA* — odbija fale według wzorca wskazanego przez kontroler systemu. Każda RMA ma swój kontroler RMA, umożlwiając zastosowanie różnych RMA lub różnych metod połączenia RMA (port szerogowy, Bluetooth).

Dodatkowo system zawiera algorytm sterowania RMA oraz schemat eksperymentu, zaimplementowane za pomocą uniwersalnych interfejsów, co umożliwia testowanie różnych algorytmów w różnych warunkach i scenariuszach. Logika działania kontrolera podzialona została na dwa etapy: faza inicjalizacji (zbieranie danych) i faza eksperymentu (faktyczne działanie algorytmu). Taka architektura zapewnia elastyczność w różnych badaniach.

3. SCHEMAT DZIAŁANIA SYSTEMU

Przepływ wiadomości między komponentami systemu pokazano na rys. 2. W fazie konfiguracji system działa synchronicznie, a urządzenia są ustawiane w trybie żądanie–odpowiedź. Po zakończeniu tej fazy system zmienia tryb działania na półasynchroniczny: kontroler wysyła polecenia do czujników, które — mimo możliwych opóźnień — przesyłają dane asynchronicznie. Kontroler systemu czeka na otrzymanie wyników z wszystkich czujników przed przejściem do kolejnego kroku.

Algorytm sterowania RMA wymaga możliwości konfiguracji RMA i generatora sygnału (tylko w fazie inicjalizacji). Przykładowo, algorytm może potrzebować danych dla wszystkich wzorców RMA przy jednej konfiguracji generatora (np. maksymalnej mocy). Schemat uwzględniający abstrakcyjny algorytm i abstrakcyjny eksperyment przedstawiono na rys. 2.



Rysunek 2: Schemat przepływu wiadomości dla systemu sterowania RMA.

Schemat działania kontrolera systemu został podzielony na dwa etapy:

- *Faza inicjalizacji* — system zasila wybrany algorytm w niezbędne dla jego funkcjonowania dane, zbierając pomiary z czujników dla różnych konfiguracji generatora i wzorców RMA. Zestaw wzorców, ustawień generatora (np. szum, sygnał użytkowy, moc) oraz strategia ich doboru definiowane są przez algorytm, który może być dostosowany do różnych scenariuszy i celów, takich jak wykrywanie widma, lokalizowania użytkowników czy redukcja zakłóceń.
- *Faza eksperymentu* — kontroler systemu wykorzystuje algorytm (posiadający niezbędne dane z poprzedniej fazy działania systemu) do wyboru wzorców RMA w konkretnym scenariuszu. Parametry eksperymentu (np. czas trwania, konfiguracja generatora w czasie trwania całego eksperymentu) są przekazywane do kontrolera. Na podstawie danych z fazy inicjalizacji system wybiera wzorce RMA na każdym kroku eksperymentu (każdorazowo tworząc zapytanie do algorytmu), a pomiary zbierane w czasie trwania eksperymentu służą późniejszej analizie skuteczności algorytmu.

System umożliwia sterowanie wzorcami RMA zgodnie z algorytmem zdefiniowanym przez użytkownika oraz automatyczne wykonanie pomiarów potrzebnych w fazie inicjalizacji.

4. OPIS EKSPERYMENTU

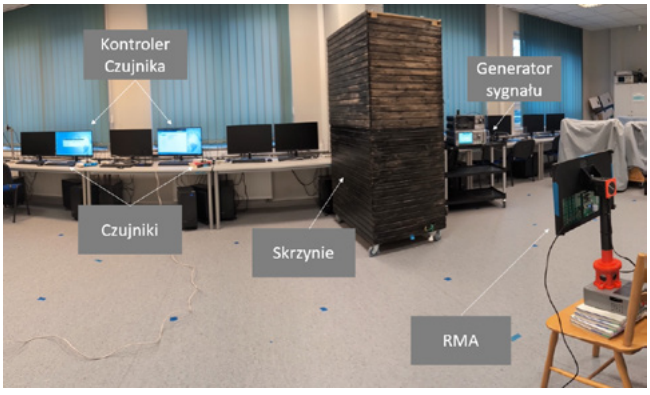
Proponowany system może być skonfigurowany w celu osiągnięcia różnych rezultatów, w zależności od założeń użytkownika. Ta praca skupia się na przypadku użycia systemu do detekcji zajętości widma (*ang. spectrum sensing*). W najprostszym wariancie każda RMA dysponuje jedynie dwoma wzorcami odbicia. W takim przypadku system musi zebrać pomiary mocy sygnału z każdego czujnika dla każdego wzorca RMA — zarówno w obecności sygnału, jak i przy jego braku.

Po zakończeniu zbierania danych, w fazie działania systemu wybierany jest najlepszy wzorzec RMA (najwyższa moc odbierana) dla każdego czujnika, w celu optymalizacji odbioru sygnału dla określonego celu - detekcji widma. Algorytm może gromadzić dane dotyczące mocy sygnału odbieranego dla różnych wzorców RMA już w fazie inicjalizacji, przy wykorzystaniu różnych poziomów mocy nadawczej generatora sygnału. Takie podejście pozwala na trafniejszy dobór wzorca RMA, uwzględniając m.in.: średnią moc szumu obserwowaną przy danym ustawieniu RMA, przy niskim SNR oraz przy wysokim SNR. W zależności od scenariusza i środowiska, najlepszy wzorzec RMA może się różnić dla każdego z czujników.

4.1 Układ pomiarowy

W celu weryfikacji działania modułowego i zdefiniowanego programowo systemu sterowania RMA, przeprowadzono eksperyment oparty na sprzęcie znajdującym się w laboratorium. Celem eksperymentu było zmaksymalizowanie mocy sygnału odbieranego przez każdy z czujników, co miało przyczynić się do poprawy skuteczności detekcji zajętości widma.

Pomiary wykonano z wykorzystaniem dwóch czujników (USRP B210), rozmieszczonych w różnej odległości od siebie i odizolowanych od nadajnika ścianą przesuwną, blokującą bezpośrednią linię widzenia (ang. Line-of-Sight, LOS). Jedna RMA, zgodna z opisem w [4], została umieszczona tak, aby odbijała transmitowany sygnał w kierunku czujników. Do transmisji sygnału wykorzystano generator R&S SMM100A.

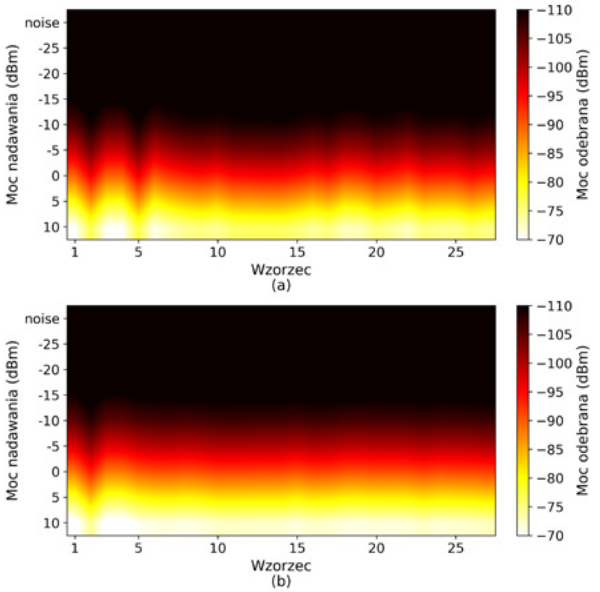


Rysunek 3: Zdjęcie przedstawiające ustawienie urządzeń podczas pomiaru

Rys. 3 przedstawia fotografię rzeczywistego rozmieszczenia komponentów systemu. Częstotliwość transmisji została ustawiona na 5 GHz, a do nadawania i odbioru użyto anten dookólnych. Moc transmisji była dostosowywana w krokach co 1 dBm, w zakresie od -30 dBm do 10 dBm. Książka kodowa dla matrycy RMA została stworzona z 27 wzorców, z których każdy zawierał ręcznie wybrane ustawienia, takie jak pionowe lub poziome linie aktywnych elementów. Po zakończeniu transmisji z wykorzystaniem wszystkich z 27 wzorców generator zwiększał moc (zgodnie z definicją algorytmu) o jeden krok, a pomiar był powtarzany dla nowych parametrów. Wybór 27 wzorców był arbitralny i dokonano go w celu uzyskania czytelnych i łatwych do interpretacji kombinacji. Przedstawione poniżej wyniki zebrane zostały dla pojedyńczego powtórzenia procedury pomiarowej.

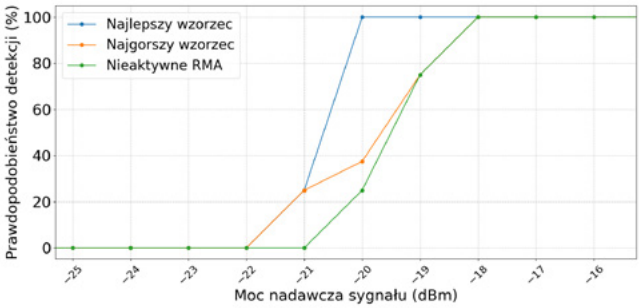
4.2 Wyniki pomiarów

Wyniki uzyskane po pierwszej fazie działania algorytmu (fazie inicjalizacji), przedstawiono na rys. 4. Można zauważyć, że zastosowanie określonych wzorców może zarówno poprawić, jak i pogorszyć skuteczność detekcji zajętości widma, ponieważ moc odbieranego sygnału zależy od wyboru wzorca. Dla czujnika 1 optymalny jest wzorzec 4, a dla czujnika 2 — wzorzec 3. Ponieważ zapewniają one najwyższą moc odbieraną, będą używane w fazie operacyjnej systemu. Różnice w mocy między wzorcami są niewielkie, więc w innych warunkach, np. przy zastosowaniu anten kierunkowych, ich wpływ może być większy.



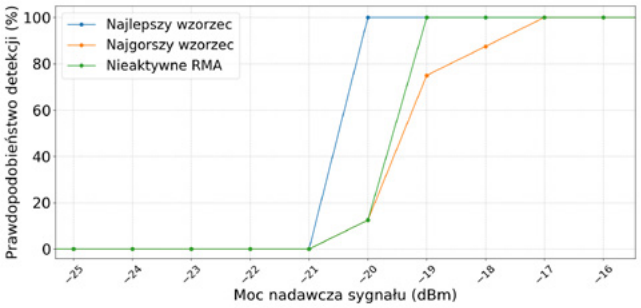
Rysunek 4: Wyniki przedstawione jako mapa ciepła dla każdego czujnika zmierzonego w etapie I działania systemu: (a) Czujnik 1, (b) Czujnik 2

Dla obu czujników prawdopodobieństwo detekcji widma przekracza 95% przy mocy transmisji ustawionej na −24 dBm. Zmierzone prawdopodobieństwo fałszywego alarmu wynosi 1,03% dla czujnika 1 oraz 0,16% dla czujnika 2. Pomimo niewielkich różnic między czujnikami w zmierzonym poziomie szumu, dostosowanym progu detekcji energii i uzyskanym prawdopodobieństwie fałszywego alarmu, charakterystyki detekcji pozostają niemal identyczne, co potwierdza wiarygodność pomiarów.



Rysunek 5: Prawdopodobieństwo detekcji dla czujnika 1 dla różnych wzorców RMA

Aby zilustrować wpływ wyboru wzorca RMA, rys. 5 przedstawia prawdopodobieństwo detekcji dla trzech przypadków: najlepszego, najgorszego oraz przy nieaktywnym RMA. W tym układzie wybór odpowiedniego wzorca RMA umożliwia detekcję sygnałów o ok. 2 dB słabszych, co jest niewielką - wartością - głównie ze względu na wybrany scenariusz pomiarowy. Jednakże, nawet w takim przypadku widać potencjał w zastosowaniu RMA do poprawy skuteczności detekcji zajętości widma. Rys. 6 pokazuje analogiczne wyniki dla czujnika 2, położonego w nieco większej odległości od RMA. W obu przypadkach wpływ wyboru wzorca RMA na skuteczność detekcji jest wyraźny. Lepszy wybór wzorca RMA umożliwia wykrywanie słabszych sygnałów, co przekłada się na wyższe prawdopodobieństwo detekcji.



Rysunek 6: Prawdopodobieństwo detekcji dla czujnika 2 dla różnych wzorców RMA

5. PODSUMOWANIE

Artykuł przedstawia projekt dedykowanego systemu sterowania RMA oraz analizuje jego wydajność na przykładzie poprawy skuteczności detekcji zajętości widma. Zaprojektowane rozwiązanie umożliwia dynamiczne, zdalne i programowalne sterowanie RMA, ułatwiając adaptację do różnych scenariuszy eksperymentalnych. Dzięki szybkiemu protokołowi komunikacyjnemu zminimalizowano opóźnienia, co jest kluczowe w zastosowaniach, gdzie opóźnienia są istotne. Pomiary wykazały, że wybór wzorca RMA istotnie wpływa na efektywność detekcji zajętości widma, umożliwiając wykrycie sygnałów o niższej mocy odbieranej. Nawet niewielkie zmiany konfiguracji mogą prowadzić do zauważalnej poprawy wydajności.

Modułowa, sterowana programowo architektura systemu zapewnia elastyczność i łatwą integrację nowych algorytmów, co czyni system atrakcyjną platformą badawczą. Może być również rozszerzony o funkcje, takie jak poprawa stosunku sygnału do zakłóceń czy zwiększenie zasięgu.

6. PODZIĘKOWANIA

Praca powstała w ramach projektu badawczego OPUS 2021/43/B/ST7/01365 finansowanego przez Narodowe Centrum Nauki w Polsce.

LITERATURA

[1] Y. Liu i in., “Reconfigurable Intelligent Surfaces: Principles and Opportunities,“ IEEE Communications Surveys & Tutorials, tom 23, numer 3, str. 1546-1577, 2021, doi: 10.1109/COMST.2021.3077737

[2] A. Araghi i in., “Reconfigurable Intelligent Surface (RIS) in the Sub-6 GHz Band: Design, Implementation, and Real-World Demonstration,“ IEEE Access, tom 10, str. 2646-2655, 2022, doi: 10.1109/ACCESS.2022.3140278

[3] L. Dai i in., “Reconfigurable Intelligent Surface-Based Wireless Communications: Antenna Design, Prototyping, and Experimental Results,“ IEEE Access, tom 8, str. 45913-45923, 2020, doi: 10.1109/ACCESS.2020.2977772

[4] M. Heinrichs, A. Sezgin, R. Kronberger, “Open Source Reconfigurable Intelligent Surface for the Frequency Range of 5 GHz WiFi,“ 2023 IEEE International Symposium On Antennas And Propagation (ISAP), str. 1-2, Kuala Lumpur, Malezja, 2023, doi: 10.1109/ISAP57493.2023.10389095

[5] X. Mu, Y. Liu, L. Guo, J. Lin, R. Schober, “Simultaneously Transmitting and Reflecting (STAR) RIS Aided Wireless Communications,“ IEEE Transactions on Wireless Communications, tom 21, numer 5, str. 3083-3098, Maj 2022, doi: 10.1109/TWC.2021.3118225

[6] A. Tishchenko i in., “The Emergence of Multi-Functional and Hybrid Reconfigurable Intelligent Surfaces for Integrated Sensing and Communications -A Survey,“ IEEE Communications Surveys & Tutorials, doi: 10.1109/COMST.2024.3519785

[7] W. Ni, A. Zheng, W. Wang, D. Niyato, N. Al-Dhahir, M. Debbah, “From Single to Multi-Functional RIS: Architecture, Key Technologies, Challenges, and Applications,“ IEEE Network, tom 39, numer 1, str. 38-46, Styczeń 2025, doi: 10.1109/MNET.2024.3481028

[8] P. Hatka, M. Garczyk, P. Placzkiewicz, D. Brzakala, K. Cichon, A. Kliks, “RIS-Measurements for Codebook Design,“ 2024 20th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), str. 437-442, Paryż, Francja, 2024, doi: 10.1109/WiMob61911.2024.10770406

[9] A. Abdallah, A. Celik, M. M. Mansour, A. M. Eltawil, “Multi-Agent Deep Reinforcement Learning for Beam Codebook Design in RIS-Aided Systems,“ IEEE Transactions on Wireless Communications, tom 23, numer 7, str. 7983-7999, Lipiec 2024, doi: 10.1109/TWC.2023.3347419

[10] Y. Liu, F. Gao, L. Zhang, “Quantization Beam Analysis and Codebook Design for One-Bit Reconfigurable Intelligent Surface,“ IEEE Wireless Communications Letters, tom 13, numer 7, str. 1793-1797, Lipiec 2024, doi: 10.1109/LWC.2024.3388273

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

WPŁYW ZJAWISK ATMOSFERYCZNYCH NA BILANS ŁĄCZA W SYSTEMACH 6G WYKORZYSTUJĄCYCH PASMA SUBTERAHERCOWE

THE INFLUENCE OF ATMOSPHERIC PHENOMENA ON THE LINK BUDGET FOR 6G SYSTEMS USING SUB-TERAHERTZ BANDS

Sławomir Hausman[1]; Piotr Korbel[2]

[1,2] Politechnika Łódzka, Instytut Elektroniki,
Al. Politechniki 8, 93-590 Łódź
[1]slawomir.hausman@p.lodz.pl; [2]piotr.korbel@p.lodz.pl

DOI: 10.15199/59.2025.4.31

Streszczenie: **Artykuł pokazuje wpływ zjawisk atmosferycznych na bilans łącza w systemach 6G wykorzystujących subterahercowe i terahercowe pasma częstotliwości. Skupia się na modelowaniu propagacji sygnałów z uwzględnieniem absorpcji przez tlen i parę wodną zależnie od lokalnych parametrów atmosfery oraz rozpraszania przez opady atmosferyczne. Uzyskane przykładowe wyniki bilansu łącza ilustrują wyzwania związane z wykorzystaniem pasm tak dużych częstotliwości i wskazują między innymi na potrzebę stosowania inteligentnych powierzchni odbijających.**

Abstract: **The article shows the impact of atmospheric phenomena on the link budget in 6G systems using sub-terahertz and terahertz frequency bands. It focuses on signal propagation modelling, taking into account absorption by oxygen and water vapour depending on local atmospheric parameters and scattering by atmospheric precipitation. The obtained sample link budget results illustrate the challenges associated with the use very high frequency bands and, among other things, indicate the need for intelligent reflective surfaces.**

Słowa kluczowe: **6G, inteligentne powierzchnie odbijające, bilans łącza, model propagacji, fale terahercowe.**

Keywords: **6G, intelligent reflecting surface, link budget, propagation model, terahertz waves.**

1. WSTĘP

W odpowiedzi na przewidywany wzrost zapotrzebowania na szybką transmisję danych w systemach łączności ruchomej, European Telecommunications Standards Institute (ETSI) uznaje, że badania nad systemami komórkowymi kolejnej, czyli szóstej generacji powinny dotyczyć możliwości uzyskania przepływności nawet rzędu 1 Tb/s na pojedynczego użytkownika. ETSI (podobnie jak inne organizacje standaryzacyjne) podaje też szereg przypadków użycia, np. immersyjne holograficzne systemy telekonferencyjne [1]. Wymagana, znacznie większa niż w systemach 5G (do 400 MHz), szerokość pasma potrzebna do uzyskania dużej szybkości transmisji wynika wprost z prawa Shannona–Hartleya, co w konsekwencji oznacza konieczność wykorzystania pasm wielkich częstotliwości w zakresie fal subterahercowych (100–300 GHz) i terahercowych (300–1000 GHz), w połączeniu z antenami o dużych wartościach zysku oraz inteligentnymi powierzchniami odbijającymi, aby uzyskać wymaganą wartość stosunku sygnału do szumu, w trudnych warunkach propagacyjnych występujących przy tych częstotliwościach. Dla tych pasm przeszkody architektoniczne stają się nieprzenikliwe, dyfrakcja pomijalna, a absorpcja w gazach atmosferycznych i opadach atmosferycznych bardzo znacząca.

Przejście do pasm sub- i terahercowych otwiera ponadto nowe perspektywy integracji komunikacji z zaawansowanymi aplikacjami, takimi jak obrazowanie trójwymiarowe (3D imaging) i detekcja gazów czy cząstek (sensing), które dotychczas wymagały odrębnych platform [6]. Wymaga to jednak opracowania nowatorskich paradygmatów architektury sieci stacji bazowych, opanowania produkcji dostatecznie tanich elementów (wzmacniaczy, modulatorów, anten) pracujących w pasmach terahercowych, czyli w tzw. przerwie terahercowej (ang. terahertz gap), a także nowych architektur sieci i algorytmów zarządzania ruchem wykorzystujących mechanizmy uczenia maszynowego, które pozwolą pokonać problemy związane z szybkimi i bardzo znaczącymi zmianami parametrów kanału radiowego.

Artykuł koncentruje się tylko na jednym aspekcie z tej grupy zagadnień – wpływie absorpcji w gazach atmosferycznych i mniejszym stopniu także wpływie opadów atmosferycznych na budżet energetyczny łącza.

2. MODEL PROPAGACJI

2.1. Absorpcja w gazach atmosferycznych

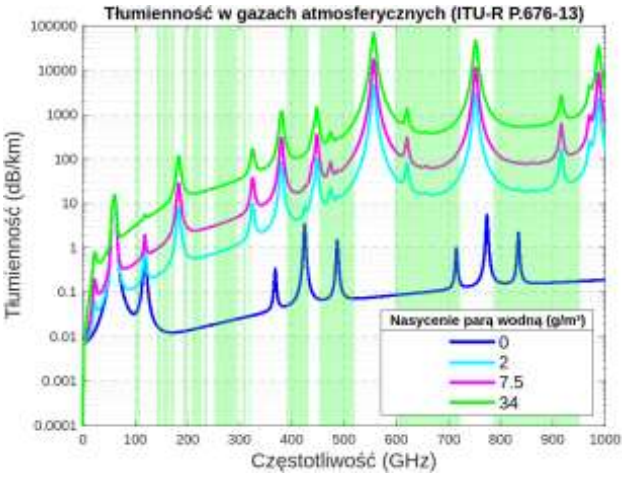
Podstawowym mechanizmem powodującym tłumienie fal o częstotliwościach od 22 GHz do 1 THz są przejścia wibracyjne i rotacyjno-wibracyjne cząsteczek tlenu O_2 i pary wodnej H_2O. Zalecenie ITU-R P.676-13 [5] uwzględnia 44 linie (częstotliwości) absorpcyjne dla tlenu cząsteczkowego i 35 linii absorpcyjnych dla pary wodnej oraz tzw. kontinuum. Kontinuum pozwala modelować absorpcję między liniami absorpcyjnymi. Spektroskopowe parametry linii (częstotliwości środkowe, wartość tłumienności, szerokości, współczynniki temperaturowe oraz współczynniki kontinuów są zestawione w załącznikach do rekomendacji [5].

Całkowita tłumienność (będąca funkcją częstotliwości) $\gamma_{\text{total}}(f)$ (1) jest w tym modelu wyznaczana jako suma

wkładów ze wszystkich pojedynczych linii rezonansowych O_2 oraz H_2O $\gamma_{line}(f)$ (z uwzględnieniem tzw. poszerzeń Dopplera i ciśnieniowego) oraz tzw. suchych i parowych (związanych z H_2O) składników kontinuum absorpcji $\gamma_{continuum}(f)$. W rzeczywistości przebieg kontinuum wynika z nakładania się znacznej liczby zjawisk fizycznych. Przykładem jest pośredni, ale istotny wpływ cząsteczek azotu N_2, które zderzają się z cząsteczkami O_2 oraz H_2O poszerzając ich linie absorpcyjne.

$$\gamma_{total}(f) = \gamma_{line}(f) + \gamma_{continuum}(f) \quad (1)$$

Model umożliwia obliczanie tłumienności w zakresie do 1 THz. Na Rys. 1 przedstawiono krzywe absorpcji, dla wybranych wartości wilgotności powietrza, wyznaczone za pomocą autorskiego programu opracowanego w języku Matlab na podstawie zalecenia ITU-R P.676-13 [5].



Rys. 1. Tłumienność w gazach atmosferycznych w funkcji częstotliwości dla wybranych wartości wilgotności bezwzględnej. Zielonym kolorem zaznaczono pasma częstotliwości które mogą być potencjalnie wykorzystane w systemach 6G.

Zielonym kolorem zaznaczono pasma częstotliwości które mogą być potencjalnie wykorzystane w systemach 6G, ponieważ nie zawierają częstotliwości linii absorpcyjnych pary wodnej oraz nie kolidują z pasmami, które zostały już przeznaczone dla innych służb [2]. W zakresie częstotliwości od 0 do 10 GHz tłumienność atmosferyczna jest poniżej 0,1 dB/km, niezależnie od zawartości pary wodnej w powietrzu. To sprawia, że ten zakres jest przydatny nawet dla systemów komunikacyjnych o dużym zasięgu, rzędu kilkudziesięciu km. W przedziale od 10 do 100 GHz zaczynają pojawiać się wyraźne maksima tłumienności, szczególnie w okolicach 22 GHz (H_2O), oraz w pobliżu 60 GHz (O_2). W tym zakresie tłumienie staje się bardziej zauważalne, zwłaszcza przy dużej wilgotności, co wymaga uwzględnienia warunków atmosferycznych już przy projektowaniu systemów 5G. Powyżej 100 GHz, ze względu na liczne rezonanse molekularne, tłumienność rośnie gwałtownie. Ten zakres częstotliwości nadaje się głównie do zastosowań krótkiego zasięgu (setki metrów w warunkach troposfery Ziemi). Wpływ wilgotności powietrza na tłumienność jest bardzo wyraźny. Przy zerowej zawartości pary wodnej tłumienie wynika głównie z obecności tlenu i jest najmniejsze. Gdy gęstość pary wodnej wzrasta do 2 g/m³, obserwuje się wzrost tłumienności ok. 22 GHz oraz powyżej 100 GHz. Przy wilgotności, rzędu 7.5 g/m³, tłumienność staje się znacząca w szerokim zakresie częstotliwości, szczególnie w pasmach powyżej 100 GHz. Oznacza to, że zmienność wilgotności silnie wpływa na stosunek sygnału do szumu w pasmach większych częstotliwości. W warunkach standardowych (dla których przyjmuje się zwykle temperaturę 288 K, ciśnienie 1013 hPa i wilgotność bezwzględną 7,5 g/m³), szczytowe wartości tłumienności przekraczają 10 dB/km przy 183 GHz oraz 100 dB/km dla 380 GHz oraz nawet 10000 dB/km na liniach pary wodnej w pobliżu 557 GHz i 752 GHz, podczas gdy okna transmisji między liniami wykazują straty <100 dB/km w pasmie do ok. 900 GHz. Model umożliwia uwzględnienie specyficznych dla danej lokalizacji geograficznej profili temperatury, ciśnienia i wilgotności w celu uzyskania realistycznych obliczeń budżetu łącza.

W niniejszej publikacji zakłada się komunikację między terminalami i stacjami bazowymi zlokalizowanymi blisko powierzchni ziemi. Dlatego pominięty będzie aspekt obliczeń dla tzw. ścieżek ukośnych, w tym refrakcji troposferycznej i tzw. nadwyżkowej długości drogi propagacji.

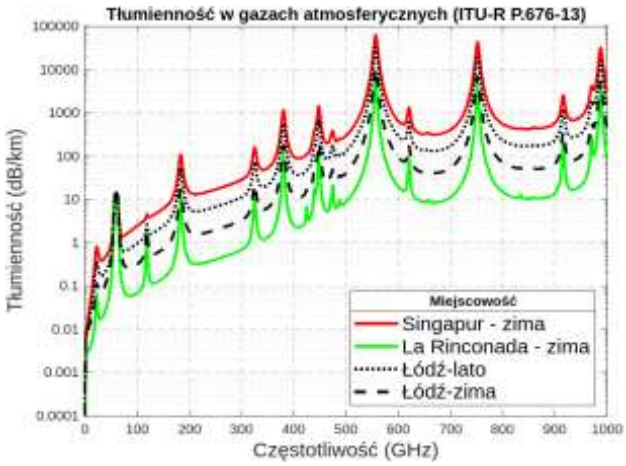
Aby zilustrować rozpiętość wartości absorpcji atmosferycznej rozważmy przykład warunków klimatycznych w Singapurze (poziom morza), w najwyżej położonym na świecie mieście La Rinconada w Peru (ok. 5100 m n.p.m) i Łodzi w Polsce. Maksymalna gęstość pary wodnej w Singapurze w temperaturze 34°C i wilgotności względnej 100% wynosi ok. 34,5 g/m³ przy ciśnieniu atmosferycznym ok. 1013 hPa. Z kolei w La Rinconada najmniejsza wilgotność bezwzględna to ok. 1,5 g/m³ przy temperaturze ok. -10 °C ciśnieniu atmosferycznym zaledwie 540 hPa. W klimacie umiarkowanym Polski bezwzględna gęstość pary wodnej wykazuje znaczącą zmienność sezonową oraz wahania krótkoterminowe. Średnie wartości miesięczne (g/m³) obliczone na podstawie wieloletnich pomiarów przedstawia Tab. 1.

Tab. 1. Średnia wartość wilgotności bezwzględnej powietrza [g m⁻³] w Polsce w różnych miesiącach

Miesiące	*Wartość*
Styczeń-luty	4–4,5
Marzec–kwiecień	5–6
Maj	8
Czerwiec	11
Lipiec-sierpień	13
Wrzesień	10

Wykres na Rys. 2. przedstawia tłumienność atmosfery w funkcji częstotliwości i uwzględnia cztery przypadki: Singapur zimą, La Rinconada zimą, Łódź latem oraz Łódź zimą. Najwyższą tłumienność obserwuje się w Singapurze, co wynika z jego tropikalnego klimatu i wysokiej wilgotności powietrza, natomiast najniższą – w La Rinconadzie, położonej wysoko w górach, gdzie niskie ciśnienie i mała zawartość pary wodnej znacząco redukują tłumienność. Łódź znajduje się pomiędzy tymi skrajnościami, a różnice sezonowe są wyraźne – zimą tłumienność jest niższa niż latem, co odpowiada niższej wilgotności bezwzględnej powietrza w tym okresie. Różnice te

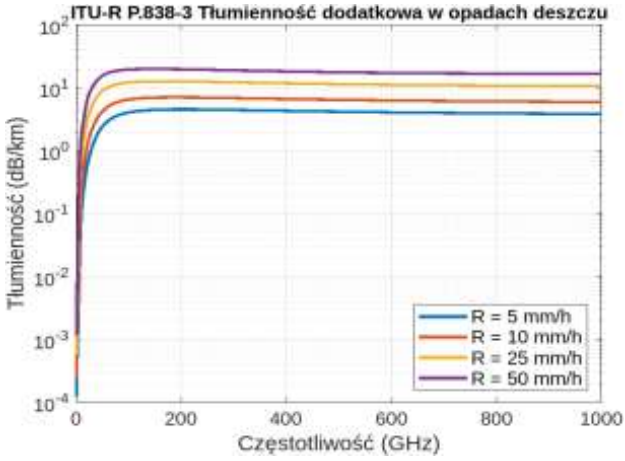
są szczególnie widoczne w pasmach częstotliwości związanych z rezonansami cząsteczek wody, takich jak okolice 22 GHz i 183 GHz. Tłumienności dla Łodzi w funkcji pory roku i warunków pogodowych różnią się kilkukrotnie (w dB), a w funkcji lokalizacji geograficznej rozważanych tu miast o ponad rząd wielkości (w dB). Systemy komórkowe pracujące w obecnie wykorzystywanych pasmach częstotliwości nie muszą sobie radzić z taką zmiennością w funkcji czasu i miejsca instalacji.



Rys. 2. Tłumienność w gazach atmosferycznych w funkcji częstotliwości dla wybranych lokalizacji i przykładowych warunków atmosferycznych: Singapur – zima: 1013 hPa, 307,15 K, 34,5 g/m³, La Rinconada – zima: 540 hPa, 263,15 K, 1,5 g/m³, Łódź-lato: 1015 hPa, 293 K, 13 g/m³, Łódź-zima: 1015 hPA, 268 K, 4 g/m³.

2.2. Absorpcja w opadach atmosferycznych

Dla częstotliwości pasm subterahercowego i terahercowego istotne znaczenie mają również opady atmosferyczne. Dla zilustrowania tego zagadnienia na Rys. 3 pokazano zależność tłumienność [dB/km] w funkcji częstotliwości, przy wybranych wartościach intensywności opadów deszczu w zakresie 5–50 mm/h.



Rys. 3. Tłumienność w opadach deszczu w funkcji częstotliwości dla wybranych wartości intensywności opadów

Powyżej 100 GHz wszystkie skumulowane składowe tłumienności mogą osiągnąć wartości przekraczające kilkadziesiąt dB/km. Wówczas transmisja wymaga kompensacji strat np. przez zwiększenie mocy nadawania, przełączenia na pasma mniejszej częstotliwości lub ograniczenia zasięgu.

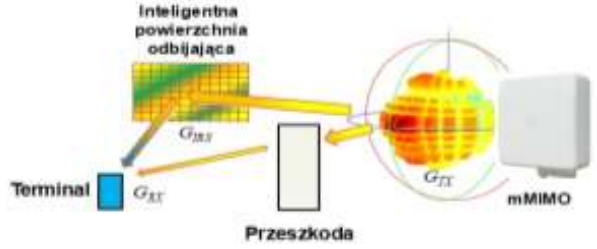
3. BILANS ŁĄCZA

Jedną z kluczowych technologii przewidywanych do wykorzystania w sieciach 6G są tzw. inteligentne powierzchnie odbijające (IRS – ang. Intelligent Reflective Surfaces). Wykorzystanie nawet stosunkowo niedużych elementów, np. o powierzchni ok. 10 cm^2 może skutkować poprawą bilansu łącza nawet o 20 dB (wartość dla idealnej powierzchni) dla częstotliwości rzędu 300 GHz. Warto dodać, że IRS są dopiero w fazie opracowywania i ich rzeczywiste parametry mogą się okazać gorsze od zakładanych.

Zależność mocy odebranej z uwzględnieniem inteligentnych powierzchni odbijających (Rys. 4) przyjmuje postać:

$$P_{\mathrm{RX}}(f,d) = P_{\mathrm{TX}} + G_{\mathrm{TX}} + G_{\mathrm{RX}} + G_{\mathrm{IRS}} - L_{\mathrm{FS}}(f,d) - L_{\mathrm{gas}}(f,d) - L_{\mathrm{precip}}(f,d) - L_{\mathrm{misc}} \quad (2)$$

$P_{\mathrm{RX}}(f,d)$ – moc odbierana [dBm] w funkcji częstotliwości f i odległości d.
P_{TX} – moc nadawania [dBm],
G_{TX} – zysk anteny nadawczej [dBi],
G_{RX} – zysk anteny odbiorczej [dBi],
G_{IRS} – zysk odbicia od inteligentnej [dB]
$L_{\mathrm{FS}}(f,d)$ – straty propagacyjne w wolnej przestrzeni [dB],
$L_{\mathrm{gas}}(f,d)$ – tłumienność gazowa [dB] zgodnie z ITU-R P.676 [5],
$L_{\mathrm{precip}}(f,d)$ – tłumienność opadowa [dB] – obejmuje deszcz, śnieg, grad – zgodnie z ITU-R P.838 [3],
L_{misc} – pozostałe straty [dB], które mogą obejmować tłumienie spowodowane przez mgłę, czy opady (ITU-R P.840 [4]), straty w torach antenowych, niedopasowanie polaryzacji, propagację wielodrogową, itp.



Rys. 4. Uwzględniane zyski anten i powierzchni odbijającej

W przypadku jednoczesnego istnienia ścieżki bezpośredniej i odbitej od powierzchni odbijającej (Rys. 4) wyznaczenie mocy odebranej jest znacznie trudniejsze i zależy między innymi od różnicy faz w punkcie odbioru dla składowych bezpośredniej i odbitej od IRS. Maksymalizacja mocy odebranej (superpozycja konstruktywna) wymaga wówczas, by wartości argumentu współczynnika odbicia dla poszczególnych meta-atomów doprowadziły do synfazowości tych składowych w antenie odbiorczej. Możliwe jest zastosowanie wielu innych scenariuszy optymalizacyjnych, które obejmują nie tylko IRS, ale także sterowanie wiązkami anten mMIMO.

4. WYNIKI

W [7] autorzy podjęli próbę oszacowania możliwych do uzyskania zasięgów działania łączy systemów 6G w pasmach subterahercowych dla kilku wybranych scenariuszy. Do rozważań przyjęto transmisję w pasmie 300 GHz (granica fal subterahercowych i terahercowych), w kanale o szerokości 30 GHz i przewidywane parametry urządzeń i systemów antenowych. Rozważane scenariusze uwzględniają łączność z terminalem mobilnym, jak również łącze typu punkt-punkt. Przedstawione wyniki miały na celu oszacowanie m.in. w jakiej odległości od punktu dostępowego możliwe jest uzyskanie zakładanej szybkości transmisji (100 Gb/s). W cytowanych analizach uwzględniono jednak wyłącznie tłumienność wolnej przestrzeni, a w przypadku łącza punkt-punkt również zgrubnie wpływ atmosfery (założono wartość 3 dB dla łącza o długości 1 km) oraz opadów o intensywności 20 mm/h (10,7 dB/km). Pominięto jednak aspekty dotyczące wpływu dobowej/rocznej zmienności parametrów atmosfery na bilans łącza, co jest przedmiotem rozważań w niniejszym artykule, przy założeniach jak w Tab. 2.

Tab. 2. Tłumienność atmosfery dla wybranych lokalizacji i częstotliwości 300 GHz [dB/km]

Lokalizacja	*Wartość*
La Rinconada	0,68
Singapur	28,72
Łódź – lato	9,37
Łódź – zima	3,30

Zakładając scenariusz łącza punkt-punkt [7] z użyciem anten o zysku 55 dBi oraz mocy EIRP nadajnika 64,4 dBm oraz dla założonej szybkości transmisji 100 Gb/s możliwe są do uzyskania zasięgi łączności jak w Tab. 3.

Tab. 3. Zasięg łącza punkt-punkt dla wybranych lokalizacji i częstotliwości 300 GHz [m]

Lokalizacja	*Zasięg [m]*
La Rinconada	6100
Singapur	770
Łódź – lato	1653
Łódź – zima	3067

Zakładając scenariusz łącza w dół w kontekście komunikacji z urządzeniem mobilnym [7] z użyciem anten MIMO o zyskach 29,1 dBi (nadajnik, 512 elementów) oraz 17,1 dBi (odbiornika, 32 elementy), mocy EIRP nadajnika 65,6 dBm oraz dla założonej szybkości transmisji 100 Gb/s możliwe są do uzyskania zasięgi łączności jak oszacowane w Tab. 4.

Tab. 4. Zasięg łącza z terminalem mobilnym dla wybranych lokalizacji i częstotliwości 300 GHz [m] bez i z wykorzystaniem IRS (20 dB zysku)

Lokalizacja	*Zasięg[m]*	*Zasięg z IRS [m]*
La Rinconada	80	762
Singapur	65	299
Łódź – lato	74	481
Łódź – zima	78	635

W rozważaniach uwzględniono także wpływ ewentualnego zastosowania IRS (o założonej wartości 20 dB zysku) przy istotnym uproszczeniu rozważań, polegającym na zaniedbaniu różnicy długości trasy propagacji wzdłuż ścieżki odbitej od IRS w stosunku do ścieżki bezpośredniej. Jak widać, uwzględnienie dodatkowych 20 dB zysku związanego z wykorzystaniem IRS pozwala na kilkukrotne zwiększenie zasięgu łącza w systemie punkt-punkt.

5. WNIOSKI I PODSUMOWANIE

Omówiona w artykule analiza łącza radiowego w pasmie 300 GHz (leżącego na granicy pasm subterahercowego i terahercowego), przeprowadzona dla przykładowych lokalizacji geograficznych i reprezentatywnych dla nich parametrów atmosfery (ciśnienie, wilgotność, temperatura) pokazuje silną zależność bilansu łącza na zmiany warunków atmosferycznych. Przykładowe obliczenia przeprowadzono dla parametrów systemowych opisanych w [7], gdzie jednak omówienie wpływu atmosfery na propagację zostało pominięte. Przeprowadzone przez nas obliczenia ilustrują skalę nowych problemów związanych z wykorzystywaniem zakresów subterahercowych i terahercowych w systemach komórkowych oraz pokazują ilościowo (choć tylko na jednym przykładzie) korzyści z zastosowania inteligentnych powierzchni odbijających.

LITERATURA

[1] European Telecommunications Standards Institute, TeraHertz Technology (THz): Identification of Use Cases for THz Communication Systems. ETSI Group Report DGR/THz-001 V1.1.1 (January 2024).

[2] European Telecommunications Standards Institute, TeraHertz Technology (THz): Identification of Frequency Bands of Interest for THz Communication Systems. ETSI Group Report DGR/THz-002 V1.1.1 (March 2024).

[3] International Telecommunication Union. Recommendation ITU-R P.838-3: Specific Attenuation Model for Rain for Use in Prediction Methods. Geneva: ITU, March 2005.

[4] International Telecommunication Union, Radiocommunication Sector. 2019. Recommendation ITU-R P.840-8: Attenuation Due to Clouds and Fog.

[5] International Telecommunication Union. Recommendation ITU-R P.676-13: Attenuation by Atmospheric Gases and Related Effects. Geneva: ITU, August 2022.

[6] Kaushik, Aryan, Rohit Singh, Shalanika Dayarathna, Rajitha Senanayake, Marco Di Renzo, Miguel Dajer, Hyoungju Ji, Younsun Kim, Vincenzo Sciancalepore, Alessio Zappone, and Wonjae Shin. *Towards Integrated Sensing and Communications for 6G: A Standardization Perspective*. arXiv preprint arXiv:2308.01227, August 2, 2023.

[7] Rikkinen, K., Kyösti, P., Leinonen, M. E., Berg, M., and Parssinen, A. 2020. "THz Radio Communication: Link Budget Analysis toward 6G." IEEE Communications Magazine 58, no. 11: 22–27.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

WPŁYW WZAJEMNYCH SPRZĘŻEŃ POMIĘDZY META-ATOMAMI NA PARAMETRY META-POWIERZCHNI NA PRZYKŁADZIE INTELIGENTCH POWIERCHNI ODBIJAJĄCYCH

INFLUENCE OF MUTUAL COUPLING BETWEEN META-ATOMS ON META-SURFACE PARAMETERS AS EXAMPLED BY INTELLIGENT REFLECTIVE SURFACES

Łukasz Jopek[1]; Sławomir Hausman[2];

[1] Instytut Elektroniki Politechnika Łódzka, Łódź, lukasz.jopek@p.lodz.pl
[2] Instytut Elektroniki Politechnika Łódzka, Łódź, slawomir.hausman@p.lodz.pl

DOI: 10.15199/59.2025.4.32

Streszczenie: **W artykule opisano jeden z możliwych problemów związanych z projektowaniem Inteligentnych Powierzchni Odbijających (IRS - Inteligent Reflective Surface) w postaci meta-powierzchni. Jest to problem związany z wzajemnymi oddziaływaniami na siebie sąsiadującymi ze sobą komórek w meta-powierzchni- tzw. sprzężeniami. W artykule posłużono się prostym modelem obwodowym. Obejmuje on model komórki elementarnej, jako układu RLC o stałych skupionych, oraz model sprzęg indukcyjnych i pojemnościowych pomiędzy dwiema sąsiadującymi ze sobą komórkami. Pozwala on na analityczne wyznaczenie charakterystyk amplitudowych obu komórek. Mają one zasadnicze znaczenie dla końcowych parametrów takiej komórki - wartości fazy i modułu. Obliczenia analityczne wsparto wynikami symulacji numerycznych przeprowadzonych metodą elementu skończonego w dziedzinie czasu (FDTD).**

Abstract: **The article describes one of the problems related to the design of Intelligent Reflective Surfaces (IRS) in the form of meta-surfaces. These problem is related to the mutual interactions of neighboring unit cells in the meta-surface. A simple unit cell can be modeled as lumped RLC circuit. Mutual interactions can be inductive and capacitive. Model allows for the analytical determination of the amplitude characteristics of both unit cells. These couplings are of fundamental importance for the final parameters of such a cell - phase delay and reflection modulus. The analytical calculations are supported by the results of numerical simulations carried out using the finite difference time-domain-method in the time domain (FDTD).**

Słowa kluczowe: **metamateriały, sprzężenia wzajemne, inteligentne powierzchnie odbijające, FDTD.**

Keywords: **4-5 metamaterials, mutual couplings, Intelligent Reflective Surfaces, FDTD.**

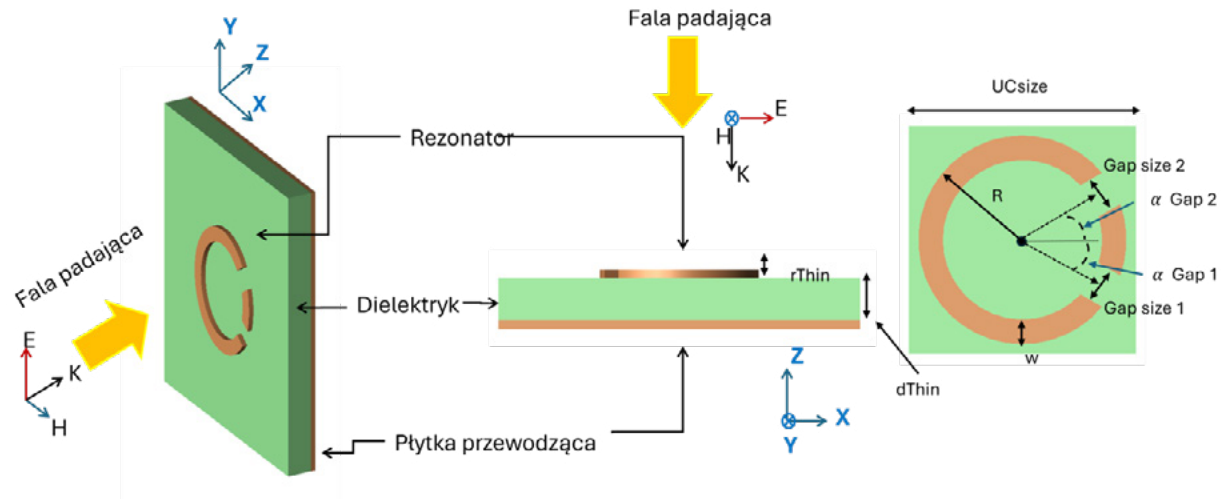
1. Metapowierzchnie

Metamateriały to periodyczne struktury, które zostały sztucznie wytworzone. Podstawowym elementem składowym jest komórka elementarna (dalej określana jako Unit Cell - UC) znana również w literaturze jako meta-atom. Typowy rozmiar komórki elementarnej UC koreluje z długością fali, dla której meta-powierzchnia została zaprojektowana, i zawiera się w przedziale 0,1 dla 0.8 λ. Są to więc materiały pasmowe, które zachowują swoje własności tylko w ograniczonym zakresie częstotliwości. Do typowym zastosowań można zaliczyć m.in.: filtry[9, 10], soczewki [12], absorbery [4, 6, 15] , czy różnego typu powierzchnie odbijające [1, 13, 14]. Szczególnie interesujące wydają się Inteligentne Powierzchnie Odbijające (IRS), które mają zdolność do kontrolowanego odbijania fali elektromagnetycznej, niekoniecznie zgodnego z prawem Snella. UC czasami nazywa się rezonatorem. Dzieje się tak dlatego, ponieważ zazwyczaj posiada on jedną bądź więcej częstotliwości rezonansowych. Układ taki bowiem zachowuje się podobnie jak skupiony układ RLC, posiada pewną indukcyjność własną (związaną z kształtem i rozmiarem przewodnika), pojemnością własną (związaną z rozmiarem i parametrami użytego dielektryka, a także pojemnościami w szczelinach [11] jeśli występują, oraz tzw. pojemnościami powierzchniowymi [11]), a także z rezystywnością, będą konsekwencją wykorzystania konkretnego typu przewodnika (miedź, złoto itp.). Podobnie jak klasyczne układy RLC posiadają charakterystykę częstotliwościową obraz dobroć, zależną od L , C oraz R. Źródłem energii dla takiego układu jest zewnętrzna fala EM. Zależnie od wzajemnej relacji powierzchni rezonatora oraz padającej fali pobudzenie pochodzić może od składowej magnetycznej (H), elektrycznej (E) lub od obydwu (EH). Częstotliwość rezonansowa zależna jest od geometrii rezonatora, gdyż od niej właśnie zależą wartości L , C oraz R [2, 16]. W niniejszej publikacji skupiono się na analizę rezonatorów ze szczeliną (Split Ring Rezonator - SRR). Przykład budowy UC ilustruje Rysunek 1.

1.1. Inteligentne powierzchnie obijające (IRS)

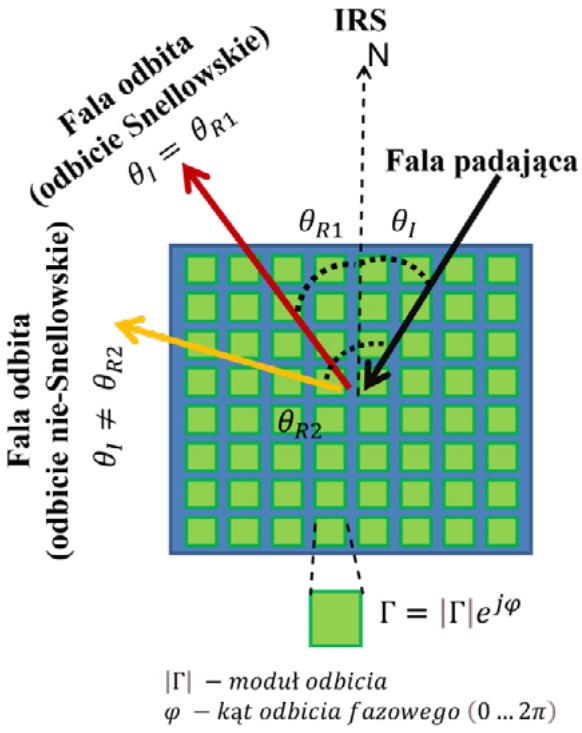
Powierzchnie IRS [3, 8, 11] mogą odbijać fale elektromagnetyczne w kierunkach innych niż wynikające z prawa Snella, a przy tym dynamicznie zmienianych zależnie od potrzeb, np. od zmian liczby i lokalizacji odbiorników. IRS mogą być modelowane na kilka sposobów, m.in: jako meta-powierzchnia, której komórki elementarne (UC) charakteryzują się jedynie dwoma parametrami: fazą odbicia (φ) oraz modułem odbicia ($|\Gamma|$), oraz jako meta-powierzchni zawierającej rezonatory lub też macierzy układów RLC. Różnica odbić, pomiędzy falą

padającą oraz falą odbitą została przedstawiona na Rysunku 2, gdzie kąt fali padającej został oznaczony jako θ_I, natomiast kąt fali odbitej jako θ_{R1}, dla przypadku odbicia zgodnego z



Rysunek 1: Typowa budowa rezonatora (wysokoimpedacyjne powierzchnie odbijające

prawem Snella oba kąty mają tę samą wartość (Rysunek 2, czerwona linia). Natomiast w przypadku odbicia anomalnego (niesnellowskiego) kąty θ_I i θ_{R2} są różne (Rysunek 2, żółta linia).



Rysunek 2: Ilustracja działania powierzchni IRS

2. Modelowanie sprzężeń pomiędzy meta-atomami

2.1. Modelowanie polowe

Pomiędzy rezonatorami może dochodzić do sprzężeń o charakterze magnetycznym i elektrycznym [7]. Najogólniej rzecz ujmując jest to efekt wzajemnego przenikania się pól generowanych przez sąsiednie rezonatory, który można przedstawić następująco [1]:

$$K = K_M + K_E = \frac{\iiint \mu \overrightarrow{H_1} \cdot \overrightarrow{H_2} dv}{\sqrt{\iiint \mu \left|\overrightarrow{H_1}\right|^2 dv \times \iiint \mu \left|\overrightarrow{H_2}\right|^2 dv}} + \frac{\iiint \varepsilon \overrightarrow{E_1} \cdot \vec{E}_2 dv}{\sqrt{\iiint \varepsilon \left|\overrightarrow{E_1}\right|^2 dv \times \iiint \varepsilon \left|\overrightarrow{E_2}\right|^2 dv}} \quad (1)$$

gdzie:

K – całkowity współczynnik sprzężenia (pomiędzy rezonatorami),

K_M – współczynnik sprzężenia magnetycznego,

K_E – współczynnik sprzężenia elektrycznego,

$\overrightarrow{E_n}$ – pole elektryczne generowane przez n-ty rezonator,

H_n- pole magnetyczne generowane przez n-ty rezonator,

ε- względna przenikalność elektryczna ośrodka, w którym umieszczono rezonatory,

μ - względna przenikalność magnetyczna ośrodka, w którym umieszczono rezonatory.

Można się spodziewać silniejszych sprzężeń dla układów rezonatorów umieszczonych blisko siebie. Jeśli wartość współczynnika K będzie większa niż tzw. wartość krytyczna wtedy, dla dwóch sprzężonych rezonatorów możemy spodziewać się charakterystyki częstotliwościowej dwugarbnej, a w przypadku układu wielu sprzężonych rezonatorów – charakterystyki wielogarbnej.

2.2. Modelowanie obwodowe

Jak wspomniano wyżej, w rezonatorze, pod wpływem zewnętrznej EM indukuje się prąd o częstotliwości zależnej od indukcyjności i pojemności własnej. Tak więc rezonator może być, w pewnym uproszczeniu, traktowany

jako obwód skupiony LC. Przyjmując więc pewnie założenia można modelować rezonator jako obwód (R)LC, korzystając z technik analitycznych o bardzo niskiej złożoności pamięciowej i obliczeniowej. Pozwala to na znaczne zredukowanie czasu potrzebnego np. w przypadku optymalizacji metapowierzchni złożonej z wielu rezonatorów. Należy dodatkowo wziąć pod uwagę charakter sprzężenia: może być ono magnetyczne (indukcyjne), elektryczne (pojemnościowe) lub mieszane. Sprzężenia wzajemne dla siatki UC mogą być przedstawione w formie macierzowej [9], poniżej jednak przedstawiona została uproszona wersja równań pozwalająca na wyznaczenie częstotliwości rezonansowych dwóch sprzęgniętych ze sobą układów LC:

$$\omega_1 = \sqrt{\frac{R_B - R_C}{R_A}} \tag{9}$$

$$\omega_2 = \sqrt{\frac{R_B + R_C}{R_A}} \tag{10}$$

$$\text{dla } \omega_1 \leq \omega_2$$

gdzie:

$$R_A = 2\left(L_1C_1L_2C_2 - L_M{}^2C_1C_2 - C_M^2L_1L_2 + L_M^2C_M^2\right) \tag{11}$$

$$R_B = (L_1C_1 + L_1L_2 - 2L_MC_M) \tag{12}$$

$$R_C = \sqrt{R_B^2 - 2R_A} \tag{13}$$

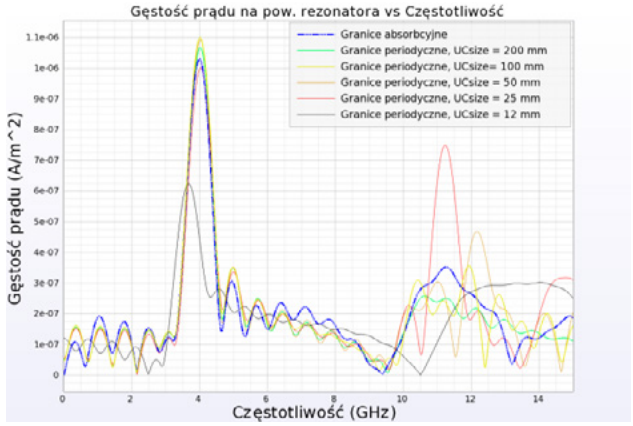
L_1, L_2 – odpowiednio indukcyjności własne UC,

C_1, C_2 – odpowiednio pojemności własne UC,

L_m, C_m – współczynniki sprzężenia: indukcyjne i pojemnościowe (0…1).

3. Rezultaty

Analizę wpływu sprzężeń pomiędzy UC podzielono na dwa aspekty: (1) analiza charakterystyki amplitudowej prądu j_c płynącego przez wybrany UC oraz analizę wpływu sprzężeń z siadującymi UC na charakterystyki fazy współczynnika odbicia φ i modułu $|\Gamma|$. Pierwszy eksperyment polegał na zbudowaniu rezonatora o dużym rozmiarze samej komórki (Rysunek 3). Po stwierdzeniu, że sam parametr UCsize nie ma wpływu ani na charakterystykę j_c ani też parametry odbiciowe *(φ i $|\Gamma|$)* wykonano dwie symulacje numeryczne metodą FDTD (oprogramowanie Remcom XFDTD™), z odpowiednio warunkami brzegowymi pochłaniającymi („Absorbing boundaries") i periodycznymi („Perioic boundaries"). Przy rozmiarze komórki UCsize=200mm charakterystyki amplitudowe j_c były praktycznie takie same, co dowodzi braku wpływu sąsiadujących elementów na pracę badanego UC (dla symulacji z granicami periodycznymi). Był to punkt wyjścia do dalszej analizy. Zgodnie z [5, 7, 10] wraz ze wzrostem wartości sprzężenia K rosnąc amplituda drgań częstotliwości rezonansowej, aż do przekroczenia tzw. Wartości krytycznej. Wtedy (dla układu dwóch sprzężonych ze sobą UC) powinno otrzymać się charakterystykę dwu-garbną, oraz wielogarbną dla układu wielu sprzężonych ze sobą UC. Ponieważ środowisko symulacyjne pozwalało jedynie na symulowanie (nieskończonej) siatki, więc rozpatrzony zostanie wpływ wszystkich sąsiadujących UC na badany UC. Zgodnie z oczekiwaniami [7], wraz ze zmniejszaniem wartości parametru UCsize, rosły sprzężenia wzajemne, ponieważ rosły natężenia strumieni E i H generowanych przez sąsiadujące elementy. Dla UCsize = 200 i 100 mm widać pewien podstawowej wzrost amplitudy rezonansowej (rezonans własny UC), później zaś (UCsize = 50, 25 i 12 mm) charakterystyka wyraźnie zmienia swój charakter na wielogarbny. Drugi aspekt analizy wpływu sprzężeń pomiędzy UC obejmował pomiar φ i $|\Gamma|$.



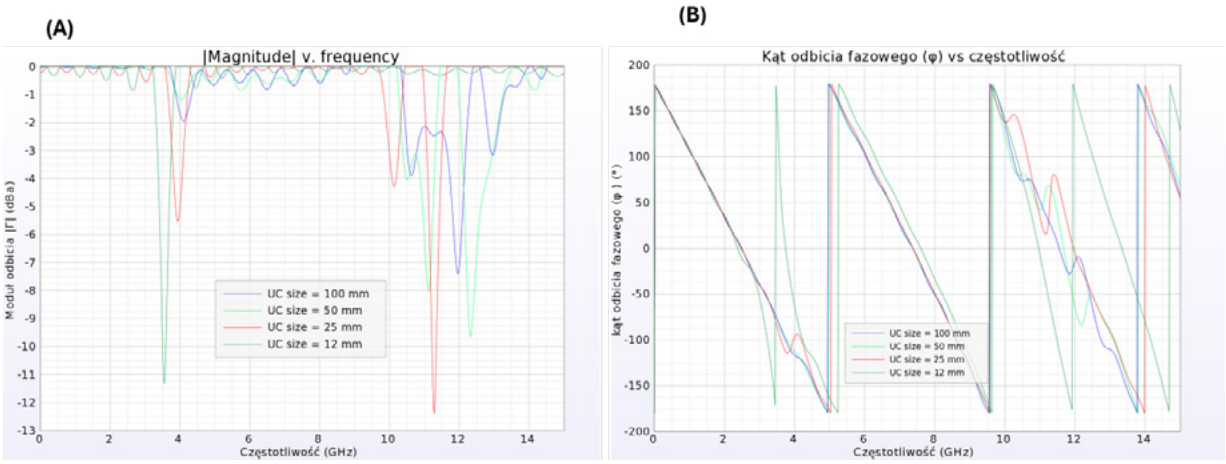
Rysunek 3: Charakterystyka gęstości prądu na powierzchni rezonatora dla różnych rozmiarów UC – parametr UCsize oznacza całkowity rozmiar komórki (badanie wpływu sprzężeń wzajemnych z otaczającymi rezonatorami).

Rysunek 4 przedstawia zestaw charakterystyk dla kąta odbicia fazowego φ orasz modułu odbicia $|\Gamma|$. Zarówno wartość φ jak i $|\Gamma|$ zmienią się pod wpływem sprzężeń pomiędzy sąsiednimi UC. Zmiany te są niestety nieliniowe w funkcji częstotliwości W przypadku kąta opóźnienia fazowego zmiany mogą być stosunkowo niewielkie, jak np. dla $f = 8GHz$, gdzie największa zmiana wynosi ok 16°, lub stosunkowo duże – np. dla $f = 13.5\ GHz$, gdzie ta różnica sięga już do ok 150°. Z kolei różnice w module odbicia mogą sięgać od kilku dB do nawet kilkunastu (przypadek f = *ok. 3.8 GHz*, *4 GHz* lub ok *11.5 GHz)*. Może to powodować trudności w projektowaniu meta-powierzchni, w tym przypadku powierzchni IRS, o ile nie weźmie się pod uwagę natury i charakteru tych sprzężeń. Pewnym wyjściem z sytuacji mogło by się wydawać zredukowanie wpływu tych sprzężeń poprzez znacznie zwiększenie rozmiaru pojedynczego UC. Jednak takie rozwiązanie ma dwie zasadnicze wady: po pierwsze, jak można choćby pokazuje Rysunek 3, wpływ sprzężeń może obejmować duże rozmiary. W rozważanych przypadku, gdzie rozmiar rezonatora wynosił ok 10 mm,

sprzężenia zanikają dopiero przy rozmiarze całkowitym UC wynoszącym ok 200 mm. Drugi zaś przeszkoda związana jest z faktem, że rozmiar całkowity UC (lub też odstęp pomiędzy UC) ma istotny wpływ na charakterystykę kierunkową powierzchni IRS. Typowo wynosi ona między *0.3* a *0.8* λ, najczęściej przyjmuje się ok 0.5 λ. Przekreśla to możliwość niwelowania wzajemnych sprzężeń poprzez zwiększanie rozmiaru całkowitego UC.

W niniejszej publikacji pokazano wpływ wzajemnych sprzężeń pomiędzy elementami metapowierzchni. zwanymi UC na jej końcowe własności na przykładzie metapowierzchni typu IRS. Wykazano wpływ sprzężeń na wartości współczynnika odbicia (φ *i* $|\Gamma|$), które decydują o charakterystyce kierunkowej IRS. To z kolei powoduje, że muszą być one brane pod uwagę podczas projektowania / modelowania tego typu struktur

4. Podsumowanie



Rysunek 4: Charakterystyki kąta odbicia fazowego oraz modułu odbicia w funkcji rozmiaru UC (wpływu sprzężeń pomiędzy sąsiadującymi UC): (A) moduł odbicia, (B) kąt odbicia fazowego.

LITERATURA

[1] Agarwal, K. et al. 2016. Wearable AMC Backed Near-Endfire Antenna for On-Body Communications on Latex Substrate. *IEEE Transactions on Components, Packaging and Manufacturing Technology*. 6, 3 (2016), 346–358.

[2] Baena, J.D. et al. 2005. Equivalent-circuit models for split-ring resonators and complementary split-ring resonators coupled to planar transmission lines. *IEEE Transactions on Microwave Theory and Techniques*. 53, 4 (Apr. 2005), 1451–1461.

[3] Cui, T.J. et al. 2014. Coding Metamaterials, Digital Metamaterials and Programming Metamaterials. arXiv.

[4] Dongju, S., Hyuk-Kee Lee, Lim, Sungjoon 2007. Flexible subterahertz metamaterial absorber fabrication using inkjet printing technology | SpringerLink. *Applied Physics B*. (2007).

[5] Gao, L. et al. 2013. Magnetic coupling resonant wireless energy transmission coil quantitative relation modeling and simulation research. *2013 10th IEEE International Conference on Control and Automation (ICCA)* (Hangzhou, China, Jun. 2013), 202–207.

[6] Gholipour, M. et al. 2019. Simple EBG surface for X-band radar cross section reduction. *Aeu-International Journal of Electronics and Communications*. 100, (2019), 47–55.

[7] Hałgas, S. et al. 2025. Analysis and Applications of Magnetically Coupled Resonant Circuits. *Electronics*. 14, 2 (Jan. 2025), 312. DOI:https://doi.org/10.3390/electronics14020312.

[8] Holloway, C.L. et al. 2012. An Overview of the Theory and Applications of Metasurfaces: The Two-Dimensional Equivalents of Metamaterials. *IEEE Antennas and Propagation Magazine*. 54, 2 (Apr. 2012), 10–35.

[9] Hong, J.-S. 2001. *Microstrip filters for RF*. John Wiley & Sons, Inc.

[10] Huang, Y. et al. 2014. Systematical analysis for the mixed couplings of two adjacent modified split ring resonators and the application to compact microstrip bandpass filters. *AIP Advances*. 4, 10 (Oct. 2014), 107119.

[11] Kliks, A. 2023. Metamaterialy, konfigurowalne matryce antenowe i komunikacja holograficzna. Wstepna analiza nowej koncepcji bezprzewodowej transmisji danych. *Przegląd Telekomunikacjny- Wiadomości Telekomunikacyjne*. 1, 4 (Aug. 2023), 31–41. DOI:https://doi.org/10.15199/59.2023.4.4.

[12] Liang, M. and Xin, H. 2015. 3D printed microwave and THz components. *2015 Asia-Pacific Microwave Conference (APMC)* (Nanjing, China, Dec. 2015), 1–3.

[13] Malhat, H.A. et al. RCS Reduction from Conformal Surfaces Using Plasma-Based AMC Arrays. *Plasmonics*.

[14] Ramos, W.T.S. et al. 2017. Design of the artificial magnetic conductors with meander line elements: reduction in the first and second resonant frequencies. *Materials Research Express*. 4, 7 (2017), 075801.

[15] Valagiannopoulos, C.A. et al. 2015. Perfect magnetic mirror and simple perfect absorber in the visible spectrum. *Physical Review B - Condensed Matter and Materials Physics*. 91, 11 (2015).

[16] Vallecchi, A. et al. 2019. Analytical model of the fundamental mode of 3D square split ring resonators. *Journal of Applied Physics*. 125, 1 (Jan. 2019), 014901. DOI:https://doi.org/10.1063/1.5053482.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

ANALIZA DYNAMICZNEJ REALOKACJI JEDNOSTEK DU W SIECIACH XHAUL Z KOMUTACJĄ PAKIETÓW PRZY ZMIENNYM RUCHU

ANALYSIS OF DYNAMIC DU REALLOCATION IN PACKET-SWITCHED XHAUL NETWORKS UNDER VARYING TRAFFIC

Mirosław Klinkowski[1];

[1] Instytut Łączności - Państwowy Instytut Badawczy, ul. Swojczycka 38, 51-501 Wrocław, M.Klinkowski@il-pib.pl

DOI: 10.15199/59.2025.4.33

Streszczenie: Artykuł prezentuje wyniki analizy dynamicznej realokacji jednostek DU w pakietowych sieciach Xhaul obsługujących zmienne wymagania ruchowe w cyklu dobowym. Badanie dotyczy sieci umożliwiającej agregacje DU w mniejszej liczbie węzłów przetwarzających wraz z deaktywacją nieużywanych węzłów w stanach mniejszego obciążenia ruchowego. Analiza realokacji DU przeprowadzona w scenariuszach sieci pierścieniowej wskazuje na możliwość redukcji liczby aktywnych w czasie węzłów przetwarzających średnio w granicach $10-20\%$.

Abstract: The article presents the results of an analysis of the dynamic reallocation of DUs in packet-switched Xhaul networks under varying traffic demands over the daily cycle. The study concerns a network that allows DU aggregation in a smaller number of processing nodes, along with the deactivation of unused nodes in states of lower traffic load. The DU reallocation analysis carried out in ring network scenarios indicates the possibility of reducing the number of active processing nodes over time by $10-20\%$, on average.

Słowa kluczowe: alokacja zasobów, optymalizacja sieci, radiowe sieci dostępowe, sieci pakietowe, Xhaul.

Keywords: network optimization, packet networks, radio access networks, resource allocation, Xhaul.

1. WSTĘP

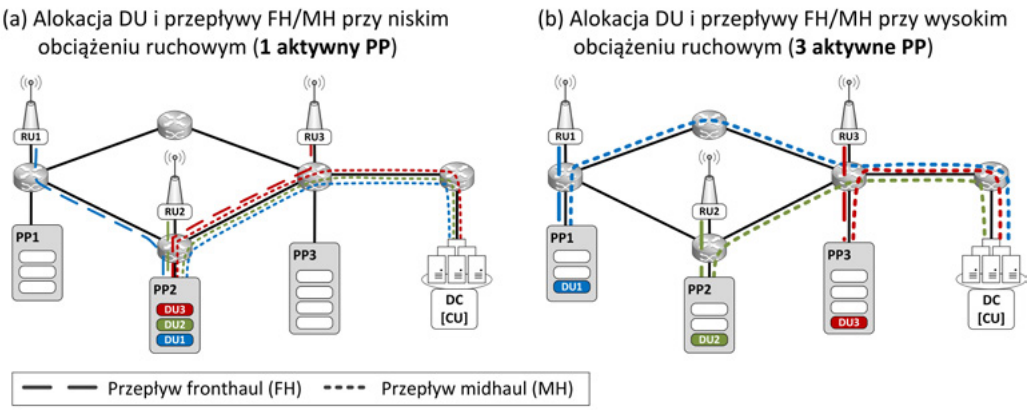
Sieci Xhaul stanowią infrastrukturę transportową łączącą elementy zcentralizowanych radiowych sieci dostępowych (C-RAN), takich jak jednostki radiowe (RU) rozproszone na pewnym obszarze, z zasobami obliczeniowymi "chmury radiowej" realizującymi funkcje przetwarzania sygnału radiowego oraz z siecią rdzeniową [1]. Przetwarzanie sygnału realizowane jest w jednostkach rozproszonych (DU) znajdujących się w sąsiedztwie RU, oraz jednostkach centralnych (CU), zazwyczaj zlokalizowanych w koncentratorze. Termin Xhaul obejmuje segmenty sieci fronthaul (FH), midhaul (MH) i backhaul w sieciach mobilnych 5G/6G.

Sieci z komutacją pakietów oparte na transmisji Ethernet, rozszerzone o mechanizmy TSN (Time-Sensitive Networking) umożliwiające obsługę przepływów danych wrażliwych na opóźnienia, stanowią jedno z podstawowych rozwiązań proponowanych do implementacji sieci Xhaul [2]. Technologia Ethernet zwiększa efektywność łączy transmisyjnych dzięki statystycznej multipleksacji pakietów przenoszących dane radiowe, umożliwiając skalowalną i elastyczną realizację połączeń w pakietowej sieci Xhaul [3].

Podstawowe zagadnienie optymalizacyjne w sieciach Xhaul dotyczy zapewnienia zasobów przetwarzania radiowego oraz łączności w sieci transportowej pomiędzy elementami architektury C-RAN [4]. W analizowanej w niniejszej pracy pakietowej sieci Xhaul problem ten sprowadza się do wyboru węzłów obliczeniowych (ang. processing pool, PP) dla alokacji jednostki DU oraz do wyznaczenia tras przepływów pakietowych FH i MH w sieci, przy zadanych wymaganiach ruchowych (związanych z jednostkami RU), które konkurują o ograniczone zasoby transmisyjne. Warunkiem jest jednoczesne zapewnienie, że opóźnienia przepływów nie przekraczają dopuszczalnych limitów. Problem wspólnej alokacji jednostek DU w węzłach PP oraz routingu przepływów FH i MH w pakietowej sieci Xhaul jest problemem trudnym [4], między innymi z powodu niedeterministycznego charakteru transmisji pakietów, prowadzącego do nieprzewidywalnych opóźnień wynikających z buforowania pakietów w przełącznikach.

Badania dotyczące alokacji zasobów przetwarzania i realizacji połączeń w pakietowych sieciach Xhaul obejmowały głównie scenariusze statyczne, w których sieć była optymalizowana przy założeniu określonego, nie zmiennego w czasie obciążenia [4, 5]. Przepływy ruchu w sieciach Xhaul nie są jednak statyczne lecz podlegają zmianom spowodowanym zmiennym wykorzystaniem sieci radiowej czy też dynamicznym zmianom jej konfiguracji lub stanu. Podobnie, powiązania między elementami RU a zasobami obliczeniowymi PP realizującymi funkcje DU mogą się zmieniać w wyniku różnych czynników, takich jak uszkodzenie elementów sieci, zmiany w obciążeniu sieci, czy też wymagania oszczędności energetycznej [6].

Autorzy [6] rozważali scenariusz całkowicie optycznej, światłowodowej sieci Xhaul, w której jednostki DU są realokowane pomiędzy węzłami PP w zależności od obciążenia sieci, umożliwiając wyłączanie nieużywanych zasobów PP dla poprawy efektywności energetycznej sieci C-RAN. Praca [6] prezentuje koncepcję sieci, jakkolwiek bez analizy jej wydajności. W niniejszej pracy, opierając się na idei realokacji jednostek DU w odpowiedzi na zmiany obciążenia sieci, analizujemy korzyści wynikające z takiego scenariusza w pakietowej sieci Xhaul, na podstawie wy-



Rysunek 1: Przykład alokacji DU w węzłach obliczeniowych PP wraz z routingiem przepływów FH i MH w pakietowej sieci Xhaul w przypadku wymagań ruchowych: (a) niskich i (b) wysokich.

ników symulacji. Liczba prac dotyczących optymalizacji dynamicznych sieci Xhaul pozostaje niewielka, a zgodnie z naszą wiedzą problem dynamicznej alokacji DU w pakietowej sieci Xhaul nie był dotąd analizowany w literaturze.

Artykuł ma następującą strukturę. Sekcja 2 przedstawia model badanej sieci. Sekcja 3 prezentuje zastosowany algorytm realokacji jednostek DU. Sekcja 4 przedstawia wyniki analizy. Sekcja 5 stanowi podsumowanie pracy.

2. MODEL SIECI

W badaniach przyjęto podobny model sieci, ruchu i opóźnień jak w [4, 5]. Komponenty sieci RAN obejmują: (a) jednostki zdalne (RU) znajdujące się w lokalizacjach anten, (b) jednostki rozproszone (DU), realizujące część funkcji przetwarzania pasma podstawowego, które mogą być zwirtualizowane i wykonywane w lokalizacjach sieci dysponujących zasobami obliczeniowymi (PP), oraz (c) jednostki centralne (CU), które finalizują przetwarzanie radiowe we wspólnym dla sieci centrum danych (DC). Jednostki RU są połączone z węzłami PP i DC za pośrednictwem sieci Xhaul z komutacją pakietów [3], wykorzystującej przełączniki Ethernet [2] do multipleksowania i kierowania różnych przepływów danych przesyłanych między komponentami RAN, jak zilustrowano na Rys. 1. Zakładamy, że węzły RU podłączone do danego przełącznika tworzą klaster wymagający lokalizacji ich DU w tym samym węźle PP do celów koordynacji między-komórkowej.

Rozważamy dwa rodzaje przepływów danych: (a) fronthaul (FH) - pomiędzy RU i PP/DU, oraz (b) midhaul (MH) - między PP/DU i DC/CU. Komunikacja odbywa się zarówno w kierunku uplink (RU→PP→DC), jak i downlink (DC→PP→RU). Każdy typ przepływu ma unikalne wymagania dotyczące przepustowości i opóźnień. Dane radiowe w przepływach FH i MH są enkapsulowane i przesyłane jako serie ramek Ethernet [7]. Zakładamy stałą długość ramki wynoszącą 1542 bajty [2]. Serie ramek są nadawane okresowo przez węzły RU/PP/DC i przesyłane przez sieć Xhaul w kierunku węzłów docelowych przepływów. W przełącznikach, serie ramek przepływu są buforowane i nadawane w całości, bez ich podziału i przemieszania z ramkami innych serii. Wybór serii do transmisji w przełącznikach odbywa się zgodnie z algorytmem ścisłego priorytetu i bez wywłaszczenia [2], gdzie serie typu FH są traktowane priorytetowo w stosunku do serii MH [2], a serie o tym samym priorytecie są obsługiwane zgodnie z zasadą FIFO (ang. First In First Out). Okres pomiędzy kolejnymi transmisjami serii, zwany oknem transmisyjnym, zależy od zastosowanej konfiguracji parametrów transmisyjnych 5G, zwanej numerologią (ang. numerology). Numerologia określa m.in. rodzaj odstępu między podnośnymi w siatce zasobów radiowych [7]. W 5G odstęp waha się od 15 kHz do 240 kHz i numerologia jest reprezentowana przez liczbę całkowitą μ, gdzie $0 \leq \mu \leq 4$. Dla danego μ okno transmisji wynosi $66.\overline{6} \times 2^{-\mu}$ μs.

Model opóźnień przepływów uwzględnia zarówno opóźnienia statyczne, jak i dynamiczne, którym ulegają serie ramek podczas transmisji w sieci [4]. Opóźnienia statyczne obejmują opóźnienia propagacji w łączach sieciowych (przy prędkości 2×10^5 km/s), opóźnienia przechowywania i przekazywania (ang. store-and-forward, S&F) w przełącznikach oraz czasy transmisji serii ramek w łączach (zależne od długości serii i przepływności łączy). Opóźnienia dynamiczne wynikają z buforowania serii w portach wyjściowych przełączników i do ich szacowania stosowany jest model estymacji opóźnienia dla "najgorszym przypadku" (WC), opisany w [2] i zastosowany m.in. w [4, 5]. Model WC szacuje maksymalne opóźnienie spowodowane buforowaniem jakiego może doświadczyć seria pakietów na ścieżce routingu. W modelu opóźnienie to jest wywołane transmisją serii należących do wszystkich innych przepływów o tym samym lub wyższym priorytecie, które czekają w kolejce FIFO portu wyjściowego przełącznika, jak i najdłuższą serią o niższym priorytecie, która jest aktualnie transmitowana.

3. DYNAMICZNA REALOKACJA DU

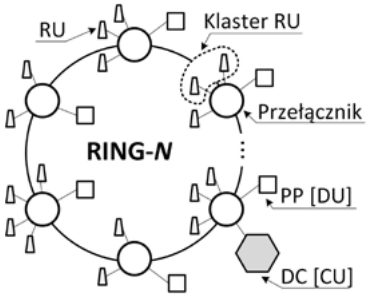
W pracy badana jest realokacja jednostek DU w węzłach PP w scenariuszu sieci o zmiennym obciążeniu, w którym przepływności FH i MH dla poszczególnych RU podlegają zmianom w czasie, jak w cyklu dobowym [9]. W szczególności, w przypadku zmniejszającego się obciążenia sieci następuje konsolidacja jednostek DU w mniejszej liczbie węzłów PP, co umożliwia wyłączenie nieaktywnych węzłów PP, jak na Rys. 1(a). Gdy ruch sieciowy wzrasta, węzły PP są ponownie aktywowane i następuje realokacja jednostek DU do tych węzłów (Rys. 1(b)). W rozważanej realokacji DU wybór węzłów PP dla poszczególnych RU jest dokonywany offline, wykorzystując model optymalizacyjny programowania mieszanego całkowitoliczbowego liniowego (MILP) [4], dla dwóch skrajnych przypadków zakładających maksymalne oraz minimalne wymagania ruchowe jednostek RU.

Algorytm realokacji DU pomiedzy optymalnymi węzłami PP przy wysokim (PPhigh) i niskim (PPlow) obciążeniu działa następująco. W chwili zaobserwowania zmiany w wymaganiach ruchowych, sprawdzane jest czy zasoby transmisyjne w łączach oraz opóźnienia przepływów umożliwiają obsługę ruchu przy aktualnej alokacji DU. Gdy ruch jest akceptowalny, sprawdzana jest możliwość dezaktywacji węzłów PPhigh, zaczynając od tych najmniej obciążonych (tzn. obsługujących najmniejszą liczbę RU), przez przeniesienie całości obsługiwanych przez nie jednostek DU do odpowiednich węzłów PPlow. Gdy ruch nie jest akceptowalny, jednostki DU realokowane są w kolejności odwrotnej na reaktywowanych węzłach PPhigh. Realokacja DU dokonywana jest w takim stopniu w jakim umożliwia to lub wymaga tego sieć, tzn. maksymalnie gdy ruch maleje lub minimalnie gdy ruch rośnie.

Scenariuszem referencyjnym w analizie jest sieć, w której alokacja jednostek DU jest stała i nie podlegająca zmianom w czasie, jak we wcześniejszych pracach [4, 5]. Wybór węzłów PP przy stałej alokacji DU jest dokonywany zakładając maksymalne zapotrzebowanie ruchowe jednostek RU. Gwarantuje to obsługę przepływów z poziomami opóźnień, które nie przekraczają dopuszczalnych limitów bez względu na obciążenie sieci.

4. WYNIKI ANALIZY

Analizę przeprowadzono dla pakietowej sieci transportowej o topologii pierścieniowej RING-N, przedstawionej na Rys. 2, rozważając różną liczbę przełączników (SW) $N \in \{6, 8, 10\}$ oraz R losowo dołączonych jednostek RU, gdzie $R \in \{10, 20, 30\}$. Tabela 1 prezentuje przepustowości i długości łączy w sieci RING-N [5].



Rysunek 2: Topologia badanej sieci RING-N.

Tabela 1: Parametry łączy.

Łącze	Długość [km]	Pojemność [Gbit/s]
SW-SW	[1 . . . 3]	100
SW-RU	[0.2 . . . 0.5]	25
SW-PP	[0.2 . . . 0.5]	400
SW-DC	[10 . . . 15]	400

Tabela 2: Maksymalne przepływności Xhaul [Gbit/s].

Przepływ	Uplink	Downlink
Fronthaul	21.624	22.204
Midhaul	3.024	4.016

Badania dotyczyły systemu radiowego z ośmioma warstwami MIMO, 32 portami antenowymi i kanałami o szerokości 100 MHz [8]. Przyjęto odpowiednio opcję 7.2 i opcję 2 dla podziału funkcjonalnego fronthaul i midhaul. W analizie uwzględniono numerologie radiowe 5G $\mu \in \{1, 2\}$. Tabela 2 prezentuje maksymalne przepływności FH i MH dla pojedynczej jednostki RU, oszacowane zgodnie z modelem opisanym w [8].

Przyjęto opóźnienie S&F przełącznika równe 5μs [2]. Badano sieci z limitem opóźnień przepływów FH (L^{FH}) równym 75μs i 100μs. Dla przypływów MH przyjęto limit opóźnienia 1 ms [4, 5]. Rozważano najkrótsze ścieżki routingu pomiędzy węzłami. Wyniki uzyskano z wykorzystaniem symulatora pakietowej sieci Xhaul, zaimplementowanego w środowisku OMNET++ v.5.6.1 [10].

Badania przeprowadzono dla 24-godzinnego cyklu dobowego przyjmując zmieniające się co godzinę wymagania ruchowe, zobrazowane czerwonymi znacznikami na Rys. 3, zgodnie z godzinowym ruchem w Europie Zachodniej w gęstej sieci miejskiej, przedstawionym w *Ericsson Mobility Report 2023* [9]. Przyjęto takie same wymagania ruchowe dla wszystkich jednostek RU w sieci, gdzie o godzinie 13:00 jednostki RU generowały maksymalny ruch, jak w Tabeli 2, odpowiadający pojemności rozważanego system radiowego, podczas gdy o godzinie 5:00 obciążenie sieci było najniższe i wynosiło 10% wartości maksymalnych. Realokacja DU realizowana była co godzinę.

Rys. 3 prezentuje przykładową liczbę aktywnych PP (z) oraz maksymalne zaobserwowane w sieci opóźnienie wśród przepływów FH w poszczególnych godzinach cyklu dobowego dla stałej oraz dla zmiennej alokacji DU. Wyniki otrzymano w scenariuszu sieci RING-8 obsługującej $R = 30$ jednostek RU, przy zastosowanej numerologii $\mu = 1$ i przy założeniu maksymalnego dopuszczalnego opóźnienia przepływów FH $L^{FH} = 75\mu$s.

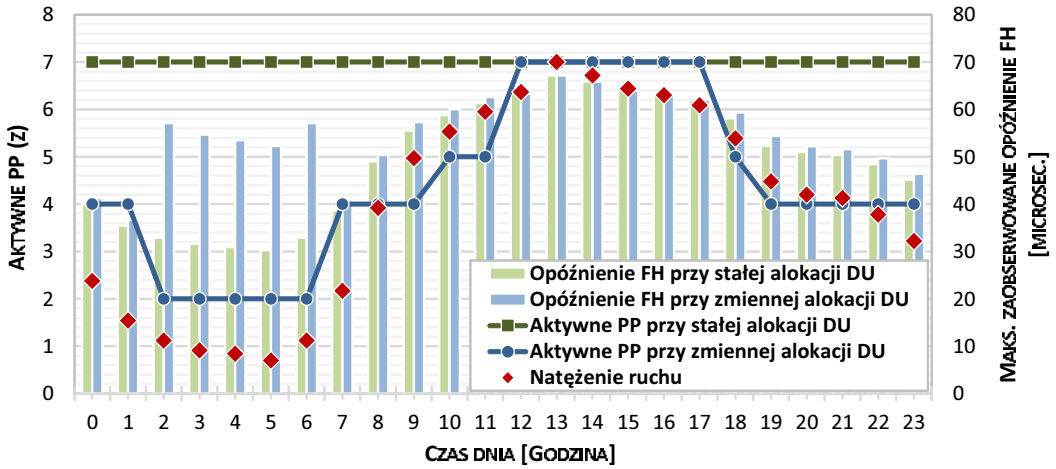
W prezentowanym na Rys. 3 przykładzie, przy stałej alokacji DU w sieci przez cały czas aktywnych jest 7 jednostek PP, które są wymagane do obsługi maksymalnego ruchu (dla godziny 13:00). Obserwowane maksymalne opóźnienia przepływów FH wahają się pomiędzy około 30μs (dla godziny 5:00) i 67μs (dla godziny 13:00), nie przekraczając dopuszczalnego poziomu 75μs. Zastosowanie zmiennej alokacji DU umożliwia zmniejszenie liczby aktywnych węzłów PP w okresach mniejszego obciążenia sieci. W skrajnym przypadku w godzinach nocnych (pomiedzy godziną 2:00 i 6:00) wystarczy aktywacja jedynie 2 węzłów PP do realizacji funkcji DU dla wszystkich jednostek RU, co stanowi redukcję o około 70% w stosunku do liczby 7 PP przy stałej alokacji DU. Maksymalne opóźnienia przepływów FH są znacznie wyższe w tym okresie w stosunku do stałej alokacji DU, wynosząc pomiędzy 53μs i 57μs, jednak w żadnym momencie nie przekraczają one granicznego poziomu 75μs.

W celu oceny wpływu zastosowania zmiennej alokacji DU na redukcję liczby aktywnych jednostek PP w sieci, zdefiniowano metrykę Δ, która jest obliczana jako:

$$\Delta = 1 - \frac{\sum_{i=0}^{i=23} z_i}{24 \cdot z^{max}}, \tag{1}$$

gdzie z_i oznacza liczbę aktywnych PP dla godziny i przy zmiennej alokacji DU, a z^{max} to liczba aktywnych PP przy stałej alokacji DU. Metryka Δ reprezentuje uśrednioną w czasie względną redukcję liczby aktywnych PP.

Tabela 3 prezentuje wartości Δ otrzymane dla różnych scenariuszy sieci RING-N. Wyniki pokazują duże zróżnicowanie wartości Δ, maksymalnie pomiędzy 1% i 36% w przypadku sieci RING-8. Uśrednione wartości Δ dla poszczególnych sieci, prezentowane w ostatniej kolumnie, wskazują na zwiększony poziom redukcji aktywnych



Rysunek 3: Liczba aktywnych PP (linie ciągłe) oraz maksymalne zaobserwowane opóźnienie wśród przepływów FH (słupki) w poszczególnych godzinach doby przy stałej oraz zmiennej alokacji DU dla sieci RING-8, $R = 30$ RUs, $\mu = 1$ i $L^{FH} = 75\mu$s. Linia przerywana ilustruje zmiany w natężeniu ruchu.

PP wraz ze wzrostem liczby obsługiwanych RU (R). Jednocześnie, można zaobserwować nieznaczne zmniejszenie wartości Δ wraz ze wzrostem rozmiaru sieci transportowej, tzn. liczby przełączników (N) w pierścieniu RING-N.

Tabela 3: Redukcja liczby aktywnych PP (Δ) wynikająca z realokacji DU, w stosunku do stałej alokacji, dla scenariuszy sieci RING-N z różną liczbą RU (R).

Sieć		Δ				
		$L^{FH} = 75$		$L^{FH} = 100$		
N	R	$\mu = 1$	$\mu = 2$	$\mu = 1$	$\mu = 2$	**Średnia**
6	10	17%	6%	13%	17%	13%
	20	17%	13%	25%	17%	18%
	30	19%	18%	26%	30%	23%
8	10	1%	10%	19%	23%	14%
	20	24%	7%	13%	21%	16%
	30	36%	24%	10%	21%	23%
10	10	4%	7%	14%	4%	7%
	20	3%	21%	24%	14%	15%
	30	19%	18%	13%	15%	16%

5. PODSUMOWANIE

Przedstawione wyniki analizy wskazują, iż zmienna alokacja (realokacja) DU realizowana w cyklu dobowym w pakietowej sieci Xhaul w wielu przypadkach jest korzystna i umożliwia zmniejszenie liczby aktywnych węzłów obliczeniowych (PP) w okresach mniejszego zapotrzebowania ruchowego. Realokacja DU umożliwia potencjalnie deaktywację (uśpienie) niewykorzystywanych zasobów PP, co może prowadzić do oszczędności zarówno kosztu działania sieci jak i zużycia energii. Nasze przyszłe badania skupią się na opracowaniu algorytmów optymalnego wyboru realokowanych jednostek DU i deaktywowanych węzłów PP oraz skutecznych mechanizmów decyzyjnych dotyczących momentu przeprowadzenia realokacji, biorąc pod uwagę wpływ na jakość i niezawodność usług (uwzględniając opóźnienia przepływów i dostępność usług). Badania uwzględnią stosowny model zużycia energii w sieci.

PODZIĘKOWANIA

Praca powstała w wyniku realizacji projektu badawczego OPUS o nr 2024/53/B/ST7/02482 finansowanego ze środków Narodowego Centrum Nauki.

LITERATURA

[1] O-RAN Alliance. 2021. „Xhaul Transport Requirements Tech. Spec. v1.0”. https://www.o-ran.org/.

[2] IEEE. 2018 Nov. „802.1CM-2018 – IEEE Standard for Local and Metropolitan Area Networks – Time-Sensitive Networking for Fronthaul”.

[3] IEEE. 2019 Nov. „1914.1-2019 - IEEE Standard for Packet-based Fronthaul Transport Networks”.

[4] M. Klinkowski. 2020. „Optimization of Latency-Aware Flow Allocation in NGFI Networks”. *Comp. Commun.*, 161 : 344-359.

[5] M. Klinkowski. 2024. „Optimized Planning of DU/CU Placement and Flow Routing in 5G Packet Xhaul Networks”. *IEEE Trans. Netw. Serv. Manag.*, 21 (1) : 232-248.

[6] IOWN Global Forum. 2024 Aug. „PoC Reference of Mobile Fronthaul over APN”. https://iowngf.org/.

[7] G. O. Perez, D. Larrabeiti, J. A. Hernandez. 2019. „5G New Radio Fronthaul Network Design for eCPRI-IEEE 802.1CM and Extreme Latency Percentiles”. *IEEE Access*, 7 : 82218-82229.

[8] S. Lagen, L. Giupponi, A. Hansson, X. Gelabert. 2021. „Modulation Compression in Next Generation RAN: Air Interface and Fronthaul Trade-offs”. *IEEE Comm. Mag.*, 59 (1) : 89-95.

[9] Ericsson Mobility Report. 2023 Jun. „Exploring how traffic patterns drive network evolution”. https://www.ericsson.com/mobility-report.

[10] A. Varga. „OMNeT++ Discrete Event Simulator”. https://omnetpp.org/.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

ANALIZA SKUTECZNOŚCI REGRESJI LOGISTYCZNEJ W KLASYFIKACJI ATAKÓW DDOS

ANALYSIS OF THE EFFECTIVENESS OF LOGISTIC REGRESSION IN CLASSIFYING DDOS ATTACKS

Norbert Roczniak[1]; Krzysztof Rusek[2];
[1] AGH University of Krakow, nroczniak@student.agh.edu.pl
[2] AGH University of Krakow, krusek@agh.edu.pl

DOI: 10.15199/59.2025.4.34

Streszczenie: Rosnące zagrożenie atakami DDoS (Distributed Denial of Service) dla dostępności sieci wymaga metod detekcji skuteczniejszych niż systemy regułowe. Uczenie maszynowe, w tym regresja logistyczna (model szybki, interpretowalny i dokładny), oferuje skuteczne narzędzia do rozwiązania tego problemu. W pracy oceniono efektywność regresji logistycznej w wieloklasowej klasyfikacji ataków DDoS oraz zidentyfikowano jej ograniczenia i możliwości rozwoju.

Abstract: The growing threat of DDoS (Distributed Denial of Service) attacks to network availability requires detection methods more effective than rule-based systems. Machine learning, including logistic regression (a fast, interpretable, and accurate model), offers effective tools to address this challenge. The aim of the work was to evaluate the effectiveness of logistic regression in multi-class classification of DDoS attacks and to identify its limitations and potential for future development.

Słowa kluczowe: Ataki DDoS, Bezpieczeństwo Sieci, Klasyfikacja, Regresja Logistyczna, Uczenie Maszynowe

Keywords: Classification, DDoS Attacks, Logistic Regression, Machine Learning, Network Security

1. WSTĘP

Ataki typu Distributed Denial of Service (DDoS) stanowią poważne zagrożenie dla dostępności sieci, przeciążając systemy złośliwym ruchem. Rosnące uzależnienie od usług internetowych i powszechne wdrożenie urządzeń Internetu Rzeczy (IoT) zwiększyły zarówno częstotliwość, jak i skutki tych ataków. Najnowsze raporty podkreślają ich rosnącą złożoność, skalę oraz pilną potrzebę adaptacyjnych strategii obronnych [9].

Ataki DDoS dzieli się zazwyczaj na trzy główne kategorie, z których każda celuje w inne zasoby i warstwy sieci, wymagając specyficznych metod przeciwdziałania. Są to: ataki wolumetryczne, które mają na celu przeciążenie dostępnej przepustowości sieci, uniemożliwiając dotarcie legalnego ruchu (ich skala mierzona jest np. w bitach na sekundę – b/s); ataki protokołowe, koncentrujące się na wyczerpywaniu zasobów serwerów, zapór sieciowych czy systemów równoważenia obciążenia poprzez wykorzystanie słabości protokołów komunikacyjnych (mierzy się je np. w pakietach na sekundę – pps); oraz ataki na warstwę aplikacji, skierowane przeciwko konkretnym aplikacjom lub usługom i dążące do ich unieruchomienia poprzez generowanie pozornie legalnych, lecz masowych żądań (mierzonych np. w żądaniach na sekundę – rps).

Precyzyjna identyfikacja typu ataku jest zatem kluczowa dla wyboru skutecznych i adekwatnych mechanizmów obronnych.

Tradycyjne systemy oparte na regułach, takie jak *Snort* [14] i *Suricata* [15], są skuteczne przeciwko znanym zagrożeniom, ale mają trudności z ewoluującymi atakami wielowektorowymi. To ograniczenie zwiększyło zainteresowanie podejściami opartymi na uczeniu maszynowym (ML), które oferują lepszą adaptacyjność dzięki wykrywaniu wzorców i anomalii w ruchu sieciowym.

Wśród metod ML wyróżnia się regresja logistyczna ze względu na swoją prostotę, wydajność i interpretowalność. Choć nie stanowi samodzielnego rozwiązania, może odgrywać kluczową rolę w wielowarstwowych systemach bezpieczeństwa, zwiększając dokładność detekcji i dostarczając wgląd w charakterystykę ruchu.

Choć zastosowanie regresji logistycznej do detekcji ataków DDoS było już przedmiotem wcześniejszych badań [12], większość dotychczasowych prac koncentrowała się na klasyfikacji binarnej – odróżnianiu ruchu złośliwego od legalnego. Niniejsza praca wypełnia lukę w literaturze poprzez szczegółową analizę klasyfikacji wieloklasowej, która umożliwia precyzyjną identyfikację konkretnych kategorii ataków DDoS. Wkładem badania jest systematyczne porównanie skuteczności klasyfikacji binarnej i wieloklasowej regresji logistycznej, analiza istotności cech dla różnych typów ataków oraz ocena praktycznych ograniczeń metody w kontekście grupowania wektorów ataków według kategorii funkcjonalnych.

W niniejszej pracy przedstawiono analizę zastosowania regresji logistycznej do klasyfikacji ataków DDoS, koncentrując się na jej skuteczności w zadaniu wieloklasowym. Dla porównania przeprowadzono również klasyfikację binarną, aby ocenić kompromisy między szczegółowością wykrywania a wydajnością. W dalszej części przegląd literatury przedstawia dotychczasowe metody ML w detekcji DDoS, następnie omówiono teoretyczne podstawy regresji logistycznej, opisano metodologię badania i zbiór danych, zaprezentowano wyniki eksperymentów, a całość zamyka podsumowanie z kierunkami przyszłych badań.

2. PRZEGLĄD LITERATURY

Ataki DDoS pozostają stałym wyzwaniem w dziedzinie cyberbezpieczeństwa, coraz częściej celując w krytyczną infrastrukturę z rosnącą złożonością. W odpowiedzi liczne badania eksplorują podejścia oparte na uczeniu maszynowym do ich wykrywania i klasyfikacji.

Singh i Gupta [13] zaproponowali taksonomię ataków DDoS opartą na metodzie wykonania i docelowej warstwie modelu odniesienia OSI (Open Systems Interconnection) – wolumetryczne, protokołowe i warstwy aplikacyjnej – co poprawia interpretowalność modeli ML i wspiera ukierunkowane przeciwdziałanie. Uzupełniająco, Sharafaldin i in. [12] wprowadzili taksonomię opartą na zachowaniu, wyróżniającą ataki oparte na odbiciu i oparte na eksploatacji, usprawniając konstrukcję zbiorów danych i przejrzystość detekcji.

Wiele badań porównawczych oceniało algorytmy ML pod kątem wykrywania DDoS. W pracy [7] regresję logistyczną zestawiono z algorytmem k-najbliższych sąsiadów, maszynami wektorów nośnych i naiwnym klasyfikatorem Bayesa, wykazując jej skuteczność zwłaszcza dla wzorców nieliniowych. Naing i Thwel również stwierdzili, że regresja logistyczna (LR) przewyższa inne modele na zbiorze APA-DDOS, osiągając 93% precyzji. Wyniki te podkreślają równowagę między wydajnością a prostotą LR.

Szersza ocena na jedenastu zbiorach danych [5] wykazała, że żaden klasyfikator nie jest uniwersalnie optymalny, co uwydatnia znaczenie dostosowanej selekcji cech. W odpowiedzi, w pracy [10] zaproponowano zespołową metodę selekcji cech opartą na głosowaniu większościowym, poprawiającą wyniki zarówno dla tradycyjnego ML, jak i modeli głębokiego uczenia.

W porównaniu z płytkimi sieciami neuronowymi (SNN), LR osiągała zbliżoną skuteczność przy znacznie niższych kosztach obliczeniowych, co czyni ją praktycznym wyborem do detekcji w czasie rzeczywistym [1]. Tymczasem Chartuni i Márquez [2] zastosowali głębokie uczenie na zbiorze CICDDoS2019, uzyskując 94% dokładności po normalizacji i balansowaniu metodą SMOTE (*Synthetic Minority Over-sampling Technique*). Ich praca nie uwzględniała jednak testów odpornościowych na niewidziane wektory ataków ani nie zaproponowała systemowej metody grupowania statystycznie podobnych ataków.

Regresja logistyczna i inne modele ML powinny uzupełniać – a nie zastępować – tradycyjne mechanizmy w wielowarstwowych systemach ochrony. Przykładem skutecznej integracji jest system FLDX opracowany przez Karpowicza [4], wykorzystujący adaptacyjne strojenie mechanizmów ograniczania ruchu z regulacją sprzężenia zwrotnego.

Warto zauważyć, że większość dotychczasowych badań koncentruje się na klasyfikacji binarnej – odróżnianiu ruchu złośliwego od legalnego – bez identyfikacji konkretnych wektorów ataków. W obliczu wzrostu wielowektorowych i zagrożeń typu *zero-day* [9], bardziej szczegółowa klasyfikacja jest kluczowa dla skutecznej, ukierunkowanej obrony. Ta luka w literaturze motywuje przeprowadzone badania nad klasyfikacją wieloklasową.

3. ATAKI DDOS I REGRESJA LOGISTYCZNA

Regresja logistyczna jest metodą klasyfikacji, która szacuje prawdopodobieństwo przynależności danej obserwacji (np. przepływu sieciowego) do określonej klasy na podstawie analizy jej cech, takich jak liczba pakietów czy czas trwania przepływu [3]. Jej prostota i interpretowalność czynią ją atrakcyjną w zastosowaniach bezpieczeństwa.

W kontekście klasyfikacji ataków DDoS, model ten może być stosowany zarówno w trybie binarnym, do odróżniania ruchu normalnego od złośliwego, jak i w bardziej zaawansowanej wersji wieloklasowej. Ta druga pozwala na precyzyjne rozróżnianie poszczególnych typów ataków, na przykład identyfikację ataków typu *SYN Flood* czy *UDP Flood*.

Istotną zaletą regresji logistycznej jest jej zdolność do efektywnej pracy z różnymi typami danych wejściowych oraz możliwość przeprowadzania analizy istotności cech, co dostarcza wglądu w charakterystykę wykrywanych anomalii. Te właściwości sprawiają, że regresja logistyczna jest praktycznym i efektywnym narzędziem do budowy systemów detekcji DDoS w rzeczywistych środowiskach sieciowych [7].

Regresja logistyczna, jako interpretowalna metoda uczenia maszynowego, stanowi cenne narzędzie do detekcji i klasyfikacji złożonych zagrożeń sieciowych, takich jak ataki DDoS, zwłaszcza w scenariuszach, gdzie tradycyjne systemy oparte na sygnaturach okazują się niewystarczające. Zaawansowane techniki analityczne pozwalają na skuteczniejsze przeciwdziałanie tym poważnym zagrożeniom dla dostępności usług.

4. ŚRODOWISKO BADAWCZE I METODOLOGIA

Przedstawiona metodologia opiera się na eksperymentach z użyciem rzeczywistych danych sieciowych. Poniżej opisano szczegóły badań.

4.1 Projekt systemu

W badaniu zastosowano oznaczone zbiory danych z próbkami ruchu normalnego i DDoS. Przetwarzanie wstępne obejmuje obsługę brakujących danych, normalizację cech i kodowanie zmiennych kategorycznych. Modele regresji logistycznej oceniano w konfiguracjach: binarnej (legalny/złośliwy) i wieloklasowej (kategorie DDoS), co umożliwiło analizę detekcji i różnicowania ataków.

Podczas wstępnych testów szczegółową klasyfikację dla poszczególnych wektorów wykluczono z powodu niskiej dokładności i ograniczonej generalizacji. Grupowanie ataków w kategorie poprawiło odporność modeli. W celu zwiększenia wiarygodności wyników, każdą konfigurację modelu uruchamiano wielokrotnie, a uzyskane wyniki uśredniono, minimalizując wpływ losowej inicjalizacji oraz fluktuacji w podziale danych.

4.2 Zbiór danych

W badaniu zastosowano autorski zbiór danych stworzony na podstawie plików pcap (packet capture) z repozytoriów CICDDoS2019 [12] oraz CICIoT2023 [8], przetworzonych za pomocą narzędzia *CICFlowmeter* [6]. Obejmują one zarówno ruch legalny, jak i ataki typu wolumetrycznego, aplikacyjnego oraz protokołowego. Pliki zostały przekształcone do formatu CSV (Comma-Separated Values) i wzbogacone o atrybuty takie jak: czas trwania przepływów, liczba pakietów, typy protokołów, flagi TCP (Transmission Control Protocol) oraz intensywność ruchu.

Większość dostępnych wektorów ataków (takich jak Distributed Reflection Denial of Service (DrDoS), SYN Flood, TCP Flood, UDP Flood) reprezentuje klasy wolumetryczne i protokołowe. Z kolei ataki aplikacyjne (*HTTP Flood*, *SlowLoris*, MSSQL, NetBIOS) są istotnie niedoreprezentowane ze względu na większą złożoność generowania ich syntetycznych odpowiedników.

W celu poprawy jakości klasyfikacji zastosowano techniki równoważenia danych, takie jak przycinanie klas nadreprezentowanych i duplikacja przykładów z klas niedoreprezentowanych. Wektory ataków pogrupowano w trzy kategorie: reprezentujące ataki aplikacyjne, protokołowe i wolumetryczne, na podstawie cech statystycznych i funkcjonalnych [2], co ograniczyło błędy wynikające z podobieństwa pomiędzy poszczególnymi wektorami w obrębie danej klasy.

4.3 Miary skuteczności klasyfikacji

Skuteczność modelu oceniano za pomocą czterech standardowych metryk: dokładności, precyzji, czułości oraz wyniku F1 [13]. Dobór tych miar umożliwia kompleksową ocenę skuteczności klasyfikatora zarówno pod kątem ogólnej trafności, jak i zdolności do rozróżniania między różnymi klasami ataków.

5. BADANIA I WYNIKI

Przeprowadzono analizę skuteczności regresji logistycznej w klasyfikacji ataków DDoS w ujęciu binarnym i wieloklasowym, uwzględniając wyniki oraz interpretację istotnych cech. Uzyskane rezultaty pozwalają na ocenę przydatności tego modelu w różnych scenariuszach detekcji.

5.1 Skuteczność klasyfikacji binarnej i wieloklasowej

Model binarny osiągnął bardzo dobre wyniki, potwierdzając swoją skuteczność w detekcji podstawowych ataków. Model wieloklasowy uzyskał nieco niższe wartości metryk ze względu na większą złożoność zadania, jednak utrzymał wysoką jakość klasyfikacji (Tabela 1). Dodatkowo, klasyfikacja wieloklasowa może dostarczać cenniejszych informacji – umożliwiając identyfikację typu ataku i dopasowanie strategii obronnej.

Tabela 1: Porównanie wyników klasyfikacji binarnej i wieloklasowej

Metryka	Klasyfikacja binarna	Klasyfikacja wieloklasowa
Dokładność	0,9494	0,9190
Precyzja	0,9544	0,9184
Czułość	0,9497	0,9201
Wynik F1	0,9631	0,9199

5.2 Wyniki modelu w klasyfikacji wieloklasowej

Ataki wolumetryczne i protokołowe zostały wykryte z niemal idealną precyzją i czułością (bliskie 0,99), podczas gdy ataki warstwy aplikacji charakteryzowały się niższą czułością (0,78), co wskazuje na tendencję do błędnej klasyfikacji jako ruch normalny (Tabela 2).

Tabela 2: Skuteczność klasyfikacji według kategorii

Kategoria	Precyzja	Czułość	Wynik F1
Aplikacyjny	0,91	0,78	0,84
Normalny	0,81	0,91	0,86
Protokołowy	0,98	0,99	0,98
Wolumetryczny	0,99	0,99	0,99

Spadek czułości w warstwie aplikacji odzwierciedla trudności w wykrywaniu ukrytego lub przypominającego użytkownika ruchu i wskazuje na możliwość przyszłych usprawnień poprzez użycie cech kontekstowych, modeli hybrydowych lub modelowania czasowego.

5.3 Istotność cech i jakość dopasowania modelu

W Tabeli 3 przedstawiono najistotniejsze cechy wpływające na klasyfikację ataków DDoS. Dla kategorii aplikacyjnej dominują zmienne związane z długością pakietów i bajtami przesyłanymi w kierunku przeciwnym, natomiast dla ataków protokołowych i wolumetrycznych – parametry czasowe przepływu i flagi TCP. Ruch normalny charakteryzuje się stabilnym rozmiarem pakietów i natężeniem transmisji. Istotność cech określono na podstawie niezerowych współczynników regresji oraz średnich efektów krańcowych.

Tabela 3: Najważniejsze cechy według kategorii ataków DDoS

Kategoria	Kluczowe cechy
Aplikacyjna	Bwd Packet Length Mean, Subflow Bwd Bytes, ECE Flag Count
Protokołowa	Flow IAT (Inter-Arrival Time) Mean, Fwd IAT Max, Bwd IAT Max
Wolumetryczna	URG Flag Count, Init Win Bytes Fwd, Packet Length Mean
Normalna	Flow Bytes/s, Fwd/Bwd Packet Length Mean, Avg. Packet Size

Z kolei Tabela 4 zawiera kluczowe statystyki dopasowania modelu wieloklasowej regresji logistycznej. Wysoka wartość pseudo-R^2 (0,8179) oraz istotność testu LLR (Log-Likelihood Ratio) [16], świadczą o wysokiej trafności modelu. Zbieżność optymalizacji i stabilność estymacji potwierdzają poprawność konstrukcji modelu [11].

Tabela 4: Kompaktowe podsumowanie dopasowania modelu regresji logistycznej dla klasyfikacji wieloklasowej

Statystyka	**Wartość**
Stopnie swobody (model)	204
Stopnie swobody (residuum)	655 409
Pseudo-R^2	0,8179
Log-likelihood	−165 510
Log-likelihood (model zerowy)	−908 880
Test LLR	$p < 0,0001$
Zbieżność optymalizacji	Tak
Typ macierzy kowariancji	niezmodyfikowana

6. PODSUMOWANIE

Choć klasyfikacja binarna zapewnia wyższą skuteczność, to podejście wieloklasowe dostarcza większej wartości operacyjnej – umożliwia precyzyjną mitygację, optymalne zarządzanie zasobami i lepsze rozpoznanie sytuacyjne. Regresja logistyczna okazała się skutecznym, lekkim i interpretowalnym modelem bazowym dla obu zadań, szczególnie w przypadku ataków wolumetrycznych i protokołowych. Jej niski koszt obliczeniowy oraz transparentność czynią ją odpowiednią do zastosowań w czasie rzeczywistym i środowiskach o ograniczonych zasobach.

Największym wyzwaniem pozostaje detekcja ataków aplikacyjnych – mimo wysokiej precyzji, czułość była ograniczona ze względu na podobieństwo do ruchu legalnego. Ograniczenia badania obejmują niezrównoważone zbiory danych, niedostateczną reprezentację ataków aplikacyjnych oraz koncentrację wyłącznie na fazie detekcji bez mechanizmów automatycznej reakcji. Istotnym ograniczeniem jest także fakt, że metody ML działające w izolacji mogą okazać się niewystarczające – skuteczna obrona przed DDoS wymaga wielowarstwowego podejścia, w którym algorytmy uczenia maszynowego stanowią jeden z elementów szerszego systemu bezpieczeństwa.

W przyszłych pracach kluczowe będzie rozwijanie metod wykrywania subtelnych ataków aplikacyjnych, tworzenie zrównoważonych zbiorów danych oraz integracja z automatycznymi systemami reakcji. Przyszłe badania powinny skupić się na budowie kompletnych platform obrony przed DDoS, łączących zaawansowaną detekcję opartą na ML z mechanizmami mitygacji w czasie rzeczywistym oraz tradycyjnymi systemami bezpieczeństwa.

LITERATURA

[1] Batchu, Raj Kumar, and Hari Seetha. 2021. "A generalized machine learning model for DDoS attacks detection using hybrid feature selection and hyperparameter tuning." *Computer Networks* 200: 108498. https://doi.org/10.1016/j.comnet.2021.108498.

[2] Chartuni, Andres, and Jose Marquez. 2021. "Multiclassifier of DDoS attacks in computer networks built on neural networks." *Applied Sciences* 11(22): 10609. https://doi.org/10.3390/app112210609.

[3] Hosmer, David W., Stanley Lemeshow, and Rodney X. Sturdivant. 2013. *Applied Logistic Regression*. 3rd ed. Hoboken, NJ: John Wiley & Sons.

[4] Karpowicz. M.P. 2021. "Adaptive tuning of network traffic policing mechanisms for DDoS attack mitigation systems." *European Journal of Control* 61: 101–118.

[5] Kishore, Dasari, and Nagaraju Devarakonda. 2021. "Detection of Different DDoS Attacks Using Machine Learning Classification Algorithms." *Ingénierie des Systèmes d'Information* 26(5): 461–468. https://doi.org/10.18280/isi.260505.

[6] Lashkari, Arash Habibi, Gerard Draper-Gil, Mohammad Saiful Islam Mamun, and Ali A. Ghorbani. 2017. "Characterization of Tor Traffic Using Time-Based Features." In *Proceedings of the 3rd International Conference on Information System Security and Privacy (ICISSP)*, Porto, Portugal. SCITEPRESS.

[7] Naing, Soe Kalayar, and Tin Thein Thwel. 2023. "A Study of DDoS Attack Classification Using Machine Learning Classifiers." In *Proceedings of the 2023 IEEE Conference on Computer Applications (ICCA)*, 108–112. IEEE. https://doi.org/10.1109/ICCA51723.2023.10182146.

[8] Neto, E. C. P., S. Dadkhah, R. Ferreira, A. Zohourian, R. Lu, and A. A. Ghorbani. 2023. "CICIoT2023: A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment." Manuscript submitted for publication. *Sensors*.

[9] Radware Ltd. 2025. *Global Threat Analysis Report: Executive Summary RWI-426*. Accessed May 19, 2025. https://www.radware.com/pleaseregister.aspx/?returnurl=ccb13ccd-2d85-4ab5-bf61-050070c49cae.

[10] Saha, S., A. T. Priyoti, A. Sharma, and A. Haque. 2022. "Towards an Optimal Feature Selection Method for AI-Based DDoS Detection System." In *Proceedings of the 2022 IEEE 19th Annual Consumer Communications and Networking Conference (CCNC)*, 425–428. IEEE. https://doi.org/10.1109/CCNC49033.2022.9700569.

[11] Seabold, Skipper, and Josef Perktold. 2010. Statsmodels: Econometric and Statistical Modeling with Python. Accessed May 20, 2025. https://www.statsmodels.org/stable/discretemod.html.

[12] Sharafaldin, Iman, Arash H. Lashkari, Samaneh Hakak, and Ali A. Ghorbani. 2019. "Developing Realistic Distributed Denial of Service (DDoS) Attack Dataset and Taxonomy." In *Proceedings of the 12th IEEE Annual Cybersecurity Conference*. https://doi.org/10.1109/CCST.2019.8888419.

[13] Singh, Anshuman, and Brij Gupta. 2022. "Distributed Denial-of-Service (DDoS) Attacks and Defense Mechanisms in Various Web-Enabled Computing Platforms: Issues, Challenges, and Future Research Directions." *International Journal on Semantic Web and Information Systems* 18(1): 1–43. https://doi.org/10.4018/IJSWIS.297143.

[14] Snort. n.d. "Snort – Network Intrusion Detection and Prevention System." Accessed April 8, 2025. https://www.snort.org/.

[15] Suricata. n.d. "Suricata – The Next Generation IDS/IPS Engine." Accessed April 8, 2025. https://suricata.io/.

[16] Wilks, S. S. 1938. "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses." Annals of Mathematical Statistics 9 (1): 60–62. https://doi.org/10.1214/aoms/1177732360.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

LEKKA METODYKA EWALUACJI CYBERBEZPIECZEŃSTWA KOMPONENTÓW OPEN RAN

LIGHT-WEIGHT CYBERSECURITY EVALUATION METHODOLOGY FOR OPEN RAN COMPONENTS

Piotr Krawiec[1]; Jordi Mongay Batalla[2]

[1] Instytut Łączności - Państwowy Instytut Badawczy oraz Politechnika Warszawska, Warszawa, p.krawiec@il-pib.pl
[2] Politechnika Warszawska, Warszawa, jordi.mongay.batalla@pw.edu.pl

DOI: 10.15199/59.2025.4.35

Streszczenie: **Artykuł prezentuje koncepcję metodyki ewaluacji cyberbezpieczeństwa modułów Open RAN, która bazuje na lekkiej metodyce FiT CEM gwarantującej predefiniowany, stały czas procesu oceny. Proponowana metodyka jest dedykowana w szczególności dla komponentów software'owych, charakteryzujących się dynamicznym cyklem życia produktu. Pozwala ona na uzyskanie wystarczającego poziomu zaufania do cyberbezpieczeństwa produktu, utrzymując zarazem obciążenie producenta procesem oceny na rozsądnym, przewidywalnym poziomie.**

Abstract: **The article presents the concept of cyber-security evaluation methodology for Open RAN modules based on the lightweight FiT CEM methodology that guarantees a predefined, fixed time for the evaluation process. It is dedicated in particular to dynamic life cycle software components. The proposed methodology provides a sufficient level of trust, while keeping the product manufacturer effort at a reasonable and predictable level.**

Słowa kluczowe: **5G, ewaluacja cyberbezpieczeństwa, O-RAN**

Keywords: **5G, cybersecurity evaluation, O-RAN**

1. WSTĘP

Od wielu lat aspekty cyberbezpieczeństwa stanowią istotną domenę rozwoju sieci telekomunikacyjnych. W ujęciu proceduralnym, działania związane z cyberbezpieczeństwem sieci możemy podzielić na reaktywne, jak np. aktywności zespołów CSIRT (ang. *Computer Security Incident Response Team*) reagujących na incydenty bezpieczeństwa, oraz proaktywne, obejmujące działania realizowane zanim wystąpi niebezpieczna sytuacja, celem zapobieżenia wystąpieniu incydentów bezpieczeństwa. Nabierającym coraz większego znaczenia działaniem proaktywnym jest certyfikacja cyberbezpieczeństwa, która zwiększa zaufanie że dany produkt IT spełnia określone wymagania bezpieczeństwa. Jednakże zaufanie do certyfikatu wymaga, aby sam system certyfikacji (określany również jako "schemat certyfikacji") był wiarygodny oraz transparentny, i bazował na powszechnie uznanych normach referencyjnych definiujących sposób przeprowadzenia procesu oceny (tj. metodykę ewaluacji) z zapewnieniem porównywalności i powtarzalności wyników.

Najpowszechniej obecnie używaną normą certyfikacji cyberbezpieczeństwa produktów IT jest międzynarodowa norma ISO/IEC 15408 określana jako *Common Criteria* (CC) [1]. Schematy certyfikacji bazujące na tej normie funkcjonują w 18-tu krajach na całym świecie, w tym w ośmiu europejskich (włączając w to Polskę) [2].

Norma CC odnosi się do każdego rodzaju produktu IT: od mikroukładów i kart inteligentnych, po urządzenia sieciowe, systemy operacyjne i aplikacje. *Common Criteria* jest normą kompletną, pozwalającą na ocenę produktu pod kątem spełniania przez niego wymagań bezpieczeństwa w całym cyklu życia, od zaprojektowania do wycofania z użytku. Z drugiej strony, norma ta jest uważana za "ciężką", tj. skomplikowaną i wymagającą istotnego wysiłku dewelopera niezbędnego do przygotowania odpowiedniej dokumentacji dowodowej, w efekcie skutkującą długim czasem trwania procesu oceny (często kilkunastomiesięcznym). Jest to istotnym ograniczeniem zastosowania normy CC, w szczególności w odniesieniu do produktów software'owych, gdzie kolejne wersje produktu często powstają w odstępach kilkumiesięcznych, powodując w efekcie konieczność ponownej realizacji procesu oceny. Z tego względu, opracowywane są inne metodyki, ukierunkowane na szybszy czas realizacji ewaluacji, bardziej dopasowany do cyklu życia produktów software'owych. Jedną z nich jest norma CEN/CENELEC EN 17640 FiT CEM: *Fixed-time cybersecurity evaluation methodology for ICT products* [3], zakładająca ustalony czas trwania ewaluacji, zwykle ograniczony do 3 - 4 miesięcy.

Niniejszy artykuł przedstawia propozycję lekkiej metodyki oceny dedykowanej dla software'owych komponentów sieci Open RAN. Proponowana metodyka stanowi adaptację ogólnej metodyki FiT CEM, dopasowanej do wymagań Open RAN. W dalszej części artykułu przedstawiono zarys technologii Open RAN oraz aktualny stan certyfikacji w obszarze sieci mobilnych. Następnie zaprezentowano założenia proponowanej metodyki. W rozdziale 4 przedstawiono proces przykładowej ewaluacji komponentu Open RAN według proponowanej metodyki, z uwzględnieniem testowania interfejsów natywnych protokołów Open-RAN. Rozdział 5 stanowi podsumowanie artykułu.

2. OPEN RAN I CERTYFIKACJA CYBERBEZPIECZEŃSTWA

Architektura sieci mobilnych zdefiniowana przez 3GPP wprowadza separację pomiędzy siecią dostępową RAN (*Radio Access Network*) i siecią szkieletową, jak

również dalszy podział RAN na odrębne komponenty: *Central Unit* (CU) oraz *Distributed Unit* (DU) wraz z *Radio Unit* (RU). Stowarzyszenie O-RAN (Open RAN) Alliance zaproponowało nowatorską koncepcję sieci RAN poprzez dalszą segmentację jej funkcji sieciowych, umożliwiającą odseparowanie warstwy software'owej od warstwy sprzętowej. Architektura O-RAN, zdefiniowana w [4], wprowadza nowe, otwarte interfejsy takie jak: A1, E2 czy O1, nowy moduł RIC (*RAN Intelligent Controller*) czyli software'owy komponent dla potrzeb orkiestracji i zarządzania segmentem RAN, czy komponent O-Cloud, tj. infrastrukturę fizyczną na której uruchamiane są funkcje O-RAN.

W efekcie, O-RAN oferuje większą elastyczność w zakresie doboru sprzętu i oprogramowania celem budowy systemu 5G. Sieć można zbudować poprzez integrację komponentów od różnych dostawców. Jednakże zapewnienie bezpieczeństwa w tak złożonym środowisku ma kluczowe znaczenie, ponieważ użytkownik nie może już polegać na bezpieczeństwie rozwiązania oferowanego całościowo przez jednego dostawcę. Dostawca O-RAN może zwiększyć zaufanie do bezpieczeństwa oferowanego przez swój produkt poprzez uzyskanie certyfikatu wystawionego przez stronę trzecią.

Obecnie dostawcy systemów mobilnych certyfikują swoje produkty w schemacie certyfikacji NESAS (*Network Equipment Security Assurance Scheme*), który ocenia spełnienie wymagań bezpieczeństwa zdefiniowane przez 3GPP. Jest to system stworzony i utrzymywany przez organizację GSMA łączącą operatorów sieci mobilnych, dostawców sprzętu i dostawców usług mobilnych. Oprócz NESAS, dostawcy komponentów O-RAN muszą przejść dodatkowe testy na otwartych interfejsach zdefiniowanych przez O-RAN Alliance [5]. Jednakże brak niezależnego organu do utrzymania schematów (zarówno NESAS, jak i testów bezpieczeństwa O-RAN) oznacza, że programy te są bardziej samoocenią bezpieczeństwa urządzeń i nie powinny być traktowane jako certyfikacja produktów przez stronę trzecią.

Powszechnie rozpoznawalną, niezależną normą certyfikacji cyberbezpieczeństwa jest *Common Criteria* [1]. Dotychczas CC opierało się na dwóch porozumieniach: globalnym CCRA (*Common Criteria Recognition Arrangement*) oraz europejskim SOG-IS MRA (*Senior Officials Group – Information Security Mutual Recognition Agreement*). Członkowie tych porozumień zobowiązują się do wzajemnego uznawania wydawanych przez swoje schematy certyfikatów.

Milowym krokiem w obszarze certyfikacji cyberbezpieczeństwa w ramach Unii Europejskiej było przyjęcie Rozporządzenia EU 2019/881, zwanego Aktem o Cyberbezpieczeństwie CSA (*Cybersecurity Act*). Jest to akt horyzontalny, umożliwiający ustanawianie niezależnych europejskich programów certyfikacji z uwzględnieniem zasad nadzoru nad rynkiem oraz zasad normalizacji (normy i specyfikacje techniczne). W oparciu o CSA opracowano pierwszy europejski schemat certyfikacji EUCC (*European Common Criteria-based cybersecurity certification Scheme*), który rozpoczął funkcjonowanie w lutym 2025 roku. EUCC jest następcą porozumienia SOG-IS, i przeznaczony jest do realizacji ewaluacji cyberbezpieczeństwa zgodnie z normą CC, w szczególności w takich domenach technicznych jak karty inteligentne oraz urządzenia sprzętowe z modułami bezpieczeństwa.

Mając na uwadze istotność certyfikacji, agencja UE ds. Cyberbezpieczeństwa (ENISA) rozpoczęła prace nad niezależnymi schematami certyfikacji w kolejnych obszarach tematycznych, w tym dla sieci 5G - tzw. schemat EU5G. Niemniej jednak, EU5G jest wciąż w trakcie opracowywania, i na chwilę obecną nie wiadomo kiedy, oraz w jakiej formie, zostanie wdrożony.

Z kolei w artykule [5] zaproponowano metodykę ewaluacji dla modułów O-RAN bazującą na *Common Criteria*, a zatem wpisującą się w schemat EUCC. Jak przedstawiono w [5], projekt ewaluacyjny zgodny z normą CC obejmuje sześć klas aktywności: ASE, AGD, ADV, ALC, ATE i AVA. W ramach tych klas, deweloper musi przygotować i dostarczyć do laboratorium ewaluacyjnego dokumentację o ściśle określonych, sformalizowanych wymaganiach, zaś ewaluatorzy przeprowadzają zdefiniowany przez metodykę szeroki zestaw czynności celem oceny spełnienia określonych wymagań. W przypadku wykrycia niezgodności pomiędzy deklarowanymi wymaganiami produktu a stanem rzeczywistym, deweloper ma możliwość przeprowadzenia działań korygujących. Proces oceny zgodny z CC może zapewniać wysoki stopień zaufania z uwagi na szeroki zakres, rygor i wnikliwość oceny bezpieczeństwa. Z drugiej strony, wszystkie ww. czynniki powodują, że ewaluacja CC może trwać nawet kilkanaście miesięcy. Jest to kwestia szczególnie problematyczna w przypadku oprogramowania, gdzie certyfikat wydawany po kilkunastomiesięcznej ewaluacji często dotyczyłby nieaktualnej już wersji produktu.

Biorąc pod uwagę powyższe mankamenty normy CC, w niniejszym artykule autorzy proponują tzw. "lekką" metodykę ewaluacji cyberbezpieczeństwa, ukierunkowaną na software'owe komponenty O-RAN. Metodyka ta wprowadza kompromis - poprzez ograniczenie dogłębności, zakresu i rygoru oceny produktu zapewnia predefiniowany czas trwania procesu ewaluacji, zazwyczaj nieprzekraczający kilku miesięcy. Taka metodyka lepiej wpasowuje się w oczekiwania producentów urządzeń sieciowych, oczekujących głębszej korelacji pomiędzy czasem niezbędnym na przeprowadzenie oceny a kolejnymi fazami cyklu życia produktu software'owego.

3. LEKKA METODYKA OCENY O-RAN

Proponowana metodyka bazuje na europejskiej normie EN 17640 FiT CEM [3]. Norma ta została opublikowana pod koniec 2022 roku, i może mieć zastosowanie do każdego typu produktu IT. Należy zaznaczyć, że podobne do FiT CEM, własne lekkie metodyki są obecnie z powodzeniem stosowane w schematach certyfikacji we Francji (schemat CSPN prowadzony przez ANSSI) i w Niemczech (schemat BSZ prowadzony przez BSI).

Proponowana metodyka adaptuje ogólne zapisy FiT CEM, definiując aktywności odnoszące się do analizy podatności oraz zaprojektowania scenariuszy testowych z wykorzystaniem hipotez istniejących podatności (*flaw hy-*

pothesis). W jej ramach zaproponowano następujące mapowanie pomiędzy Zadaniami Ewaluacyjnymi FiT CEM (oznaczonymi jako "ZE"), a odnoszącymi się do nich specyfikacjami O-RAN Alliance (oznaczonymi jako "O-RAN.") oraz 3GPP (oznaczonymi jako "TS"):

a) ZE: weryfikacja kompletności - sprawdzenie czy producent dostarczył przedmiot oceny, wymagany zestaw dokumentacji oraz elementy niezbędne do zestawienia środowiska testowego; referencyjne zastosowanie ma tutaj specyfikacja O-RAN.WG11.TS.SecTestSpecs definiująca wymagania na testy O-RAN.

b) ZE: przegląd funkcjonalności bezpieczeństwa - sprawdzenie czy dokumentacja dewelopera w zrozumiały i spójny przedstawia funkcjonalności bezpieczeństwa produktu oceny; dokumentem referencyjnym jest w tym przypadku specyfikacja O-RAN.WG11.TS.SecReqSpecs.

c) ZE: ocena dokumentu ST (*Security Target*) - jest to uproszczona wersja działań ewaluatora odpowiadających aktywności ASE według normy CC (patrz [5]); dokumentami referencyjnymi są O-RAN.WG11.TR.Threat-Modeling i O-RAN.WG11.TS.SecReqSpecs, oraz zestaw dokumentów dotyczących wymagań bezpieczeństwa 3GPP: TS 33.501, TS 33.117, TS 33.210 i TS 33.310.

d) ZE: specyfikacja techniczna przedmiotu oceny - jest to uproszczona wersja działań ewaluatora odpowiadających aktywności ADV z CC ([5]); dokumentami odniesienia są O-RAN.WG1.TS.OAD i WG1.TS.SecProtSpec, zbiór dokumentów O-RAN specyfikujących API i protokoły aplikacyjne, zbiór specyfikacji 3GPP dotyczących mechanizmów i interfejsów gNB.

e) ZE: ewaluacja instalacji przedmiotu oceny - uproszczona wersja działań ewaluatora odpowiadających aktywności AGD z CC ([5]); działania obejmują ocenę opracowanego przez producenta dokumentu SUG (*Secure User Guidance*); jako dokument referencyjny służy specyfikacja O-RAN.WG11.TS.SecTestSpecs z wymaganą konfigurację badanego produktu.

f) ZE: testowanie zgodności z deklarowaną funkcjonalnością bezpieczeństwa - uproszczona wersja działań ewaluatora odpowiadających aktywności ATE z CC ([5]); dokumentem referencyjnym jest specyfikacja O-RAN.WG11.TS.SecTestSpecs oraz specyfikacje 3GPP typu SCAS (*Security Assurance Specifications*): TS 33.523 i TS 33.117.

g) ZE: przegląd i testowanie pod kątem znanych podatności oraz testy penetracyjne; działania te odpowiadają aktywności AVA z CC ([5]). Jako podstawę ich realizacji proponujemy podejście hipotez istniejących podatności zaprezentowane w dalszej części rozdziału.

Zgodnie z założeniami FiT CEM przyjmujemy, że producent przedmiotu oceny zobowiązany jest do dostarczenia, wraz z samym przedmiotem oceny, jedynie dwóch dedykowanych dokumentów: Specyfikacji Zabezpieczeń ST (*Security Target*) oraz Instrukcji Użytkownika SUG (*Secure User Guidance*). Specyfikacja ST to dedykowany dla danej implementacji dokument zawierający oświadczenie dla określonego przedmiotu oceny o jego wymaganiach w zakresie bezpieczeństwa. Z kolei instrukcja SUG to udokumentowane informacje opisujące kroki niezbędne do skonfigurowania przedmiotu oceny do docelowego stanu, w którym przedmiot oceny spełnia określone wymagania bezpieczeństwa. Dokumentacja projektowa przedmiotu oceny przede wszystkim bazuje na referencjach do specyfikacji O-RAN/3GPP wskazanych w punkcie d) powyżej, ewentualnie uzupełniona o specyfikację dodatkowych funkcjonalności, nieujętych w dokumentach O-RAN/3GPP.

Przyjmując powyższy zestaw dokumentów jako dane wejściowe, ewaluator przeprowadza testy penetracyjne z uwzględnieniem podejścia *flaw hypothesis* obejmującego następujące kroki:

I. Ewaluator gromadzi wszelkie informacje odnoszące się do bezpieczeństwa przedmiotu oceny, uwzględniając dostępną dokumentację techniczną produktu, publicznie znane podatności które mają zastosowanie do przedmiotu oceny, jak również nietypowe zachowanie przedmiotu oceny zaobserwowane podczas testów funkcjonalnych.

II. Ewaluator stawia hipotezę podatności dla przedmiotu oceny uwzględniając, że nie jest w stanie przetestować, w określonym czasie przeznaczonym na ewaluację, wszystkich możliwych ataków jakie można zdefiniować w oparciu o pozyskaną wiedzę i doświadczenie. Z tego względu z zakresu hipotezy należy wykluczyć: a) ataki o których wiadomo że wymagają zbyt dużego nakładu czasu i środków względem zakładanego potencjału ataku; b) ataki, które nie są możliwe z uwagi na rozpatrywane środowisko operacyjne przedmiotu oceny zakładające występowanie odpowiednich środków kontroli fizycznego dostępu do urządzeń sieciowych; c) ataki wymagające informacji, które nie są publicznie dostępne (np. kody serwisowe).

III. Ewaluator określa priorytety dla wybranych scenariuszy ataków oszacowując, które podatności są łatwiejsze do wykazania albo które ze scenariuszy mogą występować częściej dla danego typu przedmiotu oceny.

IV. Ewaluator wykonuje testy penetracyjne dla wybranego scenariusza, aby określić rzeczywisty potencjał ataku wymagany do wykazania istnienia podatności w przedmiocie oceny.

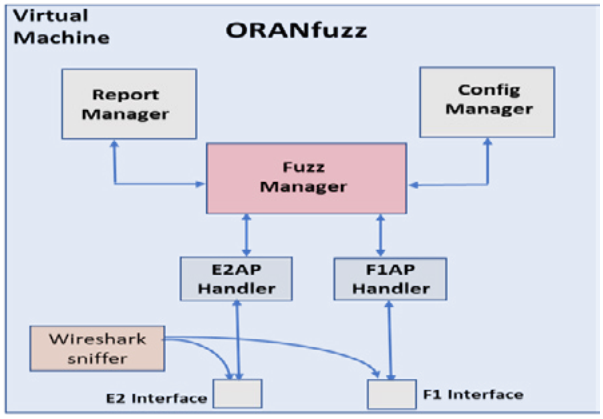
Na zakończenie procesu oceny ewaluator przygotowuje raport obejmujący listę hipotez istniejących podatności, opracowane dla nich scenariusze testowe, rezultaty

testów, zbiór podatności których istnienie udało się wykazać, a zatem które mają zastosowanie do przedmiotu oceny oraz przypisany im ich rzeczywisty potencjał ataku.

4. PROCES EWALUACJI MODUŁU O-DU

Praktyczną implementację zaproponowanej metodyki zrealizowano z wykorzystaniem otwarto-źródłowej implementacji komponentu O-DU [6]. Bazując na specyfikacjach wskazanych powyżej w rozdziale 3 pkt. c), dla oceny Specyfikacji Zabezpieczeń ST wyselekcjonowano zestaw wymagań obejmujący mechanizmy uwierzytelniania i autoryzacji, zachowania integralności i poufności, niezaprzeczalności danych oraz ochrony przez atakami powtórzeniowymi. Przyjęto wymagania na środowisko operacyjne dopuszczające sieciowy wektor ataku. Następnie bazując na specyfikacjach technicznych wskazanych w rozdz. 3 pkt. d), zidentyfikowano interfejsy F1, E2 i O1 komponentu O-DU jako posiadające hipotetyczne podatności.

Ponieważ przeszukanie publicznych baz (w tym m.in. *NIST National Vulnerabilities Database* oraz *Offensive Security Exploit Database*) nie wykazało znanych podatności jakie mogłyby mieć zastosowanie, zdefiniowano scenariusz ataku bazujący na weryfikacji mechanizmów bezpieczeństwa wskazanych interfejsów za pomocą dostępnych narzędzi *open-source*: *TLSSLed* w odniesieniu do protokołu TLS, skaner *nmap* i *Fuzzer SSH* w odniesieniu do SSH, skaner *nmap* i narzędzie *ike-scan* w odniesieniu do protokołów IKEv2/IPSec. Wyniki testów nie wykazały istnienia podatności na badanych interfejsach.



Rys. 1. Architektura narzędzia ORANfuzz.

Ponadto zaimplementowano własne narzędzie (ORANfuzz) dedykowane do badania wyspecyfikowanych przez 3GPP i O-RAN protokołów warstwy aplikacyjnej analizowanych interfejsów, tj. E2 i F1, z wykorzystaniem metody typu *fuzzing*. Architekturę narzędzia przedstawia rys. 1. Moduł *Report Manager* implementuje zestaw funkcji odpowiedzialnych za zapisywanie wszystkich wiadomości (logów) generowanych podczas pojedynczego przypadku testowego; w oparciu o przechwycone logi podejmowana jest decyzja czy test zakończył się wynikiem pozytywnym czy negatywnym. Moduł *Fuzz Manager* to główny komponent narzędzia, który pobiera konfigurację testu od modułu *Config Manager*, a następnie wywołuje moduły *E2AP/F1AP Handler* celem wysłania zmanipulowanych wiadomości protokołu E2/F1.

Przeprowadzono testy obejmujące walidację zarówno składni, tj. odporności na błędną strukturę ramek wiadomości, jak też walidację semantyczną, tj. odporności na niepoprawne wartości parametrów wiadomości. Chociaż przedmiot oceny nie obsługiwał w sposób poprawny wiadomości z zmanipulowaną strukturą nagłówka ramki, to ostatecznie obserwowane zachowanie nie wykazało podatności mogących naruszyć jego bezpieczeństwo. Końcowy rezultat ewaluacji był pozytywny.

5. PODSUMOWANIE

W artykule przedstawiono propozycję metodyki ewaluacji cyberbezpieczeństwa modułów Open RAN. Jest ona dedykowana w szczególności dla komponentów software'owych, gdyż bazuje na lekkiej metodyce FiT CEM gwarantującej predefiniowany, krótszy czas ewaluacji względem powszechnie stosowanej normy *Common Criteria*. Proponowana metodyka niesie wystarczający poziom zaufania, umożliwiając zarazem przewidywalny harmonogram procesu oceny i utrzymując obciążenie producenta ewaluowanego produktu na rozsądnym poziomie. Przedstawiona koncepcja może mieć zastosowanie przy realizacji krajowego bądź europejskiego schematu certyfikacji cyberbezpieczeństwa systemów 5G. Schematy takie mogą być ustanawiane w oparciu o rozporządzenie UE Akt o Cyberbezpieczeństwie, którego elementy dotyczące certyfikacji cyberbezpieczeństwa w odniesieniu do kwestii krajowych wdroży nadchodząca Ustawa o Krajowym Systemie Certyfikacji Cyberbezpieczeństwa.

Prace współfinansowane w ramach dotacji celowej nr 202/DC/D/2025 Ministerstwa Cyfryzacji oraz projektu KPOD.01.18-IW.03-0009/24 pt. „Krajowe laboratorium sieci i usług 5G wraz z otoczeniem” współfinasowanego w ramach Krajowego Planu Odbudowy sfinansowanego przez UE: NextGenerationEU.

LITERATURA

[1] Norma Common Criteria for Information Technology Security Evaluation 2022 (CC:2022). Revision 1. November 2022. Standard opracowany przez porozumienie CCRA.

[2] Common Criteria Recognition Arrangement (CCRA). Strona internetowa porozumienia (dostęp: 05.2025): https://www.commoncriteriaportal.org

[3] Norma EN 17640. 2022. "Fixed-time cybersecurity evaluation methodology for ICT products - FiT CEM". CEN/CENELEC

[4] O-RAN Work Group 1 (Use Cases and Overall Architecture), 2024. “O-RAN Architecture Description”, v11.0.

[5] Krawiec Piotr, Janowski Robert, Mongay Batalla Jordi [i in.]. 2025. „ On providing multi-level security assurance based on Common Criteria for O-RAN mobile network equipment. A test case: O-RAN Distributed Unit ”. *Computers & Security*, vol. 150: 104271.

[6] The O-DU-L2 project. O-RAN Software Community. Repozytorium GitHub projektu (dostęp 05.2025): https://github.com/o-ran-sc/o-du-l2

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

MACHINE LEARNING BASED DETECTION OF ATTACKS AIMED AT THE PACKET FORWARDING CONTROL PROTOCOL WITHIN THE 5G CONTROL PLANE

OPARTE NA UCZENIU MASZYNOWYM WYKRYWANIE ATAKÓW NA PROTOKÓŁ PFCP W PŁASZCZYŹNIE STEROWANIA 5GC

Andres Vejar[1]; Piotr Chołda[2]

AGH University of Krakow, Institute of Telecommunications, Kraków, Poland
[1] avejar@agh.edu.pl, [2] piotr.cholda@agh.edu.pl

DOI: 10.15199/59.2025.4.36

Streszczenie: Badania dotyczą zastosowania uczenia maszynowego w detekcji czterech klas ataków typu DoS ukierunkowanych na protokół PFCP w płaszczyźnie sterowania 5G. Modele opierają się na perceptronie wielowarstwowym i warstwach splotowych. Wykorzystują informację o pakietach na poziomie bajtów zamiast przepływów sieciowych. Użyto wyjaśnialnego uczenia maszynowego XML do wybierania cech o znaczeniu globalnym (udało się zredukować ich liczbę do 50%) przy utrzymaniu wysokiej dokładności klasyfikacji.

Abstract: This study investigates the use of machine learning methods to analyze four classes of denial-of-service attacks targeting the packet forwarding control protocol (PFCP). Neural network models were deployed with fully connected and convolutional layers, using byte-level packet information instead of network flows. Using explainable machine learning to select features of global importance; the models maintain high classification accuracy while decreasing training and evaluation times, requiring only half of the initial features.

Słowa kluczowe: gradient zintegrowany, płaszczyzna sterowania 5G, uczenie maszynowe, wyjaśnialność.

Keywords: 5G control plane, explainability, integrated gradients, machine learning.

1. INTRODUCTION

In the Fifth Generation (5G) core network, the control plane (CP) operates as a boundary element with the user plane (UP), enabling the correct handling of vital tasks such as session management and user authentication. This division between UP and CP functions facilitates rapid scalability and adaptability in network deployments, enhancing engineering flexibility to integrate cloud services and machine learning into the communication core. However, this adaptability introduces an increased risk surface with respect to network security and user privacy.

Attacks on the 5G core control plane were previously proposed to be analyzed with machine learning methods [1, 3]; however, there is a lack of studies considering an integral approach to generate valid communication traces using realistic infrastructure and real-world communication scenarios.

In this study, we deploy attacks aimed at the packet forwarding control protocol (PFCP) in a live communication core, where automated user agents perform normal tasks such as browsing the World Wide Web. We perform four classes of denial-of-service (DoS) attacks from a malicious agent infiltrated in the core and compare the results with live sessions without attacks. With the network traces, two types of detection models were trained and tested, a multilayer perceptron and a convolutional neural network.

Instead of collecting network flows, we use byte-level information as the input of the classifier, allowing for an earlier detection of these attacks with respect to flow-based classification systems. The models achieve very high classification performance. Using the integrated gradients method for model explainability, we detected that roughly half of the byte features are the most relevant for a classification decision. Therefore, a significant reduction in model training and evaluation can be obtained using half of the original input data. The reduced models achieve the same levels of classification performance as the original full models.

Our work differs from the above-mentioned work of Amponis et al. [1], given that we rely on realistic infrastructure and a very broad approach to attack emulation. This paper reports on an activity being part of the group focused on the development of software to implement and test a zero-trust security architecture in a core network. The component realized relates to the detection of anomalies in signaling data flow and is aimed at activating a multilevel authentication system upon the detection of such anomalies, as well as methods for dynamic segmentation.

2. METHODOLOGY

2.1 Studied Attacks and Data Generation

An important element of the 5G core (5GC) network is the PFCP protocol that operates on the N4 interface of 5GC. In the literature, there is continuous use of flow-

based classification [2], where the focus is on the detection of network attacks in temporal aggregates derived on the basis of well-defined data flows. To acquire labeled datasets, a common strategy involves executing targeted attack sessions. This facilitates the application of supervised classification training methods.

There are four distinct labels: one for usual traffic and three indicating PFCP-specific DoS attacks. (1) **PFCP Session Deletion Flood Attack**. This attack aims to dissociate the target UE from the data network (DN). It targets the protocol data unit (PDU) sessions between the clients and the DN to disconnect only the user equipment (UE) from the DN, not the 5G Radio Access Network (RAN) or the Core network. (2) **PFCP Session Establishment Flood Attack**. The attack aims to deplete the capacity of the user plane function (UPF) to respond to valid requests for heartbeats and session establishments. This complicates the establishment of new PDU connections between clients and the DN in the 5GC. With every request for a session creation, the session id (SEID) is generated at random. (3) **PFCP Session Modification Flood Attack** is present in two versions: (a) **DROP Apply Action Field Flags**. This attack aims to disconnect a user (UE) from the DN by discarding packet handling rules for a specific session. The Forwarding Action Rule (FAR) rules, containing the base station's IP address and TEID, are erased from the UPF if successful. The user is cut off from the Internet via the GPRS Tunneling Protocol (GTP) tunnel for their downlink data. It disconnects PDU sessions between clients and the DN, without severing their connection to the 5G RAN or Core network. (b) **DUPL Apply Action Field Flags**. The attack aims to duplicate session rules by using the DUPL flag in the Apply Action field, creating multiple paths for the same data from a single source. This could cause undefined behavior on the N6 interface or duplicate traffic upon transmission to the DN. It could be part of a larger scheme to perform distributed DoS attacks against DN hosts and exhaust UPF's resources for forwarding outgoing packets outside the 5GC. A malicious actor can generate a near-passive attack vector, affecting the traffic of numerous subscribers and exhausting UPF's packet handling resources.

The scripts specifically designed for injecting attacks were utilized. Initially, a synthetic dataset was employed, characterized by the following attributes: (a) too simplistic background traffic – absence of UE activity during attacks; (b) too few UEs present; (c) no interaction with attack traffic; (d) long capture sessions.

Long capture sessions for PFCP attacks do not have merit, since DoS attacks will be noticed almost immediately. Extended observation of the attack scenario is impractical, as users instinctively alter their behavior when noticing the attack effects. They may stop sending regular traffic and attempt to report the attack. Consequently, it is assumed that the entire infrastructure is reset following the attack incident.

It is better to repeat the attack many times with short capture sessions, for example, until the level of dozens of minutes. Therefore, new attack instances are created and attacks are observed from scratch. The question of how long to collect the data appeared. Therefore, the study took into account various levels of attack gathering time intervals: 75, 106, 150, 212, and 300 seconds. In all cases, the model was trained and tested independently for each of these time frames. Precisely distinguishing between normal and attack packets in concurrent traffic streams during an attack is crucial. Despite machine learning models usually discarding routing details like addresses and ports, layer 2 MAC addresses can tag packets as normal or anomalous. These labels, although not used by the models, are valuable for model validation, training, and testing.

Simple data is not necessarily realistic, especially with very basic background data. The first classification results were of good quality and promising. However, there were doubts about whether traffic reflects real situations. Therefore, a decision was taken to generate real background traffic. Methods were searched to generate more realistic traffic. The Python-based Noisy library was selected, intended to produce background traffic. Now, the 5G core infrastructure cooperated with hundreds of UEs that generated independent internet traffic, introducing greater entropy into the system. Unlike previous experiments in which classification achieved almost perfect results due to the lack of real system load, the new data showed more complex traffic patterns. The data distribution turned out to be more complex and non-linear, which made separation more difficult and increased classification challenges.

In all the cases, two types of model were trained and tested: one based on multilayer perceptron (MLP), that is, with dense connections between the layers, and a one-dimensional convolutional neural network (1D-CNN), i.e., with sparse interactions and filters that recognize specific patterns over many features. In order to provide a fair comparison, both MLP and 1D-CNN applied the same number of parameters.

2.2 Modeling

Models work with raw bytes from packets to find viable patterns for classification, contrary to previously tested models, where aggregated features extracted from packet flows are also studied. First, we analyze the potential to separate various types of packets, a projection is provided to show whether various classes can be separated (Figure 1).

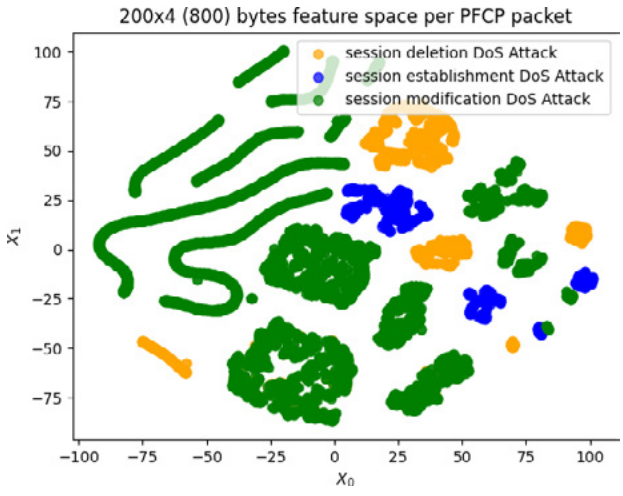


Figure 1: Projection of attack data with t-SNE.

We are able to check at least by visualizing with the t-distributed stochastic neighbor embedding method (t-SNE) [4] that there is potential for separability of normal

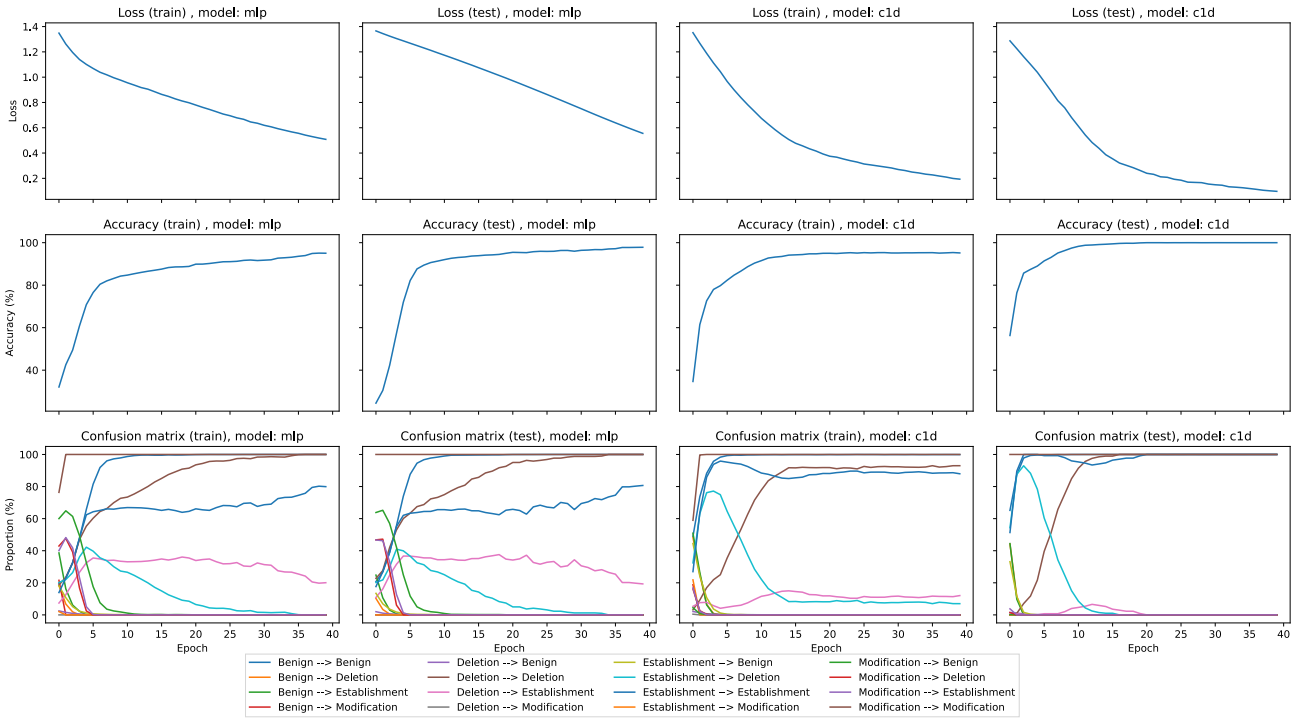


Figure 2: Performance results. First column MLP (train), second column MLP (test), third column 1D-CNN (train) and fourth column 1D-CNN (test). First row is the optimization loss, second row the accuracy, third row the confusion matrix proportions (%). For each `<original_label>` → `<estimated_label>` pair, the proportion is (100%) count`<estimated_label>`/count`<original_label>`.

and anomalous traffic, so that machine learning models can be useful for classification in the case of realistic data. The data is gathered in the form of `pfcp` packets, and only up to the first 200 bytes are being treated as input to the model. In both cases of MLP and CNN, we treat each numbered byte as the feature (attribute) of the data analyzed.

The basic hyperparameters of the models were selected in a way that they can be fairly compared, that is, MLP and CNN apply the same number of parameters (neuron weights). For selection of training / test datasets, stratified partitioning 70/30 was applied. The CNN model used applies one-dimensional filters (1D), that is, bytes of packets are analyzed in a series. During the training phases, the data sets were balanced using stratified sampling. We can observe that a rapid overfitting occurs in the case of MLP. On the other hand, CNN does not overfit. Hence, this is a better model. Generally, the results show an ideal case: performance on the test dataset shows better accuracy values than in the case of the training dataset. The training set is larger and thus more complex; thus this phenomenon can be understood. However, it shows good separability of benign / malicious cases.

3. EXPERIMENTAL RESULTS

A dense MLP model is used as a baseline to control the quality of the 1-dimensional convolutional model (Conv1D, C1D), which is treated as the proposed model. In the following, the plots summarizing the results are shown in Figure 2. The results are obtained with MLP (dense network) and CNN (convolutional neural network). The columns show: first, the results for MLP with the training set, the results for MLP with the test set, the results for CNN with the training set, and the results for CNN with the test set. The rows present: the loss function used to fit the models, the accuracy function to assess the performance of the classification task, the results related to the true negatives (Benign → Benign), true positives (e.g., Modification → Modification), omissions in detection – that is false negatives (e.g., Deletion → Benign), false alarms – that is false positives (e.g., Benign → Deletion). In addition, errors consisting of misclassifications of various attacks are also shown.

Generally, in all cases the prevalence of 1D-CNN over MLP is clearly visible. Although the convergence of 1D-CNN is quite fast obtained, and even the results for the training data set are very good (even faster from the viewpoint of training epochs than in the case of the training dataset), in this case the behavior is not coherent, and the level of errors is still high when we finish training. However, the differences between the performance of models trained with various training datasets are barely visible. The test data set is based on the subsampling of the data set related to a specific time frame.

In addition, to increase understanding and certainty about the results obtained, the explainability analysis has been performed. Generally, explainability can be useful in three ways: (1) Due to the usage of the explainability practices, it is possible to better understand the models (e.g. by analyzing false positives, negatives, etc.) and potentially improve them. (2) It is also possible to diminish the amount of the used data: we analyze which byte parts of the packets are, in fact, activating the neurons (e.g., filters in CNN layers). If we are able to state that some parts of the packets are not useful, then they can be simply omitted in the gathering process or at least can be gotten rid of during the pre-processing data and then not applied as

features to the models. (3) It is also possible by analyzing which concrete values of bytes in concrete positions (e.g., fields in the layer headers) are responsible for activations to increase the likelihood of proper classification, to construct shallow models (or even simple rule-based systems) to provide the filtering tools (potentially to be used instead of the machine learning models).

Analysis with the *Integrated Gradients* (IG) has been performed. This is a well-established tool used for interpretability purposes. Its largest advantage is that the values provided are meaningful and easy to understand; in addition, its operation is fast. The results should be read in the way that the absolute value (either positive or negative values with large amplitudes) shows the importance of a specific feature (here: byte in concrete position in the packet; the notion of feature in this context is clearer in the case of MLP).

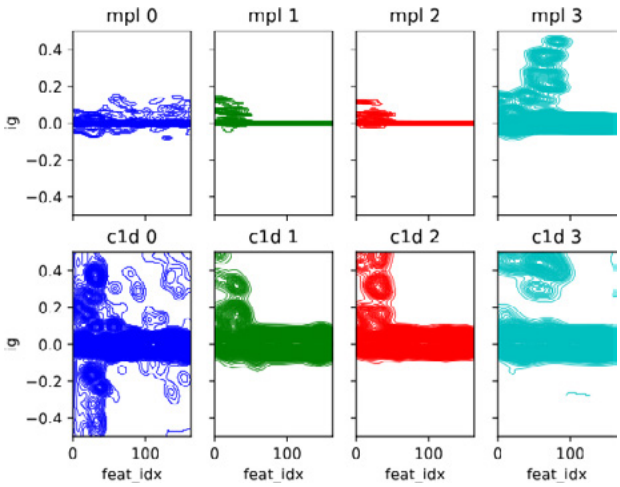


Figure 3: IG explainability per feature-byte. First row MLP, second row 1D-CNN. PFCP DoS attacks, first column: "Benign", second column: "Session Deletion", third column: "Session Modification", and fourth column: "Session Establishment".

The results are presented in Figure 3. The first column presents the importance of the IG attribution of the features to classify a label 0 sample (Benign). Similarly, the second column links to label 1 (Session deletion DoS attack), the third column links to label 2 (Session modification DoS attack), and the fourth column links to label 3 (Session establishment DoS attack). The first row of the previous figure presents the importance of IG attribution for the baseline model (MLP). The second row presents the importance of IG attribution for the CNN 1d model.

Basically, it can be stated that out of 200 bytes that are analyzed by both models, only half of them are important to provide a useful classification. In fact, mainly, the first bytes are most important. An analysis was conducted on the impact of individual characteristics on the explainability of the model. Experimental verification confirmed that crucial information is contained in the first part of the features, enabling rejection of the second part without significant loss of model quality. Training the MLP model on half of the features requires slightly more time, and for C1D, it is necessary to increase the number of epochs, but the final results remain at a comparable level. Limiting the number of features significantly simplifies the model and the structure of the neural network.

In the figure, the results related to the KDE (kernel density estimation) are shown. In practice, this means that the smoothed results related to pmf (probability mass function) are shown. The islands show where there is a considerable presence of the integrated gradient value (both with positive or negative value; in fact, only the magnitude, i.e., absolute value, matters). In these areas, apparently the model(s) are more interested in the features.

Due to initial explainability studies, it appeared that some important informative fields (e.g., header flags, especially the ones related to PFCP) do not occupy a full byte, while the model initially operated at the byte level. This resulted in the loss of part of the information. While the first idea was to focus on half-bytes, eventually it was decided to deal with bit features.

4. CONCLUSIONS

This study investigates the use of machine learning methods to analyze attacks on the fifth generation (5G) core network. It tests four classes of denial-of-service attacks targeting the packet forwarding control protocol (PFCP) on a live communication core, comparing them with live sessions without attacks. We train two detection models, a multilayer perceptron and a convolutional neural network, using byte-level packet information instead of network flows. By employing integrated gradient techniques for explainable machine learning to select features of global importance, the models maintain high classification accuracy while significantly decreasing training and evaluation times, requiring only half of the initial features. This general approach can be extended to improve network security and user privacy for different 5G core protocols.

ACKNOWLEDGMENTS

This research was supported by the National Research Institute, grant number KPOD.01.18-IW.03-0009/24 on “National Laboratory for Advanced 5G Research” as part of the National Recovery and Resilience Plan, Investment A2.4.1: Investments in the Development of Research Potential.

REFERENCES

[1] George Amponis et al. “5G Core PFCP Intrusion Detection Dataset”. In: *2023 12th International Conference on Modern Circuits and Systems Technologies (MOCAST)*. Athens, Greece, 2023.

[2] George Amponis et al. “Generating full-stack 5G security datasets: IP-layer and core network persistent PDU session attacks”. In: *International Journal of Electronics and Communications (AEÜ)* 171 (2023), p. 154913.

[3] Matthew William Anderson et al. *Machine Learning 5G Attack Detection in Programmable Logic*. Tech. rep. INL/CON-22-68438-Revision-2. Idaho Falls, Idaho: Idaho National Laboratory, 2023.

[4] Laurens van der Maaten and Geoffrey Hinton. “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

WYZNACZANIE ENERGOOSZCZĘDNYCH ŚCIEŻEK ROUTINGU W SIECIACH FANET Z WYKORZYSTANIEM MODELI LLM SONAR

DETERMINING ENERGY-EFFICIENT ROUTING PATHS IN FANET NETWORKS USING THE LLM SONAR MODELS

Tomasz Jóźwik[1]

[1] Wojskowa Akademia Techniczna, Warszawa,
tomasz.jozwik@wat.edu.pl

DOI: 10.15199/59.2025.4.37

Streszczenie: Referat przedstawia problemy oraz wyzwania sieci FANET związane z routingiem, w szczególności w zakresie energooszczędności. Przedstawiono możliwości zastosowania modeli LLM w celu wyznaczania ścieżek routingu sieci FANET. Dokonano definicji odpowiednich zapytań dla modeli LLM i zestawiono uzyskane wyniki z modelem wykorzystującym solver. Omówiono napotkane problemy na przykładzie modeli z rodziny Sonar firmy Perplexity. Przedstawiono potencjalne możliwości zastosowania wykorzystanego podejścia w dalszych badaniach.

Abstract: The paper presents problems and challenges of FANET networks related to routing, particularly in the area of energy efficiency. The possibilities of using LLM models for determining FANET network routing paths were presented. Appropriate queries for LLM models were defined and the obtained results were compared with a model using a solver. The encountered problems were discussed using examples of models from the Sonar by Perplexity family. Potential possibilities for applying the used approach in further research were presented.

Słowa kluczowe: energooszczędność, LLM, routing, sieci FANET, Sonar

Keywords: energy efficiency, FANET networks, LLM, routing, Sonar

1. WSTĘP

Sieci FANET (flying ad-hoc networks) są wykorzystywane w zastosowaniach takich jak zarządzanie kryzysowe, kontrole granic, misje ratownicze oraz nadzorowanie ruchu drogowego. Jednym z wskazywanych przez badaczy wyzwań w sieciach FANET jest dostęp do ograniczonych zasobów energii [3].

W celu efektywnego wykorzystania dostępnych zasobów energetycznych, projektowane są nowe rozwiązania, dotyczące między innymi routingu [11]. W celu wyznaczenia optymalnych ścieżek pod względem zużycia energii, utworzono model optymalizacji matematycznej dla routingu unicast [7].

Modele LLM wykazują zastosowania w obszarach takich jak tłumaczenie, streszczanie i analiza sentymentu [6]. Niniejszy referat dokonuje analizy możliwości zastosowania modeli LLM do wyznaczania ścieżek routingu w sieciach FANET poprzez dokonanie reimplementacji istniejącego modelu optymalizacji matematycznej routingu dla protokołów unicast [7] na potrzeby środowiska LLM, oraz ocenę uzyskanych wyników przy zastosowaniu takiego podejścia.

2. OPIS PROBLEMU

2.1 Routing w sieciach FANET

Badacze wskazują na główne cele protokołów routingu dla sieci FANET – poprawę stabilności, pokrycia sieci oraz wydajności routingu. Wskazywanymi problemami występującymi na drodze do powyższego celu są dynamika węzłów, koszt routingu oraz ograniczone zasoby pozostałej energii na poszczególnych węzłach [9].

Istniejące rozwiązania próbują zminimalizować problem ograniczonej dostępności energii poprzez implementację protokołów routingu wykorzystujących techniki takie jak Q-Learning [5] oraz logikę rozmytą [1].

2.2 Modele LLM

Modele LLM są modelami językowymi wykorzystującymi bardzo dużą liczbę parametrów w celu zwiększenia ich możliwości [10]. Implementowane są zróżnicowane techniki mające na celu umożliwienie modelom LLM prowadzenie procesu wnioskowania [12].

Badania prowadzone na podstawie istniejących produkcyjnych implementacji modeli LLM, takich jak ChatGPT, wskazują na możliwość zastosowania modeli LLM w takich dziedzinach jak programowanie, media społecznościowe, silniki wyszukiwania i edukacja [4].

Badacze weryfikowali modele LLM również pod kątem wykonywania zadań i analiz związanych z sieciami komputerowymi [2].

2.3 Modele Sonar firmy Perplexity

Firma Perplexity oferuje możliwość komercyjnego dostępu do stworzonych przez siebie modeli LLM [8]. Oferowany interfejs API umożliwia dostęp do modeli syntezu-

jących informacje, rozumujących i dokonujących głębokiej analizy.

Na potrzeby badań wykorzystane zostaną modele LLM dostarczane przez Perplexity [8]:

- Sonar – prosty model, dla którego jednym z kluczowych elementów jest koszt wykonania,
- Sonar Pro – zaawansowany model LLM,
- Sonar Reasoning – wydajny model rozumujący,
- Sonar Reasoning Pro – zaawansowany model rozumujący, wykorzystujący technikę Chain-of-Thought.

3. METODOLOGIA

3.1 Implementacja

Budowa zapytania oraz parsowanie odpowiedzi zostały zaimplementowane w formie skryptu w języku programowania Ruby. Stworzony skrypt wczytuje przypadek testowy zdefiniowany w pliku JSON. Przypadek testowy załączany jest jako część zapytania, zawierającego zasady wyznaczania ścieżek routingu unicast w sieciach FANET. Reguły wyznaczania ścieżek zostały zdefiniowane w formie poleceń tekstowych w języku angielskim i są równoważne uprzednio zaimplementowanemu modelowi optymalizacji, rozwiązywanemu z wykorzystaniem solvera [7].

Zdefiniowane zapytanie, wraz ze szczegółami przypadku testowego, przesyłane jest do wybranego modelu LLM z rodziny Sonar. Komunikacja z modelem odbywa się z wykorzystaniem biblioteki Typhoeus, która przekazuje zapytanie do interfejsu API, udostępnianego przez firmę Perplexity.

Odpowiedź uzyskana poprzez interfejs API ma format JSON i jest parsowana po stronie napisanego skryptu. W przypadku zapytania obsługiwanego przez modele rozumujące, odpowiedź zawiera dodatkowo informacje dotyczące przeprowadzonego procesu wnioskowania, prowadzącego do uzyskania wyniku zawartego w strukturze JSON.

3.2 Budowa zapytań

Przetwarzane zapytanie składa się z kluczy: definiującego wykorzystywany model LLM, zawierającego wiadomości wejściowe do wybranego modelu LLM oraz definiującego oczekiwany format odpowiedzi zwracanej przez model.

Klucz *model* informuje interfejs API o żądanym modelu LLM, stosowanym do rozwiązania problemu. Podczas testów wykorzystano cztery modele LLM dostarczane przez firmę Perplexity, co wiązało się z zastosowaniem poniższych wartości parametru *model*:

- *sonar* – dla modelu Sonar,
- *sonar-pro* – dla modelu Sonar Pro,
- *sonar-reasoning* – dla rozumującej wersji modelu Sonar Reasoning,
- *sonar-reasoning-pro* – dla rozumującej wersji modelu Sonar Reasoning Pro.

W tablicy *messages* zdefiniowano wiadomości przesyłane do modelu LLM. Wykorzystano wiadomość z rolą systemową, informującą model, aby zachowywał się jako pomocny asystent, specjalizujący się w rozwiązywaniu problemów matematycznych oraz optymalizacyjnych. Drugą wiadomością przesłaną do modelu jest zapytanie w roli użytkownika, zawierające tekstową definicję problemu w języku angielskim. Poinstruowano model o strukturze danych wejściowych jako przepływów w grafie.

Zdefiniowano każdy przepływ jako obiekt posiadający węzły źródłowy i docelowy, rozmiar oraz czas generacji. Przeniesiono ograniczenia zdefiniowane uprzednio w modelu optymalizacji matematycznej [7], jednocześnie redagując je w formie opisowej. Na potrzeby testów zdefiniowano koszt przesyłu danych jako funkcję wprost proporcjonalną do kwadratu odległości między komunikującymi się węzłami. Poinstruowano model, iż jego zadaniem jest zaplanowanie transmisji podanych przepływów, tak aby całkowity koszt transmisji był jak najniższy, przy jednoczesnym zachowaniu zdefiniowanych ograniczeń, co jest tożsame z funkcją celu zastosowaną w uprzednio zdefiniowanym modelu matematycznym.

Jako ostatnią część wiadomości załączono przypadek testowy do analizy, przedstawiony w formacie JSON. W opisie parametrów modelu oraz jego ograniczeń zastosowano odniesienia do nazewnictwa kluczy wykorzystywanych do definicji przypadku testowego w formie JSON, tak aby możliwe było jednoznaczne powiązanie konkretnych wartości z pliku testowego z wymaganymi do zastosowania ograniczeniami oraz oczekiwaną specyfiką przesłanych przepływów.

Ostatnim z wykorzystanych kluczy jest *response_format*, określający format odpowiedzi modelu LLM. Ustalono oczekiwany format jako JSON oraz zdefiniowano strukturę odpowiedzi zawierającą obiekty *total_cost* oraz *flow_parts*.

Wartość *total_cost* zawiera informację o całkowitym koszcie transmisji przypadku testowego. W przypadku uzyskania optymalnego planu transmisji przez model LLM, wartość *total_cost* powinna być równa wartości funkcji celu uprzednio zdefiniowanego modelu optymalizacji matematycznej, wykonującego analizę tego samego przypadku testowego.

Obiekt *flow_parts* jest tablicą zawierającą obiekty opisujące poszczególne etapy transmisji. Każdy z obiektów składa się z wymaganych atrybutów:

- *part_sending_node_id* – określający węzeł źródłowy, z którego została wysłana zadana część przepływu,
- *part_receiving_node_id* – określający węzeł docelowy, do którego wysłano zadaną część przepływu,
- *original_flow_id* – określający przynależność analizowanej transmisji do konkretnego przepływu,
- *size* – określający rozmiar przetransmitowanej części przepływu,
- *time* – określający zdyskretyzowany czas, w którym nastąpiła transmisja analizowanej części przepływu.

3.3 Definicja scenariuszy

Do definicji scenariusza testowego wykorzystano format JSON. Przeniesiono parametry wykorzystywane uprzednio w modelu optymalizacji matematycznej [7] oraz dokonano definicji struktury pliku zawierającego dane przypadku testowego w celu jego efektywnej analizy przez model LLM.

Niech plik JSON zawiera klucze: *Flows* – będący tablicą opisującą wszystkie występujące w przypadku testowym przepływy, oraz *Nodes* – tablicę opisującą węzły

oraz ich położenie w poszczególnych zdyskretyzowanych momentach przypadku testowego.

Obiekty w tablicy *Flows*, definiujące przepływy, opisano z wykorzystaniem następujących atrybutów:

- *flow_id* – liczba naturalna jednoznacznie określająca przepływ,
- *generation_timestamp* – liczba naturalna, oznaczająca zdyskretyzowany moment generacji przepływu,
- *source_node_id* – liczba naturalna identyfikująca węzeł generujący przepływ,
- *destination_node_id* – liczba naturalna identyfikująca węzeł docelowy dla przepływu,
- *size* – rozmiar przepływu, wyrażony poprzez liczbę wymierną.

Każdy z obiektów *Nodes* posiada następujące definiujące go atrybuty:

- *id* – liczba naturalna, jednoznacznie identyfikująca dany węzeł sieci,
- *positions* – tablica definiująca położenie węzła w danym momencie zdyskretyzowanego czasu.

Każdy z elementów tablicy *positions* jest parą atrybutów:

- *timestamp* – liczba naturalna, określająca moment zdyskretyzowanego czasu przypadku testowego,
- *position* – trójka liczb (x, y, z), określająca położenie węzła w trójwymiarowej przestrzeni euklidesowej.

Przypadek testowy opisywany jest przez globalne atrybuty:

- *max_timestamp* – liczbę naturalną, określającą liczbę zdyskretyzowanych próbek czasu w symulacji,
- *max_range* – liczbę wymierną, określającą maksymalny zasięg transmisji,
- *transmitter_power* – liczbę wymierną, określającą maksymalną moc nadajnika pojedynczego węzła w pojedynczej próbce zdyskretyzowanego czasu.

3.4 Scenariusz testowy

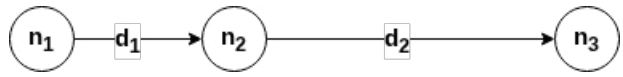
Niech zdefiniowana będzie topologia sieci składająca się z węzłów: n_1, n_2, n_3. Dany jest przepływ m, dla którego węzłem początkowym jest węzeł n_1, a węzłem docelowym węzeł n_3. Odległość między węzłami jest stała w czasie i wynosi odpowiednio d_1 dla pary węzłów (n_1, n_2) oraz d_2 dla pary węzłów (n_2, n_3).

Jednocześnie moc nadajników na poszczególnych węzłach nie pozwala na przesłanie całego przepływu m między parami węzłów (n_1, n_3) oraz (n_2, n_3) podczas pojedynczej próbki czasu.

Prawidłowe, a jednocześnie optymalne, rozwiązanie powyższego przypadku testowego zakłada:

- wykorzystanie węzła pośredniczącego n_2 do transmisji przepływu m,
- rozdzielenie przepływu m na części, których przesłanie między węzłami n_2 i n_3 nastąpi w różnych momentach zdyskretyzowanego czasu,
- transmisję całego przepływu podczas trwania testu.

Wykorzystana topologia sieci została przedstawiona na Rysunku 1.



Rysunek 1: Topologia sieci

4. WYNIKI

Dokonano 5-krotnego uruchomienia scenariusza testowego na każdym z testowanych modeli z rodziny Sonar. Zastosowano domyślne parametry konfiguracyjne modeli Sonar, zdefiniowane przez firmę Perplexity.

4.1 Uzyskane wyniki

W przypadku modeli Sonar i Sonar Pro nie udało się uzyskać poprawnego wyniku dla zdefiniowanego scenariusza testowego. Modele LLM zwracały wynik w oczekiwanym formacie JSON, jednakże nie był on możliwy do osiągnięcia. Zastosowanie proponowanych przez modele scenariuszy prowadziłoby do naruszenia zdefiniowanych ograniczeń.

W przypadku modeli rozumujących – Sonar Reasoning i Sonar Reasoning Pro, uzyskano poprawne rozwiązania zadanego przypadku testowego, tożsame z wartością funkcji celu zwracaną przez model matematyczny dla uprzednio zdefiniowanego modelu optymalizacji [7]. W większości przypadków (80% zakładanych uruchomień) uzyskano optymalny koszt transmisji. Zaobserwowano zwracanie różnych, równoważnych (pod względem całkowitego kosztu przesyłu) ścieżek routingu oraz zróżnicowanego podziału przesyłów na mniejsze części dla kolejnych uruchomień przypadku testowego.

Zaobserwowano przypadki wystąpień problemów ze zwracanymi ścieżkami przez modele rozumujące. Dokonano dodatkowych uruchomień modeli na zadanym przypadku testowym, w celu weryfikacji rodzajów potencjalnych nieprawidłowości.

4.2 Napotkane problemy

Zaobserwowano przypadek zwrócenia nieprawidłowych wyników przez model Sonar Reasoning. Zwracany koszt przesyłu był niższy od optymalnego rozwiązania w związku z nieprzesłaniem całego oczekiwanego przepływu. Zaproponowane przez model rozwiązanie zakładało, iż suma transmisji między parami węzłów (n_1, n_2) oraz (n_2, n_2) powinna być równa całkowitemu rozmiarowi przepływu m. W rzeczywistości było to podwójne zliczenie tej samej części analizowanego przepływu.

Napotkano na przypadek zignorowania części ograniczeń przez model Sonar Reasoning. Model zwrócił poprawny, optymalny koszt transmisji dla analizowanego przypadku testowego, ale dokonał transmisji całego przepływu m, między parą węzłów (n_2, n_3), w jednym zdyskretyzowanym momencie. Powyższa sytuacja jest niemożliwa ze względu na niewystarczającą moc nadajnika na węźle n_2.

Zaobserwowano przypadek zwrócenia nieoptymalnego rozwiązania przez model Sonar Reasoning Pro. Zwrócone ścieżki routingu były poprawne – wszystkie zadane ograniczenia były zachowane. Jednocześnie nastąpiła transmisja części przepływu m bezpośrednio między parą

węzłów (n_1, n_3), a w konsekwencji uzyskanie nieoptymalnego całkowitego kosztu transmisji.

Zauważono przypadek niewłaściwego parsowania uzyskanego rozwiązania dla modelu Sonar Reasoning Pro. Model dokonał poprawnego procesu wnioskowania i uzyskał optymalne rozwiązanie. Jednocześnie zwrócona odpowiedź w formacie JSON zawierała poprawną wartość całkowitego kosztu transmisji, ale niepoprawny plan transmisji przepływu. Zwrócona odpowiedź zawierała nieistniejące identyfikatory węzłów i przepływów, a także próbki zdyskretyzowanego czasu wykraczające poza okres symulacji. Opis wnioskowania zawierał wyłącznie poprawne dane, co wskazuje na problem z parsowaniem rozwiązania przez model. W efekcie, uzyskane rozwiązanie nie nadawałoby się do zastosowania w rzeczywistych systemach na podstawie samej odpowiedzi JSON.

4.3 Zestawienie z modelem matematycznym

Na podstawie przeprowadzonych badań, można stwierdzić, że zastosowanie rozumujących modeli LLM pozwala na wyznaczenie optymalnego planu routingu dla prostych przypadków testowych. Jednocześnie model LLM nie gwarantuje uzyskania optymalnego, ani nawet prawidłowego wyniku. W przypadku automatycznego wykorzystania odpowiedzi przez inne narzędzie lub zastosowania jej w rzeczywistym środowisku, podejście wykorzystujące LLM może okazać się niewystarczające i prowadzić do uzyskania nieprzewidywalnych rezultatów. Dodatkowo, wykonanie analizy pojedynczego przypadku testowego przez LLM trwa znacznie dłużej i jest bardziej kosztowne od uruchomienia modelu optymalizacji wykorzystującego solver.

Zastosowanie modelu optymalizacji matematycznej pozwala na ominięcie części powyższych problemów. O ile rozwiązanie przypadku testowego jest możliwe, solver pozwala na uzyskanie optymalnego planu routingu. Wykonanie obliczeń zajmuje znacznie krótszy czas i nie wymaga korzystania z zewnętrznych interfejsów API lub znacznych zasobów obliczeniowych w celu uruchomienia modeli LLM w środowisku lokalnym.

Ze względu na przedstawiane kroki wnioskowania, zastosowanie rozumujących modeli LLM ma potencjał przy projektowaniu protokołów i heurystyk routingu dla prostych przypadków testowych. Odpowiednie sformułowanie zapytania oraz ręczna analiza wyników zwracanych przez model LLM może stanowić dodatkowy, niezależny punkt odniesienia, możliwy do wykorzystania podczas projektowania nowych rozwiązań związanych z routingiem.

5. PODSUMOWANIE

Przeprowadzone badania wskazują, że modele rozumujące LLM mogą zostać wykorzystane do planowania energooszczędnego routingu dla prostych przypadków testowych. Jednocześnie, istnieją inne rozwiązania (np. modele optymalizacji), które pozwalają na efektywniejsze prowadzenie takich analiz.

Zastosowanie modeli LLM ma wartość szczególnie podczas początkowych etapów projektu poprzez wskazanie możliwych ścieżek wnioskowania. Wykorzystanie LLM w takim celu może pomóc w projektowaniu efektywniejszych modeli optymalizacji oraz ułatwiać walidację poprawności zwracanych przez nie rezultatów na etapie implementacji.

FINANSOWANIE

Praca powstała w ramach uczelnianego grantu badawczego nr UGB/22-059/2025/WAT, finansowanego przez Wojskową Akademię Techniczną (WAT).

LITERATURA

[1] Lee Sang-Woong, Saqib Ali, Mohammad Sadegh Yousefpoor, Efat Yousefpoor, Pooia Lalbakhsh, Danial Javaheri, Amir Masoud Rahmani, Mehdi Hosseinzadeh. 2021. "An Energy-Aware and Predictive Fuzzy Logic-Based Routing Scheme in Flying Ad Hoc Networks (FANETs)". IEEE Access 9 : 129977-130005.

[2] Donadel Denis, Francesco Marchiori, Luca Pajola, Mauro Conti. 2024. "Can LLMs Understand Computer Networks? Towards a Virtual System Administrator". 2024 IEEE 49th Conference on Local Computer Networks (LCN), Normandy, France.

[3] Hasan Soran Ahmed, Marwan Aziz Mohammed, Sazan Kamal Sulaiman. 2024. "Flying Ad-Hoc Networks (FANETs): Review of Communications, Challenges, Applications, Future direction and Open Research Topics". ITM Web of Conferences 64, 01002.

[4] Chiarello Filippo, Vito Giordano, Irene Spada, Simone Barandoni, Gualtiero Fantoni. 2024. "Future applications of generative large language models: A data-driven case study on ChatGPT". Technovation 133, 103002.

[5] Anwekar Devashri, Vikas Sakalle. 2025. "Improving Energy Consumption in Q-Learning based Routing Protocol for Flying Ad-hoc Networks (FANETs)". International Journal of Scientific Research & Engineering Trends 11(2) : 2106–2114.

[6] Hang Ching-Nam, Pei-Duo Yu, Roberto Morabito, Chee-Wei Tan. 2024. "Large Language Models Meet Next-Generation Networking Technologies: A Review". Future Internet 16(10), 365.

[7] Jóźwik Tomasz. 2024. "Model optymalizacji dla energooszczędnego routingu unicast w sieciach FANET". Przegląd Telekomunikacyjny 2024(4) : 225–228.

[8] Perplexity, "Models", Data dostępu: 01.06.2025, https://docs.perplexity.ai/models/model-cards

[9] Khan Muhammad Fahad, Kok-Lim Alvin Yau, Rafidah Md Noor, Muhammad Ali Imran. 2020. "Routing Schemes in FANETs: A Survey". Sensors 20(1), 38.

[10] Pahune Saurabh, Manoj Chandrasekharan. 2023. "Several Categories of Large Language Models (LLMs): A Short Survey". International Journal for Research in Applied Science & Engineering Technology 11(7) : 615–633.

[11] Khan Inam Ullah, Ijaz Mansoor Qureshi, Muhammad Adnan Aziz, Tanweer Ahmad Cheema, Syed Bilal Hussain Shah. 2020. "Smart IoT Control-Based Nature Inspired Energy Efficient Routing Protocol for Flying Ad Hoc Network (FANET)". IEEE Access 8 : 56371–56378.

[12] Huang Jie, Kevin Chen-Chuan Chang. 2023. "Towards Reasoning in Large Language Models: A Survey". Findings of the Association for Computational Linguistics: ACL 2023 : 1049–1065.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

ANALIZA KOMPATYBILNOŚCI ELEKTROMAGNETYCZNEJ OPARTA NA MAKROMODELACH KAWAŁKAMI LINIOWYCH PO TRAJEKTORIACH

ELECTROMAGNETIC COMPATIBILITY ANALYSIS BASED ON TRAJECTORY PIECEWISE-LINEAR MACROMODELS

Michał Rewieński[1]

[1] Politechnika Gdańska, michal.rewienski@pg.edu.pl

DOI: 10.15199/59.2025.4.38

Streszczenie: W artykule przedstawiono zastosowanie algorytmu nieliniowej redukcji rzędu modelu opartego na przybliżeniu kawałkami liniowym po trajektoriach (ang. trajectory piecewise-linear – TPWL) [3], do szybkiej analizy zniekształceń harmonicznych i intermodulacyjnych dla nieliniowych systemów technologii mieszanej (ang. mixed-technology systems). Przedstawioną metodę zweryfikowano dla przykładu mikroprzełącznika sterowanego elektrostatycznie.

Abstract: This paper presents application of an algorithm for nonlinear model order reduction (MOR) based on the trajectory piecewise-linear (TPWL) approximation [3], for fast analysis of harmonic and intermodular distortions for mixed-technology nonlinear systems. The proposed approach has been verified for an example of an electrostatically actuated micromachined switch.

Słowa kluczowe: redukcja rzędu modelu, modele kawałkami liniowe, makromodele, systemy nieliniowe, kompatybilność elektromagnetyczna

Keywords: model order reduction, piecewise-linear models, macromodels, nonlinear systems, electromagnetic compatibility

1. WSTĘP

Urządzenia i systemy mikroelektromechaniczne (ang. micro-electro-mechanical systems – MEMS), w tym np. mikroprzełączniki RF MEMS dla sygnałów o częstotliwościach radiowych i mikrofalowych, znajdują szerokie zastosowanie w telekomunikacji (np. komunikacji satelitarnej), ze względu na ich pożądane właściwości, takie jak niskie zużycie energii czy niska tłumienność [5], [8]. Tworzenie nowych systemów MEMS wymaga konstruowania prototypów numerycznych, wykorzystywanych m.in. podczas weryfikacji projektu, symulacji badających integrację z innymi układami (w tym np. z układami scalonymi) oraz przy analizie kompatybilności elektromagnetycznej. Symulacje MEMS oparte na pełnych opisach fizycznych okazują się zazwyczaj bardzo wymagające, ze względu na konieczność modelowania sprzężonego (ang. multiphysics modeling), łączącego np. mechanikę klasyczną, mechanikę płynów, modele elektrostatyczne lub elektromagnetyczne oraz silnie nielinowe zachowanie urządzeń w typowym zakresie swojego działania. W szczególności, analiza zniekształceń harmonicznych jest kosztowna obliczeniowo, ze względu na konieczność wyznaczenia okresowego stanu ustalonego za pomocą symulacji w dziedzinie czasu. W celu przyspieszenia symulacji systemów silnie nieliniowych, w [3] zaproponowano metodę TPWL wyznaczania makromodeli opartych na przybliżeniach kawałkami liniowych po trajektorii układu w przestrzeni stanów oraz redukcji rzędu modelu. Jak wykazano, okazuje się ona wydajna i skuteczna w modelowaniu m.in. urządzeń MEMS, układów elektronicznych [6], czy urządzeń mikroprzepływowych [7]. Niniejszy referat proponuje konkretną strategię, wykorzystującą metodę makromodelowania TPWL, zorientowaną na wydajne wyznaczanie parametrów, takich jak zniekształcenia harmoniczne i intermodulacyjne, wykorzystywanych podczas analizy kompatybilności elektromagnetycznej.

Kolejny rozdział opisuje metodę redukcji rzędu modelu opartą na przybliżeniach kawałkami liniowych po trajektoriach dla systemów nieliniowych. Rozdział 3 prezentuje wyniki symulacji numerycznych dla przykładu mikroprzełącznika, weryfikujące proponowaną strategię wyliczania zniekształceń harmonicznych, a Rozdział 4 zawiera wnioski.

2. REDUKCJA RZĘDU MODELU TWPL

Niniejszy rozdział zawiera streszczenie metody TPWL automatycznej redukcji rzędu modelu dla systemów nieliniowych, szczegółowo opisanej w [3] oraz [4]. Na potrzeby tej prezentacji przyjęto, że analizowany system może zostać opisany jako nieliniowy układ dynamiczny w postaci:

$$\begin{aligned} \frac{dx}{dt} &= f(x) + Bu \\ y &= C^T x \end{aligned} \tag{1}$$

gdzie $x = x(t)$ jest N-wymiarowym wektorem stanu, $u = u(t)$ jest M-wymiarowym sygnałem wejściowym (w typowym przypadku $M \ll N$), $f : R^N \to R^N$ jest nieliniową funkcją o wartościach wektorowych, B jest macierzą dla wejścia o wymiarach $N \times M$, $y = y(t)$ jest K-wymiarowym sygnałem wyjściowym, a C jest macierzą dla wyjścia o wymiarach $N \times K$.

2.1 Quasi kawałkami liniowe przybliżenie nieliniowości

Punktem wyjścia metody TWPL jest zbiór linearyzacji modelu (1) dla różnych stanów układu x_0, ..., x_{s-1}:

$$\frac{dx}{dt} = f(x_i) + A_i(x - x_i) + Bu \tag{2}$$

gdzie $A_i = \left[\frac{\partial f}{\partial x}\right](x_i)$ jest macierzą Jacobiego dla f w punkcie x_i. Model quasi kawałkami liniowy określa się jako ważoną, wypukłą kombinację s powyższych układów liniowych:

$$\frac{dx}{dt} = \sum_{i=0}^{s-1} \tilde{w}_i(x) f(x_i) + \sum_{i=0}^{s-1} \tilde{w}_i(x) A_i(x - x_i) + Bu \tag{3}$$

gdzie $\tilde{w}_i(x)$ są wagami zależnymi od stanu układu x, przyjmującymi wartości nieujemne i takimi, że $\sum_{i=0}^{s-1} \tilde{w}_i(x) \equiv 1$. Wyniki eksperymentów numerycznych [3] pokazują, że wagi dane wzorem: $w_i(x) = (\exp(||C(x - x_i)||/m))^{-25} / \sum_{i=0}^{s-1}(\exp(||C(x - x_i)||/m))^{-25}$, gdzie $m = \min_i ||C(x - x_i)||$, sprawdzają się przy modelowaniu wielu silnie nieliniowych systemów, z jednej strony zapewniając gładkość trajektorii układu przechodzących pomiędzy sąsiednimi otoczeniami modeli liniowych, a z drugiej, dzięki szybkiej zmienności, gwarantując (poza bardzo ograniczonymi obszarami) dominację pojedynczego modelu liniowego w otoczeniu zbioru punktów linearyzacji w przestrzeni stanów. Uzasadnia to określenie modelu (3) jako quasi kawałkami liniowego.

2.2 Projekcja modelu

Redukcja rzędu (wymiarowości) modelu jest realizowana w metodzie TWPL przez rzutowanie N-wymiarowej przestrzeni stanów na przestrzeń q-wymiarową ($q \ll N$), określone przez bazę ortogonalną w przestrzeni zredukowanej, zadaną w macierzy V o rozmiarze $N \times q$. Stosując projekcję $x = Vz$ w równaniu (3), gdzie z jest wektorem stanu o rozmiarze q, otrzymujemy model zredukowany:

$$\frac{dz}{dt} = \sum_{i=0}^{s-1} \left[w_i(z) A_{ir} z + V^T(f(x_i) - A_i x_i)\right] + B_r u \tag{4}$$

$$y = C_r^T z$$

gdzie $A_{ir} = V^T A_i V$, $B_r = V^T B$, $C_r = V^T C$ oraz $w_i(z)$ są wagami zdefiniowanymi analogicznie jak $\tilde{w}_i(x)$, ale zależnymi od odległości $||C_r(z - z_i)||$, gdzie punkty $z_i = V^T x_i$ odpowiadają w przestrzeni zredukowanej kolejnym punktom linearyzacji.

Kluczową zaletą modelu (4) jest bardzo niski koszt obliczeniowy wyznaczania prawej strony równania, wynoszący $O(sq^2)$, ponieważ wszystkie obliczenia wykonywane są wyłącznie dla wektorów i macierzy rzędu q, a potencjalnie drogie obliczeniowo ewaluacje nieliniowych funkcji wektorowych zostały wyeliminowane przez zastosowanie linearyzacji.

2.3 Szybka generacja modelu TPWL na potrzeby analizy kompatybilności elektromagnetycznej

Konstrukcja modelu TPWL wymaga:

1. wyznaczenia bazy projekcyjnej V. W tym celu stosowane są bardzo różne metody i strategie np. metody podprzestrzeni Kryłowa (jedno- i wielopunktowe) [3], algorytm TBR [6] lub metoda POD [2]. (Przegląd dostępnych metod projekcji wykracza poza zakres tego referatu.);
2. znajomości pewnej liczby punktów linearyzacji x_i, które są wektorami stanu układu. W podejściu przedstawionym w [3] punkty te są wybierane z stanów układu leżących na trajektorii 'treningowej', odpowiadającej pewnemu, reprezentatywnemu sygnałowi wejściowemu.

Koszt obliczeniowy wyznaczenia bazy projekcyjnej może bardzo istotnie różnić się w zależności od zastosowanego algorytmu, poczynając od jednopunktowych metod podprzestrzeni Kryłowa (dla których koszt jest rzędu $O(q^2N)$, przy założeniu że macierz Jacobiego dla nieliniowej funkcji f jest rzadka), a kończąc na metodzie TBR wymagającej rozwiązania równania macierzowego Łapunowa, w której koszt wynosi $O(N^2)$ nawet dla rzadkich macierzy układu.

Ponadto, w algorytmie TPWL, konieczne jest wyznaczenie trajektorii treningowej, co z kolei w ogólności wymaga przeprowadzenia kosztownej obliczeniowo (w typowym przypadku) symulacji pierwotnego układu nieliniowego. Poniesienie wysokiego kosztu obliczeniowego symulacji 'treningowej' może być uzasadnione, o ile model zredukowany TPWL jest następnie wielokrotnie wykorzystywany i zmniejszenie całkowitego czasu potrzebnego na wykonanie całej serii symulacji kompensuje długi czas potrzebny na symulację treningową i generację modelu zredukowanego.

W przypadku wykonywania analiz kompatybilności elektromagnetycznej założenie o możliwości wielokrotnego wykorzystania makromodelu TPWL, opartego na pojedynczej trajektorii treningowej, jest w ogólności nieprawdziwe, ze względu na konieczność przeprowadzenia symulacji w różnych zakresach częstotliwości lub zastosowania pobudzeń o istotnie różniących się amplitudach. W tej sytuacji możliwe jest skonstruowanie makromodelu TWPL opartego na kilku trajektoriach treningowych (np. dla sygnałów o istotnie różnym widmie) oraz zastosowanie bazy projekcyjnej będącej agregacją baz z wielu punktów linearyzacji. Taki makromodel mógłby zostać użyty w całej serii zróżnicowanych symulacji. Wadą tego rozwiązania jest kilkukrotnie dłuższy czas generacji modelu w porównaniu do czasu generacji modelu opartego na pojedynczej trajektorii. Ponieważ czas generacji modelu jest w wielu przypadkach wielokrotnie dłuższy niż czas symulacji (także dla modeli opartych na pojedynczej trajektorii lub nawet przy wykorzystaniu szybkiej metody do wyznaczenia przybliżonej trajektorii, opisanej w paragrafie poniżej), poniesienie dużego kosztu dla generacji makromodelu może okazać się nieuzasadnione.

W niniejszej pracy proponujemy inną strategię, polegającą na minimalizacji *łącznego* czasu generacji modelu TPWL i pojedynczej symulacji z wykorzystaniem utworonego modelu zredukowanego. Wyniki eksperymentów wydajnościowych przedstawionych w [3] sugerują, że łączny czas generacji modelu i pojedynczej symulacji zostanie zminimalizowany przy zastosowaniu następującej strategii:

1. Baza projekcyjna V jest wyznaczana metodą Arnoldiego, z wykorzystaniem pojedynczego modelu liniowego, wyznaczonego dla stanu początkowego x_0 układu.

2. Kolejne punkty linearyzacji są wybierane z pojedynczej, *przybliżonej* trajektorii treningowej, wyznaczonej z pomocą szybkiej przybliżonej symulacji [3]

Metoda szybkiej przybliżonej symulacji polega na cyklicznym wykonaniu następujących kroków (dla $i = 0 \dots (s-2)$): 1) wyznaczaniu zredukowanego modelu liniowego z wykorzystaniem bazy V, na podstawie linearyzacji układu (1) w danym punkcie (x_i); 2) wykonaniu symulacji z wykorzystaniem modelu zredukowanego z punktu 1) do wyznaczenia kolejnego punktu linearyzacji (x_{i+1}).

Możliwa jest również zastosowanie 'bogatszej' bazy projekcyjnej, powstałej w wyniku agregacji baz wyznaczonych w kolejnych punktach linearyzacji. W ten sposób uzyskamy zazwyczaj model TWPL o niższym rzędzie, niż z pomocą 'jednopunktowej' bazy projekcyjnej (przy porównywalnej dokładności obu modeli). W typowym przypadku, krótszy czas symulacji TPWL (ze względu na niższy rząd modelu) nie skompensuje jednak dłuższego czasu generacji modelu TWPL (spowodowanego koniecznością generacji bazy projekcyjnej w wielu punktach) [3].

Poprzednie prace sugerują, że zastosowanie najprostszej strategii opisanej powyżej jest najwydajniejsze. Strategia ta jest również bardzo elastyczna, ponieważ generuje 'na żądanie' model zredukowany zgodny z planowaną symulacją. Wyniki zawarte w kolejnym rozdziale pokazują również, że proponowana strategia jest rzeczywiście skuteczna w analizie zniekształceń harmonicznych dla przykładu mikroprzełącznika.

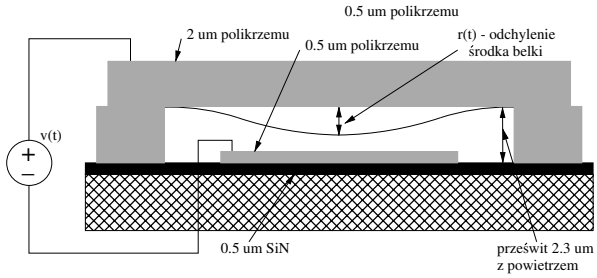
3. WYNIKI OBLICZEŃ NUMERYCZNYCH

Ten rozdział przedstawia wyniki symulacji numerycznych dla przykładu mikroprzełącznika sterowanego elektrostatycznie, pokazanego na Rys. 1. System ten może być opisany za pomocą jednowymiarowego równania belki Eulera (modelującego mechaniczne własności belki z polikrzemu) oraz dwuwymiarowego równania Reynoldsa (modelującego dynamikę cienkiej warstwy powietrza między belką a podłożem):

$$\hat{E}I\frac{\partial^4 u}{\partial x^4} - S\frac{\partial^2 u}{\partial x^2} = F_{el} + \int_0^w (p - p_a)dy - \rho\frac{\partial^2 u}{\partial t^2}, \quad (5)$$

$$\nabla \cdot ((1+6K)u^3 p\nabla p) = 12\mu\frac{\partial(pu)}{\partial t}, \quad (6)$$

gdzie I jest momentem bezwładności belki, $\hat{E}$ jest modułem Younga, S jest współczynnikiem naprężenia, ρ jest gęstością powietrza, p_a jest ciśnieniem otoczenia, μ jest lepkością powietrza, K jest liczbą Knudsena, $u = u(x,t)$ określa położenie belki nad podłożem oraz $p = p(x,y,t)$ jest rozkładem ciśnienia w warstwie powietrza pod belką. Zakładając, że elektrody są wzajemnie równoległe, siła elektrostatyczna może zostać przybliżona wzorem: $F_{el} = -\frac{\epsilon_0 w v^2}{2u^2}$, gdzie v jest przyłożonym napięciem.



Rysunek 1: Mikroprzełącznik sterowany elektrostatycznie (za [1]).

Dla powyższego układu sprzężonych równań różniczkowych cząstkowych zastosowano dyskretyzację przestrzenną (za pomocą różnic skończonych), prowadzącą do nieliniowego układu dynamicznego w postaci (1), w którym wektor stanu x ma rozmiar $N = 880$.

Następnie, dla otrzymanego pełnowymiarowego modelu nieliniowego wygenerowano model TPWL, stosując strategię opisaną w poprzednim rozdziale. Dla pojedynczej przybliżonej trajektorii treningowej układu, wyznaczonej za pomocą szybkiej symulacji, dla skokowego sygnału wejściowego o amplitudzie 9V, wybrano 33 punkty linearyzacji. Baza projekcyjna rozpinająca momenty układu liniowego dla stanu początkowego, wygenerowana z pomocą metody Arnoldiego, miała rząd $q = 41$.

W pierwszej serii testów, zastosowano pełnowymiarowy model nieliniowy oraz zredukowany model TWPL do wyznaczenia okresowego stanu ustalonego układu. W tym celu zastosowano metodę strzałów (ang. shooting method) oraz symulacje w dziedzinie czasu. Dla wyznaczonych okresowych stanów ustalonych obliczono harmoniczne. Tabela 1 porównuje pierwsze 3 harmoniczne dla obu modeli. Dobra zgodność wyników pokazuje, że model TPWL może zostać skutecznie zastosowany do analizy zniekształceń harmonicznych.

Tabela 1: Porównanie kolejnych harmonicznych dla okresowego stanu ustalonego, wyznaczonych z pomocą pełnego modelu nieliniowego rzędu $N = 880$ i makromodelu TWPL rzędu $q = 41$. Sygnał wejściowy dla układu: $u(t) = (9\cos(\pi t))^2$

.

Harmoniczna	*Model pełny*	*Model TWPL*	*Błąd* [%]
dc	1.9587	1.9525	0.4
1.	-0.1967+0.0351i	-0.1941+0.0345i	1.3
2.	-0.0289+0.0283i	-0.0271+0.0260i	7.4
3.	0.0004+0.0143i	-0.0024+0.0156i	22.1

W kolejnych testach zastosowano te same modele do wyznaczenia okresowego stanu ustalonego układu w odpowiedzi na pobudzenie będące sumą dwóch kosinusoid o różnych częstotliwościach:

$$\begin{aligned} u(t) &= \frac{1}{2}\left[(8\cos(2\pi f_1))^2 + 8\cos(2\pi f_2))^2\right] \\ &= 16\cos(2\pi(2f_1)) + 16\cos(2\pi(2f_2)) + 32 \quad (7) \end{aligned}$$

gdzie $f_1 = 2$kHz and $f_1 = 2.2$kHz.

Tabela 2: Porównanie amplitud głównych harmonicznych intermodulacyjnych okresowego stanu ustalonego dla pełnego modelu nieliniowego rzędu $N = 880$ i makromodelu TPWL rzędu $q = 41$.

Harmoniczna	*Model pełny*	*Model TWPL*	*Błąd* [%]
dc	2.09058	2.09514	-0.3
$f_2 - f_1$	0.00524	0.00663	-26.7
$2f_1 - f_2$	0.00056	0.00072	-29.8
f_1	0.05650	0.05770	-2.2
f_2	0.05585	0.05681	-1.8
$2f_2 - f_1$	0.00053	0.00056	-3.9
$2f_1$	0.00222	0.00255	-15.2

Tabela 2 porównuje amplitudy dla głównych produktów intermodulacji oraz harmonicznych sterujących wyliczone (z pomocą DFT) dla pełnowymiarowego modelu nieliniowego i modelu zredukowanego TPWL dla przykładu mikroprzełącznika. Powyższy wynik pokazuje, że modele zredukowane TPWL mogą zostać z powodzeniem zastosowane do badania zniekształceń intermodulacyjnych dla silnie nieliniowych systemów.

4. WNIOSKI

Wstępne wyniki zaprezentowane w niniejszym referacie pokazują skuteczność zaproponowanej prostej strategii symulacyjnej, polegającej na generacji 'na żądanie' makromodelu TPWL dla pierwotnego modelu nieliniowego i następującym potem wykonaniu symulacji dla układu o obniżonym wymiarze. Konieczne jest kontynuowanie prac, w celu zweryfikowania wydajności tego podejścia dla szerszego spektrum problemów w dziedzinie analizy kompatybilności elektromagnetycznej.

LITERATURA

[1] Hung E., Yang Y., Senturia S. 1997. Proceedings of the IEEE International Conference on Soild State Sensors and Actuators, vol. 2: 1101-4.

[2] Kramer Boris, Willcox Karen. 2019. „Nonlinear Model Order Reduction via Lifting Transformations and Proper Orthogonal Decomposition." AIAA Journal, 57 (6) (opublikowany online).

[3] Rewienski Michal, White Jacob. 2003. „A Trajectory Piecewise-Linear Approach to Model Order Reduction and Fast Simulation of Nonlinear Circuits and Micromachined Devices." IEEE Transactions on Computer-Aided Design, 22 (2): 155-170.

[4] Rewienski Michal, White Jacob. 2006. „Model order reduction for nonlinear dynamical systems based on trajectory piecewise-linear approximations." Linear Algebra and Its Applications, 415: 426-454.

[5] Shao B., Lu C., Xiang Y., Li F., Song M. 2024. „Comprehensive Review of RF MEMS Switches in Satellite Communications."' Sensors 2024, 24: 3135.

[6] Vasilyev Dmitry, Rewienski Michal, White Jacob. 2003. „A TBR-based trajectory piecewise-linear algorithm for generating accurate low-order models for nonlinear analog circuits and mems." Proceedings of the 40th Conference on Design Automation: 490-5.

[7] Vasilyev Dmitry, Rewienski Michal, White Jacob. 2006. „Macromodel Generation for Biomems Components Using a Stabilized Balanced Truncation Plus Trajectory Piecewise Linear Approach." w: K. Chakrabarty and J. Zheng (eds.), Design Automation Methods and Tools for Microfluidics-Based Biochips: 169-187.

[8] Wang LF., Huang QA., Han L. 2018. „RF MEMS Switch" w: Huang QA. (eds) Micro Electro Mechanical Systems. Micro/Nano Technologies. Springer, Singapore.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

SYMULACJA PROCESU REKONSTRUKCJI OBRAZU NA PODSTAWIE EMISJI UJAWNIAJĄCEJ SYGNAŁU WIDEO

SIMULATION OF THE IMAGE RECONSTRUCTION PROCESS BASED ON COMPROMISING EMANATION OF A VIDEO SIGNAL

Emilia Sakowska[1]; Leszek Nowosielski[2];

[1] Wojskowa Akademia Techniczna, Szkoła Doktorska, Warszawa, emilia.gabrielczyk@student.wat.edu.pl
[2] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, leszek.nowosielski@wat.edu.pl

DOI: 10.15199/59.2025.4.39

Streszczenie: **Artykuł przedstawia symulacje procesu przechwytywania i rekonstrukcji obrazu wideo pochodzącego od emisji ujawniającej generowanej przez urządzenia IT w warunkach laboratoryjnych. Celem jest najwierniejsze odwzorowanie zjawisk zachodzących podczas tego typu analizy z uwzględnieniem poszczególnych etapów. Symulacja umożliwia ocenę jakości obrazu i przekłada uzyskane wyniki na ocenę jakości sygnału w rzeczywistych warunkach. Zaproponowane podejście pozwala ograniczyć subiektywność oceny sygnału emisji ujawniającej przez operatora.**

Abstract: **The article presents a simulation of the process of capturing and reconstructing a video image originating from compromising emanations generated by IT devices in laboratory conditions. The aim of the work is to faithfully replicate the phenomena occurring during this type of analysis, taking into account the individual stages of the process. The simulation enables the evaluation of image quality and let results into an assessment of signal quality under real-world conditions. The proposed approach helps reduce the subjectivity of the operator's evaluation of in the laboratory.**

Słowa kluczowe: **emisja elektromagnetyczna, infiltracja elektromagnetyczna, jakość sygnału, przetwarzanie sygnałów,**

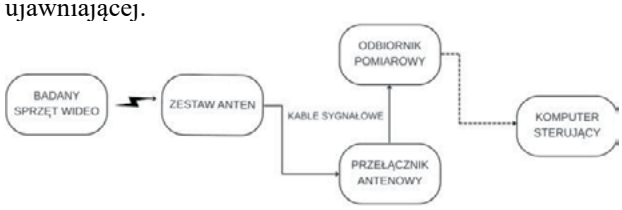
Keywords: **electromagnetic emissions, electromagnetic infiltration, image processing, signal quality**

1. WSTĘP

W literaturze technicznej termin „emisja ujawniająca" odnosi się do nieumyślnego wysyłania w postaci fal elektromagnetycznych sygnału skorelowanego z przetwarzaną informacją pochodzącego od pracujących urządzeń elektronicznych [1]. Sygnał ten po przechwyceniu i rekonstrukcji pozwala na odtworzenie skorelowanej informacji. Ataki tego typu noszą nazwę ataków typu TEMPEST. Dynamiczny rozwój technologii pozwala na łatwiejsze, szybsze oraz dokładniejsze przeprowadzenie tego typu ataków [2] [3]. W związku z tym rośnie potrzeba prowadzenia badań, które pozwolą na określenie poziomu ryzyka oraz lepsze zrozumienie mechanizmów za nimi stojących. W tym celu niezbędnym jest lepsze zrozumienie mechanizmów procesu emisji ujawniającej. Jednakże pomimo wzrostu świadomości społeczeństwa co do zabezpieczeń sprzętowych i programowych, bagatelizowane jest zagrożenie ze strony ataku typu TEMPEST. Wbrew rosnącej liczby artykułów powiązanych z bezpieczeństwem informacji czy ogólnie cyberbezpieczeństwem, zagadnienie emisji ujawniającej pozostaje wciąż bardzo słabo reprezentowane w ogólnodostępnych badaniach. Dotychczasowe prace skupiają się głównie na rodzaju zabezpieczenia jakie należy stosować np. czcionki [4], sposoby ekranowania urządzeń, a pomijają znaczenie modelowania i symulacji w warunkach laboratoryjnych. Brakuje także ujednoliconych metod oceny jakości sygnału emisji ujawniającej.

2. OPIS ŚRODOWISKA SYMULACYJNEGO

Symulacja procesu rekonstrukcji obrazu na podstawie emisji ujawniającej sygnału wideo została przeprowadzona w środowisku MATLAB. W pierwszym kroku odwzorowano rzeczywisty tor pomiarowy wykorzystywany do przeprowadzania badań w kabinie SAC. Następnie wyznaczono i zamodelowano zakłócenia pochodzące od kluczowych elementów badanego toru pomiarowego: badanego urządzenia generującego sygnał wideo (laptop, monitor), zestawu anten odbiornika, przełącznika antenowego i odbiornika pomiarowego. W celu poprawnego zamodelowania działania poszczególnych komponentów toru pomiarowego przeanalizowano ich działanie pod kątem parametrów oraz uwzględniono wpływ środowiska laboratoryjnego na propagacje sygnałów elektromagnetycznych. Takie podejście pozwoliło na zwiększenie realizmu symulacji. Schemat toru pomiarowego przedstawiono na Rys.1. Pewne elementy tj. kable sygnałowe, komputer sterujący, zostały pominięte ze względu na marginalny wpływ w ocenie końcowej odtworzonego obrazu. Uzyskany model symulacyjny stanowi podstawę do dalszych analiz i symulacji oraz oceny jakości sygnału emisji ujawniającej.



Rys. 1 Schemat toru pomiarowego emisji ujawniającej

3. ZAKŁÓCENIA I SZUMY W MODELU

Zdefiniowane zakłócenia dla poszczególnych komponentów starano się odwzorować pod względem wpływu na odbierany obraz. Skupiono się na zamodelowaniu zakłóceń elektromagnetycznych generowanych w

środowisku pracującego urządzenia IT. Poza zakłóceniami uwzględniono również istotne, z punktu widzenia odbieranego sygnału emisji ujawniającej, parametry.

3.1 Charakterystyka szumów własnych

Uwzględniając przyjęty tok postępowania, w pierwszej kolejności dodano do modelu szumy o rozkładzie normalnym (Gaussa), aby odwzorować nieunikniony wpływ szumu termicznego, generowanego przez przypadkowy ruch nośników ładunków w urządzeniach i elementach elektronicznych. Uwzględniono jego zmienny charakter oscylujący wokół wartości średniej.

3.2 Wpływ rozdzielczości obrazu

Rozdzielczość obrazu określono jako kluczowy parametr w tworzonej symulacji. Po pierwsze, rozdzielczość obrazu pozwala na dopasowanie rozdzielczości symulowanego obrazu do rozdzielczości obrazu uzyskanego w rzeczywistych warunkach, co w dalszych krokach ułatwia jego porównanie i ocenę jakości sygnału. Po drugie w warunkach rzeczywistych, wyższa rozdzielczość obrazu odpowiada za większą liczbę pikseli przetwarzanych w jednostce czasu, w związku z tym sygnały generowane są szybciej, co bezpośrednio przekłada się na szersze pasmo emisji ujawniającej.

3.3 Modelowanie anteny jako filtru

Następnie przystąpiono do symulacji działania zestawu anten odbiornika. Na podstawie rzeczywistych wyników pomiarów przyjęto, że sygnał emisji ujawniającej z największym prawdopodobieństwem występuje w paśmie (200-1000) MHz. Zakres ten pokrywa się z pasmem pracy aktywnej anteny dipolowej wykorzystywanej w pomiarach w kabinie SAC. W związku z tym skupiono się wyłącznie na symulacji działania tego typu anteny. Działanie anteny można porównać do filtru pasmowo-przepustowego, ponieważ odbiera ona sygnały w określonym paśmie częstotliwości. W uproszczeniu, na potrzeby symulacji, jej działanie można traktować jako filtr górnoprzepustowy, ponieważ z punktu widzenia detekcji emisji ujawniającej szczególne znaczenie mają wysokie częstotliwości (szybkie zmiany sygnału). Celem tej części analizy jest zbudowanie modelu filtru górnoprzepustowego, którego działanie odzwierciedlałoby efekt różniczkowania sygnału przez antenę. Poprzez różniczkowanie sygnału przez antenę należy rozumieć wykrywanie zmian pola elektromagnetycznego w czasie. Założenie to ma swoje matematyczne uzasadnienie. Wzór ogólny na pierwszą pochodną sygnału:

$$f'(t) = \lim_{\Delta t \to 0} \frac{f(t+\Delta t) - f(t)}{\Delta t} \tag{1}$$

można przedstawić wykorzystując dyskretne przybliżenie w postaci filtru górnoprzepustowego:

$$f'(t) \approx \frac{f(t) - f_{LP}(t)}{\Delta t} \tag{2}$$

,gdzie f_{LP} – to funkcja gęstości prawdopodobieństwa amplitudy sygnału po przejściu przez filtr dolnoprzepustowy. Ten sposób rozumowania przedstawia działanie analogiczne do filtru górnoprzepustowego.

3.4 Modelowanie przełącznika antenowego

Kolejnym etapem było dodanie zakłóceń pochodzących od przełącznika antenowego i przewodów pomiarowych w torze transmisyjnym zamodelowanych w postaci kwadratowego impulsu. Model ten odzwierciedla zakłócenia generowane podczas przełączania preselektorów w ramach jednej anteny oraz wpływu przewodów pomiarowych na jakość sygnału. Tego typu zakłócenia charakteryzują się krótkotrwałymi, ale intensywnymi skokami mocy o szerokim spektrum częstotliwości. W przypadku odbioru sygnału emisji ujawniającej, połączenia mechaniczne w przełącznikach antenowych oraz przewodach pomiarowych w torze transmisyjnym mogą generować produkty intermodulacyjne (zjawisko pasywnej intermodulacji PIM), co skutkuje zniekształceniami odbieranego sygnału.

3.5 Modelowanie odbiornika pomiarowego

Ostatnim zasymulowanych elementem jest odbiornik antenowy. W jego modelu uwzględniono aspekty, takie jak: korekcja gamma, filtracja dolnozaporowa, kwantyzacja oraz szum Gaussowski. Zastosowanie korekcji gamma, czyli nieliniowego przekształcenia sygnału, ma na celu zasymulowanie dostosowania jasności i kontrastu odbieranego sygnału do charakterystyki urządzenia wyświetlającego odebrany obraz w warunkach rzeczywistych. Następnie zastosowano filtr dolnoprzepustowy w celu eliminacji zakłóceń pochodzących z sieci zasilającej, które występują w rzeczywistych warunkach. Kolejnym krokiem jest dodanie kwantyzacji, która symuluje proces przekształcenia sygnału analogowego na sygnał cyfrowy, odzwierciedlając działania przetworników analogowo-cyfrowych. Na koniec dodano szumy Gaussowskie, aby uwzględnić szumy termiczne i inne losowe fluktuacje występujące w rzeczywistych systemach elektronicznych, analogicznie jak na początku budowanego układu symulacyjnego.

4 METODY ANALIZY DANYCH

Na każdym etapie przetwarzania sygnału możliwe jest przeprowadzenie podglądu oraz analizy zrastrowanego obrazu, który porównywano z oryginałem. Pozwalało to na bieżącą kontrolę i lepsze zrozumienie wpływu poszczególnych elementów toru pomiarowego na jakość zrastrowanego sygnału wideo. W ramach analizy zastosowano zarówno klasyczne metryki jakościowe, takie jak SNR i PSNR, jak również bardziej zaawansowane, w tym transformację falkową, współczynnik korelacji oraz wskaźnik SSIM. Z uwagi na to, że jest to symulacja toru pomiarowego emisji ujawniającej, ze względu na wybór sygnału wideo jako danych wejściowych, główny nacisk położono na analizę właściwości obrazu.

4.1 Klasyczne metryki oceny jakości sygnału

Jako klasyczne metryki oceny jakości sygnału zastosowano SNR (Signal-to-Noise Ratio), wyrażony wzorem:

$$SNR_{dB} = 10 \cdot log_{10}\left(\frac{P_{sygnał}}{P_{szum}}\right) \tag{3}$$

oraz jego rozszerzone warianty: RMS SNR (Root Mean Square Signal-to-Noise Ratio) i PSNR (Peak Signal-to-Noise Ratio). RMS SNR, w odróżnieniu do klasycznej postaci SNR opartej na całkowitej mocy sygnału i szumu, opiera się na wartościach skutecznych. W ten sposób jest bardziej odporny na wartości odstające w porównaniu do klasycznej metody, i lepiej oddaje średni poziom sygnału. Ma to szczególne znaczenie w określaniu jakości sygnału emisji promieniowanej, w której obraz zrastrowany zazwyczaj jest nierównomiernie zaszumiony. PSNR jest

często wykorzystywaną metryką do określenia jakości obrazów. Porównuje maksymalną możliwą wartość piksela w obrazie z poziomem szumu mierzonego jako błąd średniokwadratowy. Im wyższa wartość PSNR, tym lepsza zgodność z oryginałem.

4.2 Zaawansowane metryki jakości sygnału

Jako zaawansowane metryki oceny jakości sygnału zastosowano metody uwzględniające analizę struktury obrazu oraz percepcyjne kryteria jakości. W celu uzyskania bardziej kompleksowego obrazu wpływu zakłóceń wykorzystano:

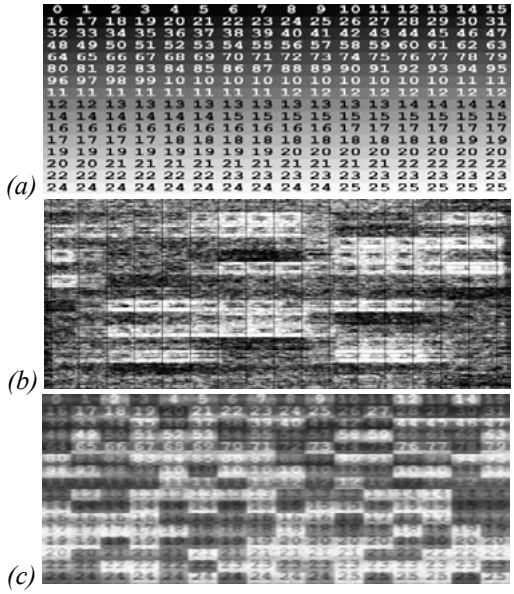
- transformację falkową,
- współczynnik korelacji,
- SSIM (Structural Similarity Index Measure),
- MSSIM (Mean SSIM),
- UQI (Universal Quality Index).

Przeprowadzono dwupoziomową dekompozycje falkową 2D z użyciem falki Daubechies 1 (db1), koncentrując się na składowych niskoczęstotliwościowych, które zawierają najwięcej informacji o obrazie. Na ich podstawie wyznaczono tzw. falkowy współczynnik SNR, który odzwierciedla stosunek sygnału do szumu w przestrzeni reprezentacji falkowej. Współczynnik korelacji służy do liniowej oceny zależności pomiędzy pikselami obrazu oryginalnego a zrastrowanego. Może być to szczególnie istotne w momencie kiedy treść zrastrowanego obrazu wydaje się zniekształcona, a jednak informacja jest silnie skorelowana z informacją użyteczną i możliwą do odszyfrowania. SSIM jest metryką jakości uwzględniająca, lokalne zmiany jasności, kontrastu i struktury, a jej poziom ma wpływ na ocenę percepcyjną przez operatora sprzętu. MSSIM, w porównaniu do SSIM ocenia jakość całego obrazu i częściowo kompensuje wpływ lokalnych anomalii. Z kolei UQI to alternatywny wskaźnik jakości obrazu, który uwzględnia korelacje, kontrast i jasność.

5 WYNIKI

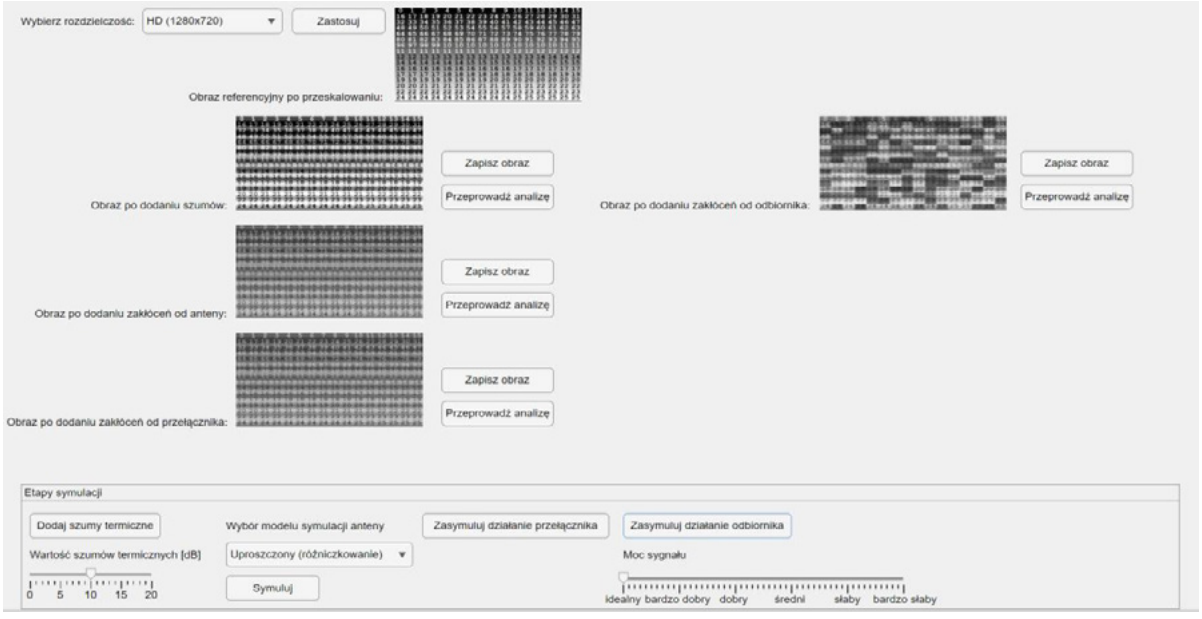
Jak wcześniej wspominano w aplikacji do symulacji, uwzględniono możliwość wyboru rozdzielczości obrazu, dodano możliwość zdefiniowania wartości szumów termicznych, zakłócenia po przejściu przez zestaw anten, przełącznik antenowy oraz odbiornik pomiarowy (Rys.2).

Ponadto na sam koniec przy odbiorniku pomiarowym zdefiniowano funkcję określenia mocy sygnału zrastrowanego w skali od „sygnał idealny" do „sygnał bardzo słaby", co zwiększało możliwość oceny wybranych parametrów na jakość sygnału odbieranego oraz dawało większą elastyczność do wykorzystania wyników symulacji.
Na Rys.3. widać obraz oryginalny - mapa luminescencyjna (a), przykładowy obraz po zrastrowaniu (b) oraz przykładowy wynik symulacji (c). Wyniki symulacji wskazują na duży potencjał uzyskanych rezultatów. Dzięki możliwości podglądu obrazu pod wpływem każdego z komponentów toru pomiarowego, możliwa była bieżąca kontrola oraz precyzyjne dostosowanie wartości parametrów dla poszczególnych elementów.



Rys. 3 Obraz oryginalny (a), obraz zrastrowany (b), wynik symulacji (c)

W ramach pracy skupiono się także na jakościowej analizie obrazu. Wyniki symulacji umożliwiły analizę parametrów dla różnych scenariuszy.

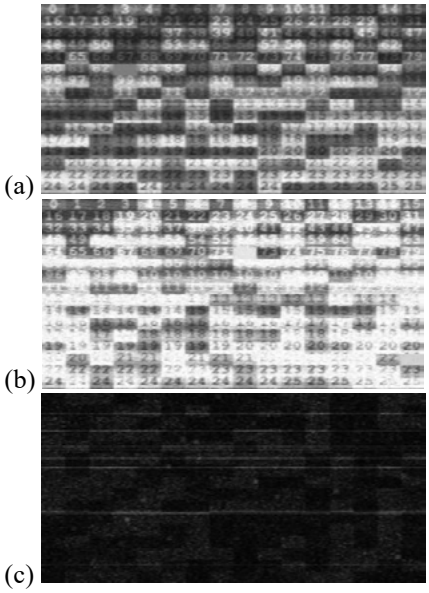


W Tab. 1 przedstawiono porównanie wartości wybranych metryk przy stałym poziomie szumów termicznych oraz zmiennej mocy sygnału – idealny, dobry, słaby (Rys.4.).

Tab. 1. Wyniki pomiarów przy zmiennej mocy sygnału

Parametr	*Idealny*	*Dobry*	*Słaby*	*Jednostka*
SNR	2,75	2,16	0,32	dB
SNR RMS	7,71	6,58	0,98	dB
PSNR	27,36	26,83	24,93	dB
SNR falkowy	6,31	2,51	1,07	dB
SSIM	0,23	0,18	0,08	-
MSSIM	0,36	0,32	0,08	-
UQI	0,34	0,29	0,00	-
Korelacja	0,40	0,39	0,03	-

Wszystkie przedstawione wartości parametrów maleją wraz ze spadkiem mocy sygnału, co jest zgodne z oczekiwaniami. Znaczne rozbieżności między wartościami SNR obliczonego w sposób klasyczny, a uzyskanymi przy wykorzystaniu transformacji falkowej, podkreślają istotność obu parametrów w ostatecznej ocenie wyniku symulacji. Bardzo niskie wartości parametrów SSIM i MSSIM, wskazujące na małe podobieństwo strukturalne miedzy obrazem referencyjnym, a końcowym wynikiem symulacji, sugerują, że nie są to odpowiednie parametry do oceny zrastrowanych obrazów w rzeczywistych warunkach, które potrafią być silniej zaszumione i rozmyte. Parametr UQI, choć teoretycznie świadczący o pogarszającej się jakości sygnału, także nie jest odpowiedni do oceny tego typu danych. Dla sygnału o słabej mocy jego wartość wyniosła 0, mimo że obraz wciąż zawierał elementy istotne z punktu widzenia zagrożeń infiltracji elektromagnetycznej. Pomimo niskich wartości współczynnika korelacji, wskazujących na degradacje jakości przetwarzanego obrazu, wskazuje on duży potencjał jako odpowiedni parametr przy dalszej analizie. Kluczowym będzie tutaj opracowanie odpowiedniej skali interpretacyjnej, która odpowiadałaby percepcyjnym możliwościom oceny przez człowieka.



Rys. 4. Obraz o mocy sygnału: idealnej (a), dobrej (b), słabej (c)

6 WNIOSKI

Przeprowadzone symulacje potwierdziły, że opracowane środowisko pozwala na kompleksową analizę wpływu poszczególnych komponentów toru pomiarowego na jakość odbieranego obrazu. Ponadto, środowisko daje możliwość do stworzenia bazy obrazów wzorcowych, co jest niezbędne do dalszej automatyzacji procesu badań emisji ujawniającej i wykrywania podobieństwa między zrastrowanych obrazem, a obrazem referencyjnym. Dzięki możliwości obserwacji efektów działania każdego elementu oraz regulacji parametrów, system ten stanowi efektywne narzędzie do weryfikacji i optymalizacji konfiguracji toru pomiarowego. Wszystkie zastosowane metryki jakości obrazu poprawnie wskazywały spadki wartości wraz z obniżeniem mocy sygnału. Jednakże niskie wartości SSIM, MSSIM i UQI w warunkach symulacyjnych, wykluczają ich zastosowanie w warunkach rzeczywistych charakteryzujących się wysokim poziomem szumów i rozmycia. Pomimo niskich wartości współczynnika korelacji w przypadku słabego sygnału, wskaźnik ten wykazuje duży potencjał jako parametr do dalszej analizy jakości obrazu. Wyniki badań wskazują na potrzebę dalszego rozwoju środowiska symulacyjnego oraz wykorzystania go także w bardziej rozbudowanej wersji. Ponadto widać potrzebę znalezienia nowych i skuteczniejszych metryk jakości obrazu, które byłyby bardziej odporne na zakłócenia charakterystyczne dla rzeczywistych warunków pomiarowych lub opracowania nowych wskaźników percepcyjnych mogących przyczynić się do lepszej oceny jakości obrazów w kontekście zagrożeń związanych z infiltracją elektromagnetyczną.

Praca została sfinansowana przez Wojskową Akademię Techniczną w ramach projektu nr UGB/22-055/2025/WAT pt. ”Badanie odporności wybranych smartfonów na promieniowane pole elektromagnetyczne w aspekcie wymagań normy PN-EN 17025”

LITERATURA

[1] National Security Agency (NSA). 1972. „TEMPEST Fundamentals”. NACSIM 5000 (odtajnione).

[2] Bulut, Fatih, Veyis Solak i Hasan Seçkin Efendioğlu. 2024. „A Novel Approach to Measure Shielding Effectiveness in TEMPEST-Protected Buildings”. 2024 International Symposium on Electromagnetic Compatibility – EMC Europe, Brugia, Belgia, s. 279–283.

[3] Fernández, Santiago, Emilio Martínez, Gabriel Varela, Pablo Musé i Federico Larroca. 2024. „DeepTEMPEST: Using Deep Learning to Eavesdrop on HDMI from its Unintended Electromagnetic Emanations”. Proceedings of the 13th Latin-American Symposium on Dependable and Secure Computing (LADC '24).

[4] Kubiak, Ireneusz, Alexandru Boitan i Simona Halunga. 2020. „Assessing the Security of TEMPEST Fonts against Electromagnetic Eavesdropping by Using Different Specialized Receivers”. Applied Sciences 10(8): 2828.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

WYTWARZANIE ZNIEKSZTAŁCEŃ MULTIPLIKATYWNYCH ANTENĄ OBCIĄŻONĄ ZMIENNĄ IMPEDANCJĄ
MULTIPLICATIVE INTERFERENCES GENERATION USING ANTENNA WITH VARIABLE LOAD

Olga Błaszkiewicz; Alicja Olejniczak; Jarosław Sadowski, Jacek Stefański

Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki
email: {olga.blaszkiewicz, alicja.olejniczak, jarsadow1, jacstefa1}@pg.edu.pl

DOI: 10.15199/59.2025.4.40

Streszczenie: **W referacie przedstawiono wynik eksperymentu polegającego na umieszczeniu w pobliżu łącza radiowego dodatkowej anteny podłączonej do zmiennego obciążenia. Rezultatem było wystąpienia konwersji częstotliwości sygnału analizowanego łącza. Referat zawiera także komentarz odnośnie przełożenia wyników tego eksperymentu na potencjalne problemy wynikające ze stosowania w radiokomunikacji rekonfigurowalnych powierzchni odbijających.**

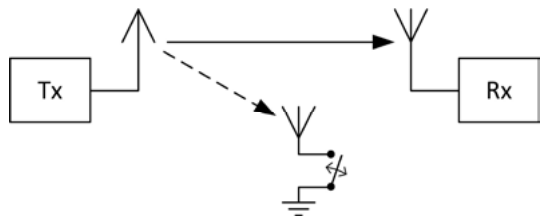
Abstract: **The paper presents the results of an experiment consisting in placing an additional antenna connected to a variable load near a radio link. The result was the occurrence of a frequency conversion of the analyzed link signal. The paper also contains a commentary on the impact of the results of this experiment on potential problems resulting from the use of reconfigurable reflective surfaces in radio communication.**

Słowa kluczowe: **interferencje, przemiana częstotliwości, rekonfigurowalne powierzchnie odbijające**
Keywords: **frequency conversion, radio interferences, reconfigurable intelligent surfaces**

1. WSTĘP

Urządzenia radiowe nadawcze, a nawet urządzenia nieprzeznaczone do emisji i odbioru sygnałów radiowych przechodzą obowiązkowe badania kompatybilności elektromagnetycznej. Ich celem jest zapewnienie, że wprowadzane do obrotu i oddawane do użytku urządzenia nie wytwarzają nadmiernych zaburzeń elektromagnetycznych oraz że mogą pracować prawidłowo w obecności zaburzeń elektromagnetycznych o określonych poziomach [7]. Prowadzenie badań kompatybilności elektromagnetycznej jest szczególnie istotne w kontekście możliwego zakłócania pracy odbiorników radiowych, które są najpowszechniej spotykanymi urządzeniami elektronicznymi o dużej wrażliwości na nawet niewielkie poziomy sygnałów zakłóceń przewodzonych czy promieniowanych. Badania nie muszą być wykonywane w przypadku urządzeń niezdolnych do wytwarzania zaburzeń elektromagnetycznych oraz urządzeń odbiorczych. Oczywistym jest, że urządzenia radiowe takie jak nadajniki, zbudowane z nieliniowych elementów aktywnych (tranzystory), mogą wytwarzać na wyjściach antenowych sygnały o częstotliwościach niepożądanych, wpływających negatywnie na pracę innych urządzeń radiowych. Mogą one także wytwarzać niepożądaną konwersję częstotliwości sygnału własnego z sygnałami obcymi, docierającymi do nich przez port antenowy (intermodulacje). Zatem testy kompatybilnościowe aktywnych urządzeń nadawczych są bezsprzecznie uzasadnione [8].

Rozważmy proste łącze radiowe nadajnik-odbiornik, w pobliżu którego znajduje się antena (może to być także w ogólności przewodnik, niekoniecznie pierwotnie zaprojektowany i przeznaczony do pracy jako antena) niebędąca częścią tego łącza (rys. 1.). Emisja sygnału radiowego z nadajnika powoduje indukowanie w tej antenie prądów. Antena jest podłączona do idealnego, bezstratnego przełącznika, powodującego jej czasowe zwieranie. W myśl przepisów, anteny, które nie są zintegrowane z urządzeniami nadawczymi, nie wymagają wykonywania badań kompatybilnościowych [8]. Przełącznik nie wytwarza energii elektrycznej i nie dostarcza energii do anteny, więc antena z przełącznikiem nie może być uznana za część urządzenia radiowego nadawczego. Zatem także tak obciążona antena nie jest objęta obowiązkowymi badaniami zgodności.



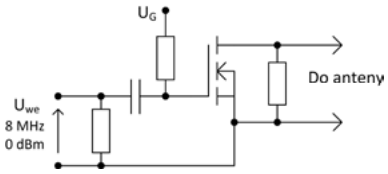
Rys. 1. Uproszczony schemat blokowy analizowanego łącza radiowego.

Przy rozwartym przełączniku antena jest nieobciążona i cała indukowana w niej energia jest zwrotnie wypromieniowywana w postaci wtórnie generowanej fali elektromagnetycznej. Niemal identyczna sytuacja występuje przy zwarciu przełącznika, tylko zmianie ulega faza odbicia sygnału elektrycznego od przełącznika, skutkując zmianą fazy emitowanej wtórnie fali elektromagnetycznej. Zatem nie jest to instalacja radiowa nadawcza. Antena jest elementem liniowym, przełącznik także spełnia zasadę liniowości. Nie ma żadnych podstaw do wymagania wykonania badań kompatybilności elektromagnetycznej tego układu. Rodzi się więc pytanie

czy przełączanie przełącznika wpłynie na funkcjonowanie łącza radiowego?

2. EKSPERYMENT

Przełącznik do zmiany impedancji obciążenia anteny zrealizowano z użyciem tranzystora MOSFET małej mocy, zgodnie ze schematem przedstawionym na rys. 2.

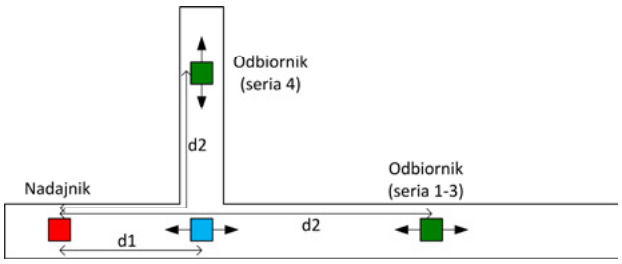


Rys. 2. Schemat przełącznika do przełączania obciążenia anteny.

Ponieważ obwód drenu nie ma zasilania, nie jest to wzmacniacz a kanał dren-źródło stanowi rezystancję sterowaną potencjałem bramki, o dobrej liniowości dla sygnałów o małych amplitudach – można go traktować w praktyce jak liniowy rezystor o zmiennej rezystancji, więc nie jest to obwód stacjonarny. Układ ten nie może być traktowany jako nadajnik radiowy i z tego powodu nie wymaga obowiązkowych badań kompatybilności elektromagnetycznej. Napięcie U_G to stałe napięcie polaryzacji bramki dobrane dla uzyskania dużej proporcji rezystancji tranzystora wyłączonego do włączonego, do którego dodawana jest składowa zmienna. W wykonanych eksperymentach sygnałem sterującym był sygnał z generatora niemodulowanej nośnej o częstotliwości 8 MHz. Tranzystor w stanie włączonym miał rezystancję ok. 6 Ω, więc zamiast pełnego odbicia sygnału z anteny stanowił obciążenie stratne, niedopasowane impedancyjnie. W stanie wyłączenia tranzystora rezystancja układu była ograniczona dodatkowym rezystorem do ok. 10 kΩ. Niewielka pojemność bramka-kanał tranzystora (ok. 8 pF) powodowała przenikanie sygnału z bramki do drenu z tłumieniem przekraczającym 30 dB, które nie miało wpływu na przebieg eksperymentu.

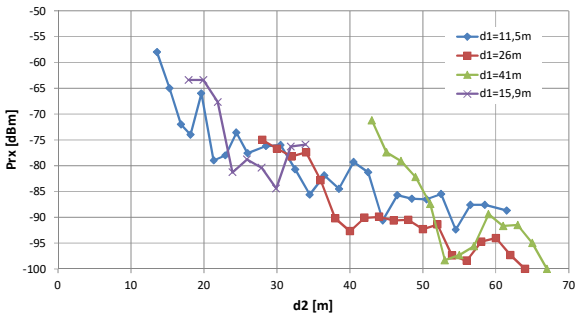
Układ przełącznika był podłączony do anteny krótkim przewodem koncentrycznym o długości ok. 5 cm, ale nieznana budowa wewnętrzna zastosowanej anteny OA2-0.3-10V, w tym w szczególności nieznana długość wewnętrznej linii zasilającej między złączem koncentrycznym N a punktem zasilania właściwego promiennika anteny powodują, że z całą pewnością wystąpił efekt transformacji zmiennej impedancji tranzystora MOSFET na również zmienną, ale nieznaną w takich warunkach impedancję obciążenia struktury anteny. Wpływ tej transformacji impedancji na przebieg eksperymentu może wymagać dalszych badań.

W trakcie eksperymentów stanowisko nadawcze modelowego łącza radiowego wytwarzało niemodulowaną nośną o częstotliwości 438,025 MHz o mocy 10 W. Sygnał był nadawany przez antenę dookólną OA2-03-10V zainstalowaną na korytarzu budynku Wydziału Elektroniki, Telekomunikacji i Informatyki Politechniki Gdańskiej. Stanowisko odbiorcze składało się z takiej samej anteny podłączonej do analizatora widma pracującego jako selektywny miernik mocy sygnału odbieranego. Pomiar poziomu mocy odbieranego sygnału wykonywano na częstotliwości 430,025 MHz. Pomiędzy nadajnikiem i odbiornikiem umieszczono antenę ze zmiennym obciążeniem, zgodnie ze schematem pokazanym na rys. 3.



Rys. 3. Szkic rozmieszczenia elementów łącza radiowego w pierwszym scenariuszu pomiarowym.

Odległość d1 od nadajnika do anteny z przełączanym obciążeniem oraz odległość d2 od nadajnika do odbiornika były zmieniane w kilku seriach pomiarowych. Wyniki pomiarów poziomu mocy sygnału na częstotliwości będącej różnicą częstotliwości sygnału nadajnika i częstotliwości przełączania obciążenia dodatkowej anteny przedstawiono na wykresie na rys. 4. Trzy serie pomiarowe (d1=11,5m; 26m i 41m) zrealizowano przy umieszczeniu wszystkich trzech anten w linii prostej, w czwartym przypadku (d1=15,9m) antena z przełączanym obciążeniem była umieszczona na łączeniu dwóch korytarzy pod kątem 90°, nadajnik umieszczono w jednym korytarzu a stanowisko odbiorcze przemieszczano wzdłuż drugiego korytarza.



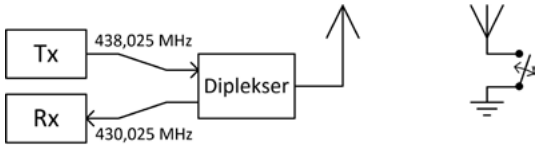
Rys. 4. Wyniki pomiaru poziomu mocy sygnału na częstotliwości 430,025 MHz w funkcji odległości punktu odbioru od anteny nadawczej dla kilku pozycji dodatkowej anteny.

W celu weryfikacji mechanizmu generowania przemiany częstotliwości wykonano następujące testy:

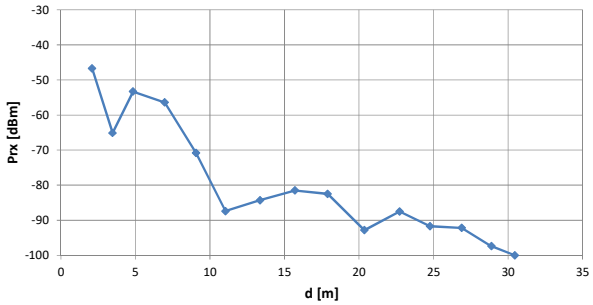
- zamiast sterowania bramką tranzystora MOSFET, sygnał 8 MHz doprowadzono przez tłumik 30 dB i sumator do drenu tego tranzystora, symulując przesłuch sygnału przez pojemność bramka-kanał. Nie zaobserwowano żadnych widocznych efektów przemiany częstotliwości na nieliniowości pasożytniczej diody między źródłem a drenem w strukturze tranzystora, więc za efekt przemiany częstotliwości odpowiada sterowanie zmienną rezystancją kanału;
- sygnał 8 MHz o pełnym poziomie mocy 0 dBm doprowadzono bezpośrednio do anteny z pominięciem tranzystora. Częstotliwość 8 MHz leży znacznie poniżej dolnej częstotliwości granicznej pracy anteny

(300 MHz) a sama antena nie ma podanej w specyfikacji charakterystyki pracy na tak niskiej częstotliwości. Gdyby w stanowisku pomiarowym występowały warunki do emisji sygnału o częstotliwości 8 MHz to pominięcie układu tranzystora zwiększyłoby poziom emisji tego sygnału o ponad 30 dB. Nie zaobserwowano jednak w takim przypadku żadnych efektów przemiany częstotliwości w badanym łączu radiowym, więc za przemianę częstotliwości nie może odpowiadać nieliniowość układów wyjściowych nadajnika lub układów wejściowych analizatora widma.

W drugim scenariuszu analizator widma i nadajnik zostały podłączone do wspólnej anteny nadawczo-odbiorczej przez diplekser zapewniający ponad 80 dB separacji pomiędzy torem nadawczym, pracującym nadal na częstotliwości 438,025 MHz i odbiorczym, pracującym na częstotliwości 430,025 MHz. Taka konfiguracja pozwala zamodelować pracę stacji bazowej lub terminala użytkownika w systemie z dupleksem częstotliwościowym (FDD, ang. Frequency Division Duplex), a wybrane w tym eksperymencie wartości częstotliwości wynikają z parametrów dostępnego dipleksera i uwarunkowań prawnych emisji sygnałów w paśmie amatorskim. Antena z przełączaną impedancją była umieszczana w różnej odległości od anteny stanowiska nadawczo-odbiorczego, a wyniki pomiaru poziomu mocy sygnału po konwersji częstotliwości przedstawia wykres na rys. 6.



Rys. 5. Schemat stanowiska pomiarowego do pomiarów wpływu anteny z przełączanym obciążeniem na pracę urządzenia z dupleksem częstotliwościowym.



Rys. 6. Wyniki pomiaru poziomu mocy sygnału na częstotliwości 430,025 MHz w stanowisku emulującym pracę urządzenia z dupleksem częstotliwościowym.

Podczas eksperymentu obserwowano tylko konwersję częstotliwości sygnału z jednego łącza radiowego. Ale ponieważ zastosowane anteny są bardzo szerokopasmowe, układ przełącznika impedancji obciążenia anteny z całą pewnością spowodował jednocześnie generowanie podobnego efektu konwersji częstotliwości, czyli zakłóceń multiplikatywnych, na nieznanej liczbie sygnałów innych łączy i sieci radiowych aktywnych w pobliżu miejsca prowadzenia eksperymentu. Zmiana sposobu sterowania przełącznikiem obciążenia anteny może pozwolić na zrealizowanie dodatkowej modulacji sygnału radiowego, wtórnie emitowanego przez tę dodatkową antenę, ale z uwagi na brak selektywności częstotliwościowej oprócz pożądanej modyfikacji sygnału właściwego łącza radiowego może wystąpić niekontrolowany wpływ na sygnały innych sieci radiowych. W opisanym eksperymencie nie projektowano w żaden sposób impedancji obciążenia anteny, gdyż impedancja przełączanego tranzystora polowego podlegała konwersji w linii koncentrycznej i w wewnętrznej linii zasilającej w użytej antenie dookólnej. Zarówno impedancja obciążenia widoczna bezpośrednio na zaciskach promiennika anteny, jak i kształt sygnału sterującego tranzystorem kluczującym impedancję obciążenia anteny mogą mieć wpływ nie tylko na sprawność konwersji częstotliwości, ale na widmo wynikowego sygnału wypromieniowanego wtórnie przez antenę z kluczowanym obciążeniem. Wszystko to stanowi szerokie pole do dalszych badań w tym zakresie.

3. REKONFIGUROWALNE POWIERZCHNIE ODBIJAJĄCE

Przedstawiony w rozdziale 2 eksperyment pokazał możliwość wygenerowania konwersji częstotliwości sygnału radiowego przez proste sterowanie impedancją obciążenia jednej anteny, umieszczonej w pobliżu łącza radiowego. Ale ten sam mechanizm zmiany impedancji obciążenia promienników jest stosowany w inteligentnych (rekonfigurowalnych) powierzchniach odbijających RIS (ang. Reconfigurable Intelligent Surfaces) do modyfikacji parametrów tych powierzchni, w tym w szczególności do zmiany kąta odbicia lub transmisji sygnału elektromagnetycznego przez te powierzchnie. Rekonfigurowalne powierzchnie odbijające to powierzchnie z metamateriałów, których właściwości interakcji z sygnałami radiowymi mogą być sterowane [2]. Sterowanie to obejmuje najczęściej możliwość zmiany kąta odbicia sygnałów radiowych oraz zmiany kierunku propagacji sygnałów przy transmisji przez RIS [1][4]. Oba te zjawiska mogą pomóc w poprawie funkcjonowania sieci radiowych w miejscach, w których bezpośrednia propagacja sygnałów ze stacji bazowych jest utrudniona, np. w kanionach ulic w środowiskach miejskich [3][6]. Dodatkowo rozważa się możliwość modulowania parametrów odbijanych sygnałów radiowych przez zmianę właściwości RIS w czasie.

W literaturze można znaleźć wiele przykładów wskazujących na zagrożenie wynikające z niskiej selektywności częstotliwościowej RIS, która powoduje, że powierzchnia odbijająca sterowana przez jednego operatora tak, by poprawić jakość komunikacji w jednej sieci radiowej, może odbijać lub uginać sygnał innej sieci radiowej na tej samej lub innej częstotliwości, w sposób będący poza kontrolą operatora tej sieci [5]. Przedstawiony eksperyment wskazuje, że odpowiednio sterowany RIS może dodatkowo wykonywać konwersję częstotliwości sygnałów radiowych i o ile dany operator obsługujący daną powierzchnię może kontrolować do pewnego stopnia parametry konwersji częstotliwości, to sama powierzchnia odbijająca będzie w stanie wywołać niepo-

żądaną przemianę częstotliwości sygnałów także innych sieci na innych częstotliwościach.

Warto wskazać, że nawet jeśli potraktuje się RIS jako bezstratną powierzchnię odbijającą, która w żaden sposób nie zmienia energii padającego sygnału (a tylko ją odbija w różnych, sterowanych kierunkach), to z punktu widzenia nieruchomego obserwatora, zmienny kierunek promieniowania sygnału odbitego od RIS będzie równoważny modulacji amplitudy i/lub fazy odbieranego sygnału, a więc wystąpią techniczne możliwości wytworzenia zakłóceń o charakterze multiplikatywnym. W obecnym stanie rozwoju techniki zmiana parametrów odbić/transmisji sygnałów przez RIS jest wykonywana zwykle stosunkowo wolno, z tempem odpowiadającym całym ramkom nadawanych sygnałów, ale skracanie czasów emisji sygnałów w usługach o niskich opóźnieniach URLLC (ang. Ultra Reliable Low Latency Communications) do pojedynczych symboli i jednoczesne skracanie czasów trwania symboli przy poszerzaniu pasma emisji sygnałów w sieci 5G powodują, że tempo sterowania RIS może w przyszłości być znacznie wyższe co potencjalnie zwiększy ryzyko wystąpienia obserwowalnych zakłóceń w postaci pasożytniczej konwersji częstotliwości sygnałów.

4. PODSUMOWANIE

Antena obciążona zmienną impedancją, umieszczona w pobliżu łącza radiowego, nawet jeśli ma znikomy wpływ na funkcjonowanie tego łącza, może wytworzyć niepożądaną, niekontrolowaną konwersję częstotliwości sygnałów radiowych generując interferencje w pracy odbiorników na innych częstotliwościach. Brak jakiejkolwiek filtracji i zwykle bardzo szerokopasmowa praca anten powoduje, że takiej konwersji częstotliwości mogą zostać poddane sygnały z szerokiego zakresu częstotliwości, także nie powiązane z częstotliwością badanego łącza radiowego czy częstotliwością sygnałów sterowania zmienną impedancją obciążenia tej dodatkowej anteny. Mając na uwadze, że zmiana impedancji obciążenia elementów antenowych jest podstawą działania inteligentnych powierzchni odbijających, możliwość wytworzenia pasożytniczej konwersji częstotliwości sygnałów sieci obcych jest potencjalnym czynnikiem ryzyka stosowania RIS w radiokomunikacji. Z uwagi na planowane użycie RIS przede wszystkim w pasmach FR2 (24 GHz i powyżej), wskazana jest kontynuacja badań na wyższych częstotliwościach i w miarę możliwości z użyciem rzeczywistych powierzchni RIS, z których znaczna część jest projektowana pod kątem zmiany kierunku odbicia sygnałów radiowych a nie zmiany fazy odbijanych sygnałów.

LITERATURA

[1] Deng M. et al. 2024. „Reconfigurable Intelligent Surfaces Enabled Vehicular Communications: A Comprehensive Survey of Recent Advances and Future Challenges", IEEE Transactions on Intelligent Vehicles, doi: 10.1109/TIV.2024.3476934.

[2] Hong, Ic-Pyo. 2023. "Reviews Based on the Reconfigurable Intelligent Surface Technical Issues" Electronics 12, no. 21: 4489. https://doi.org/10.3390/electronics12214489

[3] Liang Renjie, Liang Junjie. 2025. „A Survey of Multiple Reconfigurable Intelligent Surface Aided Systems", IEEE Access, doi: 10.1109/ACCESS.2025.3547562

[4] Liu Y. et al. 2021. „Reconfigurable Intelligent Surfaces: Principles and Opportunities", IEEE Communications Surveys & Tutorials, vol. 23, no. 3, pp. 1546-1577, doi: 10.1109/COMST.2021.3077737.

[5] Miridakis N. I., T. A. Tsiftsis, P. A. Karkazis, H. C. Leligou, P. Popovski. 2024. „Impact of Inter-Operator Interference via Reconfigurable Intelligent Surfaces", IEEE Wireless Communications Letters, vol. 13, no. 9, pp. 2536-2540, doi: 10.1109/LWC.2024.3425728.

[6] Syed, Moazzam Shah Bukhari, Hafiz Muhammad Attaullah, Sundus Ali, and Muhammad Imran Aslam. 2023. „Wireless Communications beyond Antennas: The Role of Reconfigurable Intelligent Surfaces", Engineering Proceedings 32, no. 1: 10. https://doi.org/10.3390/engproc2023032010

[7] Ustawa z dnia 13.04.2007 r. o kompatybilności elektromagnetycznej, Dz. U. 2007 nr 82 poz. 556 z późniejszymi zmianami

[8] Ustawa z dnia 12.07.2024 r. Prawo komunikacji elektronicznej, Dz. U. 2024 poz. 1221

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

A COMPARISON OF QAOA VS. VQE ALGORITHMS

PORÓWNANIE ALGORYTMÓW QAOA i VQE

Mikita Vostry[1]; Piotr Chołda[2]

AGH Akademia Górniczo-Hutnicza, Instytut Telekomunikacji, al. Mickiewicza 30, 30-059 Kraków
[1] mikvos@student.agh.edu.pl, [2] piotr.cholda@agh.edu.pl

DOI: 10.15199/59.2025.4.41

Abstract: This work presents a comparative study of the QAOA and VQE algorithms that have important cryptanalityc potential in the context of lattice-based system. Both algorithms are applied to Quadratic Unconstrained Binary Optimization (QUBO) problems within the `Qrisp` quantum framework. While QAOA-based QUBO optimization is natively supported in `Qrisp`, this work extends the framework by implementing a custom VQE-based approach. The paper explains the underlying algorithmic principles, describes their implementation, and compares the performance of both methods.

Streszczenie: W pracy przedstawiono porównanie dwóch istotnych algorytmów o zastosowaniach kryptoanalitycznych w odniesieniu do systemów kratowych: QAOA i VQE. Oba algorytmy zostały zastosowane do problemów binarnej kwadratowej optymalizacji bez ograniczeń (QUBO) w środowisku kwantowym `Qrisp`. Podczas gdy optymalizacja QUBO oparta na QAOA jest natywnie wspierana przez `Qrisp`, w niniejszej pracy rozszerzono to środowisko o własną implementację podejścia VQE. Artykuł omawia podstawy algorytmiczne obu metod, sposób ich realizacji programowej oraz przeprowadza analizę porównawczą ich efektywności.

Słowa kluczowe: kryptoanaliza, obliczenia kwantowe, QAOA, VQE

Keywords: cryptanalysis, QAOA, quantum computing, VQE

1. INTRODUCTION

Lattice-based cryptography is widely considered as one of the most promising foundations for post-quantum secure systems. Recently, NIST published standards FIPS 203 [6] and FIPS 204 [7], selecting the CRYSTALS-Kyber scheme [2] for key encapsulation and CRYSTALS-Dilithium for digital signatures. Both algorithms are built upon the hardness of lattice problems. Although these schemes are considered quantum-resistant, researchers explore how quantum-assisted strategies could potentially speed up classical cryptanalysis. Among such methods, the *variational quantum eigensolver* (VQE) [9] and the *quantum approximation optimization algorithm* (QAOA) [3] are especially promising for near-term quantum devices. These hybrid algorithms aim to combine quantum sampling with classical optimization to solve hard computational problems more efficiently.

Some works, such as [1] and [5], propose using VQE and QAOA to solve lattice-based cryptographic challenges such as the shortest vector problem (SVP). Due to these approaches, the cryptographic problem is first reformulated as a *Quadratic Unconstrained Binary Optimization* (QUBO) task, which is a well-known model for combinatorial optimization. Once in QUBO form, the problem can be encoded into a quantum algorithm and tackled using VQE or QAOA.

Motivated by this direction, this work compares the performance of QAOA and VQE on QUBO problems within the `Qrisp` [8] quantum programming framework. While `Qrisp` includes native support for solving QUBO problems using QAOA, a custom implementation of the VQE approach was developed for this study. The goal is to assess how both algorithms perform under the same conditions and to explore their effectiveness in solving QUBO-formulated binary optimization problems.

2. METHODOLOGY

2.1 QUBO Problem Formulation

In general, quadratic unconstrained binary optimization problems form a broad class of combinatorial optimization tasks that appear in diverse fields, including logistics, machine learning, computational biology, and recently in cryptanalysis, where its importance gained weight with serious introduction of lattice-based cryptosystems. A typical QUBO problem can be expressed as the task of minimizing a binary quadratic form:

$$\min_{\mathbf{x}\in\{0,1\}^n} \mathbf{x}^\top \mathbf{Q}\mathbf{x}, \tag{1}$$

where $\mathbf{x}$ is a binary vector and $\mathbf{Q}$ is a real symmetric matrix.

2.2 QAOA-Based QUBO Optimization in `Qrisp`

2.2.1 Algorithm Overview

The quantum approximate optimization algorithm is a hybrid quantum-classical variational algorithm designed to solve general combinatorial optimization problems. It alternates between quantum circuit layers and classical optimization to approximate low-cost solutions to a given problem.

The process begins with an initial quantum state, often chosen to be equal superposition state:

$$|s\rangle = \frac{1}{\sqrt{2^n}} \sum_{\mathbf{z}\in\{0,1\}^n} |\mathbf{z}\rangle, \tag{2}$$

where n is the number of qubits.

QAOA then applies alternating layers of two unitary operators:

- The phase separator $U_P(C,\gamma) = e^{-i\gamma C}$, where C is the cost operator representing the objective function of the problem.
- The mixer operator $U_M(B,\beta) = e^{-i\beta B}$, where B promotes transitions between computational states, encouraging exploration of the solution space.

These operations are applied sequentially to evolve the initial state into a variational state:

$$|\psi_p\rangle = U_M(B,\beta_p)U_P(C,\gamma_p)\cdots U_M(B,\beta_1)U_P(C,\gamma_1)|s\rangle.$$

The goal is to find the set of parameters $\gamma = (\gamma_1,\ldots,\gamma_p)$ and $\beta = (\beta_1,\ldots,\beta_p)$ that minimize the expected value of the cost function.

2.2.2 `Qrisp`-Specific Implementation

The implementation of QAOA in `Qrisp` is based on the *quantum alternating operator ansatz* [4], a generalization that broadens the set of available mixers and cost encodings.

To solve a QUBO problem with QAOA in `Qrisp`, the following components must be defined:

- The original QUBO objective function (1) is transformed into a *quantum Hamiltonian* using the standard binary-to-qubit mapping:

$$x_i = \frac{1-Z_i}{2} \Rightarrow H_C = \sum_{i,j} Q_{ij}\frac{(1-Z_i)(1-Z_j)}{4}. \tag{3}$$

  This results in a cost Hamiltonian H_C expressed as a sum of Pauli-Z terms and pairwise Z_iZ_j interactions.
- The cost unitary $U_C = e^{-i\gamma H_C}$ is constructed based on the Hamiltonian.
- A standard RX-based mixer is used to explore the solution space.
- The quantum state is initialized as an equal superposition (2).

Once the components are prepared, the execution of the QAOA in `Qrisp` follows these steps iteratively:

1. A quantum circuit is assembled by applying alternating layers of the cost and mixer operators.
2. The circuit is compiled and executed.
3. For each measured bitstring x, the classical cost (1) is computed. The expected value is then estimated as:

$$C = \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x}^\top\mathbf{Q}\mathbf{x}, \tag{4}$$

   where $p(\mathbf{x})$ is the measured probability of bitstring $\mathbf{x}$ in the measurement output.
4. A classical optimizer, such as COBYLA (constrained optimization by linear approximation) or the one based on the Nelder–Mead approach, updates the parameters γ,β to minimize the expected cost.

This iterative process continues until convergence or a maximum number of iterations is reached.

2.3 VQE-Based QUBO Optimization in `Qrisp`

2.3.1 Algorithm Overview

The variational quantum eigensolver is a hybrid quantum-classical algorithm to find the eigenvalues of a quantum Hamiltonian. It provides an alternative to purely quantum approaches such as quantum phase estimation, offering a more practical implementation on near-term quantum devices. In this approach, the problem is first represented by a Hamiltonian expressed as a weighted sum of Pauli strings:

$$H = \sum_j \alpha_j P_j,$$

where each P_j is a product of Pauli operators $\sigma_i^j \in I,X,Y,Z$ acting on qubit i, and α_j are real coefficients.

A parameterized quantum circuit (ansatz) prepares a trial quantum state $|\psi(\theta)\rangle$, where $U(\theta)$ consists of p layers and each layer depends on m variational parameters:

$$|\psi(\theta)\rangle = U(\theta)|\psi_0\rangle, \quad U(\theta) = \prod_{l=1}^{p}\tilde{U}(\theta_l).$$

The expected energy is then evaluated as:

$$E(\theta) = \langle\psi(\theta)|H|\psi(\theta)\rangle = \sum_j \alpha_j\langle\psi(\theta)|P_j|\psi(\theta)\rangle.$$

The goal of VQE is to find parameters θ that minimize $E(\theta)$, which approximate the ground-state energy E_0 of H.

2.3.2 `Qrisp`-Specific Implementation

Solving a QUBO problem using VQE in `Qrisp` proceeds through a sequence of structured steps that alternate between quantum state preparation and classical parameter optimization.

Before execution, three main components must be defined:

- The QUBO cost function (1) is transformed into a quantum Hamiltonian using the same relation as in QAOA (3).
- To prepare the quantum state, we use a hardware-efficient ansatz shown in Fig. 1 composed of multiple layers of parameterized quantum gates.
- The quantum state is initialized as an equal superposition (2).

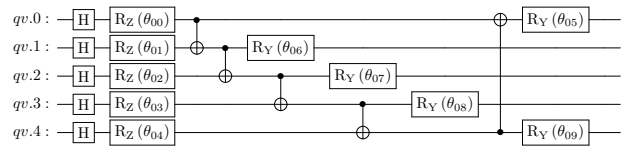


Figure 1: One layer of hardware-efficient ansatz for 5 qubits

Once the components are prepared, the execution of the VQE in `Qrisp` follows these steps:

1. The defined ansatz is used to construct a quantum circuit consisting of repeated layers.

2. The quantum circuit is executed, and the expected value of the energy is estimated using the observable decomposition of the Hamiltonian:

$$E(\theta) = \sum_k c_k \cdot \langle \psi(\theta) | P_k | \psi(\theta) \rangle$$

where each P_k is a Pauli term and c_k its coefficient. Each $\langle P_k \rangle$ is measured by executing the circuit multiple times in the appropriate basis.

3. A classical optimizer, such as COBYLA or Nelder-Mead, receives the measured energy and updates the ansatz parameters θ.

3. RESULTS

3.1 Energy Optimization Trajectory Analysis

Figs. 2 and 3 illustrate the optimization trajectories for energy minimization using QAOA and VQE, respectively, applied to the same randomly generated 5-qubit instance of a QUBO problem with randomly initialized optimization parameters. In both cases, the optimization settings were identical (`depth` $= 2$, `max_iter` $= 250$). Each plot presents the evolution of the energy across 10 independent trials, with the red line indicating the known optimal energy.

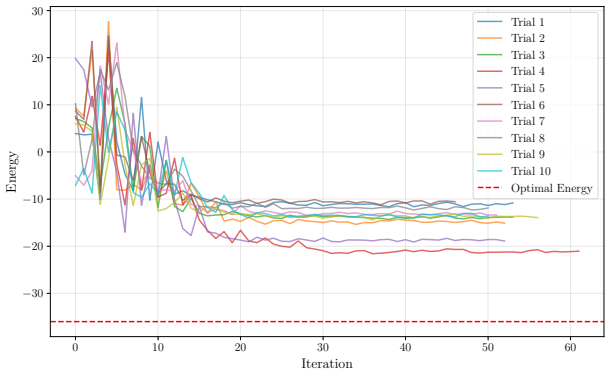


Figure 2: Energy optimization trajectory for QAOA on a 5-qubit QUBO instance

In Fig. 2 (QAOA), most trials start with relatively high energy values, followed by a rapid drop within the first 10–15 iterations. However, further optimization yields minimal improvement. It is also notable that the final energy of the system remains noticeably above the optimum. This behavior arises from the nature of QAOA-based QUBO optimization implementation: the algorithm minimizes expected value over the probability distribution generated by the quantum circuit.

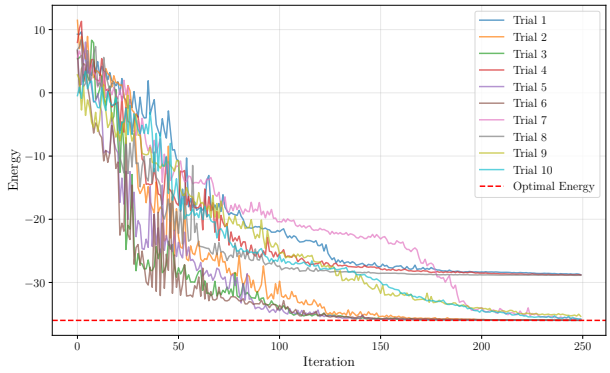


Figure 3: Energy optimization trajectory for VQE on a 5-qubit QUBO instance

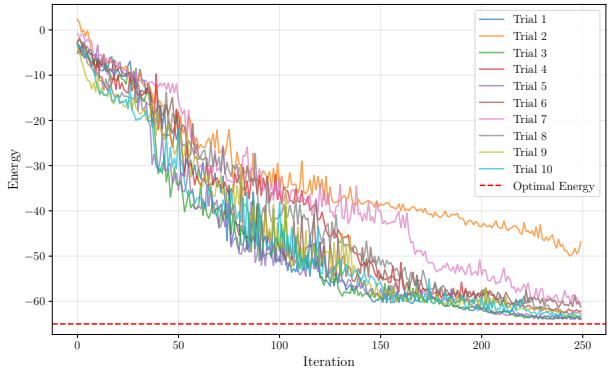


Figure 4: Energy optimization trajectory for VQE on a 9-qubit QUBO instance

In contrast, Fig. 3 (VQE) shows a more gradual and consistent reduction in energy over a longer number of iterations. The decrease is smoother, and the runs converge close to the optimal energy with minor deviations. This is because VQE directly minimizes the expected energy by measuring the expectation values of individual quantum operators in the Hamiltonian.

In the case of a larger QUBO instance, shown in Fig. 4, the overall optimization behavior remains similar. However, a notable distinction is that several curves continue to exhibit downward trends even in the later stages of optimization. This suggests that convergence has not yet plateaued for all runs within the predefined number of iterations. To further enhance solution quality in such cases, we can increase the maximum number of iterations to allow the optimizer to reach lower-energy configurations.

3.2 Accuracy Evaluation over Problem Instances

Fig. 5 illustrates how the probability of obtaining the correct solution varies with the size of the problem for both QAOA and VQE. For each problem size, ranging from 2 to 10, ten independent trials were conducted using randomly generated QUBO instances.

The results demonstrate that VQE consistently outperforms QAOA in terms of accuracy across all tested problem sizes. This performance gap is large due to the way the energy is computed in each approach. In VQE, the algorithm directly estimates the expectation value of the

Hamiltonian by averaging the contributions of each term, allowing it to converge more accurately to low-energy configurations. In contrast, QAOA relies on sampling from a distribution and evaluating the classical cost function (4), which focuses more on identifying groups of low-cost bitstrings rather than concentrating probability on the optimal solution.

However, it is also evident that the error bars (95% confidence intervals, computed from 10 runs using Student's t-distribution) for VQE are relatively large across different problem sizes. This behavior can be attributed to how the final measurement step works: the quantum state often collapses to a specific bitstring that is close to the minimum energy state found during optimization. However, due to the randomness in initial parameter selection and the variability across different problem instances, the minimum energy found is not always the global optimum. As a result, while the algorithm consistently converges toward low-energy regions, the probability of retrieving the exact optimal solution varies significantly between runs, leading to broader error margins.

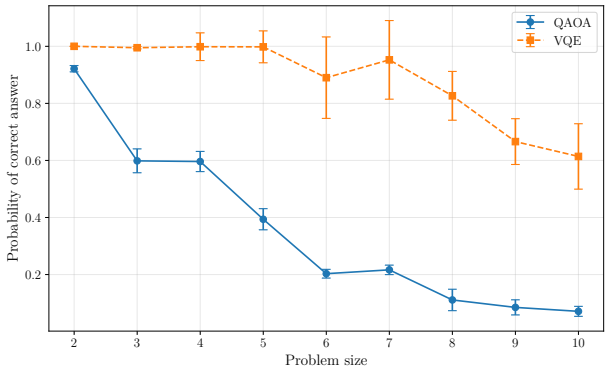


Figure 5: Probability of retrieving the optimal solution as a function of problem size

3.3 Runtime and Scalability Analysis

Although QAOA circuits are more complex because of encoding the cost function directly into quantum gates, they require only one circuit per iteration. The cost is evaluated by sampling bitstrings and computing their classical cost values. In contrast, VQE circuits are more shallow, but much more expensive to evaluate. Energy is estimated using the expectation values of individual Pauli terms. Each group must be measured separately, which requires multiple distinct circuit executions per iteration. As a result, while QAOA is limited primarily by circuit depth, the computational cost of VQE increases rapidly with the number of qubits due to the large number of measurements. This makes VQE significantly more time-consuming for larger QUBO problems.

4. CONCLUSION

This work compares the QAOA and VQE algorithms for solving QUBO problems using the `Qrisp` quantum programming framework. The altorithm performance is important from the viewpoint of cryptanalysis of lattice-based cryptosystems. Both methods showed strengths in different areas: VQE focuses on minimizing the energy of a Hamiltonian by preparing a quantum state whose final measurement typically collapses to the bitstring closest to the lowest energy solution found during optimization. In contrast, QAOA produces a distribution over a group of low-cost solutions, allowing for post-selection — one can examine several candidate outputs and classically identify the best among them.

Although VQE achieves more accurate convergence due to direct energy estimation, it is significantly slower, because it requires estimating the expectation values of individual Hamiltonian terms, each of which involves repeated measurements. For small problem instances where high precision is important, this method performs well. However, for larger problem sizes, QAOA becomes more practical. Its ability to reuse measurement samples and compute classical costs from bitstrings makes it more efficient, enabling faster exploration of the solution space with fewer resources.

REFERENCES

[1] Martin R. Albrecht et al. "Variational quantum solutions to the Shortest Vector Problem". In: *Quantum* 7 (Mar. 2023), p. 933. ISSN: 2521-327X. DOI: 10.22331/q-2023-03-02-933. URL: http://dx.doi.org/10.22331/q-2023-03-02-933.

[2] Joppe Bos et al. "CRYSTALS - Kyber: A CCA-Secure Module-Lattice-Based KEM". In: *2018 IEEE European Symposium on Security and Privacy (EuroSP)*. 2018, pp. 353–367. DOI: 10.1109/EuroSP.2018.00032.

[3] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. *A Quantum Approximate Optimization Algorithm*. 2014. arXiv: 1411.4028 [quant-ph]. URL: https://arxiv.org/abs/1411.4028.

[4] Stuart Hadfield et al. "From the Quantum Approximate Optimization Algorithm to a Quantum Alternating Operator Ansatz". In: *Algorithms* 12.2 (Feb. 2019), p. 34. ISSN: 1999-4893. DOI: 10.3390/a12020034. URL: http://dx.doi.org/10.3390/a12020034.

[5] Lihui Lv et al. "Using Variational Quantum Algorithm to Solve the LWE Problem". In: *Entropy* 24.10 (2022). ISSN: 1099-4300. DOI: 10.3390/e24101428. URL: https://www.mdpi.com/1099-4300/24/10/1428.

[6] National Institute of Standards and Technology. *FIPS 203: Module-Lattice-Based Key-Encapsulation Mechanism Standard*. Federal Information Processing Standards Publication FIPS 203. U.S. Department of Commerce, 2024. URL: https://csrc.nist.gov/pubs/fips/203/final.

[7] National Institute of Standards and Technology. *FIPS 204: Module-Lattice-Based Digital Signature Standard*. Federal Information Processing Standards Publication FIPS 204. U.S. Department of Commerce, 2024. URL: https://doi.org/10.6028/NIST.FIPS.204.

[8] Raphael Seidel et al. *Qrisp: A Framework for Compilable High-Level Programming of Gate-Based Quantum Computers*. 2024. arXiv: 2406.14792 [quant-ph]. URL: https://arxiv.org/abs/2406.14792.

[9] Jules Tilly et al. "The Variational Quantum Eigensolver: A review of methods and best practices". In: *Physics Reports* 986 (Nov. 2022), 1–128. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2022.08.003. URL: http://dx.doi.org/10.1016/j.physrep.2022.08.003.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

Algorytm kryptografii postkwantowej oparty na izogeniach krzywych eliptycznych: implementacja praktyczna

Post-Quantum Cryptography Algorithm Based on Isogenies of Elliptic Curves: Practical Implementation

Beniamin Jankowski[1]; Kamil Szydłowski[2]; Piotr Chołda[3];

AGH Akademia Górniczo-Hutnicza, Instytut Telekomunikacji, Kraków
[1] benjankowski@student.agh.edu.pl, [2] kramil208@gmail.com, [3] piotr.cholda@agh.edu.pl

DOI: 10.15199/59.2025.4.42

Streszczenie: Referat przedstawia wyniki opracowania oraz implementacji komunikatora internetowego wykorzystującego algorytmy kryptografii postkwantowej oparte na *izogeniach krzywych eliptycznych*. Zestaw funkcjonalności obejmuje szyfrowanie wiadomości typu *end-to-end* przy użyciu *X3DH* oraz *Double–Ratchet*. Uzgadnianie kluczy dla szyfrowania *end-to-end* odbywa się za pomocą *CSIDH*, gdzie do inicjowania rozmowy wykorzystano *SeaSign*. Wydajność CSIDH i SeaSign jest oceniana w zależności od wybranego poziomu bezpieczeństwa.

Abstract: The paper presents development and implementation of an Internet communicator employing post-quantum cryptography algorithms based on *isogenies of elliptic curves*. The communicator's functionality set includes transmissions encrypted end-to-end using the *X3DH* and *Double–Ratchet* algorithms. Key agreement used for end-to-end encryption is performed with the *CSIDH*, while the *SeaSign* was used to initiate a conversation. The performance of the CSIDH and SeaSign algorithms is evaluated depending on the chosen security level.

Słowa kluczowe: CSIDH, kryptografia oparta na izogeniach, kryptografia postkwantowa, SeaSign

Keywords: CSIDH, isogeny-based cryptography, post-quantum cryptography, SeaSign

1. WSTĘP

Celem pracy jest wykazanie użyteczności wybranych postkwantowych protokołów kryptograficznych opartych na *izogeniach krzywych eliptycznych* w praktycznym zastosowaniu, jakim jest komunikator internetowy zapewniający szyfrowanie w trybie od końca do końca (ang. *end-to-end encryption*). Szyfrowanie zostało zapewnione dzięki zastosowaniu algorytmu X3DH oraz Double-Ratchet. W pracy przeanalizowano wydajność opracowanej implementacji pod kątem praktyczności zaproponowanego rozwiązania.

Dotychczas opublikowano tylko jeden artykuł [8] opisujący możliwość praktycznego wykorzystania kryptografii opartej na izogeniach z użyciem wykorzystanego przez nas protokołu CSIDH. Jego autorzy również opisują projekt aplikacji do wymiany wiadomości. W odróżnieniu od naszej pracy ich projekt nie implementuje techniki szyfrowania od końca do końca. Jako protokół podpisu cyfrowego wykorzystuje Falcon [5], czyli kryptosystem, który również jest postkwantowy, lecz oparty na kratach (ang. *lattices*).

2. APARAT MATEMATYCZNY

2.1 Krzywe eliptyczne

W krzywych eliptycznych operuje się nad pewnym ciałem skończonym. $\mathbb{F}_p$. Interesują nas krzywe eliptyczne nad ciałem K o genusie 1 zawierające specyficzny punkt zwany punktem w nieskończoności. W naszych algorytmach stosujemy krzywe eliptyczne w postaci Montgomery'ego:

$$E_{A,B} : By^2 = x^3 + Ax^2 + x \tag{1}$$

gdzie $A, B \in \mathbb{F}_q \wedge B(A^2 - 4) \neq 0$. Taką postać stosuje się również w algorytmie ECDH (*Elliptic-curve Diffie-Hellman*) z krzywą *Curve25519*.

Z kolei krzywą w postaci (krótkiej) Weierstrassa nazywamy taką krzywą, którą można opisać poniższym równaniem:

$$E_{a,b} : y^2 = x^3 + ax + b \tag{2}$$

gdzie $a, b \in \mathbb{F}_q$ oraz $4a^3 + 27b^2 \neq 0$.

Dla skręcenia kwadratowego (por. niżej) stosowana jest postać Weierstrassa, która pozwala na łatwiejsze operowanie na postaci krzywej. Skręceniem kwadratowym nazywamy taki izomorfizm między krzywymi eliptycznymi, w którym otrzymujemy nową krzywą E^t w postaci $y^2 = x^3 + d^2x + d^3$, gdzie d nie jest pierwiastkiem kwadratowym w $\mathbb{F}_{p^2}$. Z każdej krzywej w postaci Montgomery'ego jest możliwe przejście do postaci Weierstrassa.

2.2 Izogenie

Izogenie to homomorfizmy między dwiema krzywymi eliptycznymi, których jądro generują wszystkie punkty, które – po przejściu przez izogenię – stają się punktami w nieskończoności. Formalnie, *izogenią* między krzywymi eliptycznymi nazywamy taki morfizm, że:

$$\phi : E_1 \to E_2, \quad \phi(\mathcal{O}_\infty) = \mathcal{O}_\infty \tag{3}$$

Jądrem izogenii jest zatem zbiór takich punktów, że $\ker \phi = \{P \in E_1 : \phi(P) = \mathcal{O}_\infty\}$. Stopień izogenii rozumiemy jako rozmiar jej jądra, zatem $\deg \phi = \# \ker \phi$.

Nazwa używanego protokołu wymiany klucza CSIDH (*Commutative Supersingular Isogeny Diffie–Hellman*) zawiera pojęcie supersingularności. Krzywa ma tę właściwość, gdy liczba punktów należących do niej wynosi $\#E(\mathbb{F}_p) = p + 1$. Warunek ten dotyczy tylko takich p, które są liczbami nieparzystymi (w omawianym przypadku pierwszymi). Pierścieniem endomorfizmów krzywej eliptycznej E nazywamy pierścień wszystkich izogenii $\phi\colon E \to E$ wraz z mnożeniem przez zero ($0 \cdot P = P \cdot 0 = \mathcal{O}_\infty$) oraz z działaniami dodawania oraz ich składania.

Dla krzywych w postaci Montgomery'ego istnieją wzory pozwalające obliczyć omawiane izogenie. Zostały one zawarte w pracy [9]. Niech $\phi : E \to \tilde{E}$ to izogenia, przez którą przejdzie punkt $P = (x, y)$. Nowa odcięta takiego punktu to $f(x)$, gdzie:

$$f(x) = x \prod_{T \in \ker \phi \setminus \{\mathcal{O}_\infty\}} \frac{x x_T - 1}{x - x_T} \tag{4}$$

gdzie x_T to odcięte każdego punktu T. Dodatkowo:

$$\pi_T = \prod_{T \in \ker \phi \setminus \{\mathcal{O}_\infty\}} x_T,\ \sigma_T = \sum_{T \in \ker \phi \setminus \{\mathcal{O}_\infty\}} \left(x_T - \frac{1}{x_T}\right) \tag{5}$$

Współczynnik A nowej krzywej jest równy $\pi(a - 3\sigma)$.

3. ALGORYTM CSIDH

3.1 Wstęp

Protokół CSIDH, tj. przemiennej wymiany kluczy Diffiego–Hellmana opartej na izogeniach supersingularnych został zaproponowany jako postkwantowy (oparty na trudnym problemie obliczeniowym, którego złożonośc jest nadwielomianowa nawet w przypadku obliczeń na komputrze kwantowym), statyczny (czyli zapewniający długi okres ważności kluczy) schemat ustalania współdzielonego klucza. Opiera się na problemie znajdowania izogenii między danymi krzywymi eliptycznymi [2]. W celu ustalenia wspólnego klucza strony protokołu przemieszczają się po grafie, którego wierzchołkami są supersingularne krzywe eliptyczne, a krawędziami izogenie między nimi. Krzywa startowa jest jednym z parametrów uzgodnionym przed rozpoczęciem algorytmu. Ścieżka pokonywana przez uczestnika wymiany wyznaczana jest przez jego klucz prywatny, który ma postać wektora o współczynnikach ustalanych w sposób losowy. Kluczem publicznym oraz kluczem współdzielonym jest współczynnik A krzywej eliptycznej w postaci Montgomery'ego, który jednoznacznie ją identyfikuje. Protokół CSIDH zapewnia efektywną walidację kluczy publicznych oraz cechuje się ich kompaktowym rozmiarem.

Mnożenie punktu przez skalar $[n] : P \to nP$ służy do wprowadzenia endomorfizmu Frobeniusa danego jako $\pi : P = (x, y) \to P' = (x^p, y^p)$, gdzie $\pi^2 = \pi \circ \pi := [-p]$. Rozważamy pierścień endomorfizmów w postaci:

$$\mathbb{Z}[\sqrt{-p}] = \{n + m\sqrt{-p}; n, m \in \mathbb{Z}\} \tag{6}$$

Grupę klas ideałów pierścienia $\mathbb{Z}[\sqrt{-p}]$ będziemy oznaczać jako $\mathcal{C}l(\mathbb{Z}[\sqrt{-p}])$. Ideał $\mathcal{I} \subseteq \mathbb{Z}[\sqrt{-p}]$ zapisujemy jako:

$$\mathcal{I} = \langle \alpha_1, \alpha_2, ..., \alpha_n \rangle \subseteq \mathbb{Z}[\sqrt{-p}] \tag{7}$$

Ostatecznie, działanie grupy na zbiorze w CSIDH można zapisać następująco:

$$(\mathcal{I}, E) \to \mathcal{I} * E := \phi_{\mathcal{I}}(E) \tag{8}$$

CSIDH można zaatakować algorytmem zaproponowanym przez Kuperberga, który bierze za cel atak na problem *hidden-shift* w grupie abelowej [6]. Jego złożoność jest nadwielomianowa: wynosi $2^{O\left(\sqrt{\log_2 N}\right)}$, gdzie N to rząd grupy abelowej.

3.2 Schemat działania na przykładzie implementacji

Poniżej w list. 1 przedstawiono używany kod algorytmu CSIDH. Całość implementacji została napisana przez nas w języku Python.

Listing 1: Algorytm CSIDH

```
class CSIDH_Tools:
  @staticmethod
  def isogeny(E, P, R):
    p = E.p
    A = E.A
    B = E.B
    G = E.generate_subgroup(R)
    G = [point for point in G if not point.
        is_infinity]
    pi = prod([T.x for T in G])
    sigma = sum([(T.x - Modular_Math.inverse_of(T.x
        , p)) for T in G])
    new_A = pi * (A - 3*sigma) % p
    new_E = MontgomeryEllipticCurve(p, new_A, B)
    x = P.x
    product = [(x*T.x - 1)*(Modular_Math.inverse_of
        (x - T.x, p)) for T in G]
    f_x = x * prod(product)
    new_x = f_x
    new_y = new_E.lift_x(new_x).y
    if new_x == 0 and new_y == 0:
      new_P = Montgomery_Point(p, new_E)
    else:
      new_P = Montgomery_Point(p, new_E, new_x,
          new_y)
    return new_E, new_P

  @staticmethod
  def CSIDH(E, priv, p, ell):
    L_side = (p+1)*E.random_point()
    R_side = Montgomery_Point(p, E)
    assert L_side == R_side
    for cl in ([max(0, e) for e in priv],[max(0, -e
        ) for e in priv]):
      while any(cl) != 0:
        j = [n for n in zip(ell, cl)]
        P = E.random_point()
        S = [j.index(ei) for ei in j if ei[1]!=0]
        if S:
          k = prod(ell[i] for i in S)
          Q = ((p+1)//k)*P
          for i in S:
            l = ell[i]
            R = (k//l)*Q
            if not (R.x is None and R.y is None):
              E, Q = CSIDH_Tools.isogeny(E, Q, R)
              k = k//l
              cl[i] -= 1
      E = E.quadratic_twist()
    return E
```

Najpierw prostym warunkiem sprawdzamy, czy dana krzywa jest supersingularna. Zastępujemy ujemne wartości ich wartościami bezwzględnymi – to one będą wymagały później skręcenia krzywej. Algorytm wykonuje się dopóki zostały jakieś działania klasy ideału. W przypadku, gdy zostały działania do wykonania, przechodzimy do odwzorowania liczb pierwszych z odpowiadającymi liczbami z wektora klucza prywatnego. Ten krok pozwala nam na ustalenie ile razy należy wykonać e_i jako działanie klasy ideału.

Wybieramy losowy element z krzywej. Funkcja `random_point()` będzie wykonywana tak długo, póki nie znajdziemy punktu należącego do krzywej E, gdyż nie każda odcięta będzie miała rzędną na zadanej krzywej.

Następnie tworzymy listę indeksów działań S, które muszą zostać jeszcze wykonane. W przypadku pustej listy bierzemy następny element cl i cofamy się do początku pętli `for`. W przypadku niepustej listy S obliczamy iloczyn liczb pierwszych o zadanych indeksach. Mnożymy punkt P przez liczbę $(p+1)//k$ (symbol // oznacza dzielenie, gdzie pod uwagę bierzemy tylko część całkowitą wyniku takiego działania). Następnym krokiem jest wzięcie liczb pierwszych, które będą nam wskazywać, jakiego stopnia izogenię mamy obliczyć – jest to zbiór liczb w `S` w linii 37 podanego kodu. Obliczamy zatem punkt R, który jest rzędu l. Stanie się on generatorem jądra izogenii. Dzięki wzorom Vélu zaimplementowanym w funkcji `isogeny()` jesteśmy w stanie znaleźć obraz punktu Q po przejściu przez izogenię o zadanym jądrze. Otrzymujemy również nowo powstałą krzywą.Obliczyliśmy izogenię jądra l, więc zmniejszamy k według wzoru $k = k//l$. Zmniejszamy przez to również liczbę działań potrzebnych do wykonania (`cl[i] -= 1`).

Po przebiegu algorytmu, w którym wykonujemy wszystkie działania klas ideałów, zostaje jeszcze podmiana krzywej na jej skręcenie kwadratowe na podstawie właściwości: $[\mathfrak{a}]^{-1}E = ([\mathfrak{a}]E^t)^t$. Pozwala nam to na działania również w przypadku, gdy w kluczu prywatnym występują liczby ujemne. Na koniec całej sekwencji otrzymujemy krzywą, która może zostać przekazana drugiej stronie lub będzie to właśnie współdzielony sekret.

4. ALGORYTM SEASIGN

SeaSign [4] to opracowany w 2018 roku przez Lucę De Feo i Stevena D. Galbraitha protokół podpisu cyfrowego, oparty na wcześniejszej pracy Antona Stolbunowa [11], którą autorzy postanowili dostosować do wykorzystania aparatu matematycznego kryptosystemu CSIDH. Schemat podpisu cyfrowego opiera się na wykorzystaniu pewnej odmiany transformacji Fiata-Shamira opracowanej przez Vadima Lyubashevsky'ego [7] (ang. *Fiat-Shamir with aborts*). Podpisy generowane przez SeaSign mają wielkość 978 bajtów dla parametrów podanych przez autorów CSIDH jako spełniające poziom bezpieczeństwa NIST I, czyli $p = 4 \cdot \ell_1 \cdot \ell_2 \cdot \ell_3 \cdot ... \ell_{73} \cdot 587 - 1$ gdzie $\ell_1, ..., \ell_{73}$ są to kolejne liczby pierwsze, a wartości w kluczu prywatnym zawierają się w przedziale od -5 do 5. Klucze publiczne dla takich parametrów są wielkości 4 megabajtów, a czas potrzebny na podpisanie wiadomości i weryfikację podpisu to około kilku minut.

Istnieje prawdopodobieństwo, że wiadomości nie uda się podpisać. Według autorów [4] prawdopodobieństwo, że algorytm wygeneruje prawidłowy podpis wynosi co najmniej e^{-1}. W przeciwnym przypadku należy powtarzać algorytm do czasu, aż wygeneruje on prawidłowy podpis. Idea odrzucania niepoprawnych podpisów została zaczerpnięta ze wspominanej już rozwiniętej przez Vadima Lyubashevsky'ego [7] transformacji Fiata-Shamira.

5. APLIKACJA INTERNETOWA

5.1 Założenia nt. szyfrowania *end-to-end*

W celu praktycznego wykorzystania omawianych kryptosystemów, opracowaliśmy komunikator tekstowy, który pozwala na przesyłanie wiadomości tekstowych oraz plików binarnych. Na uwadze przy implementacji mieliśmy podejście szyfrowania od końca do końca. Dzięki temu eliminujemy możliwość podejrzenia wiadomości przez serwer i jedynie nadawca / odbiorca potrafią wydobyć tekst jawny z zaszyfrowanej wiadomości. Największą zaletą takiego rozwiązania jest przede wszystkim zagwarantowanie użytkownikowi prywatności.

Na podstawie analizy literaturowej, w tym artykułów [3] oraz [1], a także skali, na jaką Signal Protocol jest praktycznie używany, zdecydowaliśmy się skorzystać właśnie ze wzorców, które proponuje Signal Foundation. Wybraliśmy dwa algorytmy do zintegrowania w naszej implementacji: Double-Ratchet algorithm oraz X3DH, które rozszerzą bezpieczeństwo samej aplikacji. Używamy zatem [10]: (a) X3DH – algorytm pozwalający na rozpoczęcie wymiany wiadomości w taki sposób, abyśmy mogli w prawidłowy sposób zidentyfikować rozmówcę. (b) algorytm Double-Ratchet – pozwala zachować właściwości *perfect backward secrecy* oraz *perfect forward secrecy*.

Algorytm X3DH służy do ustalenia wspólnego, ale jednego, sekretu. Algorytm Double-Ratchet służy do zapewnienia bezpieczeństwa przy kompromitacji kluczy szyfrujących – nie jesteśmy w stanie rozszyfrować już wysłanych wiadomości, jak również przyszłych wiadomości. Oba algorytmy uzupełniają się, gdyż np. dla algorytmu Double-Ratchet potrzeba początku wymiany wiadomości.

5.2 Stos technologiczny aplikacji

Najważniejszym założeniem, jakie miała spełniać nasza aplikacja, było zapewnienie szyfrowania w trybie od końca do końca. Materiałem kryptograficznym pozwalającym na odszyfrowanie przesyłanych wiadomości mają dysponować wyłącznie strony komunikacji. Serwer, będący pośrednikiem w komunikacji, ma nie posiadać możliwości podejrzenia treści wiadomości. W celu zrealizowania tego założenia nasz komunikator jest aplikacją desktopową – wszystkie funkcje kryptograficzne wykonują się lokalnie u klienta, a wiadomości trafiają do serwera już w postaci zaszyfrowanej. Dodatkowo, klucze prywatne użytkowników przechowywane są wyłącznie w lokalnym systemie plików klientów i nigdy nie opuszczają ich urządzenia.

6. TESTY WYDAJNOŚCI

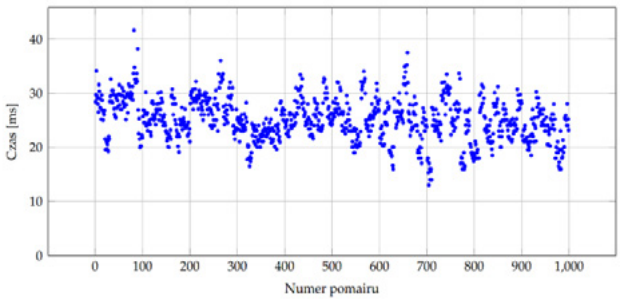
Aplikacja, która powstała, powinna być traktowana raczej jako środowisko utworzone na potrzeby zaprezentowania możliwości kryptosystemów opartych na izogeniach i sprawdzenia ich wydajności niż jako równorzędna modułom kryptograficznym. Stąd większy nacisk został położony na implementację oraz testy wspomnianych modułów niż na samą aplikację.

W celu porównania osiągów wydajnościowych naszych algorytmów kryptograficznych, chcieliśmy wyznaczyć zależność między bezpieczeństwem a szybkością danego rozwiązania.

6.1 Wydajność CSIDH

Najważniejszym wskaźnikiem jakości jest czas ustalenia współdzielonego klucza. W tym celu zbadaliśmy sytuację, w której tylko jedna ze stron będzie brana pod uwagę pod względem czasowym. Ostatecznie, mierzymy dwa zsumowane czasy: jak długo zajmie stronie zadziałanie kluczem prywatnym na krzywą startową oraz jak długo zajmie zadziałanie kluczem prywatnym na krzywą otrzymaną przez drugą stronę (druga strona wykonuje

krok wcześniej, a następnie przekazuje tę krzywą pierwszej stronie). Na rys. 1 znajduje się wykres przedstawiający, ile czasu zajmuje ustalenie współdzielonego klucza.



Rysunek 1: Pomiary czasowe dla CSIDH-64

W tab. 1 przedstawiono ponadto czasy dla różnych parametrów CSIDH. W oparciu o przedstawione wyniki zdecydowaliśmy się właśnie na wariant CSIDH-64, gdyż chcieliśmy uzyskać jak najlepszy kompromis między bezpieczeństwem a wydajnością (dlatego mniejsze wartości parametrów CSIDH nie były brane pod uwagę w obliczeniach).

Tabela 1: Pomiary czasowe dla kolejnych parametrów CSIDH

Poziom bezp.	*CSIDH*	*Średnia [s]*
—	64	0.026217
—	128	0.288998
—	256	8.831445
NIST I	512	16.785786
NIST III	1024	212.711439
NIST V	1792	2131.080233

7. PODSUMOWANIE

Głównym celem naszej pracy było sprawdzenie, w jaki sposób można w praktyce zastosować algorytmy postkwantowe oparte na izogeniach. Wynik połączenia ze sobą algorytmów CSIDH oraz SeaSign może posłużyć do utworzenia kanału, dzięki któremu będziemy mogli przesyłać zaszyfrowane dane. Będą one trafiały jedynie do osób, które w odpowiedni sposób przeszły autoryzację oraz posiadają odpowiedni materiał kryptograficzny. Za adekwatny przykład, w którym algorytmy CSIDH i SeaSign miałyby praktyczne zastosowanie, uznaliśmy komunikatory tekstowe z szyfrowaniem od końca do końca.

REFERENCES

[1] Joël Alwen, Sandro Coretti, and Yevgeniy Dodis. "The Double Ratchet: Security Notions, Proofs, and Modularization for the Signal Protocol". In: *Advances in Cryptology – EUROCRYPT 2019*. Ed. by Yuval Ishai and Vincent Rijmen. Cham: Springer International Publishing, 2019, pp. 129–158. ISBN: 978-3-030-17653-2.

[2] Wouter Castryck and Thomas Decru. "An Efficient Key Recovery Attack on SIDH". In: *Advances in Cryptology — EUROCRYPT 2023*. Ed. by Carmit Hazay and Martijn Stam. Cham: Springer Nature Switzerland, 2023, pp. 423–447. ISBN: 978-3-031-30589-4.

[3] Katriel Cohn-Gordon et al. "A Formal Security Analysis of the Signal Messaging Protocol". In: *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*. 2017, pp. 451–466. DOI: 10.1109/EuroSP.2017.27.

[4] Luca De Feo and Steven D. Galbraith. "SeaSign: Compact Isogeny Signatures from Class Group Actions". In: *Advances in Cryptology – EUROCRYPT 2019*. Ed. by Yuval Ishai and Vincent Rijmen. Cham: Springer International Publishing, 2019, pp. 759–789. ISBN: 978-3-030-17659-4.

[5] Pierre-Alain Fouque et al. "Falcon: Fast-Fourier lattice-based compact signatures over NTRU". In: *Submission to the NIST's post-quantum cryptography standardization process* 36.5 (2018), pp. 1–75.

[6] Greg Kuperberg. *A subexponential-time quantum algorithm for the dihedral hidden subgroup problem*. 2004. arXiv: quant-ph/0302112 [quant-ph].

[7] Vadim Lyubashevsky. "Fiat-Shamir with Aborts: Applications to Lattice and Factoring-Based Signatures". In: *Advances in Cryptology – ASIACRYPT 2009*. Ed. by Mitsuru Matsui. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 598–616. ISBN: 978-3-642-10366-7.

[8] Luc Nhu Quynh et al. "Secure Messaging Application Development: Based on Post-Quantum Algorithms CSIDH, Falcon, and AES Symmetric Key Cryptosystem". In: *Programming and Computer Software* 50 (July 2024), pp. 322–333. DOI: 10.1134/S0361768824700130.

[9] Joost Renes. "Computing isogenies between Montgomery curves using the action of $(0,0)$". In: https://eprint.iacr.org/2017/1198. 2017.

[10] Signal Foundation. *Signal Protocol Documentation*. https://signal.org/docs/.

[11] Anton Stolbunov. "Cryptographic Schemes Based on Isogenies". PhD thesis. Norwegian University of Science, Technology, Faculty of Information Technology, Mathematics, and Electrical Engineering, Department of Telematics, Jan. 2012. DOI: 10.13140/RG.2.2.20826.44488.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

ANALIZA WYBRANYCH IMPLEMENTACJI ALGORYTMÓW KRYPTOGRAFII POSTKWANTOWEJ

ANALYSIS OF SELECTED IMPLEMENTATIONS OF POST-QUANTUM CRYPTOGRAPHIC ALGORITHMS

Zuzanna Pietrzak; Filip Opiłka; Marcin Niemiec

AGH Akademia Górniczo-Hutnicza, Instytut Telekomunikacji
Mickiewicza 30, 30-059 Kraków

DOI: 10.15199/59.2025.4.43

Streszczenie: Artykuł przedstawia wyniki badań wydajności wybranych algorytmów postkwantowych – ML-DSA oraz ML-KEM – w dwóch projektach kryptograficznych: Terra Quantum Public i Open Quantum Safe. Badania obejmowały pomiar czasów generowania kluczy, enkapsulacji, dekapsulacji dla ML-KEM oraz podpisywania i weryfikacji podpisów dla ML-DSA, w zależności od rozmiaru danych. Wyniki wykazują, że biblioteka liboqs oferuje większą wydajność i stabilność czasów wykonania poszczególnych operacji w porównaniu do biblioteki TQ42.

Abstract: The article presents the results of a performance analysis of selected post-quantum algorithms (ML-DSA and ML-KEM) within two cryptographic frameworks: Terra Quantum Public and Open Quantum Safe. The study involved measuring key generation, encapsulation, and decapsulation times for ML-KEM, as well as signinature generation and verification times for ML-DSA, with respect to varying data sizes. The findings indicate that the liboqs library provides higher performance and greater consistency in execution times.

Słowa kluczowe: algorytmy postkwantowe, generowanie kluczy, kryptografia, podpisy cyfrowe, szyfrowanie

Keywords: post-quantum algorithms, key generation, cryptography, digital signatures, encryption

1. WSTĘP

Kryptografia postkwantowa od kilkunastu lat stanowi istotny temat badań w środowisku naukowym. Już w 1994 roku Peter Shor [1] opublikował algorytm kwantowy, który umożliwia rozkładanie liczb na czynniki pierwsze w czasie wielomianowym przy wykorzystaniu komputera kwantowego. Odkrycie to zasygnalizowało poważne zagrożenie dla szeroko stosowanych asymetrycznych algorytmów kryptograficznych, takich jak RSA, które opiera się na trudności rozkładu liczby na czynniki. Mimo że obecnie komputery kwantowe nie są jeszcze dostępne w formie komercyjnej, ich pojawienie się na rynku w przyszłości wydaje się być kwestią czasu. W związku z tym istnieje pilna potrzeba opracowywania algorytmów kryptograficznych odpornych na złamanie bezpieczeństwa kluczy publicznych przy wykorzystaniu technologii kwantowej.

Od wielu lat naukowcy na całym świecie rozwijają algorytmy kryptografii postkwantowej, które są implementowane w różnych językach programowania. Każda implementacja charakteryzuje się odmienną jakością kodu i wydajnością działania, co stanowi ważny obszar badań i analiz w kontekście przyszłych zastosowań praktycznych.

2. STAN WIEDZY

W odpowiedzi na zagrożenia wynikające z rozwoju komputerów kwantowych, amerykański Narodowy Instytut Standaryzacji i Technologii (NIST) rozpoczął w 2016 roku inicjatywę mającą na celu standaryzację algorytmów kryptografii postkwantowej. Po trwających kilka lat badaniach, w 2022 roku NIST ogłosił wyniki trzeciej rundy konkursu i wytypował algorytmy, które miały być podstawą przyszłych standardów. W sierpniu 2024 roku opublikowano nowe standardy kryptograficzne FIPS 203, 204 oraz 205 [2], które obejmują m.in. algorytmy Kyber, Dilithium oraz SPHINCS+. Standaryzacja tych algorytmów ma kluczowe znaczenie dla zapewnienia bezpieczeństwa danych w erze postkwantowej. Ich analiza i implementacja pozwalają nie tylko na ocenę wydajności w różnych środowiskach, ale również na zrozumienie wyzwań związanych z ich wdrożeniem w systemach informatycznych. W niniejszym artykule podjęto próbę porównania różnych implementacji algorytmów postkwantowych w celu zbadania ich efektywności i praktycznego zastosowania w kontekście współczesnych potrzeb bezpieczeństwa cyfrowego.

W artykule [3] dokonano przeglądu kryptografii postkwantowej. Omówiono w nim różne platformy sprzętowe, na których implementowane są algorytmy kryptografii postkwantowej i podkreślono potrzebę optymalizacji algorytmów w zależności od zastosowań. W artykule konferencyjnym [4] porównano wydajność algorytmów postkwantowych z tradycyjnymi algorytmami kryptograficznymi. Do pomiaru czasu obliczeń w przypadku tradycyjnych algorytmów wykorzystano bibliotekę Crypto++, natomiast dla algorytmów postkwantowych zastosowano bibliotekę liboqs. Warto także wspomnieć o badaniach zaprezentowanych w artykułach [5] i [6]. Ukazują one potencjał wybranych algorytmów postkwantowych w kontekście efektywności obliczeń. Jednak przedstawione tam wyniki także bazują tylko na jednej bibliotece (liboqs).

3. ŚRODOWISKO BADAWCZE

Niniejszy artykuł przedstawia wyniki badań efektywności algorytmów postkwantowych dwóch różnych bibliotek programistycznych. Na maszynie wirtualnej zainstalowano obie testowane biblioteki: Terra Quantum Public [7] oraz Open Quantum Safe [8]. Każda z nich posiada wrapper w języku Python, który został wykorzystany w przeprowadzonych badaniach.

Środowisko badawcze skonfigurowano na maszynie wirtualnej działającej w infrastrukturze OpenStack. Uruchomiony system operacyjny Ubuntu 22.04 LTS dysponował następującymi zasobami: 16 GB pamięci RAM, 8 vCPU i dyskiem o pojemności 160 GB. Do maszyny przypisano dedykowany rdzeń CPU, co zapewnia izolację zasobów oraz stabilność wydajności podczas przeprowadzania testów.

Do badań wybrano algorytmy, które mają implementacje w obu bibliotekach. Są to ML-KEM do szyfrowania oraz ML-DSA do podpisów cyfrowych. Każdy z algorytmów testowano w trzech trybach (na różnych poziomach bezpieczeństwa), dla których rozmiary kluczy zostały przedstawione w tabeli 1. Dokonano pomiarów czasu generowania par kluczy, odpowiednio enkapsulacji i dekapsulacji dla ML-KEM, oraz czasów podpisywania cyfrowego dla ML-DSA, a także weryfikacji poprawności działań.

Tabela 1: Rozmiary kluczy dla wybranych trybów ML-KEM oraz ML-DSA. *Źródło: Terra Quantum [9]*

Tryb ML-KEM	*Klucz enkapsulacji*	*Klucz dekapsulacji*
512	800 [B]	1632 [B]
768	1184 [B]	2400 [B]
1024	1568 [B]	3168 [B]
Tryb ML-DSA	*Klucz publiczny*	*Klucz prywatny*
44	1312 [B]	2560 [B]
65	1952 [B]	4032 [B]
87	2592 [B]	4896 [B]

Symulacje obejmowały wstępne powtórzenia rozgrzewkowe, charakteryzujące się zauważalnie dłuższym czasem wykonania. Wszystkie uwzględnione w wynikach operacje wykonano 1000 razy. Wyniki zostały uśrednione, a następnie obliczono dla nich 95% przedziały ufności. Badania przeprowadzono dla różnych rozmiarów danych, które były analizowane jako potęgi liczby 2 w bajtach, począwszy od 2 i kończąc na 256.

4. WYNIKI

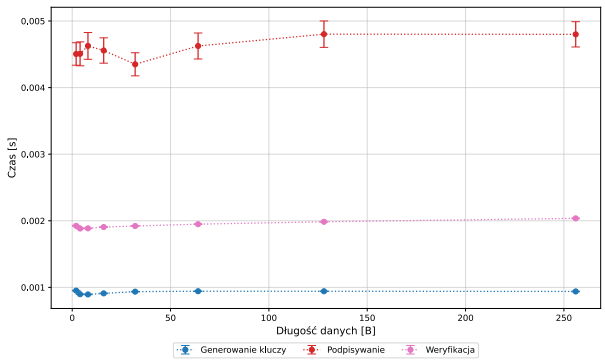
Autorzy przeprowadzili dokładne badania wydajności algorytmów postkwantowych zaimplementowanych w bibliotekach TQ42 od Terra Quantum Public (TQP) [7] oraz liboqs z Open Quantum Safe Project (OQS) [8]. Jednak ze względu na limit stron, w artykule zaprezentowano jedynie wybrane wykresy z uzyskanymi wynikami.

4.1 Terra Quantum Public

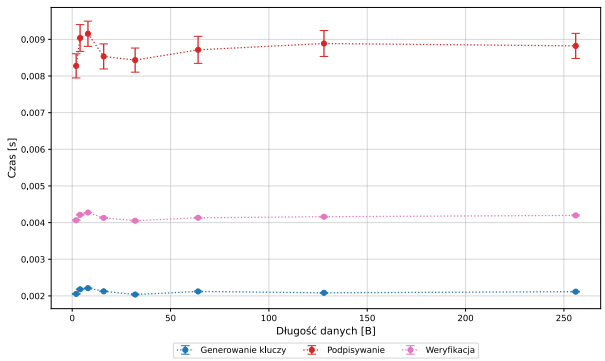
Biblioteka TQ42 [7] to biblioteka napisana w języku C++, ale oferuje również wrapper dla Pythona, co umożliwia łatwe wykorzystanie jej funkcji w różnych środowiskach programistycznych. W przyszłości biblioteka ma zostać rozszerzona o wsparcie dla dodatkowych języków programowania, a także o nowe algorytmy postkwantowe [9].

Pierwszym z badanych algorytmów jest ML-DSA, dla którego przetestowano funkcje generowania pary kluczy, podpisywania wiadomości oraz weryfikacji podpisu. Na rysunkach 1 i 2 przedstawiono czasy wykonania tych operacji dla algorytmu w wariantach 44 oraz 87, w zależności od rozmiaru danych. Z przeprowadzonych badań wynika, że generowanie kluczy jest najszybszą operacją, a czas jej wykonania jest stabilny we wszystkich trybach algorytmu. Wartości czasów generowania kluczy rosną z przejściem na wyższy wariant algorytmu od około 0,0009 s w trybie 44, do około 0,0021 s w trybie 87.

Podpisywanie wiadomości jest najdłuższą operacją i wykazuje największe fluktuacje. Dla rozmiarów danych od 100 B do 256 B czas tej operacji jest stosunkowo stały w trybach 44 i 87. Największe zmiany występują przy mniejszych rozmiarach wiadomości. Przedziały ufności dla tej operacji są szerokie, szczególnie w trybach 65 i 87, co sugeruje mniejszą przewidywalność procesu. Dla trybu 44 czas podpisywania waha się od 0,00435 s do 0,00480 s, w trybie 65 od 0,00680 s do 0,00787 s, a w trybie 87 od 0,00827 s do 0,00904 s. Weryfikacja podpisu jest stabilna niezależnie od rozmiaru wiadomości, ale jej czas wykonania jest dłuższy niż czas generowania kluczy.



Rysunek 1: Czasy operacji algorytmu ML-DSA-44 z biblioteki TQP w zależności od rozmiaru wiadomości.

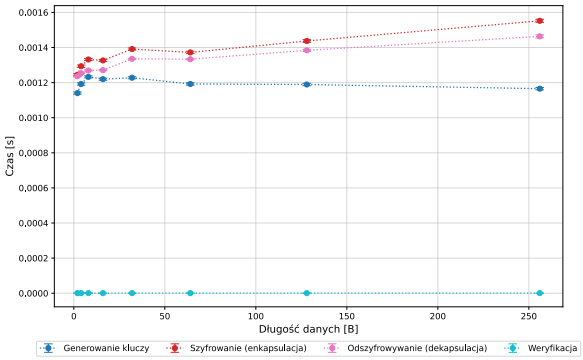


Rysunek 2: Czasy operacji algorytmu ML-DSA-87 z biblioteki TQP w zależności od rozmiaru wiadomości.

Dla drugiego algorytmu, ML-KEM, przeprowadzono pomiary czasów generowania par kluczy, enkapsulacji, dekapsulacji oraz weryfikacji poprawności wiadomości. Na rysunkach 3 oraz 4 przedstawiono czasy wykonania poszczególnych operacji w zależności od rozmiaru wiadomości dla wersji 512 i 1024 algorytmu ML-KEM.

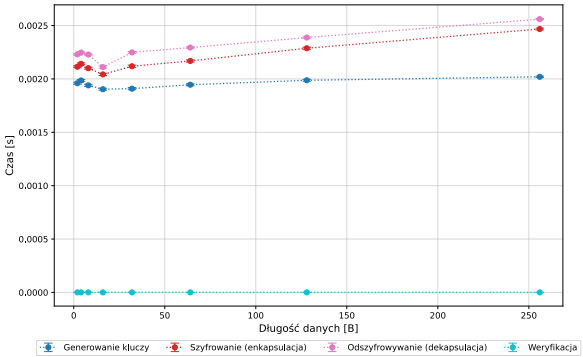
Wyniki wskazują, że średni czas generowania kluczy we wszystkich trybach jest stosunkowo stabilny i nieznacznie zależny od rozmiaru wiadomości. Dla wszystkich wariantów algorytmu największe różnice w czasie wykonania pojawiają się dla wiadomości o rozmiarach od 2 do 32 B. W przypadku większych rozmiarów danych wpływ na czas generowania kluczy jest minimalny. Przedziały ufności są małe, co potwierdza wysoką powtarzalność procesu.

Operacje szyfrowania i deszyfrowania, w każdym przypadku, wykazują wzrost czasu wykonania w miarę zwiększania długości danych. Średnie czasy tych operacji są zbliżone, co sugeruje podobną złożoność obliczeniową obu procesów. Czas szyfrowania w badanych trybach wzrasta od około 0,0012 s do 0,0025 s, a czas deszyfrowania od 0,0012 s do 0,0026 s. Przedziały ufności są ponownie minimalne, co potwierdza wiarygodność uzyskanych wyników. Czas wykonania operacji weryfikacji odszyfrowanej wiadomości dla każdego z wariantów pozostaje praktycznie stały i stosunkowo mały.



Rysunek 3: Czasy operacji algorytmu ML-KEM-512 z biblioteki TQP w zależności od rozmiaru wiadomości.

Analizując oba algorytmy, można zaobserwować, że trendy dla poszczególnych operacji pozostają podobne w przypadku różnych wariantów każdego z nich. Wyniki wskazują jednak, że kluczowa operacja algorytmu ML-DSA jest znacznie mniej stabilna i powtarzalna w badanych warunkach w porównaniu do algorytmu ML-KEM. Ponadto, czas generowania kluczy jest krótszy w przypadku algorytmu ML-KEM.

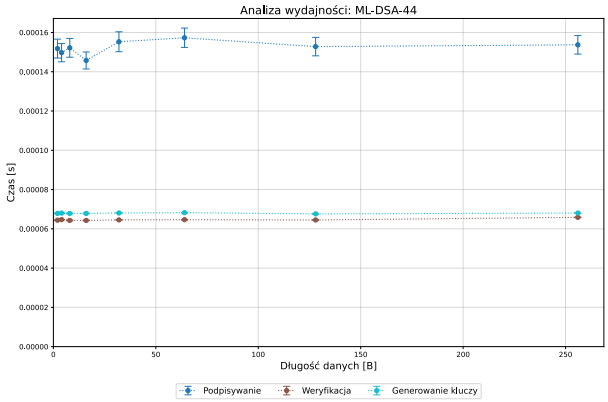


Rysunek 4: Czasy operacji algorytmu ML-KEM-1024 z biblioteki TQP w zależności od rozmiaru wiadomości.

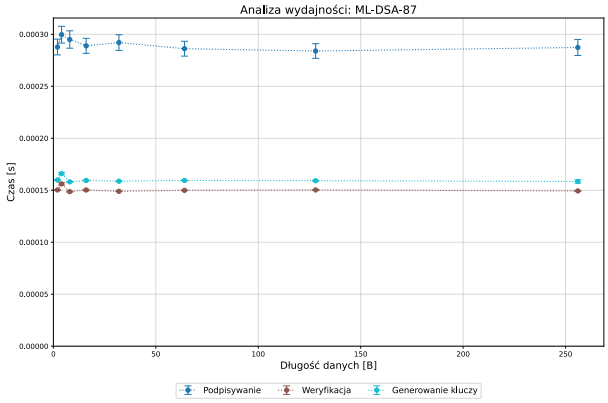
4.2 Open Quantum Safe

Biblioteka liboqs [8] jest napisana w języku C, lecz dla szerszej dostępności, twórcy udostępniają również wrappery w innych językach, w tym Python. Liboqs stanowi część projektu Open Quantum Safe i jest starsza od omawianej w poprzedniej sekcji biblioteki TQP [10].

W przypadku algorytmu ML-DSA ponownie przetestowano funkcje generowania pary kluczy, podpisywania oraz weryfikacji podpisu. Wyniki pomiarów czasu wykonania tych operacji w zależności od rozmiaru podpisywanych danych zostały przedstawione na rysunkach 5 i 6. Wykresy ilustrują dane dla dwóch wersji algorytmu: ML-DSA-44 oraz ML-DSA-87.



Rysunek 5: Czasy operacji algorytmu ML-DSA-44 z biblioteki OQS w zależności od rozmiaru wiadomości.

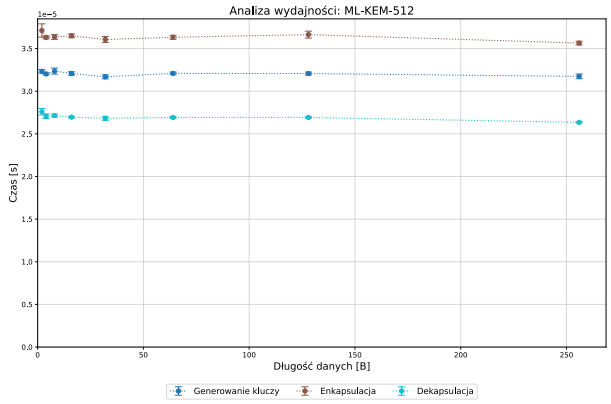


Rysunek 6: Czasy operacji algorytmu ML-DSA-87 z biblioteki OQS w zależności od rozmiaru wiadomości.

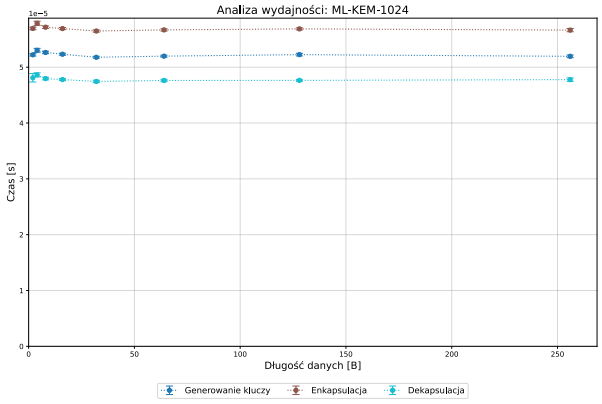
Na obu wykresach widoczny jest podobny trend: operacją zajmującą najwięcej czasu jest podpisywanie, podczas gdy generowanie kluczy oraz weryfikacja podpisu charakteryzują się zbliżoną wydajnością czasową i są około dwa razy szybsze niż tworzenie podpisu cyfrowego. Dla dwóch mniej czasochłonnych operacji słupki błędów są minimalne, natomiast w przypadku generowania podpisu są one wyraźnie większe, co może wynikać z większej złożoności tej operacji. Warto wspomnieć, że wyniki są niemal niezależne od wielkości danych.

Drugim testowanym przypadkiem był algorytm enkapsulacji klucza ML-KEM. Podobnie jak w przypadku ML-DSA, badania przeprowadzono w zależności od rozmiaru danych, w zakresie do 256 bajtów. Uwzględniono dwie wersje algorytmu, z których każda kolejna oferuje deklarowany wyższy poziom bezpieczeństwa. W ramach

testów analizowano operacje generowania pary kluczy, enkapsulacji klucza oraz dekapsulacji. Wyniki badań dla tego algorytmu przedstawiono na rysunkach 7 oraz 8.



Rysunek 7: Czasy operacji algorytmu ML-KEM-512 z biblioteki OQS w zależności od rozmiaru wiadomości.



Rysunek 8: Czasy operacji algorytmu ML-KEM-1024 z biblioteki OQS w zależności od rozmiaru wiadomości.

Z wykresów wynika, że każda z operacji wykazuje stosunkowo niewielką zależność od rozmiaru danych, a pomiary charakteryzują się niskim poziomem błędów, na co wskazują wąskie przedziały ufności. Badania potwierdzają, że enkapsulacja jest najbardziej czasochłonną operacją, podczas gdy dekapsulacja zajmuje najmniej czasu.

4.3 Elastyczność integracji

Podczas badań zwrócono także uwagę na proces integracji obu bibliotek w rzeczywistych scenariuszach. Biblioteka liboqs okazuje się prostsza we wdrożeniu i integracji w dużych projektach programistycznych. Udostępnia wrappery w wielu językach programowania, nie tylko w Pythonie, oraz oferuje szerszy wybór algorytmów kryptograficznych. Ponadto umożliwia integrację z innymi protokołami, takimi jak OpenSSL, za pomocą oficjalnych wtyczek, podczas gdy TQ42 dostarcza jedynie własne API wraz z przykładami użycia. Ponadto, biblioteka TQ42 jest trudniejsza do uruchomienia, gdyż wymaga samodzielnej kompilacji i uruchamiania skryptów, choć proces ten jest dobrze udokumentowany. Wymaga również licencji komercyjnej lub AGPL v3, natomiast liboqs jest dostępna na licencji MIT. Obie biblioteki są zgodne ze standardami FIPS 203/204/205 [7, 8, 9].

5. PODSUMOWANIE

W przeprowadzonych badaniach porównano dwie biblioteki kryptograficzne w języku Python pod kątem wydajności i stabilności algorytmów postkwantowych ML-KEM i ML-DSA. Analiza wyników wykazała, że biblioteka liboqs przewyższa TQ42 pod względem czasu wykonania operacji. Sprawia to, że jest ona bardziej efektywna w kontekście zastosowań praktycznych. Dodatkowo, liboqs charakteryzuje się mniejszą wrażliwością na zmiany rozmiaru danych, co przyczynia się do stabilności wyników. Ponadto, w przypadku obu bibliotek czas wykonania operacji rósł wraz ze wzrostem poziomu bezpieczeństwa ze względu na większy rozmiar kluczy i związaną z tym większą złożoność obliczeniową. W bibliotece TQ42 różnice między poszczególnymi trybami były bardziej zauważalne niż w liboqs, co wskazuje na nieco mniejszą przewidywalność wyników przy zmianie poziomu bezpieczeństwa.

Podsumowując, przeprowadzone badania wykazały, że biblioteka od Open Quantum Safe zapewnia wyższą wydajność i stabilność czasów wykonania operacji w porównaniu do projektu Terra Quantum Public, szczególnie w kontekście algorytmu ML-KEM. Obie biblioteki wykazują jednak stosunkowo wysoką efektywność, więc wybór zależy od specyficznych wymagań w konkretnych zastosowaniach postkwantowych. Biblioteka TQ42 jest młodsza, co może sugerować, że wciąż znajduje się w fazie rozwoju.

LITERATURA

[1] P. W. Shor. 1994. "Algorithms for Quantum Computation: Discrete Logarithms and Factoring." Symp. on Foundations of Computer Science, Santa Fe, USA.

[2] NIST. 2024. "FIPS 203: Module-Lattice-Based Key-Encapsulation Mechanism Standard", "FIPS 204: Module-Lattice-Based Digital Signature Standard", "FIPS 205: Stateless Hash-Based Digital Signature Standard".

[3] DT. Dam, TH. Tran, VP. Hoang, CK. Pham, TT. Hoang. 2023. "A Survey of Post-Quantum Cryptography: Start of a New Race." Cryptogr. 7 (3): 40.

[4] F. Lauterbach, P. Burdiak, F. Richter, M. Voznak. 2021. "Performance Analysis of Post-Quantum Algorithms." Tel. Forum (TELFOR), Belgrad, Serbia.

[5] F. Opiłka, M. Niemiec, M. Gagliardi, MA. Kourtis. 2024. "Performance Analysis of Post-Quantum Cryptography Algorithms for Digital Signature" Applied Sciences 14, 12: 4994.

[6] D. Dziechciarz, M. Niemiec. 2025. "Efficiency Analysis of NIST-Standardized Post-Quantum Cryptographic Algorithms for Digital Signatures in Various Environments" Electronics 14, 1: 70.

[7] Terra Quantum Public. "TQ42-PQC-OSS" wer. 1.0.0. Dostęp 05.05.2025. https://github.com/terra-quantum-public/tq42-pqc-oss.

[8] Open Quantum Safe. "liboqs" GitHub, Wersja 0.8.0 (2024). Dostęp 05.05.2025. https://github.com/open-quantum-safe/liboqs.

[9] Terra Quantum. "TQ42 Cryptography Library." Dostęp 05.05.2025. https://terra-quantum-public.github.io/tq42-pqc-oss/.

[10] D. Stebila, M. Mosca. 2017. "Post-quantum key exchange for the Internet and the Open Quantum Safe project". Lecture Notes in Comp. Science, 10532.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

Efektywność algorytmów szyfrowania w symulowanym środowisku chmurowym: optymalizacja i dynamiczne zarządzanie zasobami.

Performance of Encryption Algorithms in a Simulated Cloud Environment: Resource Optimization and Dynamic Allocation

Kacper Góra[1]; Jakub Zięba[2]; Aniela Cabak[3] Paweł Rydz[4] Arina Pashkevich[5] Jan Derkacz[6]

[1] AGH, Kraków, kgora@student.agh.edu.pl
[2] AGH, Kraków, ziebajakub@student.agh.edu.pl
[3] AGH, Kraków, acabak@student.agh.edu.pl
[4] AGH, Kraków, rydzpawel@student.agh.edu.pl
[5] AGH, Kraków, pashkevich@student.agh.edu.pl
[6] AGH, Kraków, derkacz@agh.edu.pl

DOI: 10.15199/59.2025.4.44

Streszczenie: Algorytmy szyfrujące stanowią podstawę bezpieczeństwa w sieciach publicznych. Ich dobór wpływa na poziom ochrony transmisji danych. W chmurze, poza bezpieczeństwem, liczy się też efektywność wykorzystania zasobów. Algorytmy różnią się zapotrzebowaniem na zasoby przy różnych obciążeniach.

W pracy przeanalizowano popularne algorytmy szyfrujące i opracowano klasyfikator, który dynamicznie przydziela zasoby (CPU, RAM), zapewniając wydajność i bezpieczeństwo przetwarzania.

Abstract: Encryption algorithms are the foundation of security in public networks. Their selection affects the level of protection for data transmission. In cloud environments, resource efficiency is just as important as security. Different algorithms have varying resource requirements under different workloads.

This work analyzes popular encryption algorithms and presents a classifier that dynamically allocates resources (CPU, RAM), ensuring both processing efficiency and security.

Słowa kluczowe: wydajność przetwarzania, algorytmy szyfrujące, środowisko chmurowe, dynamiczne przydzielanie zasobów

Keywords: processing performance, encryption algorithms, cloud environment, dynamic resource allocation

1. WSTĘP

Obecnie algorytmy szyfrowania pełnią kluczową rolę w zapewnianiu bezpieczeństwa danych przetwarzanych w środowiskach chmurowych [3]. Z uwagi na ograniczenia zasobów wirtualnych instancji, istnieje potrzeba optymalizacji ich przydziału, co jest wielokrotnie wskazywane w artykułach [4, 5, 6]. Celem niniejszej pracy jest zaprezentowanie klasyfikatora, który na podstawie danych wejściowych dobiera optymalne zasoby, zapewniające niskie opóźnienie szyfrowania. Ze względu na zakres pracy, pominięto wpływ wielowątkowości, operacji I/O oraz nie porównywano zastosowanego modelu z innymi podejściami ML, co stanowi kierunek dalszych badań.

2. METODYKA BADAŃ

Badania przeprowadzono w środowisku kontenerowym Docker [7], uruchomionym na maszynie wirtualnej z systemem Kali Linux, wyposażonej w czterowątkowy procesor (2,5 GHz) oraz 8 GB pamięci RAM.

Celem eksperymentu była analiza wpływu przydzielonych zasobów CPU i RAM na opóźnienie operacji, sumy czasu szyfrowania i deszyfrowania danych przy użyciu wybranych algorytmów kryptograficznych [9], a także zebranie próbek niezbędnych do budowy klasyfikatora [1]. W badaniu uwzględniono trzy grupy algorytmów:

- **AES** w trybie GCM: warianty 128-, 192- i 256-bitowe,
- **ECC**: krzywe secp256k1, secp256r1, secp384r1,
- **RSA**: klucze o długości 1024, 2048 i 4096 bitów.

Każda konfiguracja była testowana 100-krotnie w celu oszacowania średnich czasów przetwarzania oraz odchyleń standardowych. Zakres testowanych parametrów obejmował: CPU — od 25 MHz do 3x2500 MHz oraz RAM — od 128 MB do 4096 MB. Dla uproszczenia, w eksperymentach przyjęto sumaryczne wartości mocy obliczeniowej CPU wyrażone w MHz (np. 5000MHz, 7500MHz), niezależnie od liczby rdzeni czy rzeczywistej architektury sprzętowej, traktując je jako logiczny ekwiwalent łącznej dostępnej mocy procesora. Symulację różnych wartości taktowania CPU realizowano poprzez manipulację liczbą rdzeni oraz ograniczeniem czasu dostępu do procesora za pomocą parametrów kontenera Docker. Ze względu na wysokie koszty obliczeniowe algorytmu RSA przy większych plikach, jego testy zostały ograniczone, co można zaobserwować na Rysunku 3.

Zebrane dane zawierały nazwę algorytmu, rozmiar pliku, przydzielone zasoby oraz zmierzone czasy szyfrowania i deszyfrowania. Za konfiguracje optymalne uznawano te, które spełniały kryteria zarówno niskiego opóźnienia, jak i stabilności czasowej, co jest opisane w sekcji 4.1.

Dane (2490 rekordów) zostały przetworzone w Pythonie. Wstępnie zakodowane algorytmy oraz parametry

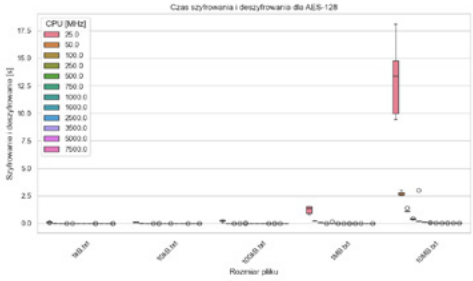
wejściowe (CPU, RAM, STD) wykorzystano do budowy klasyfikatora **RandomForestClassifier** [1, 8], który osiągnął 97,39 % skuteczności na zbiorze testowym.

3. WYNIKI I ANALIZA DANYCH

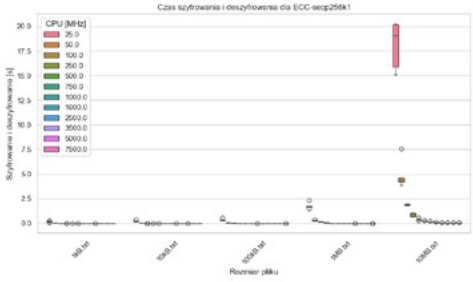
3.1 Zebrane próbki

Wykresy na Rysunkach 1–3 przedstawiają czasy szyfrowania i deszyfrowania dla wszystkich analizowanych algorytmów w zależności od rozmiaru pliku oraz przydzielonych zasobów CPU. Dodatkowo ukazują one, jak zmienia się stabilność przetwarzania, wyrażona poprzez rozrzut wyników (odchylenie standardowe).

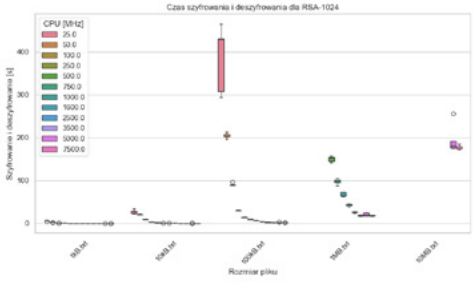
Na Rysunkach 4—9 zastosowano skalę logarytmiczną, aby poprawić czytelność danych, których wartości opóźnień różniły się o kilka rzędów wielkości w zależności od przydzielonych zasobów.



Rysunek 1: Średnie opóźnienie dla algorytmu AES-128 w zależności od rozmiaru pliku i przydzielonego CPU



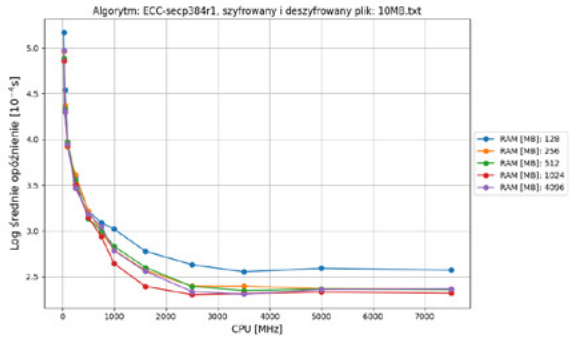
Rysunek 2: Średnie opóźnienie dla algorytmu ECC-secp256k1 w zależności od rozmiaru pliku i przydzielonego CPU



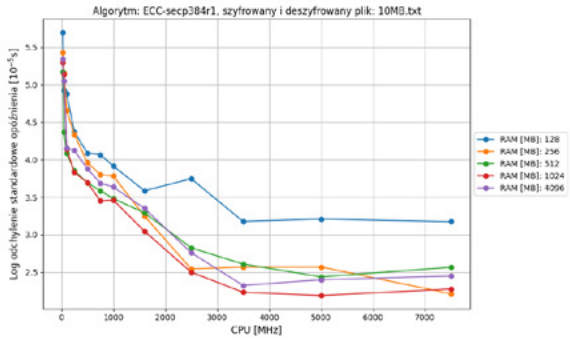
Rysunek 3: Średnie opóźnienie dla algorytmu RSA-1024 w zależności od rozmiaru pliku i przydzielonego CPU

3.2 Pomiary przy zmiennym CPU

Pomiary przedstawione na Rysunkach 4 i 5 ilustrują, jak zmienia się wydajność algorytmu ECC-secp384r1 wraz ze wzrostem przydzielonych zasobów CPU, przy różnych wartościach pamięci RAM i stałym rozmiarze pliku (10MB). Jak pokazuje Rysunek 4, od częstotliwości około 2500 MHz opóźnienie przetwarzania przestaje ulegać istotnej redukcji, co wskazuje na osiągnięcie maksymalnej efektywności algorytmu. Z kolei Rysunek 5 wskazuje, że najlepszą stabilność działania (niskie odchylenie standardowe) uzyskano od wartości CPU około 3500 MHz. Warto również zauważyć, że konfiguracje z pamięcią RAM o wielkości 128 MB znacząco odbiegają od pozostałych i nie są rekomendowane do użytku w kontekście tego algorytmu i rozmiaru danych.



Rysunek 4: Zależność średniego opóźnienia dla algorytmu ECC-secp384r1 od przydzielonych zasobów CPU, przy różnych wartościach pamięci RAM i stałym rozmiarze pliku (10MB.txt).

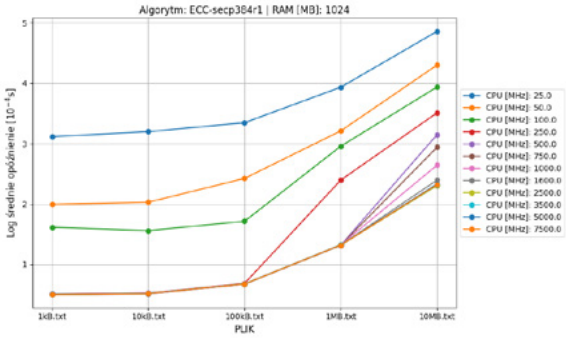


Rysunek 5: Zależność odchylenia standardowego średniego opóźnienia dla algorytmu ECC-secp384r1 od przydzielonych zasobów CPU, przy różnych wartościach pamięci RAM i stałym rozmiarze pliku (10MB.txt).

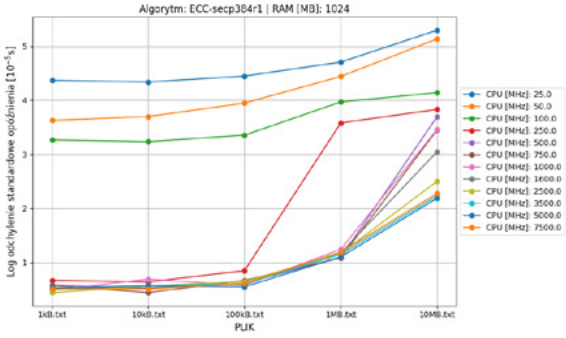
3.3 Pomiary, przy różnych wielkościach plików

Rysunek 6 przedstawia zależność średniego opóźnienia algorytmu ECC-secp384r1 od rozmiaru pliku przy stałej pamięci RAM (1024 MB) i różnych wartościach CPU. Zauważyć można, że dla małych plików ($\leq$ 100 kB) już przy 250 MHz uzyskuje się bardzo niskie opóźnienia. W przypadku większych plików (1,10 MB) wpływ CPU staje się wyraźniejszy — wzrost jego wartości znacząco

poprawia wydajność, aż do momentu nasycenia, gdzie dalsze zwiększanie CPU nie przynosi zauważalnych korzyści. Rysunek 7 obrazuje stabilność działania algorytmu, wyrażoną odchyleniem standardowym opóźnienia, w zależności od rozmiaru pliku i wartości CPU. Już od około 500 MHz obserwuje się relatywnie stabilne wyniki niezależnie od rozmiaru danych. Dla największego pliku (10 MB) różnice w stabilności między konfiguracjami CPU są stosunkowo niewielkie.



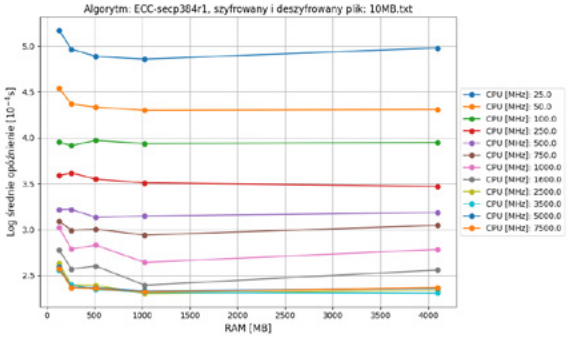
Rysunek 6: Wpływ rozmiaru pliku na średni czas opóźnienia dla algorytmu ECC-secp384r1 przy stałej wartości przydzielonej pamięci RAM i różnych wartościach CPU.



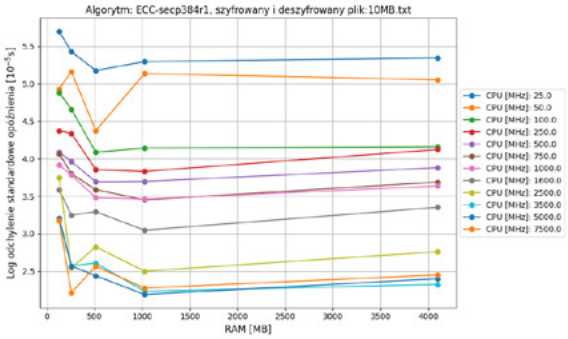
Rysunek 7: Wpływ rozmiaru pliku na odchylenie standardowe czasu opóźnienia dla algorytmu ECC-secp384r1 przy stałej wartości przydzielonej pamięci RAM i różnych wartościach CPU.

3.4 Pomiary przy zmiennym RAM

Rysunki 8 i 9 przedstawiają wpływ wielkości przydzielonej pamięci RAM na średnie opóźnienie (Rysunek 8) oraz jego odchylenie standardowe (Rysunek 9) dla algorytmu ECC-secp384r1, przy różnych wartościach CPU i stałym rozmiarze pliku (10MB). Analiza wyników wskazuje, że zwiększanie ilości pamięci RAM powyżej 1024 MB nie przynosi istotnych korzyści w zakresie poprawy czasu ani stabilności przetwarzania. Różnice w wydajności dla wyższych wartości RAM są minimalne, co sugeruje, że alokacja większej pamięci nie jest konieczna dla tego scenariusza.



Rysunek 8: Średnie opóźnienie dla algorytmu ECC-secp384r1 w funkcji ilości przydzielonej pamięci RAM, przy różnych przydzielonych wartościach CPU i stałym rozmiarze pliku (10MB.txt).



Rysunek 9: Odchylenie standardowe opóźnienia dla algorytmu ECC-secp384r1 w funkcji ilości przydzielonej pamięci RAM, przy różnych przydzielonych wartościach CPU i stałym rozmiarze pliku (10MB.txt).

W ten sposób przeprowadzono również analizę algorytmów AES i RSA. Wizualizacja wyników była niezbędna do podjęcia decyzji dotyczącej wyboru modelu uczenia maszynowego.

4. MODEL KLASYFIKATORA

W celu automatycznego doboru zasobów obliczeniowych dla procesów szyfrowania w środowisku kontenerowym, opracowano klasyfikator oparty na uczeniu maszynowym. Model wyznacza minimalne wartości CPU i RAM, które zapewniają optymalną wydajność dla zadanego algorytmu i wielkości danych.

4.1 Opis danych wejściowych

Zbiór danych wejściowych zawierał 2490 rekordów z polami:

- **Algorytm** – nazwa algorytmu,
- **PLIK** – wielkość pliku (np. 1kB.txt), przekształcona do bajtów,
- **CPU** – przydzielone zasoby procesora (wartość numeryczna),
- **RAM** – przydzielona pamięć operacyjna w MB,
- **MEAN** – średnie opóźnienie (w sekundach),
- **STD** – odchylenie standardowe opóźnienia.

Dane zostały wstępnie przetworzone, na podstawie wartości **MEAN** i **STD** dodano etykietę binarną **OPTYMALNY** z wartością **1**, jeśli spełnione zostały następujące warunki:

- Dla algorytmów AES, ECC przy plikach < 10MB: **MEAN** < 0,01 s oraz **STD** < 0,001 s
- Dla algorytmów AES, ECC przy plikach ≥ 10MB: **MEAN** < 0,1 s oraz **STD** < 0,01 s
- Dla algorytmu RSA: **MEAN** < 1,0 s oraz **STD** < 0,1 s.

Dobór tych progów został dokonany na potrzeby eksperymentu w taki sposób, aby zapewnić możliwie niskie opóźnienie i stabilność działania algorytmów. Zbyt duże wartości mogłyby negatywnie wpłynąć na jakość usług (QoS) świadczonych przez system [2].

4.2 Model i trening

Skorzystano z **RandomForestClassifier** (biblioteka **scikit-learn**) zawierający 100 drzew decyzyjnych.

Model został wytrenowany na zbiorze treningowym (80 % danych) i przetestowany na pozostałych 20 %. Uzyskano dokładność klasyfikacji równą 97,39 % na zbiorze testowym.

Zastosowano klasyfikator Random Forest jako model dobrze dopasowany do charakterystyki zbioru danych, działający w ograniczonej przestrzeni cech bez potrzeby ich transformacji. Wybór ten umożliwił zachowanie interpretowalności wyników oraz kontrolę nad ryzykiem przeuczenia. Porównanie z innymi modelami, takimi jak XGBoost, SVM czy sieci neuronowe, stanowi potencjalny kierunek dalszych badań.

4.3 Dobór optymalnych parametrów

Opracowano również funkcję **predict_optimality()**, która dla danego algorytmu i wielkości pliku przeszukuje przestrzeń dostępnych zasobów CPU i RAM w celu znalezienia minimalnej konfiguracji, która zostanie zaklasyfikowana jako optymalna.

Dla analizowanych algorytmów określa się wielkość pliku, maksymalną wartość CPU oraz maksymalną ilość pamięci RAM, którą można przydzielić. Przykładowe wywołanie funkcji przedstawiono na listingu 1.

```
1 predict_optimality("ECC-secp384r1", 10000000, 7500.0, 4096)
2 Output:
3 {'status': 'OPTYMALNY', 'cpu': np.float64(2500.0), 'ram': np.int64(256)}
```

Listing 1: Wywołanie funkcji i jej wynik

Zwrócony wynik oznacza, że dla danego obciążenia (tj. wielkości pliku) optymalne parametry to:

- CPU: 2,5 GHz
- RAM: 256 MB

5. WNIOSKI I KIERUNKI DALSZYCH BADAŃ

5.1 Wnioski

Przeprowadzone badania pozwoliły zaobserwować, że każdy z analizowanych algorytmów szyfrowania charakteryzuje się odmiennymi właściwościami wydajnościowymi i wymaganiami zasobowymi. Algorytm AES okazał się najszybszym rozwiązaniem, dobrze skalującym się przy różnych rozmiarach plików i przydziałach CPU, co czyni go najbardziej uniwersalnym. Algorytmy oparte na krzywych eliptycznych (ECC) zapewniają wysokie bezpieczeństwo przy relatywnie niskim zużyciu zasobów, jednak są wolniejsze niż AES i bardziej wrażliwe na konfigurację środowiska. RSA, ze względu na znaczne obciążenie obliczeniowe, nadaje się przede wszystkim do obsługi małych plików, takich jak podpisy elektroniczne lub certyfikaty.

Istotnym rezultatem pracy było opracowanie funkcji predykcyjnej, umożliwiającej automatyzację przydzielania zasobów CPU i RAM do kontenerów w zależności od rodzaju algorytmu i charakterystyki obciążenia. Zastosowanie takiego podejścia może znacząco poprawić efektywność przetwarzania w systemach chmurowych.

5.2 Kierunki dalszych badań

W przyszłości planowane są następujące rozszerzenia pracy:

- przeprowadzenie testów z większymi plikami (powyżej 10 MB), w celu rozszerzenia zakresu predykcji i poprawy dokładności modelu;
- weryfikacja skuteczności modelu i funkcji predykcyjnej w rzeczywistym środowisku chmurowym z większą liczbą dostępnych zasobów oraz zmiennym obciążeniem sieci i systemu w czasie rzeczywistym;
- rozbudowa klasyfikatora o dodatkowe cechy, takie jak obciążenie I/O lub wielowątkowość, oraz zastosowanie bardziej zaawansowanych modeli uczenia, np. XGBoost lub sieci neuronowych.

LITERATURA

[1] Kuhn Max , Johnson Kjell. 2013. „Applied Predictive Modeling". Springer.

[2] Silva Natassya B.F., Pigatto Daniel F., Martins Paulo S., Branco Kalinka R.L.J.C. 2016. „Case studies of performance evaluation of cryptographic algorithms for an embedded system and a general purpose computer". Journal of Network and Computer Applications, 60: 130–143.

[3] Singh Ashish, Chatterjee Kakali . 2017. „Cloud security issues and challenges: A survey". Journal of Network and Computer Applications, 79: 88–115.

[4] Li Kun, Xu Gaochao, Zhao Guangyu, Dong Yushuang, Wang Dan. 2011. „Cloud Task Scheduling Based on Load Balancing Ant Colony Optimization". 2011 Sixth Annual Chinagrid Conference, 3–9.

[5] Rodriguez Maria Alejandra , Buyya Rajkumar. 2014. „Deadline Based Resource Provisioning and Scheduling Algorithm for Scientific Workflows on Clouds". IEEE Transactions on Cloud Computing, 2, (2): 222–235.

[6] Beloglazov Anton , Abawajy Jemal , Buyya Rajkumar . 2012. „Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing". Future Generation Computer Systems, 28 (5): 755–768.

[7] Schenker Gabriel N. 2020. „Learn Docker – Fundamentals of Docker 19.X: Build, Test, Ship, and Run Containers with Docker and Kubernetes". Birmingham: Packt Publishing.

[8] Breiman Leo. 2001. „Random Forests". Machine Learning, 45 (1): 5–32.

[9] Paar Christof, Pelzl Jan. 2010. „Understanding Cryptography: A Textbook for Students and Practitioners". Springer: 87 - 115, 173 - 197, 239 - 253.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

IMPLEMENTACJA I PORÓWNANIE ATAKÓW KLASYCZNYCH NA KRYPTOGRAFIĘ POSTKWANTOWĄ OPARTĄ NA KODACH

IMPLEMENTATION AND COMPARISON OF CLASSICAL ATTACKS ON CODE-BASED POST-QUANTUM CRYPTOGRAPHY

Jakub Łomzik[1], Piotr Chołda[2]

AGH Akademia Górniczo-Hutnicza, Instytut Telekomunikacji, al. Mickiewicza 30, 30-059 Kraków
[1] jakublomzik@student.agh.edu.pl, [2] piotr.cholda@agh.edu.pl

DOI: 10.15199/59.2025.4.45

Streszczenie: Niniejszy artykuł bada działanie niekwantowych (klasycznych) ataków na postkwantowe systemy kryptograficzne oparte na kodach korekcyjnych. Obejmuje analizę kilku algorytmów do przeprowadzania takich ataków. Wyniki wygenerowane przez program porównano z teoretycznym zachowaniem algorytmów. Rezultaty analizy obrazują sposób działania algorytmów i potencjał dalszej rozbudowy narzędzi.

Abstract: This article examines the performance of non-quantum (classical) attacks on post-quantum cryptographic systems based on correction codes. It provides a comprehensive overview of the field, including an analysis of several algorithms used to perform such attacks. The results generated by the implementation were compared with the theoretical behavior of the algorithms. Findings from the analysis illustrate the functioning of the algorithms and the potential for further development.

Słowa kluczowe: kody korekcyjne, kryptografia postkwantowa

Keywords: error correction codes, post-quantum cryptography

1. WSTĘP

W czasie trwania wyścigu kwantowego mającego na celu jak najszybsze skonstruowanie użytecznego komputera kwantowego, odbywa się również wyścig w kierunku przeciwnym. Rozwój kryptografii postkwantowej ma za zadanie uchronić przed atakami z wykorzystaniem takich zaawansowanych jednostek obliczeniowych. Amerykański Narodowy Instytut Standardów i Technologii NIST wyróżnił już dwa algorytmy kryptograficzne — Kyber i Dilithium — które zostały zintegrowane z *Protokołem Signal* Fundacji Signal [6] i zapewniają bezpieczeństwo w aplikacji o tej samej nazwie. Mechanizmy te opierają się na obliczeniowej trudności rozwiązywania określonych problemów w strukturach algebraicznych znanych jako kraty; istnieją jednak również metodologie oparte na innych, podobnie złożonych wyzwaniach. Wariant jednego z najstarszych (bo zaproponowanych już w latach 70.), dobrze ugruntowanego kryptosystemu klucza publicznego autorstwa McEliece'a, zwany w literaturze po prostu *Classic McEliece*, również wykazuje istotny potencjał w odpieraniu zagrożeń kwantowych [2]. Ostatnio, bo w marcu 2025 r., metoda *Hamming Quasi-Cyclic* HQC — mechanizm enkapsulacji klucza — została przyjęta przez NIST do dalszej standaryzacji [8]. W przeciwieństwie do podejść opartych na kratach, konstrukcje te oparte są na teorii kodowania (w szczególności jej aspektach związanych z liniowymi kodami korekcyjnymi), dyscyplinie dobrze znanej specjalistom z zakresu radiokomunikacji i teleinformatyki.

Celem badań była analiza i porównanie charakterystyk algorytmów typu „ISD". Podyktowane było to potrzebą uzyskania empirycznych danych dotyczących złożoności czasowej i skuteczności ataków realizowanych z wykorzystaniem tych algorytmów. Zakres prac obejmował implementację trzech algorytmów dekodowania, generowanie kodów Hamminga i analizę statystyczną uzyskanych wyników.

Poniżej przedstawiono organizację artykułu. Sekcja 2 nt. metod opisuje narzędzie testowe specjalnie opracowane na potrzeby tego badania, wraz z zaimplementowanymi algorytmami i ich parametrami. Omówienie wyników zawarte w sekcji 3 przedstawia i ocenia empiryczne ustalenia w zestawieniu z oczekiwaniami teoretycznymi. Wreszcie, podsumowanie zawarte w końcowej sekcji 4 zbiera główne rezultaty badania i nakreśla perspektywiczne kierunki dalszego rozwoju oraz lepszego zrozumienia badanej dziedziny.

2. METODY

Poniżej przedstawiono przede wszystkim badane algorytmy, jak również pewne istotne informacje nt. użytej implementacji.

2.1 Algorytmy

System kryptograficzny McEliece (1978) zaproponowany przez Roberta McEliece'a, jest schematem szyfrowania opartym na kodach korekcyjnych, którego bezpieczeństwo względne opiera się na trudności dekodowania kodu liniowego na podstawie syndromu. System obejmuje użycie kodów korekcyjnych do szyfrowania i deszyfrowania wiadomości. Schemat jest skonstruowany następująco [7]:

Generowanie klucza
Odbiorca tworzy następujące tajne komponenty:

- kod korekcyjny: binarny kod korekcyjny (w oryginale: Goppy, ale można stosować dowolny) z parametrami (n, k, t), gdzie n to długość słowa kodowego, k to wymiar kodu, a t to zdolność korekcji błędów bitowych polegających na zamianie symbolu (j. ang. *bit flipping*). Kod ten jest reprezentowany przez macierz generującą $\mathbf{G}'$ nad ciałem binarnym $GF(2)$,
- macierz mieszająca $\mathbf{S}$: losowo wybrana odwracalna macierz $k \times k$ nad $GF(2)$, używana do ukrywania struktury $\mathbf{G}'$,
- macierz permutacji $\mathbf{P}$: macierz permutacji $n \times n$, która miesza kolejność współrzędnych kodu.

Publiczna macierz generująca $\mathbf{G}$ jest skonstruowana następująco:

$$\mathbf{G} = \mathbf{S}\mathbf{G}'\mathbf{P} \tag{1}$$

Macierz $\mathbf{G}$ jest publikowana jako klucz publiczny, podczas gdy macierze $\mathbf{S}$, $\mathbf{G}'$ i $\mathbf{P}$ są przechowywane w tajemnicy jako klucz prywatny.

Szyfrowanie
Aby zaszyfrować wektor wiadomości $\mathbf{m} \in GF(2)^k$, nadawca:

1. oblicza słowo kodowe $\mathbf{c}_0 = \mathbf{m}\mathbf{G}$;
2. celowo wprowadza wektor błędu $\mathbf{e} \in GF(2)^n$ o wadze t lub mniejszej;
3. tworzy szyfrogram $\mathbf{c} = \mathbf{c}_0 + \mathbf{e}$, który jest następnie wysyłany do odbiorcy.

Deszyfrowanie
Po otrzymaniu szyfrogramu $\mathbf{c}$, odbiorca:

1. stosuje permutację odwrotną $\mathbf{P}^{-1}$ do $\mathbf{c}$: $\mathbf{c}' = \mathbf{c}\mathbf{P}^{-1}$;
2. używa opartego na regule największego prawdopodobieństwa algorytmu dekodowania dla oryginalnego kodu Goppy (opisywanego z użyciem macierzy $\mathbf{G}'$) do odzyskania permutowanej wiadomości $\mathbf{m}\mathbf{S}$ i wektora błędu $\mathbf{e}\mathbf{P}^{-1}$;
3. usuwa efekt mieszania $\mathbf{S}$ przez obliczenie: $\mathbf{m} = (\mathbf{m}\mathbf{S})\mathbf{S}^{-1}$.

Information Set Decoding ISD [1] to rodzina technik algorytmicznych zaprojektowanych do rozwiązania ogólnego problemu dekodowania liniowych kodów korekcyjnych. Jest szczególnie istotna w kryptoanalizie, ponieważ stanowi podstawę ataków na systemy kryptograficzne oparte na kodach, takie jak opisany wcześniej system McEliece'a. ISD ma na celu zdekodowanie zadanego zaszumionego słowa kodowego $\mathbf{c}$ przez wykorzystanie struktury kodu. Obejmuje to znalezienie oryginalnego wektora wiadomości $\mathbf{m}$ pomimo błędów wprowadzonych przez wektor $\mathbf{e}$. Fundamentalnym założeniem ISD jest, że choć dekodowanie jest generalnie trudne, może stać się efektywne dzięki skupieniu się na starannie wybranym podzbiorze pozycji słowa kodowego, zwanym *zbiorem informacji*, ponieważ błędy w wybranych pozycjach są łatwiejsze do wyizolowania i skorygowania.

Algorytm Prange'a [9] (1962) jest pionierskim algorytmem ISD, który wprowadził wspomnianą koncepcję zbioru informacji. Wykorzystując atak siłowy, rozwiązanie obejmuje losowy wybór podzbioru i zastosowanie do dekodowania eliminacji Gaussa. Metoda ta ma charakter probabilistyczny ze względu na stochastyczny wybór zbioru informacji. Pomimo wysokiej efektywności pod względem wykorzystania pamięci, jej złożoność czasowa wzrasta wykładniczo [1] i jest opisywana jako $\mathcal{O}(2^{\beta n})$.

Algorytm Sterna [10] (1989) jest udoskonaleniem metody Prange'a przez integrację kompromisów probabilistycznych opartych na losowym podziale pozycji kodu. Ta metoda skutecznie zmniejsza przestrzeń poszukiwań, jednocześnie wykazując wadę w postaci istotnie zwiększonego zużycia pamięci [1]. W konsekwencji złożoność czasowa zostaje zredukowana do $\mathcal{O}(2^{0,05563n})$.

Algorytm Lee-Brickella [4] (1988) ulepsza poprzednie algorytmy ISD dzięki integracji częściowych weryfikacji zbiorów informacji wraz z zaawansowanymi metodami kombinatorycznymi w celu zmniejszenia złożoności czasowej. Główna innowacja tkwi w implementacji bardziej wyrafinowanej probabilistycznej eksploracji wzorców błędów. Chociaż złożoność czasowa algorytmu jest nieznacznie gorsza niż w przypadku alg. Sterna [1] i wynosi $\mathcal{O}(2^{0,05563n})$, to jednak wymaga mniejszego zużycia pamięci.

Ball-collision decoding [3] (2011) zmniejsza asymptotyczną złożoność dekodowania w porównaniu do klasycznych metod ISD, takich jak alg. Sterna czy Lee-Brickella, wprowadzając krok wykrywania kolizji. Dzieli wektor błędu na mniejsze rozłączne podzbiory i szuka częściowych rozwiązań w sferach Hamminga (j. ang. *Hamming balls*), wykrywając prawidłowe rozwiązania poprzez kolizje. Jego złożoność czasowa jest ponownie nieco niższa [1], na poziomie $\mathcal{O}(2^{0,05558n})$.

Algorytm May-Meurer-Thomae MMT [5] (2011) jest zaawansowaną techniką wykorzystującą znaczne zasoby pamięci do zmniejszenia czasu dekodowania. Podobnie jak algorytm ball-collision decoding, MMT koncentruje się na znajdowaniu kolizji w ustrukturyzowanych podzbiorach rozwiązań potencjalnych (kandydatów). Algorytm rozszerza jednak to podejście przez iteracyjne udoskonalanie procesu wyszukiwania, wykorzystując nakładanie się częściowych wektorów błędu w przestrzeni rozwiązań w celu maksymalizacji efektywności. MMT zawiera zaawansowane techniki kombinatoryczne, takie jak metody przesiewania, aby wcześnie w procesie wyeliminować kandydatów niedopuszczalnych. Wykorzystując te techniki, algorytm skuteczniej zawęża możliwe wzorce błędów. Jego optymalizacje dodatkowo zmniejszają asymptotyczną złożoność czasową najgorszego przypadku [1] do $\mathcal{O}(2^{0,05363n})$, wymagając za to najwięcej pamięci spośród opisywanych algorytmów.

2.2 Narzędzie informatyczne

W celu zaprezentowania i porównania efektywności badanych metod opracowano w języku programowania Rust dedykowane narzędzie obliczeniowe. Wykorzystany język Rust jako podstawa zapewnia bezpieczną pod względem zarządzania pamięcią implementację operacji kryptograficznych. Główna funkcjonalność obejmuje wykonywanie klasycznych ataków przy użyciu probabilistycznych algorytmów dekodowania zbioru informacji ISD z parametrami wejściowymi określonymi przez użytkownika. Obecnie istnieje możliwość wykorzystania losowych kodów liniowych, kodów Hamminga lub kodów Goppy. Zbierane są wszystkie istotne informacje dotyczące kodów wykorzystywanych w operacjach, a także rezultaty tych operacji (wskazujące na sukces lub niepowodzenie deszyfrowania) oraz czas trwania ataku mierzony w mikrosekundach

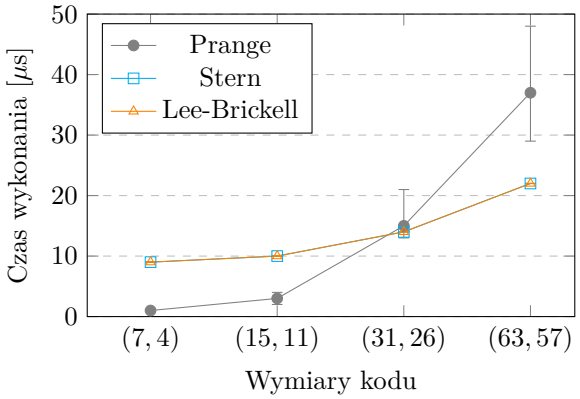
[μs]. Aby ocenić jakość algorytmu i jego implementacji, niezbędne jest zmierzenie dwóch fundamentalnych aspektów jego wykonania: czasu działania oraz zużycia pamięci. Ze względu na rozległy zakres implementacji, zaprezentowane tutaj badania ograniczają się do porównywania algorytmów wyłącznie na podstawie zużycia czasu. Nawet sam precyzyjny pomiar czasu wykonania dostarcza wielu cennych informacji na temat charakterystyki ataków.

3. OMÓWIENIE WYNIKÓW PORÓWNAWCZYCH

Precyzyjny pomiar czasu wykonania pojedynczej iteracji wybranego algorytmu jest stosunkowo prosty, ponieważ program wykonuje funkcję odpowiedzialną za kroki algorytmu tylko raz. Wyzwania pojawiają się przy próbie sekwencyjnego wykonania tego zestawu działań wielokrotnie: środowisko ma tendencję do znaczącego wpływania na wynik programu, co sugeruje, że optymalizacja zachodzi już po pierwszych kilku wykonaniach. Prawdopodobnie dzieje się tak z powodu wartości pozostałych w pamięci procesu wykorzystywanego przez kolejne wywołania. Takie podejście może znacząco zniekształcić postrzeganie wydajności algorytmu i podkreśla konieczność utworzenia hermetycznego i izolowanego środowiska do testowania na dużą skalę. Biorąc pod uwagę probabilistyczny charakter algorytmów, ocena ich rzeczywistej wydajności wymaga testowania z wykorzystaniem większej liczby próbek.

3.1 Skalowanie parametrów kodu

W badaniu reakcji algorytmów na skalowanie parametrów kodu wydajność algorytmów Prange'a, Sterna i Lee-Brickella była oceniana przy użyciu podstawowych kodów, w szczególności kodów Hamminga, przy zachowaniu spójnej typowej wagi 1 (rys. 1). Szczególnie godne uwagi jest to, że dla niższych wartości parametrów n i k, uproszczone początkowe podejście Prange'a daje znacznie korzystniejsze czasy wykonania, do około 80% szybsze niż bardziej zaawansowane algorytmy. Jednakże przewaga algorytmu Prange'a szybko zanika w związku z wykładniczym wzrostem jego złożoności obliczeniowej. W konsekwencji algorytmy wykorzystujące techniki optymalizacyjne — metody Sterna i Lee–Brickella — zyskują na efektywności, wykazując łagodniejszy przyrost czasu wykonania. Dodatkowo zauważalne są stopniowe różnice pomiędzy tymi dwoma podejściami — algorytm Lee–Brickella okazuje się korzystniejszy od metody Sterna, szczególnie w przypadku kodów o mniejszych parametrach. Niemniej jednak, wraz ze wzrostem rozmiarów kodów (co ma miejsce w praktyce kryptograficznej), metryki wydajności algorytmów wykazują tendencję do zbieżności.



Rysunek 1: Czas wykonania alg. korekcji w zależności od parametrów kodu (n, k)

Algorytm Prange'a wykazuje współczynnik powodzenia zbliżony do 100% dla wszystkich testowanych rozmiarów kodów (tab. 1). Można to przypisać stosowanemu przez niego wyczerpującemu przeszukiwaniu, która zapewnia identyfikację prawidłowego rozwiązania (pod warunkiem dysponowania odpowiednimi zasobami obliczeniowymi). Nałożyliśmy dodatkowe ograniczenie wymagające, aby algorytm dostarczył rozwiązanie w ciągu pierwszych 100 prób. To ograniczenie zostało wprowadzone w celu zilustrowania limitów algorytmu podczas skalowania parametrów wejściowych. Zastosowanie małych, stosunkowo prostych, kodów Hamminga dodatkowo przyczynia się do obserwowanego spójnego powodzenia.

Tabela 1: Współczynnik sukcesu w zależności od parametrów (n, k) użytego kodu korekcyjnego

Parametry kodu	*Prange*	*Stern*	*Lee-Brickell*
$(7, 4)$	100%	84%	80%
$(15, 11)$	100%	67%	72%
$(31, 26)$	99%	59%	61%
$(63, 57)$	83%	54%	53%

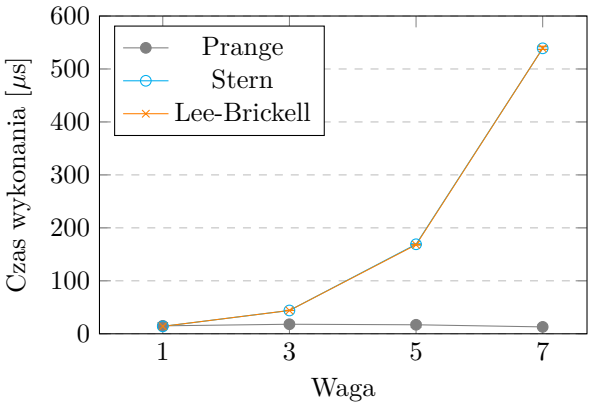
Chociaż algorytm Sterna jest szybszy niż algorytm Prange'a, jego probabilistyczny charakter prowadzi do sporadycznych niepowodzeń przy większych kodach, tym samym zmniejszając wskaźnik sukcesu do niewiele ponad 50%. Algorytm Lee-Brickella wykazuje podobną wydajność, ponieważ naturalne mocne strony obu algorytmów nie są optymalnie dopasowane do małych wymiarów i prostej struktury kodów Hamminga.

3.2 Skalowanie wagi błędu

Zachowanie algorytmów zmienia się, gdy kod wejściowy pozostaje stały (z $n = 31$ i $k = 26$), a waga, czyli liczba wprowadzonych błędów w ciągu $\mathbf{e}$, jest stopniowo zwiększana. W kontekście kodów Hamminga scenariusz ten można uznać za nietypowy ze względu na ich nieodłączne ograniczenie polegające na korygowaniu tylko pojedynczego wprowadzonego błędu (tutaj używamy tego dobrze znanego kodu jedynie ilustracyjnie). Niemniej jednak prowadzi to do godnych uwagi wyników.

W przypadku algorytmu Prange'a czas wykonania pozostaje prawie niezmienny w stosunku do zmian wagi (rys. 2). Jest to spodziewany rezultat, ponieważ algo-

rytm wyczerpująco przeszukuje całą przestrzeń potencjalnych wektorów błędów, nie wykorzystując jawnie wagi do zwiększenia efektywności wyszukiwania. W tym przypadku złożoność czasowa jest głównie zdeterminowana przez wymiarowość kodu, a nie przez wagę błędu. W konsekwencji podejście Prange'a nie jest szczególnie korzystne w zastosowaniach praktycznych, gdzie oczywiście używane będą kody o dużo większych parametrach.



Rysunek 2: Czas wykonania w zależności od liczby wprowadzonych błędów, tj. wagi ciągu **e**

W przeciwieństwie do tego, zarówno algorytm Sterna, jak i Lee-Brickella wykazują istotny i niemal identyczny wzrost czasu wykonania, odzwierciedlając ich faktyczną zależność od parametru wagi. W tym przypadku optymalizacje zaimplementowane w nowszych metodach wydają się nieskuteczne.

Wskaźnik powodzenia w tym scenariuszu wyraźnie ilustruje cechy strukturalne wszystkich testowanych algorytmów (tab. 2). Podczas gdy algorytm Prange'a utrzymuje stałą wydajność niezależnie od zwiększonej wagi, algorytmy Sterna i Lee-Brickella osiągają doskonałe wskaźniki powodzenia, demonstrując ich zależność od liczby wprowadzonych błędów w celu ułatwienia optymalizacji.

Tabela 2: Współczynnik sukcesu w zależności od liczby wprowadzonych błędów przez ciąg **e**

Liczba błędów	*Prange*	*Stern*	*Lee-Brickell*
1	99%	59%	61%
3	98%	100%	100%
5	96%	100%	100%
7	98%	100%	100%

Wskazane wyniki porównawcze dotyczące jakości działania istotnie pokazują uzyskiwanie wydajności zgodnej z przewidywaniami teoretycznymi.

4. PODSUMOWANIE

Opracowane wyniki pokazują potencjał w wyjaśnianiu klasycznych ataków na kryptosystemy oparte na kodach korekcyjnych oraz w określaniu ich kluczowych różnic. Przedstawione wyniki podkreślają znaczenie zarówno kryptosystemów odpornych na kryptoanalizę z użyciem komputerów kwantowych, jak i potencjalnych zagrożeń wynikających z użycia badanych tutaj algorytmów niekwantowych (klasycznych). Podstawy systemu kryptograficznego McEliece'a pozostają bezpieczne mimo pojawienia się bardziej współczesnych metodologii mających na celu ochronę prywatności przed rosnącym zagrożeniem ze strony komputerów kwantowych.

Przyszłe kierunki badań nad wykorzystaniem tego narzędzia skupiają się głównie na bardziej zaawansowanych kodach, takich jak kody Goppy, stosowane w *Classic McEliece*, czy kody quasi-cykliczne, będące podstawą HQC.

Narzędzie zostało również wzbogacone o stosunkowo niedawno zaproponowane metody ataku, takie jak *ball-collision decoding* czy algorytm May-Meurer-Thomae, co umożliwia bardziej kompleksowe analizy porównawcze. Ważnym obszarem badań jest także ocena wydajności probabilistycznych ataków w zestawieniu z dekodowaniem bezpośrednim, opartym na przykład na klasycznym algorytmie Pattersona.

Nie można pominąć kwestii związanych z wykorzystaniem pamięci, która odgrywa kluczową rolę w rozwoju nowoczesnych technik ataku. Optymalizacja pamięci pozostaje istotnym wyzwaniem, które może znacząco wpłynąć na efektywność przyszłych rozwiązań, a której badania są przedmiotem kolejnych kroków.

LITERATURA

[1] Baldi, M., Barenghi, A., Chiaraluce, F., Pelosi, G., Santini, P. 2019. „A Finite Regime Analysis of Information Set Decoding Algorithms" Algorithms 12, no. 10: 209.

[2] Bernstein, D.J., Chou, T., Lange, T., von Maurich, I., Misoczki, R., Niederhagen, R., Persichetti, E., Peters, C., Schwabe, P., Sendrier, N., Szefer, J., Wang, W. 2022. „Classic McEliece: conservative code-based cryptography: cryptosystem specification" Classic McEliece Documentation.

[3] Bernstein, D.J., Lange, T., Peters, C. 2011. „Smaller Decoding Exponents: Ball-Collision Decoding." Advances in Cryptology – CRYPTO 2011, vol. 6841: 743–760.

[4] Lee, P.J., Brickell, E.F. 1988. „An Observation on the Security of McEliece's Public-Key Cryptosystem." Advances in Cryptology — EUROCRYPT '88, vol. 330: 275–280.

[5] May, A., Meurer, A., Thomae, E. 2011. „Decoding Random Linear Codes in $\tilde{\mathcal{O}}(2^{0.054n})$." Advances in Cryptology – ASIACRYPT 2011, vol. 7073: 107–124.

[6] Marlinspike, M., Perrin, T. 2016. „Post-Quantum Cryptography for the Next-Generation Signal Protocol." Signal. https://signal.org/blog/pqxdh/.

[7] McEliece, Robert J. 1978. „A Public-Key Cryptosystem Based on Algebraic Coding Theory." DSN Progress Report 42-44, Communications Systems Research Section.

[8] National Institute of Standards and Technology (NIST). 2025. „HQC Announced as a 4th Round Selection." NIST Computer Security Resource Center. https://csrc.nist.gov/news/2025/hqc-announced-as-a-4th-round-selection.

[9] Prange, E. 1962. „The Use of Information Sets in Decoding Cyclic Codes." IRE Transactions on Information Theory 8 (5): 5–9.

[10] Stern, J. 1989. „A Method for Finding Codewords of Small Weight." Coding Theory and Applications 388 (1): 106–113.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

NEURONOWY MECHANIZM SZYFROWANIA RUCHU SIECIOWEGO UŻYTKOWNIKA W SIECIACH TELEKOMUNIKACYJNYCH

NEURAL MECHANISM OF ENCRYPTING USER NETWORK TRAFFIC IN TELECOMMUNICATION NETWORKS

Maciej Kaczyński[1]; Zbigniew Piotrowski[2]

[1] Wojskowa Akademia Techniczna, Warszawa, maciej.kaczynski@wat.edu.pl
[2] Wojskowa Akademia Techniczna, Warszawa, zbigniew.piotrowski@wat.edu.pl

DOI: 10.15199/59.2025.4.46

Streszczenie: **Szyfrowanie ruchu sieciowego użytkownika w sieciach telekomunikacyjnych jest niezbędne do ochrony prywatności i zapewnienia bezpieczeństwa przesyłanych danych. Celem niniejszego artykułu jest przedstawienie neuronowego mechanizmu szyfrowania ruchu sieciowego użytkownika w sieciach telekomunikacyjnych, bazującego na głębokiej sieci neuronowej pełniącej rolę klucza, który jednocześnie działa jako generator kluczy, oraz na tokenie zmieniającym klucz. Zapewnia to dwustopniową weryfikację tożsamości użytkownika.**

Abstract: **Encrypting user network traffic in telecommunication networks is essential for protecting privacy and ensuring the security of transmitted data. This article presents a neural-based mechanism for encrypting user network traffic in telecommunication networks, based on a deep neural network that functions as a key and simultaneously acts as a key generator, along with a token that modifies the key. This approach ensures two-factor user authentication.**

Słowa kluczowe: **kryptografia, szyfrowanie, utwardzanie systemu, ruch sieciowy, sieć neuronowa.**

Keywords: **cryptography, encryption, system hardening, network traffic, neural network.**

1. WSTĘP

Kryptografia neuronowa (ang. neural cryptography) to interdyscyplinarna dziedzina łącząca kryptografię z sieciami neuronowymi (SN) (ang. neural networks) w celu zabezpieczenia komunikacji. Oferuje ona obiecujące perspektywy dla przyszłości cyberbezpieczeństwa, łącząc teorię SN z praktycznymi zastosowaniami kryptografii. Głównym celem tej techniki jest tworzenie bezpiecznego mechanizmu wymiany kluczy, który zazwyczaj nie wymaga dzielenia się informacją o sekretnych danych (np. kluczu prywatnym) a priori między komunikującymi się stronami. Znaczenie kryptografii neuronowej wynika z jej potencjału tworzenia systemów bezpieczeństwa odpornych na tradycyjne metody ataków, takie jak ataki siłowe lub kryptoanaliza. Wykorzystując adaptacyjne właściwości SN, systemy te mogą dynamicznie dostosowywać się do zmieniających się warunków i zagrożeń.

Siłą kryptografii neuronowej jest jej zdolność do generowania bezpiecznych kluczy w sposób adaptacyjny, co jest trudne do osiągnięcia w tradycyjnych systemach kryptograficznych. Ponadto ta metoda zazwyczaj nie wymaga żadnych wcześniej uzgodnionych sekretnych danych między stronami, co zwiększa jej użyteczność w niezabezpieczonych środowiskach. Kryptografia neuronowa ma jednak również swoje słabości. Metody te mogą wymagać większych zasobów obliczeniowych i czasu na synchronizację kluczy, co może być niepraktyczne w przypadku niektórych zastosowań. Ponadto, jako stosunkowo nowa dziedzina, posiada niezbadane obszary, które mogą ukrywać nieznane podatności.

Jednym z najbardziej znanych zastosowań kryptografii neuronowej jest protokół Tree Parity Machine (TPM) [1], który wykorzystuje proces uczenia się sieci neuronowych (SN) do synchronizacji kluczy między stronami. Każda strona inicjuje swoją SN losowo wybranymi wagami. Strony wymieniają się wynikami sieci, ale nie samymi wagami, które pozostają tajne. Sieci dostosowują swoje wagi na podstawie otrzymywanych danych wejściowych i wyjściowych. Wagi są synchronizowane tylko wtedy, gdy dane wyjściowe sieci są identyczne. W przeciwnym razie pozostają niezmienione. W tym protokole dwie SN uczą się synchronicznie, osiągając punkt, w którym ich wagi sieciowe są identyczne i mogą służyć jako wspólny klucz tajny. Więcej informacji na temat wymiany tajnych kluczy opartej na synchronizacji neuronowej przez kanały publiczne znajduje się w przeglądzie napisanym przez Chakraborty i in. [2].

Inne techniki stosowane w kryptografii neuronowej często obejmują wykorzystanie generatywnych sieci adwersarskich (ang. generative adversarial networks) [3]. Istnieją również mniej standardowe rozwiązania związane z kryptografią i SN [4]. W tym rozwiązaniu przedstawiono koncepcję systemu polegającą na zastosowaniu SN jako elementu zmieniającego znaki ASCII na wejściu SN (ciąg 8-bitowy) na inny ciąg 8-bitowy w taki sposób, aby każdemu znakowi była przypisana nowa, niepowtarzalna wartość w zakresie ⟨0; 255⟩. Następnie, poprzez zastosowanie operacji modulo (mającej na celu ochronę przed analizą częstotliwości występowania liter w tekście) oraz przetwarzania etapowego (gdzie wektor danych, np. 16-bitowy, dzielony jest po 8 bitów i podawany na wejście różnych SN zmieniających ciągi bitowe w inny sposób), następuje wzmacnianie algorytmu szyfrującego. W tej metodzie SN można traktować jako tablicę konwer-

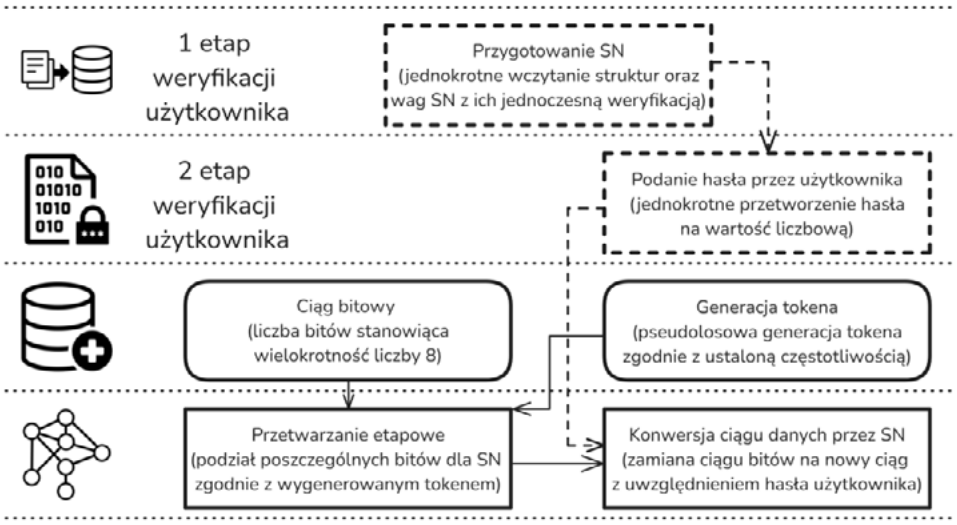
sji. Metoda ta wspiera zrównoleglenie obliczeń, umożliwiając szybkie wykonanie obliczeń na wielowątkowych jednostkach obliczeniowych. Kolejną zaletą tego rozwiązania jest jego niestandardowość oraz to, że nieznajomość architektury sieci i posiadanie jedynie pliku z wagami (lub odwrotnie – tylko z architekturą) nie umożliwia bezpośredniego odtworzenia działania SN, a tym samym sposobu konwersji ciągów bitów.

Proponowana w tym artykule metoda ma na celu przedstawienie neuronowego mechanizmu szyfrowania ruchu sieciowego użytkownika w sieciach telekomunikacyjnych, bazującego na głębokiej sieci neuronowej jako kluczu będącym jednocześnie generatorem kluczy oraz na tokenie zmieniającym klucz. Przedstawiana metoda zapewnia dwustopniową weryfikację tożsamości użytkownika oraz stanowi swoiste rozwinięcie idei omówionej wcześniej w artykule [4].

Wraz z powstawaniem nowych metod kryptografii korzystających z SN, pojawiły się również nowe metody łamania zabezpieczeń, wspierane przez sztuczną inteligencję. W literaturze dostępne są badania dotyczące podatności i strategii ataków na zabezpieczenia kryptografii neuronowej [5].

2. KONCEPCJA MECHANIZMU

Proponowany neuronowy mechanizm szyfrowania ruchu sieciowego użytkownika przedstawiony jest na schemacie ideowym (Rys. 1).



Rys. 1. Schemat ideowy proponowanej metody

W założeniach metody zakłada się istnienie tokena stanowiącego pseudolosowy generator kombinacji kluczy. Generator ten może być powiązany z bieżącą datą, godziną lub inną zmienną w czasie wartością, której częstotliwość zmian wpływać będzie na częstotliwość zmiany klucza. Zadaniem generatora jest modyfikacja przetwarzania etapowego. Modyfikacja polega na zamianie pozycji bitów wektora danych, które przetwarza SN. Przedstawiając to na przykładzie: jeśli dla 16-bitowego wektora danych (do którego przetwarzania będą wykorzystane 2 SN, gdzie każda z SN przetwarza 8 bitów) wybrana SN przetwarza bity od 8 do 16, to po zamianie tokena będzie przykładowo przetwarzać bity od 1 do 4 oraz od 12 do 15 (łącznie 8 bitów).

Mechanizm szyfrowania składa się z pięciu zasadniczych etapów, które zostaną omówione na przykładzie hipotetycznej aplikacji do przetwarzania danych użytkownika w kontekście bankowości elektronicznej. Pierwszy etap stanowi wczytanie struktur SN oraz ich wag, będących osobnymi plikami. Liczba plików zależy od długości klucza stosowanego do bazowego szyfrowania po stronie banku. Dla klucza 256-bitowego (zalecanego w kontekstach wymagających najwyższego poziomu bezpieczeństwa) będą to 32 pliki ze strukturami SN (jeśli dla różnych sieci są różne struktury, a w przypadku jednej struktury 1 plik) oraz 32 pliki z wagami SN. Każdy plik, korzystając z algorytmu skrótu (np. SHA 512), przekształca całą zawartość pliku w skrót, czyli unikalny ciąg bitów o określonej długości. Wszystkie skróty lub ich wybrane fragmenty tworzą klucz weryfikujący użytkownika (o ustalonej wcześniej długości) poprzez dedykowaną aplikację lokalną lub webową. Klucz ten przesyłany jest protokołem zabezpieczeń banku i weryfikowany po stronie banku. Protokoły zabezpieczeń banku zazwyczaj stanowią kombinację metod szyfrowania symetrycznego i asymetrycznego, takich jak AES (Advanced Encryption Standard), RSA (Rivest Shamir Adleman) oraz ECC (Elliptic Curve Cryptography). W przypadku pozytywnej weryfikacji następuje dalsza wymiana komunikacji, a w przypadku negatywnej weryfikacji zwracana jest informacja do użytkownika. W ten sposób realizowany jest pierwszy etap weryfikacji użytkownika. Rozwiązanie to przypomina posiadanie klucza USB lub pliku, którego ciąg bitów stanowi klucz do odszyfrowania danych. Takie rozwiązanie dostępne jest np. w oprogramowaniu VeraCrypt [6].

Drugi etap weryfikacji użytkownika stanowi podanie przez użytkownika hasła. Konwersję hasła na wartość liczbową przedstawia zależność:

$$\sum_{i=1}^{n+1} \frac{Sign_{Val}}{((Sign_{ValMax} + 1) \times n)} \quad (1)$$

Gdzie i to kolejne pozycje znaków hasła użytkownika; n to liczba znaków hasła; $Sign_{Val}$ to liczbowa re-

prezentacja znaku w przyjętym systemie kodowania znaków; $Sign_{ValMax}$ to największa liczbowa reprezentacja znaku w przyjętym systemie kodowania znaków (dla ASCII jest to 128, natomiast dla rozszerzonego ASCII wartość ta wynosi 256).

Trzeci etap złożony jest z dwóch komponentów, które w sposób ciągły podlegają zmianom. Pierwszym komponentem jest wartość wyjściowa tokena, determinująca sposób mieszania klucza podczas przetwarzania etapowego. Częstotliwość zmian tokena oraz sposób ich dokonywania muszą być takie same po stronie nadawczej i odbiorczej. Drugim komponentem jest przyjmowanie danych do przesłania w ustalonej kolejności i w blokach o wielkości będącej wielokrotnością liczby 8 (sumarycznie składających się na długość klucza stosowanego do bazowego szyfrowania po stronie banku).

Czwartym etapem jest wykonanie przetwarzania etapowego, polegającego na mieszaniu przydzielanych bitów na wejścia poszczególnych SN według wygenerowanego tokena. SN posiada dwa wejścia: pierwsze przyjmuje 8-bitowy ciąg danych do konwersji, drugie stanowi wartość liczbową hasła użytkownika. SN dokonuje procesu konwersji, biorąc pod uwagę wartości obydwu wejść, oraz zwraca zmodyfikowany 8-bitowy ciąg. Ze względu na stosowanie w tym etapie wielu SN, wskazane jest zrównoleglenie obliczeń w celu ich przyspieszenia.

Ostatnim etapem jest weryfikacja poprawności komunikacji po stronie odbiorczej. Jeśli bank stwierdził w pierwszym etapie prawidłowość klucza tworzonego na podstawie plików oraz po testowym zapytaniu otrzymał prawidłowo zakodowaną odpowiedź, drugi etap weryfikacji tożsamości użytkownika uznaje się za zakończony pozytywnie. W przypadku negatywnej weryfikacji tożsamości w drugim etapie, wynikającej z niewłaściwego hasła użytkownika, zależnie od przyjętych protokołów bezpieczeństwa możliwe są różne opcje. Pierwszą jest przekazanie użytkownikowi informacji o błędnym haśle. Drugim rozwiązaniem jest poinformowanie użytkownika przez alternatywny środek komunikacji, np. SMS, o nieprawidłowym podaniu hasła oraz zalogowanie na podstawie posiadanych plików z ograniczoną funkcjonalnością konta z powodu podejrzenia nieautoryzowanego dostępu. W ten sposób, w razie kradzieży z dysku twardego lub pendrive-a plików z wagami i strukturami SN, można przeciwdziałać próbom nieautoryzowanego dostępu. Komplementarnym rozwiązaniem jest przeciwdziałanie w razie lokalnego przymuszenia wobec użytkownika. W takim wypadku, po zalogowaniu specjalnym hasłem, informacja o przymuszeniu przekazywana jest do banku, który może podjąć wcześniej określone procedury.

3. WYNIKI BADAŃ

Badania przeprowadzono na systemie operacyjnym Linux (Ubuntu 22.04) z wykorzystaniem biblioteki uczenia maszynowego TensorFlow w wersji 2.19.0. Kody źródłowe zostały napisane w języku Python w wersji 3.10.12. W tej sekcji przedstawiono proces uczenia SN oraz jej architekturę, jak i wyniki uczenia.

3.1. Proces uczenia sieci neuronowej

Proces uczenia realizowany był na bazie algorytmu wstecznej propagacji błędów. W celu realizacji procesu uczenia przyjęto następujące zależności funkcji celu:

$$BCE_{SingleQuery} = BCE\left(SN_{Output}, Table_{True}\right) \quad (2)$$

$$L_{NNLossArray} = \frac{\sum_{i=0}^{Sign_{ValMax}} BCE_{SingleQuery}(i)}{Sign_{ValMax}} \quad (3)$$

Gdzie BCE to binarna entropia krzyżowa (ang. Binary Cross-entropy); i to indeks elementu; $BCE_{SingleQuery}$ to binarna entropia krzyżowa wyjściowego ciągu bitów SN w porównaniu do oczekiwanego ciągu danych zgodnie z tablicą konwersji; SN_{Output} to wartość wyjściowa sieci; $Table_{True}$ to docelowa wartość wyjściowa SN z tablicy konwersji; $Sign_{ValMax}$ to największa liczbowa reprezentacja znaku w przyjętym systemie kodowania znaków; $L_{NNLossArray}$ to błąd imitacji przez SN tablicy konwersji.

Zbiór uczący składał się z tablic konwersji generowanych losowo przed procesem uczenia. Generacja tablicy polegała na uzupełnieniu dwóch kolumn tablic od zera do docelowej maksymalnej wartości liczbowej reprezentacji znaku w przyjętym systemie kodowania znaków (rozszerzonego ASCII). Następnie druga kolumna losowo podlegała wielokrotnemu (w przyjętym scenariuszu dziesięciokrotnemu) mieszaniu. W ten sposób tworzona była tablica konwersji. Zastosowany optymalizator w procesie uczenia oparty był na algorytmie Adam o współczynniku 10^{-3}. Liczba epok uczenia wynosiła 3000.

3.2. Architektura sieci neuronowej

Struktura SN składa się z podstawowych typów warstw zdefiniowanych w bibliotece uczenia maszynowego TensorFlow (Tabela 1). Funkcję aktywacji dla poszczególnych warstw podstawowych stanowi ReLU, z wyjątkiem ostatniej warstwy, dla której zastosowano funkcję aktywacji Sigmoid.

Tab. 1. Struktura SN

Lp.	*Warstwa*	*Lp.*	*Warstwa*
1	Input(8-bitowy ciąg)	11	Dense(units=64)(10)
2	Input(Klucz)	12	Add(11, Multiply(2,2))
3	Dense(units=8)(1)	13	Dense(units=64)(12)
4	Dense(units=8)(3)	14	Subtract(13, 2)
5	Add(4, 2)	15	Dense(units=128)(14)
6	Dense(units=16)(5)	16	Dense(units=128)(15)
7	Dense(units=16)(6)	17	Subtract(16, 2)
8	Add(7, 2)	18	Dense(units=256)(17)
9	Dense(units=32)(8)	19	Dense(units=256)(18)
10	Dense(units=32)(9)	20	Dense(units=8)(19)

3.3. Wyniki działania sieci neuronowej

Wszystkie wytrenowane SN odtwarzały działanie wygenerowanych tablic konwersji w sposób bezbłędny

(bez powtórzeń kombinacji na wyjściu dla różnych wartości ciągu wejściowego). Wynik działania wytrenowanych SN przedstawia Tabela 2.

Tab. 2. Wynik działania pojedynczej SN

Czas CPU [ms]	*Czas GPU [ms]*	*Rozmiar na dysku [MB]*	*Liczba parametrów*
4,01	4,13	1,61	133 992

W Tabeli 2 dla 8-bitowej tablicy konwersji (256 kombinacji), umieszczono czasy przetwarzania danych wejściowych przez pojedynczą SN z wykorzystaniem procesora (ang. Central Processing Unit, CPU) 12th Gen Intel® Core™ i9-12900K oraz procesora graficznego (ang. Graphics Processing Unit, GPU) NVIDIA® GeForce RTX™ 3090. W tabeli znajdują się dane dotyczące potrzebnej pamięci do zapisu struktury SN wraz z jej wagami oraz całkowita liczba parametrów dla SN. Czasy przetwarzania danych przez SN dla CPU i GPU są zbliżone ze względu na czas potrzebny na przesłanie danych do GPU (dane muszą być najpierw przesłane z pamięci RAM do pamięci GPU). Wyraźne przyspieszenie obliczeń z zastosowaniem GPU widoczne jest przy bardziej złożonych obliczeniowo problemach, szczególnie przy przetwarzaniu równoległym.

Zastosowanie wielu warstw Add oraz Subtract komplikuje i wydłuża proces uczenia SN, jednak wpływa korzystnie na bezpieczeństwo przetwarzania wartości liczbowej hasła. Dla SN uczonej dla wartości hasła zgodnie z zależnością (1), wynoszącą 0,0038, SN wykonuje wszystkie konwersje poprawnie. Dla wartości liczbowej hasła 0,01 liczba poprawnych przekształceń zostaje zmniejszona z 256 do 209, natomiast dla wartości 0,1 liczba poprawnych przekształceń wynosi tylko 2.

Krytycznym elementem omawianego mechanizmu jest właściwe wytrenowanie SN, tak aby występowała jednoznaczność w sposobie interpretacji ciągu bitów. SN posiadają właściwości kompresyjne, jednak zgodnie ze stanem dzisiejszej wiedzy zdolności te są ograniczone. Dlatego też SN muszą posiadać odpowiednią liczbę parametrów, aby w sposób prawidłowy (jednoznaczny) odtwarzać działanie tablic konwersji [4].

4. PODSUMOWANIE

Wyniki badań praktycznej implementacji SN potwierdzają możliwość implementacji SN jako kombinowanej tablicy konwersji. Przedstawiony mechanizm może być stosowany jako komplementarny element wzmacniający bezpieczeństwo lub samoistnie stanowić protokół szyfrowanej wymiany informacji. Proponowany neuronowy mechanizm szyfrowania charakteryzuje się niestandardowością sposobu szyfrowania oraz wbudowanym dwuetapowym mechanizmem weryfikacji tożsamości użytkownika. Dodatkowym atutem prezentowanego rozwiązania jest przetwarzanie etapowe, które zapewnia zmienność szyfru w czasie w sposób określony przez generator tokena, utrudniając kryptoanalizę. Proponowany mechanizm posiada potencjał przyspieszenia obliczeń poprzez ich zrównoleglenie oraz jest elastyczny modułowo (umożliwia prostą wymianę pojedynczych elementów mechanizmu, np. wymiana generatora tokena umożliwia modyfikację sposobu mieszania przetwarzanych bitów oraz częstotliwości zmian sposobu mieszania bez potrzeby modyfikacji pozostałych elementów mechanizmu).

Potencjalnymi miejscami aplikacji proponowanego mechanizmu są urządzenia telekomunikacyjne oraz zdalne usługi wymagające autoryzowanego dostępu (między innymi: bankowość elektroniczna, usługi urzędowe, usługi VPN).

Dalsze badania powinny uwzględniać określenie wartości liczbowej hasła użytkownika z wykorzystaniem więcej niż jednej wartości podawanej na wejście SN, jak i nowe struktury SN. Badania powinny również uwzględnić czas opóźnień wymiany informacji w zależności od platformy sprzętowej oraz możliwość wykorzystania algorytmu VMPC [7].

PODZIĘKOWANIA

Praca badawcza została sfinansowana w ramach projektu Wojskowej Akademii Technicznej pt. „Nowe architektury sieci neuronowych do przetwarzania sygnałów i danych w radiokomunikacji i multimediach”, numer projektu: UGB 22-054/2025.

LITERATURA

[1] Ruttor, Andreas, Wolfgang Kinzel, Lanir Shacham, and Ido Kanter. 2004. “Neural Cryptography with Feedback.” *Physical Review E* 69 (4): 046110. https://doi.org/10.1103/PhysRevE.69.046110.

[2] Chakraborty, Sandip, Jiban Dalal, Bikramjit Sarkar, and Debaprasad Mukherjee. 2015. “Neural Synchronization Based Secret Key Exchange over Public Channels: A Survey.” arXiv. https://doi.org/10.48550/arXiv.1502.05153.

[3] Meraouche, Ishak, Sabyasachi Dutta, Haowen Tan, and Kouichi Sakurai. 2023. “Learning Asymmetric Encryption Using Adversarial Neural Networks.” *Engineering Applications of Artificial Intelligence* 123 (August):106220. https://doi.org/10.1016/j.engappai.2023.106220.

[4] Kaczyński, Maciej. 2021. “Data Encryption System with a Variable Key Generated Using Artificial Neural Networks.” In *ALFA I OMEGA*, 2:234–42. Poland, Kraków: Mateusz Weiland Network Solutions. https://www.konferencjenaukowe.online/monografie/monografia-alfa-i-omega-tom-2/.

[5] Beshaj, Lubjana, and Gaurav Tyagi. 2024. “Neural Cryptography: Vulnerabilities and Attack Strategies.” In *Disruptive Technologies in Information Sciences VIII*, 13058:261–68. SPIE. https://doi.org/10.1117/12.3013669.

[6] “VeraCrypt - Free Open Source Disk Encryption with Strong Security for the Paranoid.” n.d. Accessed April 2, 2025. https://veracrypt.eu/en/Keyfiles.html.

[7] Zoltak, Bartosz. 2004. “VMPC One-Way Function and Stream Cipher.” In Fast Software Encryption, edited by Bimal Roy and Willi Meier, 210–25. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-25937-4_14.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

NIEŁATWA DROGA DO 256 BITOWEGO BEZPIECZEŃSTWA W INTERFEJSIE RADIOWYM SIECI 5G NA PRZYKŁADZIE RODZINY ALGORYTMÓW SNOW

A CHALLENGING PATH TO 256-BIT SECURITY IN THE 5G RADIO INTERFACE: THE CASE OF THE SNOW ALGORITHM FAMILY

Mariusz Borowski

Wojskowy Instytut Łączności - PIB, Zegrze Płd., m.borowski@wil.waw.pl

DOI: 10.15199/59.2025.4.47

Streszczenie: **Zagrożenie poufności danych w interfejsie radiowym sieci 5G wymusza, w kolejnych wersjach tej technologii (5G Advanced), stosowanie algorytmów zapewniających 256 bitowe bezpieczeństwo, zamiast dotychczas używanych algorytmów o co najwyższej 128 bitowym bezpieczeństwie. W dokumencie pokazano trwającą ponad sześć lat drogę kryptografów i kryptoanalityków, którzy na podstawie algorytmów z rodziny SNOW, opracowali szyfr SNOW 5G o 256 bitowym bezpieczeństwie i wydajności ponad 50 Gb/s w zastosowaniach programowych, jak i sprzętowych.**

Abstract: **The threat to data confidentiality in the 5G radio interface necessitates the adoption of algorithms providing 256-bit security in future versions of this technology (5G Advanced), instead of the currently used algorithms offering at most 128-bit security. This document presents the more than six-year journey of cryptographers and cryptanalysts who, based on the SNOW family of algorithms, developed the SNOW 5G cipher. It provides 256-bit security and performance exceeding 50 Gbit/s in both software and hardware implementations.**

Słowa kluczowe: **FSM, kryptoanaliza liniowa, wielomian generujący, złożoność obliczeniowa.**

Keywords: **computational complexity, FSM, generating polynomial, linear cryptanalysis.**

1. WSTĘP

Rodzina szyfrów strumieniowych SNOW odegrała ważną rolę w zapewnianiu poufności i integralności w różnych generacjach systemów łączności mobilnej, co przedstawiono w Tab.1. SNOW 3G został zestandaryzowany w technologiach 3G i 4G, a także 5G, zapewniając w intencji twórców bezpieczeństwo na poziomie 128 bitów, i jest zazwyczaj implementowany sprzętowo w celu osiągnięcia wymaganych tam prędkości.

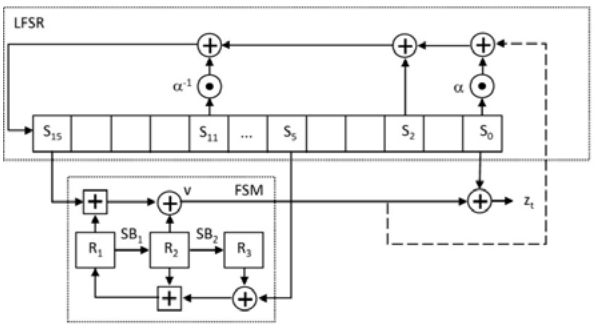
Tab. 1. Użycie algorytmu SNOW 3G do ochrony danych w różnych generacjach systemów łączności mobilnej

	SNOW 3G	
	Poufność	Integralność
3G (UMTS)	UEA2	UIA2
4G (LTE)	EEA1	EIA1
5G NR	128-NEA1	128-NIA1

W przypadku sieci 5G, fundamentalne zmiany w architekturze systemu oraz nowe wymagania dotyczące bezpieczeństwa stawiają przed algorytmami kryptograficznymi nowe wyzwania. Przede wszystkim szyfry te muszą osiągać prędkość co najmniej 20 Gb/s (maksymalna przepustowość transmisji danych w dół w 5G) w środowiskach zwirtualizowanych oraz zapewniać bezpieczeństwo na poziomie 256 bitów, także ze względu na zagrożenia ze strony komputerów kwantowych[1], [3].

2. ALGORYTMY SNOW 3G I SNOW 3G 256

SNOW 3G to strumieniowy szyfr słowowy, który generuje sekwencję 32-bitowych słów pod kontrolą 128-bitowego klucza i 128-bitowej zmiennej inicjalizacyjnej. Algorytm ten został opracowany przez Thomasa Johanssona oraz Patrika Ekdahla ze szwedzkiego Uniwersytetu Lund, a w 2006 roku został wybrany przez naukowców z grupy 3GPP do stosowania w algorytmach UxA2 sieci 3G. Algorytm SNOW składa się z dwóch warstw: LFSR oraz FSM [5]. Warstwę LFSR tworzy 16 komórkowy rejestr, którego każda komórka ma długość 32 bitów, a zatem jego długość wynosi 512 bitów. Każda wartość w komórce jest traktowana jako element ciała GF(2^{32}), a sekwencja LFSR jest określona przez wielomian generujący. Warstwę FSM tworzą trzy 32 bitowe rejestry, taktowane w sposób nieliniowy (R_1, R_2, oraz R_3); oraz para S-boxów: S1-box (Rijndael) i S2-box (Dickson). Najpierw wykonywana jest inicjalizacja (przerywana linia na Rys.1 jest aktywna), czyli szyfr jest taktowany bez generowania wyjścia. Następnie, przy każdym takcie zegara, wytwarza 32-bitowe słowo wyjściowe, wykorzystywane do zabezpieczenia tekstu jawnego.



Rys. 1. Schemat blokowy algorytmu SNOW 3G [5]

Bezpieczeństwo algorytmu SNOW 3G, wykorzystującego klucz o długości 128 bitów zostało dokładnie zbadane i uznane za niewystarczające [4], co pokazano w Tab.2. Wynika z niej istnienie ataków o złożoności dużo mniejszej niż wynikająca z długości użytego klucza (czyli 2^{128}). W celu zapobieżenia tym atakom ograniczono do 20.000 bitów długość zabezpieczanych bloków danych.

Tab. 2. Bezpieczeństwo algorytmu SNOW 3G

Typ ataku	Złożoność		
	Czas	Pamięć	Dane
Zgadnij i Określ	2^{320}	-	$9 \cdot 2^{32}$
Różnicowy atak resynchronizacyjny	2^{57}	2^{25}	2^{33}
Atak resynchronizacyjny z wybranym IV	2^{53}	-	2^{57}

W związku z potrzebą stosowania kluczy kryptograficznych o długości 256 bitów, powstał algorytmu SNOW 3G 256 bit. SNOW 3G został zaprojektowany z myślą o 128-bitowych kluczach, ale w 2018 roku rozważano jego rozszerzenie do użycia z 256-bitowymi kluczami. Algorytmy SNOW 3G i SNOW 3G 256 bit nie różnią się swoją wewnętrzną strukturą. Podstawowa architektura, w tym rejestr przesuwający ze sprzężeniem zwrotnym (LFSR), maszyna stanów skończonych (FSM) oraz stosowane wielomiany, pozostają takie same. Różna jest jedynie inicjalizacja, gdy do komórek rejestru jest podawany 256 bitowy klucz kryptograficzny dla szyfru SNOW 3G 256 bit zamiast klucza 128 bitowego (SNOW 3G) oraz 128 bitowa wartość IV.

Niestety, tak prosta zmiana nie doprowadziła do uzyskania szyfru, który mógłby być wykorzystany do zapewniania 256 bitowego bezpieczeństwa w interfejsie radiowym sieci 5G. Przedstawione w Tab.3 złożoności ataków na szyfr SNOW 3G 256 bit pokazują, że istnieją aproksymacje liniowe układu FSM umożliwiające przeprowadzenie ataków o złożoności czasowej rzędu 2^{172} i 2^{177} [9], co jest istotnie mniejszą złożonością niż złożoność pełnego przeszukiwania przestrzeni kluczy (2^{256}).

Tab. 3. Bezpieczeństwo algorytmu SNOW 3G 256 bit

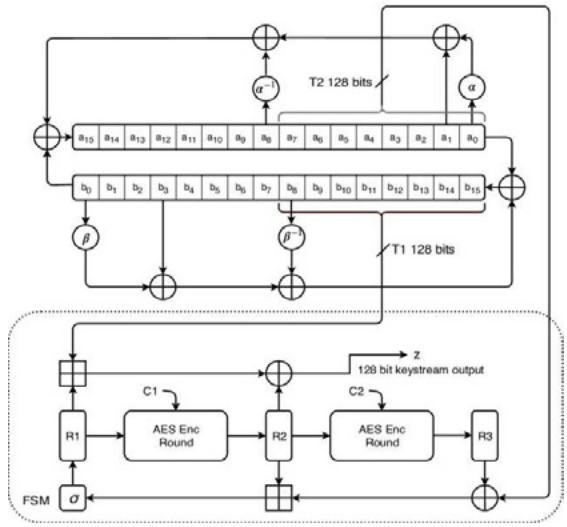
Typ ataku	Złożoność		
	Czas	Pamięć	Dane
Zgadnij i Określ	2^{384}	-	2^{3}
Atak rozróżniający	2^{172}	2^{41}	2^{172}
Atak korelacyjny	2^{177}	2^{177}	2^{176}

Ponadto algorytm SNOW 3G 256 bit nie jest odpowiedni dla technologii 5G, ponieważ aktualne testy wydajnościowe wskazują, że osiąga on jedynie szybkość około 9 Gb/s w implementacjach programowych.

3. ALGORYTM SNOW-V

W związku z nieudaną próbą adaptacji istniejącego algorytmu, organizacja 3GPP zażądała opracowania nowej konstrukcji algorytmu strumieniowego o 256 bitowym bezpieczeństwie, zaprojektowanej z myślą o dużej wydajności na współczesnych procesorach, które wspierają instrukcje SIMD (Single Instruction, Multiple Data). W 2019 roku kryptolodzy z Uniwersytetu w Lund ze Szwecji oraz firmy Ericsson zaproponowali szyfr SNOW-V [6], przeznaczony do zastosowania w technologii 5G, osiągający prędkości do 58 Gb/s w aplikacjach programowych (do 140 Gb/s w aplikacjach sprzętowych) i mający zapewnić 256 bitowy poziom bezpieczeństwa.

Algorytm SNOW-V składa się z warstwy liniowej LFSR oraz nieliniowej FSM, i został ukazany na Rys.2. Część LFSR ma postać konstrukcji pierścieniowej składającej się z dwóch rejestrów przesuwnych, z których każdy zasila drugi. Rejestry składają się z 16 komórek o długości 16 bitów czyli długość każdego z nich wynosi 256 bitów. Każda wartość w komórce jest traktowana jako element ciała GF(2^{16}), a sekwencja każdego LFSR jest określona przez własny, różny wielomian generujący. Uzyskana konstrukcja pierścieniowa zapewnia maksymalną długość cyklu równą 2^{512}-1. FSM zawiera trzy 128-bitowe rejestry oraz dwie instancje pojedynczej rundy funkcji szyfrującej AES z kluczem zerowym. Na wyjściu algorytmu otrzymujemy słowa o długości 128 bitów, które są wykorzystywane do zabezpieczania danych. Tak długie słowo (cztery razy dłuższe względem algorytmu SNOW 3G) na wyjściu algorytmu SNOW-V, ułatwia uzyskanie wymaganych przez standard 5G prędkości pracy, zarówno w aplikacjach programowych, jak i sprzętowych.



Rys. 2. Schemat blokowy algorytmu SNOW-V [6]

Bezpieczeństwo szyfru SNOW-V było poddawane wielu publicznym analizom, obejmującym ataki typu zgadnij i określ, kryptoanalizę liniową a także kryptoanalizę różnicową na zredukowaną wersję szyfru. Złożoności ataków na pełny algorytm zawarto w Tab.4 [7].

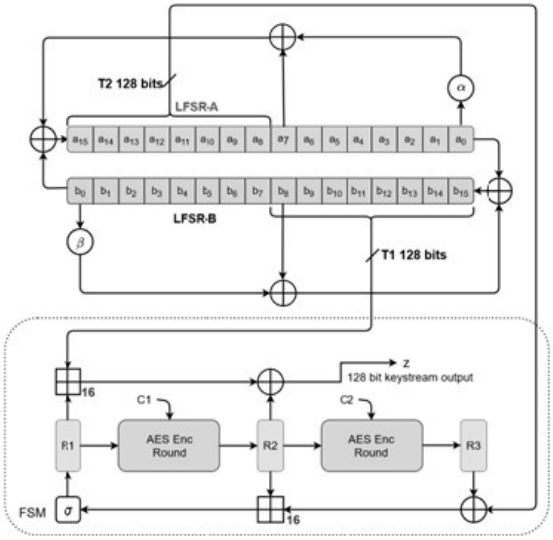
Tab. 4. Bezpieczeństwo algorytmu SNOW-V

Typ ataku	Złożoność		
	Czas	Pamięć	Dane
Zgadnij i Określ	2^{384}	-	896
Atak liniowy	2^{247}	2^{239}	2^{237}
Atak liniowy (LAT)	2^{241}	2^{240}	2^{237}

Istnienie ataków liniowych o złożoności 2^{241} oraz 2^{247} w praktyce oznacza osłabienie gwarantowanego poziomu bezpieczeństwa, co może mieć negatywny wpływ w przypadku zastosowań specjalnych. Organizacja 3GPP zażądała poprawienia oraz wzmocnienia bezpieczeństwa tego dobrze rokującego i wysoce wydajnego algorytmu kryptograficznego.

4. ALGORYTM SNOW 5G

Nowy algorytm, nazwany SNOW 5G, oparto na ogólnej strukturze algorytmu SNOW-V, która nie budziła zastrzeżeń [2], [7]. Jest on pokazany na Rys.3. Utrzymano warstwę LFSR w postaci konstrukcji pierścieniowej składającej się z dwóch rejestrów przesuwnych: LFSR-A, LFSR-B. Rejestry przesuwne składają się z 16 komórek o długości 16 bitów. Każda wartość w komórce jest traktowana jako element ciała GF(2^{16}). Układ FSM zawiera trzy 128-bitowe rejestry (R1, R2, R3) oraz dwie instancje pojedynczej rundy funkcji szyfrującej AES z kluczem zerowym (oznaczone jako E). Symbol σ to permutacja bajtów o wartościach od 0 do 15, identyczna jak w algorytmie SNOW-V. Podobnie jak w przypadku SNOW-V, maksymalna długość wygenerowanego klucza strumieniowego dla jednej pary klucza kryptograficznego i wektora inicjalizacyjnego (IV) wynosi 2^{64} bitów. Jeden klucz kryptograficzny może być użyty z maksymalnie 2^{64} różnymi wektorami IV.



Rys. 3. Schemat blokowy algorytmu SNOW 5G [7]

Na podstawie wyników kryptoanalizy liniowej algorytmu SNOW-V, w 2022 roku grupa SAGE (Security Algorithms Group of Experts) zaproponowała usunięcie wykrytych podatności, wykorzystywanych w atakach liniowych, poprzez:

- zastąpienie 32-bitowych operacji dodawania modularnego (modulo 2^{32}) użytych w SNOW-V poprzez operacje 16-bitowe (oznaczane jako $\boxplus_{16}$) podczas:
  - generowanie 128 bitowych słów na wyjściu z

$$z^{(t)} = (R_1^{(t)} \boxplus_{16} T_1^{(t)}) \oplus R_2^{(t)} \quad (1)$$

  - aktualizacji wartości rejestrów FSM

$$R_1^{(t+1)} = \sigma(R_2^{(t)} \boxplus_{16} (R_3^{(t)} \oplus T_2^{(t)}) \quad (2)$$

$$R_2^{(t+1)} = E(R_1^{(t)}) \quad (3)$$

$$R_3^{(t+1)} = E(R_2^{(t)}) \, . \quad (4)$$

- zmianę sposobu inicjalizacji szyfru. Podczas zmienionej fazy inicjalizacji, 256-bitowy klucz kryptograficzny oraz 128-bitowy wektor inicjalizacyjny (IV) są ładowane do rejestrów LFSR, a rejestry w FSM są inicjalizowane zerami. Inicjalizacja składa się z 16 kroków, podczas których szyfr jest aktualizowany w taki sam sposób jak w trybie pracy z kluczem bieżącym (ang. running-key mode), z tą różnicą, że wyjście z FSM jest zwracane do rejestrów LFSR zamiast generować słowo strumienia klucza. Dodatkowo, w dwóch ostatnich krokach fazy inicjalizacji, do rejestru R1 zostaje dodany (poprzez operację XOR) klucz.
- zmianę przyporządkowania komórek rejestrów LFSR-A oraz LFSR-B do 128 bitowych zmiennych T1 oraz T2.
- optymalizację realizacji szyfru w aplikacjach programowych. W tym celu zmieniono wielomiany generujące komórek rejestrów LFSR-A oraz LFSR-B.

Istnienie ataków o akademickiej złożoności na szyfr SNOW-V wymusiło potrzebę pogłębionej analizy bezpieczeństwa algorytmu SNOW 5G, członka rodziny algorytmów SNOW-V. W 2023 roku Ma Sudong i Jin Chenhui pokazali użycie automatycznego narzędzia do wyszukiwania ścieżek liniowych [8]. Przy jego pomocy uzyskano aproksymację liniową o korelacji $2^{-75.8}$ i użyto ją do przeprowadzenia ataku korelacyjnego o złożoności czasowej 2^{312}. Z kolei chińscy naukowcy z Information Engineering Institute dokonali, najlepszej do tej pory, kryptoanalizy liniowej algorytmu SNOW 5G [7]. Na podstawie teoretycznej analizy aproksymacji liniowych dla małych komponentów SNOW 5G wyprowadzono zależności i ograniczenia pomiędzy różnymi maskami liniowymi. Następnie zastosowano metodę podziału i zwycięstwa (divide-and-conquer), aby efektywnie wyszukiwać wysokiej jakości maski liniowe, analizując różne przypadki krytycznej pośredniej maski liniowej. W wyniku tych badań znaleziono ścieżkę liniową o korelacji $-2^{-67.7}$ i eksperymentalnie potwierdzono, że jest to dominująca ścieżka wśród zewnętrznych masek liniowych, które zostały wyznaczone. W szczególności ustalono zewnętrzne maski liniowe i obliczono korelacje dla różnych możliwych pośrednich masek liniowych. Otrzymana ścieżka wykazywała znacznie wyższą korelację niż pozostałe, co pozwoliło uznać ją za reprezentatywną dla całkowitej aproksymacji liniowej z pomijalnym błędem. Wyszukaną aproksymację liniową o korelacji $-2^{-67.7}$ wykorzystano do przeprowadzenia ataku korelacyjnego o złożoności czasowej 2^{280}.

Złożoność powyższych ataków pozwala uznać (Tab.5), że algorytm SNOW 5G zapewnia 256 bitowe bezpieczeństwo względem adwersarza klasycznego. SNOW 5G został zalecony przez grupę SAGE jako standardowy algorytm ochrony poufności i integralności oraz uwierzytelnienia danych w interfejsie radiowym sieci 5G (w ramach 5G Advanced) [2], jednak jego pełna specyfikacja nadal jest w fazie zatwierdzania [7].

Tab. 5. Bezpieczeństwo algorytmu SNOW 5G

Typ ataku	Złożoność		
	Czas	Pamięć	Dane
Zgadnij i Określ	2^{384}	-	896
Atak korelacyjny z przybliżeniem liniowym	2^{312}	2^{297}	2^{297}
Atak korelacyjny z NOWYM przybliżeniem liniowym	2^{280}	2^{265}	2^{265}

5. PODSUMOWANIE

Wydaje się, że trwające od 2018 roku prace kryptografów i kryptoanalityków z międzynarodowych ośrodków akademickich i firm telekomunikacyjnych zakończyły się sukcesem. Rozpoczynając prace od algorytmu 3G 256 bit uzyskanego poprzez proste podanie dłuższego klucza do dobrze znanego i wykorzystywanego algorytmu SNOW 3G, poprzez nową konstrukcję SNOW-V i jej modyfikacje poprawiające odporność na ataki korelacyjne, uzyskano symetryczny algorytm strumieniowy SNOW 5G, który zapewnia 256 bitowe bezpieczeństwo względem adwersarza klasycznego. Atak z użyciem algorytmu Grovera na algorytm SNOW 5G, wykonywany na komputerze kwantowym, ma złożoność $O(2^{128})$. Efektywność ataku z użyciem algorytmu Grovera nie zależy od aktualnie wykorzystywanego algorytmu symetrycznego, a tylko od bitowej długości wykorzystywanego klucza kryptograficznego [4]. Także wydajność tego algorytmu jest więcej niż zadowalająca, gdyż zapewnia prędkość pracy do 58 Gb/s w aplikacjach programowych oraz do 140 Gb/s w zastosowaniach sprzętowych.

Przyszłe wersje standardu 5G, w ramach 5G Advanced, będą zawierać ulepszenia obecnie stosowanych algorytmów 128-NEAx oraz 128-NIAx wykorzystujących klucz kryptograficzny o długości 128 bitów [3]. W szczególności możliwe jest standaryzowanie wariantów 256-bitowych algorytmów zabezpieczających interfejs radiowy: SNOW 5G, AES-256 oraz ZUC-256, które będą podstawowymi przekształceniami kryptograficznymi dla algorytmów [1],[2]:

- 256-NEAx zapewniających poufność,
- 256-NIAx zapewniających integralność (i uwierzytelnienie),
- 256-NCAx zapewniających poufność oraz uwierzytelnienie.

LITERATURA

[1] 3GPP TS 33.501 v.17 (2022-01) Security architecture and procedures for 5G system.

[2] 3GPP TS 35.240 v.19 (2024-03, under development) Specification of the SNOW 5G based 256-bits algorithm set: specification of the 256-NEA4 encryption, the 256-NIA4 integrity, and the 256-NCA4 authenticated encryption algorithm for 5G.

[3] Borowski Mariusz. „Ocena bezpieczeństwa algorytmów zapewniających poufność oraz uwierzytelnienie danych w sieci 5G". KSTiT, 2022.

[4] Borowski Mariusz, Wicik Robert. 2020. „Cryptographic protection of classified information in military radio communication faced with threats from quantum computers", SPIE 11442.

[5] Ekdahl Patrik, Johansson Thomas. „A new version of the stream cipher SNOW", in Selected Areas in Cryptology (SAC 2002), LNCS 2595, Springer-Verlag.

[6] Ekdahl Patrik, Johansson Thomas, et al. „New SNOW stream cipher called SNOW-V*", IACR Transactions on Symmetric Cryptology*, 2019.

[7] Liu Yinuo, Yang Jing. „Theoretical Linear Cryptanalysis of the 5G Standard Candidate SNOW 5G". Fast Software Encryption, 2025.

[8] Sudong Ma, Chenhui Jin. „Correlation attacks on SNOW-V-like stream ciphers based on a heuristic MILP model". *IEEE Transactions on Information Theory*, 2023.

[9] Yang Jing, Johansson Thomas, Maximov Alexander. „Vectorized linear approximations for attacks on SNOW 3G". Fast Software Encryption, 2019.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

PORÓWNANIE ALGORYTMÓW SHORA I REGEVA

COMPARISON OF SHOR'S AND REGEV'S ALGORITHMS

Przemysław Pawlitko[1]; Natalia Moćko[2]; Marcin Niemiec[3]; Piotr Chołda[4];

AGH Akademia Górniczo-Hutnicza, Instytut Telekomunikacji, al. Mickiewicza 30, 30-059 Kraków
[1] przemyslaw.pawlitko@tele.agh.edu.pl, [2] natalia.mocko@tele.agh.edu.pl, [3] marcin.niemiec@agh.edu.pl,
[4] piotr.cholda@agh.edu.pl

DOI: 10.15199/59.2025.4.48

Streszczenie: Praca porównuje klasyczny algorytm Shora służący do faktoryzacji liczb z opracowanym niedawno algorytmem Regeva. Przedstawia również pierwszą implementację algorytmu Regeva dla komputerów kwantowych, pokazując że algorytm działa z jednorodną superpozycją kubitów wejściowych zamiast pierwotnie zaproponowaną superpozycją Gaussa. Na podstawie tej implementacji przeprowadzono analizę porównawczą czasu działania i wydajności algorytmów dla małych liczb półpierwszych w zależności od wybranych parametrów.

Abstract: This work compares the classical Shor's algorithm aiming at factorization with the recently proposed Regev's algorithm. It presents the first implementation of the latter for quantum computers, demonstrating that the algorithm operates with a uniform superposition of input qubits instead of the Gaussian superposition originally proposed. Using this implementation, the both algorithms runtime and efficiency for small semiprime numbers, depending on the chosen parameters, are compared.

Słowa kluczowe: algorytmy kwantowe, algorytm Regeva, algorytm Shora, faktoryzacja, kryptografia

Keywords: cryptography, factorization, quantum algorithms, Regev's algorithm, Shor's algorithm

1. WSTĘP

Kryptografia klucza publicznego (asymetryczna) jest fundamentalnym filarem nowoczesnego bezpieczeństwa cybernetycznego. Jej zasady leżą u podstaw zabezpieczeń licznych systemów, w tym bezpiecznej komunikacji internetowej, ochrony danych w chmurze, weryfikacji integralności oprogramowania, szyfrowania danych oraz wielu innych zastosowań. Jednym z najpowszechniej stosowanych algorytmów kryptografii klucza publicznego jest RSA [7]. Jego bezpieczeństwo względne opiera się na złożoności algorytmicznej problemu rozkładu na czynniki (faktoryzacji) dużej liczby półpierwszej N, będącej iloczynem dwóch liczb pierwszych p i q. Według obecnego stanu wiedzy, dla klasycznych komputerów rozkład takiej liczby na czynniki jest obliczeniowo nieosiągalny w praktycznym (tj. wielomianowym) czasie. Kompromitacja RSA miałaby daleko idące konsekwencje dla innych algorytmów kryptograficznych opartych na podobnych problemach obliczeniowych (np. opartych na logarytmach dyskretnych). Chociaż złamanie szyfrowania RSA jest obecnie niemożliwe dla klasycznych maszyn, w 1997 roku amerykański informatyk i matematyk Peter Shor opublikował pracę „Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer" [8], w której dowiódł, że efektywny rozkład dużych liczb na czynniki jest możliwy przy użyciu komputerów kwantowych. Algorytm Shora miał przełomowe znaczenie dla kryptografii, ponieważ uświadomiono sobie, że dotychczasowe metody szyfrowania z użyciem klucza publicznego nie zachowają bezpieczeństwa w przypadku utworzenia dużego komputera kwantowego.

Jednakże algorytm Shora charakteryzuje się dużą liczbą bramek kwantowych, co przy obecnym stopniu zaawansowania komputerów kwantowych skutkuje szybką dekoherencją kwantową zaburzającą efekty kwantowe potrzebne do prawidłowego działania algorytmu.

Aby sprostać tym wyzwaniom, Oded Regev zaproponował alternatywny algorytm kwantowy w swojej ostatniej pracy „An Efficient Quantum Factoring Algorithm" [6]. Algorytm ten można postrzegać jako wielowymiarowe rozszerzenie klasycznej metody Shora. Algorytm Regeva zmniejsza liczbę bramek kwantowych potrzebnych do faktoryzacji w porównaniu z algorytmem Shora. Autor najnowszego algorytmu sam jednak przyznaje, że nie jest jasne, czy ta redukcja przełoży się na praktyczne korzyści w przypadku fizycznej realizacji komputerów kwantowych o odpowiednich rozmiarach. Pomimo tej niepewności, jego praca stanowi istotny krok w kierunku przezwyciężenia barier związanych z przetwarzaniem kwantowym i rozwoju kryptografii kwantowej.

W niniejszej pracy zaprezentowano porównanie algorytmów Shora i Regeva. Przedstawiono zarysy koncepcyjne obu algorytmów oraz wyniki implementacji porównawczej w środowisku symulacyjnym. Ze względu na ścisły limit stron, w niniejszym dokumencie przedstawiono jedynie wybrane wyniki, ale szerszy raport zawierający szczegółowe rezultaty badań znajduje się w [5].

2. OPIS ALGORYTMÓW FAKTORYZACJI KWANTOWEJ

Oba wspomniane algorytmy kwantowe mają na celu faktoryzację półpierwszej, n-bitowej liczby całkowitej N, będącej iloczynem liczb pierwszych $p, q \in P$ (tj. $N = pq$). Jednakże proces faktoryzacji jest przeprowadzany w odmienny sposób.

2.1 Idea algorytmu Shora

Algorytm Shora zasadniczo znajduje najmniejszy parzysty okres r modularnej funkcji wykładniczej, zdefiniowanej następująco:

$$z \mapsto a^z \bmod N \tag{1}$$

gdzie a jest losową liczbą całkowitą należącą do grupy multiplikatywnej $(\mathbb{Z}/N\mathbb{Z})^\times$, zaś $z \in \mathbb{Z}$. Po znalezieniu okresu tej funkcji faktoryzacja N staje się trywialna i przebiega w następujących krokach:

$$a^r \equiv 1 \bmod N \tag{2}$$

$$a^r - 1 \equiv 0 \bmod N \tag{3}$$

$$\left(a^{\frac{r}{2}} - 1\right)\left(a^{\frac{r}{2}} + 1\right) \equiv 0 \bmod N \tag{4}$$

$$\left(a^{\frac{r}{2}} - 1\right)\left(a^{\frac{r}{2}} + 1\right) \equiv 0 \bmod (pq) \tag{5}$$

Musi więc istnieć takie $k \in \mathbb{Z}$, że:

$$\left(a^{\frac{r}{2}} - 1\right)\left(a^{\frac{r}{2}} + 1\right) = pqk \tag{6}$$

W efekcie bardzo prawdopodobne jest, że p dzieli $\left(a^{\frac{r}{2}} - 1\right)$, a q dzieli $\left(a^{\frac{r}{2}} + 1\right)$[1]:

$$p = \gcd\left(\left(a^{\frac{r}{2}} - 1\right), N\right) \tag{7}$$

$$q = \gcd\left(\left(a^{\frac{r}{2}} + 1\right), N\right) = \frac{N}{p} \tag{8}$$

W implementacji najwięcej czasu pochłania realizacja wcześniej opisanej modularnej funkcji wykładniczej (2). Wynika to z faktu, że wykrycie okresu tej funkcji wymaga, aby z osiągnęło wystarczająco dużą wartość. Dla n-bitowej liczby N, $z \leq 2^n$. Ponadto funkcja ta musi być stosowana w superpozycji w części kwantowej algorytmu. Zatem wielkość obliczanych liczb znacznie spowalnia działanie samego algorytmu.

Istnieje też poważniejszy problem: efekty kwantowe, takie jak superpozycja i splątanie, mogą ulec zniszczeniu w trakcie działania algorytmu. Im więcej splątanych cząstek i im więcej jest bramek w obwodzie kwantowym, tym bardziej podatne na szum i dekoherencję są stany kwantowe, co prowadzi do ich szybszego zniszczenia [4]. Z tego powodu pożądane jest zminimalizowanie liczby bramek kwantowych.

2.2 Idea algorytmu Regeva

W celu zmniejszenia liczby bramek kwantowych, Oded Regev zaproponował podejście oparte na rozszerzeniu algorytmu Shora do większej liczby wymiarów [6]. Funkcja okresowa w przypadku jego algorytmu jest zdefiniowana nieco inaczej niż w algorytmie Shora:

$$(z_1, z_2, \ldots, z_d) \mapsto a_1^{z_1} a_2^{z_2} \ldots a_d^{z_d} \bmod N \tag{9}$$

gdzie liczby $a_1, a_2, \ldots, a_d$ to kolejne kwadraty liczb pierwszych (tj. $4, 9, 25, \ldots$), zaś $z_i \in \mathbb{Z}$. Po znalezieniu wektora okresowego $(r_1, r_2, \ldots, r_d)$ tej funkcji, faktoryzacja N jest trywialna i, podobnie jak wyżej (2), rozpoczyna się od równania:

$$a_1^{r_1} a_2^{r_2} \cdots a_d^{r_d} \equiv 1 \bmod N \tag{10}$$

Teraz, niech: $a_i = b_i^2$ z uwagi na fakt, że każda liczba a_i jest kwadratem liczby pierwszej b_i. Otrzymujemy:

$$b_1^{2r_1} b_2^{2r_2} \cdots b_d^{2r_d} \equiv 1 \bmod N \tag{11}$$

$$b_1^{2r_1} b_2^{2r_2} \cdots b_d^{2r_d} - 1 \equiv 0 \bmod N \tag{12}$$

$$\left(b_1^{r_1} b_2^{r_2} \cdots b_d^{r_d} - 1\right)\left(b_1^{r_1} b_2^{r_2} \cdots b_d^{r_d} + 1\right) \equiv 0 \bmod N \tag{13}$$

Algorytm Regeva opiera się na heurystycznym założeniu dotyczącym teorii liczb, które mówi, że przynajmniej połowa wektorów okresowych, po zastosowaniu ich w analizowanej funkcji i dwustronnym wyciągnięciu pierwiastka, daje nietrywialne dzielniki 1 mod N, tj. różne od 1 i -1. W konsekwencji w co najmniej połowie przypadków:

$$b_1^{r_1} b_2^{r_2} \cdots b_d^{r_d} \notin \{1, -1\} \bmod N \tag{14}$$

co eliminuje trywialne rozwiązania $(b_1^{r_1} \ldots b_d^{r_d} - 1) = 0$ lub $(b_1^{r_1} \ldots b_d^{r_d} + 1) = 0$, pozwalając tym samym na wyznaczenie p i q, gdyż w dalszym ciągu rozumowanie już pokrywa się z tym użytym w przypadku algorytmu Shora:

$$p = \gcd((b_1^{r_1} \cdot \cdots \cdot b_d^{r_d} - 1), N) \tag{15}$$

$$q = \gcd((b_1^{r_1} \cdot \cdots \cdot b_d^{r_d} + 1), N) = \frac{N}{p} \tag{16}$$

Poniżej przytaczamy szkic dowodu przedstawionego w pracy Regeva, że takie podejście faktycznie prowadzi do znalezienia wektora okresowego. Aby znaleźć wektor okresowy, muszą istnieć co najmniej dwie różne kombinacje wektorów $(z_1, z_2, \ldots, z_d)$, które dają ten sam wynik funkcji. Wtedy wektor okresowy można skonstruować, analizując te dwa wektory. Analizowana funkcja (9) jest iloczynem d kwadratów liczb pierwszych podniesionych do potęgi sięgającej $2^{n/d}$. W związku z tym łączna liczba wyników analizowanej funkcji wynosi $(2^{n/d})^d = 2^n$. Ponadto okres r analizowanej funkcji spełnia następujące nierówności:

$$r \leq \phi(N) = (p-1)(q-1) = pq - p - q + 1 < N \tag{17}$$

$$N = pq \leq 2^n - 1 < 2^n \tag{18}$$

Łącząc te fakty i stosując zasadę szufladkową Dirichleta, wnioskujemy, że muszą istnieć dwie różne wartości wektorów dające ten sam wynik funkcji, co gwarantuje istnienie wartości okresu umożliwiającej faktoryzację liczby N. Dzięki temu każde z_i rośnie do $2^{n/d}$, a nie do 2^n jak w podejściu przedstawionym w algorytmie Shora. W ten sposób zostaje zmniejszony zakres obliczeń i w efekcie obciążenia obwodu kwantowego. Ponadto w oryginalnym algorytmie Regeva złożoność bramek w części kwantowej ogranicza się do $\widetilde{O}(n^{3/2})$, co redukuje szum i dekoherencję. W naszym przypadku złożoność bramek wynosi $\widetilde{O}(n^{5/2} \log n)$, natomiast szerokość obwodu kwantowego jest mniejsza, rzędu $\widetilde{O}(n)$.

[1] Jak podkreśla Shor [2], prawdopodobieństwo to wynosi przynajmniej $1 - 1/2^{k-1}$, gdzie k to liczba różnych nieparzystych czynników pierwszych N.

2.3 Działanie algorytmu Regeva

W podrozdz. 2.2 została opisana idea algorytmu, której celem jest znalezienie wektora okresowego $(r_1, r_2, \ldots, r_d)$ dla wcześniej zdefiniowanej funkcji (9). Jednakże w samej implementacji algorytmu należy wprowadzić modyfikację dotyczącą zakresu wektorów $(z_1, z_2, \ldots, z_d)$: aby otrzymać wektor okresowy, część klasyczna algorytmu tworzy kratę na podstawie $d + 4$ wyników części kwantowej (tj. $d + 4$ wektorów, każdy uzyskany w jednym przebiegu części kwantowej — dlatego na rzeczywistym komputerze kwantowym trzeba ją uruchomić $d + 4$ razy). Po zbudowaniu kraty część klasyczna uruchamia algorytm LLL (Lenstry–Lenstry–Lovásza) i odzyskuje wektor okresowy. Regev [6] jednak zauważa, że dla d-wymiarowej kraty algorytm LLL daje przybliżenie ze współczynnikiem 2^d. Ponieważ wartości wektorów (wykładniki) muszą sięgać co najmniej tej wartości współczynnika przybliżenia, to każde $z_i \in \{0, 1, \ldots, 2^{n/d+d}\}$.

Część kwantowa algorytmu generuje wspomniane wektory, a część klasyczna — tworząc kratę, wykonując algorytm LLL i znajdując wektor okresowy — kończy obliczenia.

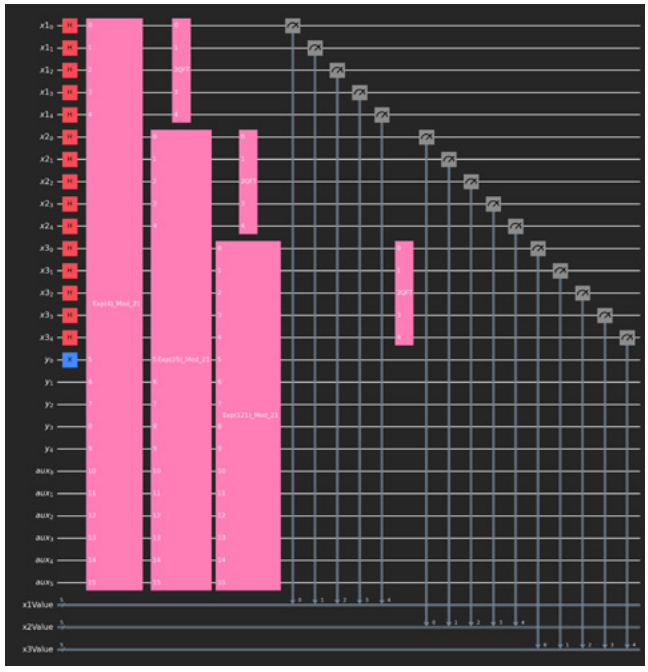
Kwantowa część algorytmu Shora tworzy superpozycję po wszystkich elementach grupy $(\mathbb{Z}/N\mathbb{Z})^\times$, podczas gdy w algorytmie Regeva wystarczy superpozycja nad niewielkim podzbiorem wartości blisko zera [3]. Aby ułatwić obliczenia dla używanej w obu przypadkach kwantowej transformaty Fouriera (QFT), Regev proponuje dyskretną superpozycję Gaussa [6, 3], ale ze względu na złożoność implementacji w naszym podejściu zastosowaliśmy superpozycję równomierną (opartą jedynie na bramkach Hadamarda), która — jak zauważa Kiebert w podrozdziale 5.4.1 swojej pracy [3] — również działa poprawnie. Potwierdziliśmy, że algorytm Regeva działa z takim ustawieniem.

Na początku podprocedura kwantowa przygotowuje stan superpozycji kwantowej na d rejestrach wejściowych, czyli superpozycję $|z\rangle$ dla $z \in \mathbb{Z}^d$. W naszym przypadku jest to superpozycja równomierna. Następnie na tym stanie superpozycji przeprowadza się operację modularnego potęgowania, uzyskując odwzorowanie:

$$|z\rangle\,|1\rangle \mapsto |z\rangle\,|a^z\rangle$$

Oznacza to, że przed odwzorowaniem $|z\rangle$ reprezentuje superpozycję rejestrów wejściowych, a $|1\rangle$ stan początkowy rejestru wyjściowego; po odwzorowaniu $|z\rangle$ nadal oznacza superpozycję rejestrów wejściowych, natomiast $|a^z\rangle$ odpowiada superpozycji wyników modularnego potęgowania dla $a \in \mathbb{Z}^d$, gdzie a jest wektorem pierwszych d kwadratów liczb pierwszych[2]. W kontekście modularnego potęgowania Regev proponuje szybki algorytm mnożenia całkowitego [2], jednak sam przyznaje, że dla małych n nie przynosi on korzyści, dlatego w takich przypadkach lepiej stosować wysoko zoptymalizowane warianty „mnożenia szkolnego". W naszej implementacji wykorzystaliśmy bramkę potęgowania modularnego zaproponowaną przez Stępnia [9], opartą na ulepszeniu klasycznej metody mnożenia w oparciu o bramkę modularnego dodawania Hänera [1].

W kolejnym kroku algorytm stosuje transformatę QFT na każdym rejestrze wejściowym. Głównym celem QFT jest przekształcenie amplitud stanów kwantowych w superpozycję stanów o różnych fazach[3]. Mechanika QFT jest szczególnie użyteczna w rozwiązywaniu problemów związanych z periodycznością. W algorytmie Regeva QFT wspomaga także identyfikację wektorów skorelowanych z wektorem okresu. Przykładowy używany obwód kwantowy dla $N = 21$ prezentujemy na rysunku 1. Na koniec pomiar wszystkich rejestrów wejściowych daje konkretny wektor, będący przybliżeniem wektora kraty dualnej $\mathcal{L}^*$ względem kraty związanej ze wspomnianą reprezentacją $\mathcal{L}$. To umożliwia części klasycznej odzyskanie ostatecznego wektora okresu.



Rysunek 1: Obwód kwantowy dla $N = 21$

3. PRZYKŁADOWE WYNIKI BADAŃ LICZBOWYCH

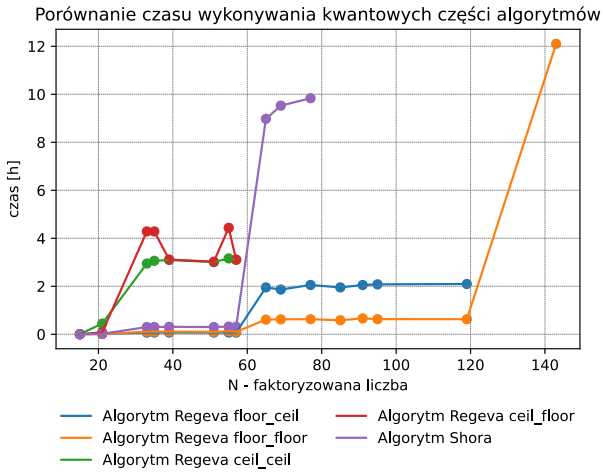
Celem naszych badań była analiza czasu działania oraz skuteczności algorytmu Regeva, a także porównanie uzyskanych wyników z algorytmem Shora. Dodatkowo skupiliśmy się na ocenie wpływu wybranych parametrów na wspomniane metryki. Jednym z aspektów, który został przez nas szczegółowo przeanalizowany, była zależność między liczbą wymiarów (d) i wartością wykładnika potęgi (qd) a szybkością działania algorytmu. Jako d przyjęliśmy wartość $\sqrt{n}$, a jako qd przyjęliśmy wartość $\left(d + \frac{d}{n}\right)$, obydwie zaokrąglając w górę (określane w nazwie parametru jako `ceil`) lub w dół (określane w nazwie parametru jako `floor`) do najbliższej liczby całkowitej. Wybraną kombinację oznaczaliśmy jako `ceil_ceil`, `ceil_floor`, `floor_ceil` lub `floor_floor`, gdzie pierwszy człon oznacza wybrane zaokrąglenie dla parametru d, a drugi dla parametru qd. Wybór zaokrągleń definiuje szerokość i głębokość obwodu, tym samym bezpośrednio wpływając na szybkość działania algorytmu.

Implementację algorytmu Regeva tworzyliśmy w oparciu o implementację algorytmu Shora opracowaną przez Bartłomieja Stępnia [9]. Następnie obydwie implementacje wykorzystaliśmy do porównania szybkości oraz

[2] Jak zauważa Kiebert [3], a może być wektorem małych, parami względnie pierwszych kwadratów liczb.

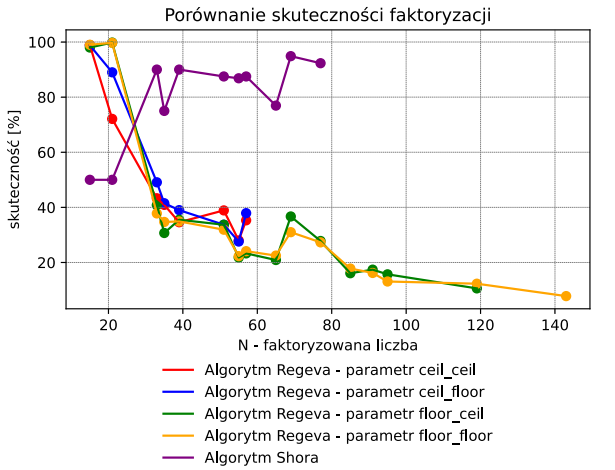
[3] Faza jest właściwością funkcji falowej opisującej stan kwantowy. Choć sama funkcja falowa dostarcza informacji o prawdopodobieństwie znalezienia cząstki w określonym miejscu, to faza odgrywa kluczową rolę w odniesieniu do interferencji kwantowej.

skuteczności badanych algorytmów. Uzyskane rezultaty prezentujemy na rysunku 2.



Rysunek 2: Wpływ wartości parametrów d i qd na szybkość działania algorytmu Regeva i porównanie ich z algorytmem Shora [5]

Z wykresu można wywnioskować, że większy wpływ na szybkość algorytmu Regeva ma liczba wymiarów niż wartość wykładnika potęgi. Ponadto dla kombinacji `floor_ceil` i `floor_floor` uzyskaliśmy lepsze wyniki w szybkości działania algorytmu Regeva w porównaniu z algorytmem Shora. Warto jednak podkreślić, że stało się to kosztem skuteczności rozkładu. Dla wartości d zaokrąglonej w górę otrzymaliśmy wyższą skuteczność faktoryzacji niż przy zaokrągleniu jej w dół. W pierwszym przypadku wraz ze wzrostem N uzyskiwana skuteczność oscylowała w okolicach 35%. Natomiast w drugim przypadku skuteczność ta malała do około 10%. Badania wykazały, że wybór zaokrąglenia dla liczby qd miał mniejszy wpływ na skuteczność faktoryzacji niż dla liczby d. Uzyskane wyniki prezentujemy na rysunku 3.



Rysunek 3: Porównanie skuteczności faktoryzacji liczby N przez algorytm Shora i algorytm Regeva dla różnych parametrów d i qd

4. PODSUMOWANIE

Stworzenie komputera kwantowego, który realnie zagrozi algorytmom kryptograficznym takim jak RSA, stanowi cały czas osiągnięcie, którego realność nie jest pewna. Niemniej jednak ryzyko związane z rozwojem algorytmów efektywnej faktoryzacji kwantowej jest traktowane poważnie, czego dowodem są aktywności standaryzacyjne NIST w zakresie opracowania metod tzw. kryptografii postkwantowej. Obecnie zagrożeniem od strony kryptoanalitycznej nie jest już tylko algorytm Shora, ale także niedawno opracowany algorytm Regeva, których porównanie jest przedmiotem niniejszej pracy.

Algorytm Regeva jest rozszerzeniem idei algorytmu Shora do przestrzeni wielu wymiarów. Dzięki temu jest on asymptotycznie szybszy niż algorytm Shora. Przeprowadzone przez nas badania pokazują, że dobór parametrów zarówno w części klasycznej, jak i kwantowej ma bardzo duży wpływ na czas działania i skuteczność algorytmu Regeva. Ponadto nasza implementacja udowadnia, że można użyć równomiernej superpozycji zamiast dyskretnej superpozycji Gaussa przy uzyskiwaniu satysfakcjonujących wyników.

Dalsze prace mogą skupić się m.in. na zaimplementowaniu dyskretnej superpozycji Gaussa i porównaniu z otrzymanymi wynikami dla superpozycji równomiernej.

LITERATURA

[1] Häner, Thomas, Martin Roetteler, and Krysta M. Svore. 2017. „Factoring using $2n+2$ qubits with Toffoli based modular multiplication". Preprint. URL: https://arxiv.org/abs/1611.07995.

[2] Harvey, David, and Joris Van Der Hoeven. 2021. „Integer multiplication in time $O(n \log n)$". Annals of Mathematics 193 (2).

[3] Kiebert, Midas. 2024. „Oded Regev's Quantum Factoring Algorithm". Mathematics and Computer Science at Informatics Institute, Korteweg-de Vries Institute for Mathematics, Faculty of Sciences, University of Amsterdam. Amsterdam.

[4] Mutter, P. M., and G. Burkard. 2023. „Theory of qubit noise characterization using the long-time cavity transmission". Physical Review A 107 (2).

[5] Pawlitko, P., Moćko, N., Niemiec M., Chołda, P. 2025. „Implementation and Analysis of Regev's Quantum Factorization Algorithm". arXiv. https://arxiv.org/abs/2502.09772.

[6] Regev, Oded. 2025. „An Efficient Quantum Factoring Algorithm". Journal of the ACM 72 (1): 1–13.

[7] Rivest, Ronald L., Adi Shamir, and Leonard Adleman. 1978. „A method for obtaining digital signatures and public-key cryptosystems". Communications of the ACM 21 (2).

[8] Shor, Peter W. 1997. „Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer". SIAM Journal on Computing. 26 (5).

[9] Stępień, Bartłomiej. 2021. „Algorytm Shora dla IBM Qiskit". MSc thesis, AGH University of Science and Technology, Kraków.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

PRAKTYCZNE METODY GENEROWANIA I TESTOWANIA BINARNYCH CIĄGÓW LOSOWYCH NA POTRZEBY SYSTEMÓW KRYPTOGRAFICZNYCH

PRACTICAL METHODS OF GENERATING AND TESTING BINARY RANDOM SEQUENCES FOR CRYPTOGRAPHIC SYSTEMS

Robert Wicik[1]; Marek Leśniewicz[1]; Piotr Komorowski[1]; Marcin Grzonkowski[1]

[1] Wojskowy Instytut Łączności - PIB, Zegrze, {r.wicik, m.lesniewicz, p.komorowski, m.grzonkowski}@wil.waw.pl

DOI: 10.15199/59.2025.4.49

Streszczenie: **Binarne ciągi losowe są podstawowym materiałem wykorzystywanym do generacji kluczy kryptograficznych, które są zasadniczym czynnikiem gwarantującym bezpieczeństwo działania algorytmów i protokołów kryptograficznych. W systemach niejawnych wymagana jest sprzętowa generacja i kontrola losowości ciągów. W artykule przedstawiono metody generowania i testowania losowości ciągów binarnych, opracowane na potrzeby zarządzania kluczami w specjalnych systemach kryptograficznych.**

Abstract: **Binary random sequences are the basic material used to generate cryptographic keys, which are a fundamental factor guaranteeing the security of cryptographic algorithms and protocols. In classified systems, hardware generation and control of the randomness is required. In this article, we present methods for generating and testing the randomness of binary sequences developed for electronic key management systems.**

Słowa kluczowe: **kryptografia, zarządzanie kluczami, sprzętowa generacja ciągów losowych, testowanie losowości.**

Keywords: **cryptography, electronic key management system, hardware random number generator, randomness testing.**

1. WSTĘP

Klucze kryptograficzne wykorzystywane w algorytmach szyfrowania, uwierzytelniania, podpisu cyfrowego oraz protokołach kryptograficznych są najważniejszym elementem, zapewniającym bezpieczeństwo informacji, realizowanym przy pomocy metod kryptograficznych. Klucze te, w zależności od danego algorytmu lub protokołu, stanowią ciągi znaków lub bitów [1]. Klucze mogą być symetryczne – takie same po obu stronach wykonywania algorytmu lub protokołu kryptograficznego lub asymetryczne – różne dla tych stron. Występują także klucze sieciowe – wspólne dla większej grupy urządzeń lub ich użytkowników. Istnieją też systemy o bezwarunkowym bezpieczeństwie, w których wykorzystywane są klucze jednorazowe. Klucze dzieli się ze względu na sposób ich ochrony – na tajne (prywatne) i publiczne. To właśnie tajne klucze w systemach ochrony informacji niejawnych są traktowane szczególnie, począwszy od procesów ich wytwarzania i weryfikacji ich jakości.

W artykule zostaną przedstawione sprzętowe metody generacji ciągów losowych na potrzeby generacji kluczy dla różnorodnych systemów ochrony informacji. Opracowane w Wojskowym Instytucie Łączności (WIŁ-PIB) Sprzętowe Generatory Ciągów Losowych (SGCL) uzyskały certyfikaty służb uprawniające do zastosowania w systemach kryptograficznych o klauzuli do ściśle tajne włącznie. Generatory te są używane w systemach ochrony informacji niejawnych w traktach grupowych (międzycentralowych), w sieciach ISDN, w militarnych sieciach radiowych oraz w specjalnych sieciach IP. Na potrzeby projektowania, produkcji i eksploatacji tych generatorów zostały opracowane zestawy testów statystycznych wykonywanych sprzętowo wewnątrz generatorów oraz programowo na zewnątrz, we współpracującym sprzęcie. Testy losowości pozawalają na ocenę jakości ciągów i bieżącą kontrolę sprawności generatorów.

Zagrożenia ze strony komputerów i algorytmów kwantowych [2] oraz ze strony klastrów komputerowych wysokiej wydajności (HPC) powodują, że szczególną uwagę przywiązuje się do generacji odpowiedniej długości kluczy, które będą nierozróżnialne od losowych. Ponadto ciągle wzrasta rola kryptografii symetrycznej, wymagającej generacji dużej ilości kluczy, ze względu na zmiany zachodzące w kryptografii asymetrycznej i przechodzenie na algorytmy post-kwantowe (PQC).

2. SPRZĘTOWE GENERATORY CIĄGÓW LOSOWYCH

Wyróżnia sią dwa rodzaje generatorów ciągów losowych. Prawdziwie losowe wykorzystują naturalnie występujące źródła entropii i zapewniają niepowtarzalność realizacji. Działanie generatorów pseudolosowych opiera się na algorytmach, których wykonanie można powtórzyć zadając ponownie te same dane wejściowe (ang. *seed*).

2.1. Metody sprzętowej generacji ciągów

Do generacji kluczy na potrzeby systemów ochrony informacji niejawnych wykorzystywane są przede wszystkimi źródła sprzętowe. Jest wiele zjawisk fizycznych uznawanych za losowe, ale niewiele jest takich, które w praktyce można użyć, jako źródła prawdziwej losowości i mogą być przetworzone do postaci cyfrowej. Tego typu zjawiska fizyczne to m.in.:

- szumy wytwarzane przez elementy półprzewodnikowe;
- niestabilne zjawiska elektryczne występujące w układach elektronicznych, np. hazardy;
- szumy mikrofalowe;
- kwantowe zjawiska optyczne;

- zjawiska promieniotwórcze.

W generatorach opracowanych w WIŁ-PIB jako źródła losowości wykorzystywane są szumy elementów półprzewodnikowych i szumy mikrofalowe.

2.2. Rodzina generatorów SGCL

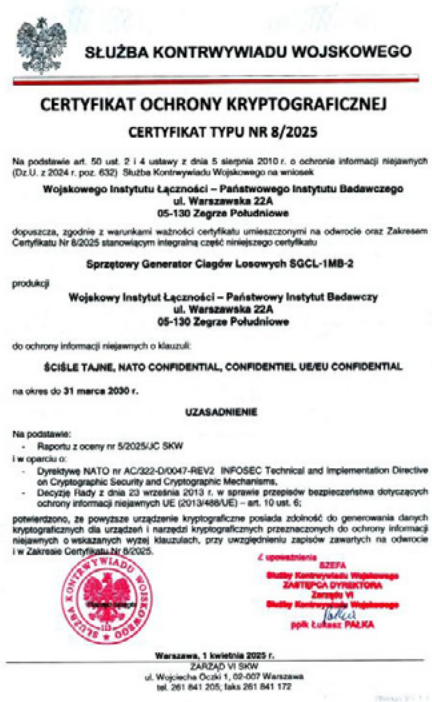
Wojskowy Instytut Łączności od ponad 30 lat pracuje nad zaawansowanymi, sprzętowymi rozwiązaniami umożliwiającymi generację binarnych ciągów losowych na potrzeby niejawnych systemów kryptograficznych. Pierwsza konstrukcja Sprzętowego Generatora Ciągów Losowych, którego dobre właściwości [3] pozwalały na wykorzystanie do generowania kluczy kryptograficznych, to SGCL-1. Generator ten uzyskał w SKW certyfikat typu nr 29/2005 uprawniający do jego zastosowania do ochrony informacji do klauzuli ściśle tajne. Generator ten umożliwiał generację binarnych ciągów losowych z wydajnością około 80 kbit/s, które oddawał poprzez interfejs szeregowy. Na bazie tego generatora zostały zbudowane systemy generacji kluczy dla urządzeń utajniających, które uzyskały certyfikaty ochrony kryptograficznej do tajne – działające w oparciu o klucze symetryczne i do ściśle tajne – wykorzystujące klucze jednorazowe.

W odpowiedzi na rosnące potrzeby systemów zarządzania kluczami w powstających systemach ochrony informacji niejawnych w sieciach: radiowych, ISDN i IP, w WIŁ-PIB został opracowany nowy typ Sprzętowego Generatora Ciągów Losowych: SGCL-1MB. Generator ten uzyskał w SKW certyfikaty typu nr 300/2011 i 143/2016 uprawniające do zastosowań do ściśle tajne włącznie. SGCL-1MB ma źródłową wydajność 1 MB/s i efektywną na interfejsie USB około 5 Mbit/s. W celu zapewnienia odpowiedniej jakości ciągów losowych, w generatorze pracuje jednocześnie 8 niezależnych źródeł losowości.

W ostatnich latach w WIŁ-PIB opracowana została rozwojowa wersja tego generatora SGCL-1MB-2, w którym zachowano metodę generacji ciągów losowych i weryfikacji jakości ciągów, protokół komunikacyjny na interfejsie USB i zabezpieczenie integralności ciągu. Wobec rozwoju interfejsów USB zastosowano w nim najnowszy układ FTDI i interfejs USB 2.0 HS 480 Mbit/s. Zwiększono dwukrotnie wynikową wydajność generacji do około 10 Mbit/s (2 MB/s źródłowo). Widok tego generatora przedstawia rysunek 1.

Rys. 1. Widok SGCL-1MB-2

Wszystkie funkcje generatora są realizowane sprzętowo w części analogowej i cyfrowej. Część cyfrowa została zaimplementowana w układzie programowalnym. SGCL-1MB-2 został poddany badaniom i certyfikacji zgodnie z ustawą o ochronie informacji niejawnych i uzyskał w SKW certyfikat typu nr 8/2025 uprawniający do jego zastosowania w systemach o klauzulach do ściśle tajne i NATO/EU confidential. Kopię certyfikatu SGCL-1MB-2 przedstawiono na rysunku 2.



SŁUŻBA KONTRWYWIADU WOJSKOWEGO

CERTYFIKAT OCHRONY KRYPTOGRAFICZNEJ

CERTYFIKAT TYPU NR 8/2025

Na podstawie art. 50 ust. 2 i 4 ustawy z dnia 5 sierpnia 2010 r. o ochronie informacji niejawnych (Dz.U. z 2024 r. poz. 632) Służba Kontrwywiadu Wojskowego na wniosek

Wojskowego Instytutu Łączności – Państwowego Instytutu Badawczego
ul. Warszawska 22A
05-130 Zegrze Południowe

dopuszcza, zgodnie z warunkami ważności certyfikatu umieszczonymi na odwrocie oraz Zakresem Certyfikatu Nr 8/2025 stanowiącym integralną część niniejszego certyfikatu

Sprzętowy Generator Ciągów Losowych SGCL-1MB-2

produkcji

Wojskowy Instytut Łączności – Państwowy Instytut Badawczy
ul. Warszawska 22A
05-130 Zegrze Południowe

do ochrony informacji niejawnych o klauzuli:

ŚCIŚLE TAJNE, NATO CONFIDENTIAL, CONFIDENTIEL UE/EU CONFIDENTIAL

na okres do **31 marca 2030 r.**

UZASADNIENIE

Na podstawie:
- Raportu z oceny nr 5/2025/JC SKW

i w oparciu o:
- Dyrektywę NATO nr AC/322-D/0047-REV2 INFOSEC Technical and Implementation Directive on Cryptographic Security and Cryptographic Mechanisms,
- Decyzję Rady z dnia 23 września 2013 r. w sprawie przepisów bezpieczeństwa dotyczących ochrony informacji niejawnych UE (2013/488/UE) – art. 10 ust. 6;

potwierdzono, że powyższe urządzenie kryptograficzne posiada zdolność do generowania danych kryptograficznych dla urządzeń i narzędzi kryptograficznych przeznaczonych do ochrony informacji niejawnych o wskazanych wyżej klauzulach, przy uwzględnieniu zapisów zawartych na odwrocie i w Zakresie Certyfikatu Nr 8/2025.

Z upoważnienia
SZEFA
Służby Kontrwywiadu Wojskowego
ZASTĘPCA DYREKTORA
Zarządu VI
Służby Kontrwywiadu Wojskowego
ppłk Łukasz PAŁKA

Warszawa, 1 kwietnia 2025 r.

ZARZĄD VI SKW
ul. Wojciecha Ościki 1, 02-007 Warszawa
tel. 261 841 205, faks 261 841 172

Rys. 2. Certyfikat SGCL-1MB-2

2.3. Mikrofalowy Generator Ciągów Losowych

W odpowiedzi na zapotrzebowanie ze strony systemów ochrony informacji dla dużych prędkości (Gigabity/s) i o wysokich klauzulach tajności (ściśle tajne) WIŁ-PIB od kilku lat pracuje nad znaczenie bardziej wydajnymi metodami generacji ciągów losowych. Jako źródło szumu zostało wybrane zjawisko mikrofalowe [4]. Prototyp MGCL-1 posiada wydajność wewnętrzną 1 Gbit/s i poprzez interfejs Ethernet może dawać efektywnie blisko 125 GB ciągu losowego na sekundę. Blisko, gdyż oprócz danych użytecznych, interfejs wymaga danych ramkowania oraz danych zapewniających możliwość weryfikacji integralności pobranego ciągu. Widok prototypu MGCL-1 przedstawia rysunek 3.

Rys. 3. Widok MGCL-1

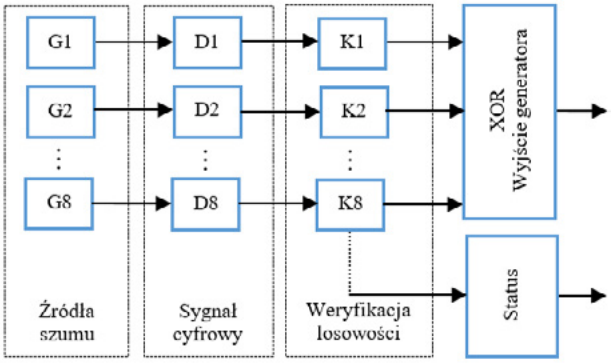
2.4. Zasada działania SGCL i MGCL

W SGCL i MGCL sygnały ze źródeł szumu są przetwarzane analogowo i przekształcane do postaci cyfrowej. Następnie sygnały cyfrowe są próbkowane i kontrolowane pod względem losowości. Wiele źródeł jest łączonych przez funkcję XOR w celu wytworzenia binarnego ciągu losowego na wyjściu generatora: w SGCL na interfejsie USB, w MGCL na interfejsie Ethernet.

Zasadnicze bloki funkcjonalne SGCL i MGCL przedstawiono na rysunku 4. Są to:

- blok przetwarzania sygnałów analogowych na bazie szumu (G1..G8),
- blok przetwarzania sygnałów do postaci cyfrowej (D1..D8),

- blok weryfikacji entropii uzyskiwanych ciągów binarnych (K1..K8),
- blok wytwarzania binarnego ciągu losowego (XOR) i przesyłania go na wyjście generatora,
- blok decyzji o jakości ciągu (Status).



Rys. 4. Schemat działania generatorów SGCL i MGCL

Łączenie wielu źródeł losowości ma na celu zwiększenia zrównoważenia binarnego ciągu wyjściowego. Pojedyncze źródła (oznaczone G→D na rysunku 4) są obarczone stosunkowo dużymi błędami próbkowania:

$$s = \left|\frac{1}{2}[P(1) - P(0)]\right| \quad (1)$$

gdzie $P(1)$ i $P(0)$ to wyznaczone prawdopodobieństwa wystąpień bitów 1 i 0.

W poprawnie skonstruowanych źródłach s nie przekracza wartości 0,01. Łączenie wielu niezależnych binarnych źródeł losowości zmniejsza tę nierównowagę. Prawdopodobieństwo nierównowagi rozkładu bitów 0 i 1 zmniejsza się wraz ze wzrostem liczby łączonych niezrównoważonych funkcji boolowskich [7]. Nierównowaga s_d na wyjściu połączonych przez sumę XOR wyjść d niezależnych funkcji boolowskich wynosi:

$$s_d = 2^{d-1}s^d \quad (2)$$

Dla $d = 4$ funkcji i $s = 0{,}01$ uzyskuje się znacznie mniejsze $s_4 = 8 \times 10^{-8}$. Jak wskazuje praktyka, w przypadku binarnych źródeł losowych, wraz ze wzrostem wielkości prób od 1 MB do 1 GB, nierównowaga 0 i 1 maleje na wyjściu generatora od około 10^{-4} do około 10^{-6}. To wystarcza, by źródła losowości wraz z funkcją XOR tworzyły poprawnie działający generator bez konieczności stosowania algorytmów przetwarzania końcowego (ang. *post-processing*). W generatorach SGCL łączy się w ten sposób 8 źródeł losowości, natomiast w MGCL-1 wystarczy tylko 6. Wyjściowe ciągi z tych generatorów spełniają rygorystyczne testy losowości.

3. TESTOWANIE LOSOWOŚCI CIĄGÓW BINARNYCH

Testy losowości ciągów binarnych są wykonywane na etapie projektowania generatorów, ich produkcji oraz podczas eksploatacji w systemach generacji kluczy kryptograficznych. Wyniki tych testów pozwalają potwierdzić, że generowane bity i sekwencje bitów pojawiają się na wyjściach generatorów składowych i na wyjściu generatora z równym prawdopodobieństwem i niezależne od siebie.

3.1. Testy wewnętrzne w SGCL i MGCL

W generatorach SGCL i MGCL zostały zaimplementowane sprzętowo 2 testy losowości pozwalające na ciągłą weryfikację generatorów składowych na podstawie stwierdzonych błędów próbkowania bitów (1) i korelacji pomiędzy kolejnymi wystąpieniami wartości bitów:

$$K = P(0,0) + P(1,1) - P(0,1) - P(1,0) - (2s)^2 \quad (3)$$

W celu uzyskania dobrej jakości ciągu losowego na wyjściu generatora założono, że każdy z generatorów składowych powinien spełniać warunki: $s < 0{,}01$ i $0 < K < 0{,}02$. Wtedy binarny ciąg wyjściowy z generatora spełnia także bardziej rozbudowane testy statystyczne.

3.2. Pakiet testów statystycznych *Pstat*

W WIŁ-PIB na potrzeby prac związanych z oceną generowanych ciągów przez sprzętowe generatory ciągów losowych została opracowana metoda i zestaw testów statystycznych [8] *Pstat*, w którym jako miary zgodności rozkładów teoretycznych i empirycznych stosowane są statystyki [5] chi-kwadrat (χ^2) o ustalonej liczbie stopni swobody oraz statystyki rozkładu normalnego N(0,1). Zamiast oceny hipotez [6] dla zadanego poziomu istotności α i prawdopodobieństwa testowego *p-value*, obliczone statystki dla prób ciągów binarnych grupuje się przedziałach [9], których granicami są kwantyle rozkładów, jak to przedstawiono dla statystyk χ^2 w tablicy 1.

Tab. 1. Oczekiwane udziały klas statystyk

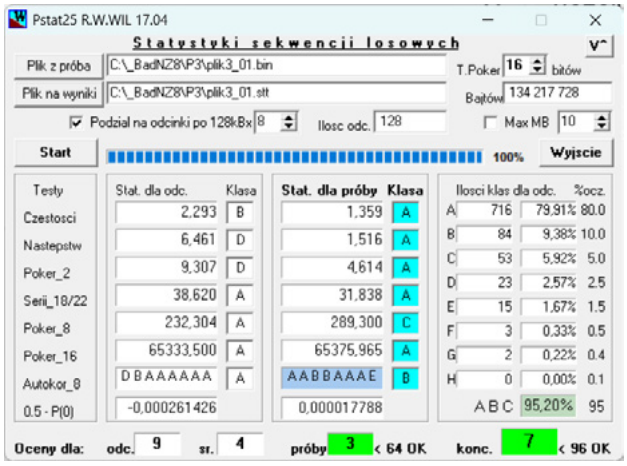
Przedział ufności	*Klasa statystki*	*Oczekiwany udział [%]*
$(1\text{-}\alpha) < 0{,}800$	A	80
$0{,}800 \le (1\text{-}\alpha) < 0{,}900$	B	10
$0{,}900 \le (1\text{-}\alpha) < 0{,}950$	C	5
$0{,}950 \le (1\text{-}\alpha) < 0{,}975$	D	2,5
$0{,}975 \le (1\text{-}\alpha) < 0{,}990$	E	1,5
$0{,}990 \le (1\text{-}\alpha) < 0{,}995$	F	0,5
$0{,}995 \le (1\text{-}\alpha) < 0{,}999$	G	0,4
$0{,}999 \le (1\text{-}\alpha)$	H	0,1

Podobnie, jak dla statystyk χ^2, przydzielane są klasy obliczonym statystykom mającym mieć rozkład normalny, pamiętając, że $N(0,1)$ jest rozkładem dwustronnym. Klasy A, B i C sumarycznie reprezentują wartości statystyk dla poziomu istotności $\alpha = 0{,}05$.

Pakiet testów *Pstat* oblicza statystyki globalne dla całej próby badanego ciągu (min. 100 MB) oraz lokalne dla odcinków o ustalonej długości (1 MB). Do automatycznej oceny ciągu z badanego źródła, klasom statystyk globalnych i lokalnych przyporządkowano wagi liczbowe, tym wyższe, im gorsza ich klasa. Pod pojęciem „gorsza" należy rozumieć taką, która powinna pojawiać się rzadziej, zgodnie z oczekiwanymi odsetkami podanymi w tablicy 1.

Rysunek 5. przedstawia pulpit sterowania i prezentacji wyników najnowszej wersji programu *Pstat25* i wyniki testów częstości, następstw i serii bitów oraz pokerowych i autokorelacji dla próby binarnego ciągu losowego pobranego ze sprzętowego generatora SGCL-1MB-2. Dla

tej i innych prób pobranych z tego generatora można zaobserwować, że uzyskiwane odsetki wystąpień klas statystyk są zbliżone do oczekiwanych (oznaczone %ocz.).



Rys. 5. Wyniki działania Pstat25 dla SGCL-1MB-2

Podobne wyniki uzyskuje się dla prób rzędu 1 GB pobranych z generatora mikrofalowego MGCL-1.

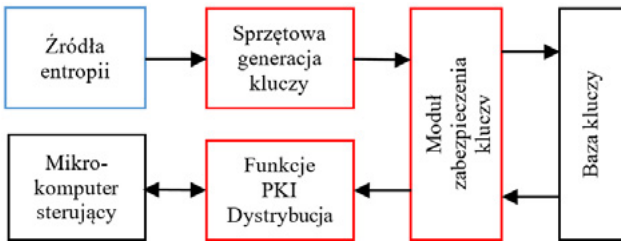
4. MODUŁ SPRZĘTOWEJ GENERACJI KLUCZY

W współczesnych systemach kryptograficznych wykorzystywane są moduły sprzętowej generacji kluczy (ang. Hardware Security Module). Moduły te, oprócz generowania ciągów losowych, kluczy i innych danych kryptograficznych (DK), zabezpieczają i przechowują klucze w bezpiecznej pamięci oraz potrafią wykonywać dodatkowe funkcje kryptograficzne.

Nowy moduł tego typu zaprojektowany w WIŁ-PIB umożliwia:

- sprzętowe generowanie ciągów losowych;
- sprzętowe generowanie tajnych kluczy symetrycznych oraz prywatnych i publicznych;
- kryptograficzne zabezpieczanie przechowywanych kluczy i na czas ich dystrybucji;
- realizację funkcji infrastruktury klucza publicznego (PKI) i dystrybucji kluczy.

Zasadnicze funkcje są realizowane sprzętowo w układzie programowalnym, który jest wspomagany przez mikrokomputer. Schemat funkcjonalny modułu sprzętowej generacji kluczy przedstawiono na rysunku 6.



Rys. 6. Schemat modułu sprzętowej generacji kluczy

Najważniejsze klucze odpowiadające za bezpieczeństwo infrastruktury kryptograficznej nie opuszczają modułu. Klucze generowane na potrzeby współpracujących urządzeń kryptograficznych są zabezpieczane w module na czas dystrybucji. Moduł sprzętowej generacji kluczy umożliwia budowanie systemów kryptograficznych i infrastruktury klucza publicznego przeznaczonych do ochrony informacji niejawnych o wysokich klauzulach tajności, wykorzystujących klucze symetryczne (tajne) oraz klucze asymetryczne (prywatne i publiczne). Połączenie dwóch rodzajów systemów kryptograficznych jest odpowiedzią na zagrożenia ze strony komputerów kwantowych i obniżonego bezpieczeństwa klasycznych algorytmów, w szczególności z kluczem publicznym, jak RSA, DSA, DH, także wykorzystujących EC.

5. PODSUMOWANIE

W artykule przedstawiono praktyczne metody i urządzenia przeznaczone do generowania binarnych ciągów losowych i kluczy kryptograficznych. Zaproponowano także metody i oprogramowanie do weryfikacji losowości ciągów binarnych. Wyniki badań i uzyskane certyfikaty potwierdzają skuteczność działania tych metod i urządzeń. Zaprojektowane i wykonane generatory ciągów losowych są efektywnie wykorzystywane w systemach zarządzania kluczami na potrzeby kryptograficznej ochrony informacji niejawnych. Dalsze prace dotyczą rozwoju bardziej wydajnej, sprzętowej generacji ciągów losowych i kluczy kryptograficznych.

LITERATURA

[1] Borowski Mariusz, Leśniewicz Marek, Grzonkowski Marcin, Wicik Robert. 2012. "Generation of random keys for cryptographic systems". *Annales UMCS Informatica*, AI XII (3) : 61–74.

[2] Borowski Mariusz, Wicik Robert. 2020. "Cryptographic protection of classified information in military radio communication faced with threats from quantum computers". *Radioelectronic Systems Conference*, SPIE Digital Library, vol. 11442, ch. 7.2.

[3] Leśniewicz Marek. 2009. Sprzętowa generacja losowych ciągów binarnych. Wydawnictwo WAT.

[4] Leśniewicz Marek. 2023. „Mikrofalowy generator losowych ciągów binarnych – konstrukcja, oprogramowanie i dowód bezpieczeństwa". *Przegląd Elektrotechniczny*, R. 99 (5) : 68–76.

[5] Menezes Alfred, Oorschot Paul, Vanstone Scott. 1997. Handbook of Applied Cryptography. CRC Press.

[6] Szreder Mirosław, Kozłowski Arkadiusz. 2024. Wnioskowanie na podstawie prób losowych i nielosowych. Wydawnictwo Uniwersytetu Gdańskiego.

[7] Wicik Robert. 1999. "Utilization of boolean bent functions for the construction of large S-boxes". *Regional Conference of Military Communications and Information Systems*, 225–228.

[8] Wicik Robert. 2024. "Generating and Testing Binary Random Sequences for Classified Information Protection Systems". *New Trends in Signal Processing* (IEEE Explore) : 175-182.

[9] Wicik Robert, Borowski Mariusz. 2008. "Randomness testing of some random and pseudorandom sequences". 2008. *Military Communications and Information Systems Conference*.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

O INTELIGENTNYM PROJEKTOWANIU SIECI KONFERENCYJNYCH W PAŚMIE FAL MILIMETROWYCH W ŚRODOWISKACH ZABUDOWANYCH

ABOUT INTELLIGENT DESIGN OF mmWAVE CONFERENCE NETWORKS IN BUILT ENVIRONMENTS

Monika Drozdowska[1]; Narcis Cardona[1]

[1] iTEAM Research Institute, Universitat Politécnica de Valencia, Walencja, Hiszpania, mdrozdo@iteam.upv.es

DOI: 10.15199/59.2025.4.50

Streszczenie: W niniejszym referacie przedstawiono kluczowe wnioski wynikające z mojej rozprawy doktorskiej poświęconej zagadnieniu inteligentnego projektowania sieci konferencyjnych działających w paśmie fal milimetrowych w środowiskach wewnętrznych. Szczególną uwagę poświęcono nielicencjonowanemu pasmu 60 GHz. Przeanalizowano wpływ środowiska wewnętrznego na propagację sygnału radiowego oraz oceniono potencjał zastosowania symulatorów opartych na metodzie śledzenia promieni (*ang. ray tracing*) do predykcji kanału radiowego.

Abstract: This paper presents key findings from a doctoral thesis on the intelligent design of millimetre-wave conference networks in built environments. The main focus is on the unlicensed 60 GHz band. The influence of the indoor environment on radio wave propagation and the possibility of employing ray tracing-based simulation tools for accurate modelling and prediction of the wireless channel were analysed.

Słowa kluczowe: pasmo fal milimetrowych, bezprzewodowe sieci konferencyjne, propagacja wewnątrzbudynkowa, symulator śledzenia promieni

Keywords: millimeter wave band, wireless conference networks, indoor propagation, RT simulator

1. WSTĘP

Wewnątrzbudynkowe sieci bezprzewodowe oparte na technologii Wi-Fi stanowią integralną część dzisiejszej komunikacji. Jednak w związku z rosnącym zagęszczeniem użytkowników, dla systemów takich jak systemy konferencyjne, wymagających wysokiej niezawodności i działania w czasie rzeczywistym, rozważane jest wykorzystanie zakresu fal milimetrowych (mmW).

Aby powstanie bezprzewodowych sieci konferencyjnych w paśmie mmW było możliwe, należy określić wymagania i zbadać standardy dla zastosowań w dużych środowiskach wewnętrznych. Istotne jest zidentyfikowanie elementów środowiska, które przyczyniają się do zmian w kanale radiowym. Na podstawie przeprowadzonych analiz, pomiarów, i symulacji powstały wskazówki dotyczące dystrybucji elementów sieciowych oraz wykorzystania środowiska do optymalizacji sieci konferencyjnych.

W dalszej części wstępu omówiono sieci konferencyjne, spektrum fal milimetrowych ze szczególnym uwzględnieniem pasma 60 GHz, wpływ środowiska propagacyjnego na projektowanie i instalację sieci oraz potencjał symulatora śledzenia promieni (*ang. ray tracing, RT*). Niezbędne referencje zawarte są w pracy doktorskiej [4].

1.1 Przypadek użycia: konferencja

Zgodnie z obowiązującą normą ISO dla systemów konferencyjnych [2], system konferencyjny jest to system składający się z urządzeń technicznych wykorzystywanych do prowadzenia konferencji. Jego podstawową funkcją jest wzmacnianie sygnałów audio od uczestników i innych źródeł audio oraz dostarczanie ich pozostałym uczestnikom.

Urządzenia konferencyjne opisane w normie ISO to m.in. punkt dostępowy (*ang. access point*, AP) i jednostka dyskusyjna (*ang. discussion unit*, DC). W dalszej części tej pracy antena nadawcza (TX) będzie odpowiadała za emulację AP, a antena odbiorcza (RX) za emulację DC, określanego również jako stacja (*ang. station, STA*) w standardzie IEEE 802.11.

Bezprzewodowe sieci konferencyjne działają głównie na częstotliwościach Wi-Fi, tj. 2,4 GHz i 5 GHz, ponieważ są one dostępne na całym świecie. Jest to ważny argument dla firm zajmujących się produkcją urządzeń i aplikacji do zastosowań konferencyjnych.

1.2 Spektrum fal milimetrowych

Jedną z interesujących częstotliwości jest pasmo 60 GHz, które jest pasmem nielicencjonowanym na całym świecie. Co również jest istotne, zaproponowano dla tego pasma poprawki do standardu IEEE 802.11, który jest podstawą rodziny certyfikacji Wi-Fi. Zatem, wykorzystanie standardu 802.11ad lub 802.11ay [3] umożliwiłoby łatwiejsze przejście bezprzewodowych sieci konferencyjnych z częstotliwości Wi-Fi w kierunku mmW.

Tłumienie w paśmie 60 GHz jest wyższe niż w pasmach 2,4 lub 5 GHz. Wynika to z tłumienia sygnału w wolnej przestrzeni, które wzrasta wraz z częstotliwością oraz z efektu absorpcji sygnału radiowego o częstotliwości 60 GHz przez tlen w atmosferze. Przez to zasięg sygnału jest ograniczony, więc rzeczone pasmo jest odpowiedniejsze do bezpośredniej komunikacji krótkiego zasięgu.

Warto również zbadać wykorzystanie częstotliwości innych niż 60 GHz, t.j. 26 GHz, 28 GHz i 40 GHz, które są już używane w zakresie częstotliwości 2 (FR2) sieci komórkowych 5G [1]. Jednak korzystanie z nich wymagałoby stworzenia dedykowanego protokołu.

1.2.1 Standard IEEE 802.11ay

Oprócz wytycznych dotyczących projektowania i wdrażania sieci wynikających ze standardu konferencyjnego, trzeba wziąć pod uwagę między innymi typ i strukturę przesyłanych danych. Dlatego istotnym zagadnie-

niem jest transmisja danych i struktura ramki wykorzystywana w zaproponowanym standardzie.

Ramka EDMG PPDU (*ang. Enhanced Directional Multi-Gigabit Physical Protocol Data Unit*) w standardzie IEEE 802.11ay wprowadza dodatkowe pola nagłówka i preambuły w porównaniu do DMG PPDU z IEEE 802.11ad. Wraz z użyciem większej liczby strumieni przestrzennych oraz trybu multi-user MIMO (MU-MIMO) wydłuża się czas trwania części ramki, która nie jest związana z długością przesyłanych danych.

Na czas trwania pola danych (*ang. Data*) wpływa długość odstępu ochronnego (*ang. guard interval*, GI) oraz wykorzystany profil schematów modulacji i kodowania (*ang. modulation and coding schemes*, MCS). Zdefiniowano trzy długości GI: krótki (32 symbole, 18,18 ns), normalny (*ang. normal GI*, NGI): (64 sym., 36,36 ns) i długi (128 sym., 72,72 ns). Jeden blok danych składa się z 512 symboli przeznaczonych na dane i GI. Krótszy GI pozwala na wykorzystanie większej liczby symboli na dane a co za tym idzie umożliwia szybszą transmisję tej samej informacji. Dodatkowo, standard IEEE 802.11ay definiuje 21 schematów MCS. Wyższy indeks MCS zakłada wyższy stopień modulacji i/lub kodowania.

1.3 Wpływ środowiska propagacyjnego

Ze względu na krótszą długość fali, zjawiska wpływające na propagację sygnału w paśmie fal milimetrowych muszą zostać szczegółowo zbadane. Np. obiekty uważane za gładkie dla niższych częstotliwości mogą być uznane za szorstkie i powodować efekt rozpraszania. Na propagację ma wpływ zarówno środowisko (jego kształt i wielkość) jak i konfiguracja anten oraz ich kierunkowość. Dlatego struktura środowiska, właściwości materiałów i wpływ konfiguracji anten powinny być rozpatrywane równolegle.

Na potrzeby omawianej pracy doktorskiej przeprowadzono kampanie pomiarowe w różnych środowiskach, takich jak sala audytoryjna, duża sala wykładowa, sala konferencyjna i sala seminaryjna. Badany był m.in. efekt wielodrogowy i charakterystyki tłumienne materiałów.

1.4 Ray Tracing

W paśmie mmW, gdzie długość fali jest znacznie krótsza niż wymiary obiektów znajdujących się w pomieszczeniach, powszechnie stosuje się symulacje oparte na technice śledzenia promieni, wykorzystującej prawa optyki geometrycznej.

Symulacje służą pogłębieniu zrozumienia zachowania mierzonych kanałów radiowych. Dzięki nim można zidentyfikować pochodzenie komponentów wielodrogowych i obiekty, które najbardziej wpływają na propagację wprowadzając dodatkowe odbicia lub blokując sygnał.

W komentowanej pracy do symulacji wykorzystano narzędzie opracowane w Instytucie iTEAM: symulator RT bazujący na silniku gier Unreal, który pozwala na precyzyjne symulowanie promieni świetlnych, które imitują ścieżki propagacji sygnału radiowego. W symulatorze stworzono repliki pomieszczeń, w których przeprowadzono pomiary oraz, dla znajdujących się w nich obiektów, wprowadzono pomierzone wartości tłumienia transmisji i odbicia uzyskane dla częstotliwości 26-40 GHz i 55-65 GHz.

2. WYNIKI

Na podstawie uzyskanych wyników powstały wytyczne, które mogą posłużyć jako wsparcie w projektowaniu bezprzewodowych sieci konferencyjnych.

2.1 Dystrybucja

Przy planowaniu rozmieszczenia urządzeń nadawczych i odbiorczych kluczowe znaczenie mają kształt i wielkość pomieszczenia. Aby poznać oczekiwane parametry kanału radiowego, takie jak np. średniokwadratowy rozkład opóźnienia (*ang. root-mean-square delay spread*, RMS DS), można skorzystać z modelu zaproponowanego przez ITU, który bierze pod uwagę pole powierzchni pomieszczenia (*ang. floor plan*, FP). We wzorze (1) RMS DS jest oznaczone jako S:

$$10 \times log_{10}(S) = 2.3 \times log_{10}(FP) + 11.0. \qquad (1)$$

Pomiary wykonane na wydziale ETI Politechniki Gdańskiej sugerują, że samo pole powierzchni nie wystarcza, aby poprawnie przewidzieć RMS DS. W Tabeli 1 przedstawiono wymiary trzech pomieszczeń, w których przeprowadzono pomiary odpowiedzi impulsowej dla częstotliwości 27 GHz oraz dwa parametry obliczone dla wartości progu 20 dB. Dla Audytorium 2 (Aud2), które jest największe pod względem pola powierzchni, mediana RMS DS uzyskana z pomiarów jest najmniejsza. Równie zaskakująca może być najmniejsza maksymalna liczba komponentów wielodrogowych (*ang. multipath components*, MPC). Nie jest to jednak błąd, a wpływ kształtu pomieszczenia w osi wertykalnej. Audytorium ma układ amfiteatrealny.

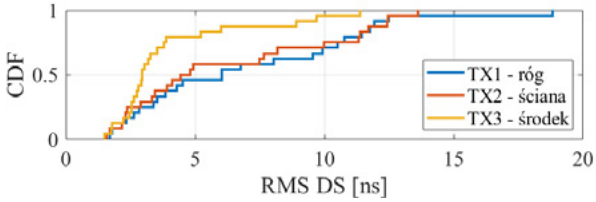
Tabela 1: Parametry środowisk pomiarowych.

Sala	*Wymiary* $[m \times m \times m]$	*Me S* $[ns]$	*max MPC*
Aud 2	12 x 16 x 3.8	2.1	6
NE308	8.2 x 20.8 x 3	9.4	18
EA427	4.5 x 7.1 x 3	3.9	9

Innym ciekawym aspektem Aud2 jest jego symetria. Ściany, drzwi i krzesła są rozlokowane w ten sam sposób po lewej i prawej stronie. Można się spodziewać, że sygnał odebrany na końcu sali w prawym i lewym rogu będzie taki sam. Tak nie jest. Jedną z przyczyn mogło być niecentralne ustawienie anteny nadawczej, jednak symulacje RT wskazują, że wpływ na tę różnicę mają niewielkie zmiany w środowisku: m.in. projektor zawieszony tylko po lewej stronie sali oraz typ materiału, z którego stworzone są suchościeralna tablica po prawej i ekran projektora po lewej stronie frontowej ściany.

Sala wykładowa NE308 jest pomieszczeniem o najwyższym współczynniku proporcji. Ze względu na to różnica między odległością od anteny nadawczej do najbliższego i najdalszego odbiornika jest dużo większa niż w Aud2 i sali seminaryjnej EA427. Jest to jedna z przyczyn, dla których RMS DS osiąga większe wartości.

Kampania pomiarowa przeprowadzona w sali konferencyjnej na Politechnice w Walencji (UPV) została wykorzystana do analizy wpływu położenia, wysokości i typu TX na odebrany sygnał. Podczas gdy RX znajdował się na stałej wysokości 1 m, TX we wszystkich trzech pozycjach (na środku pomieszczenia, w rogu i przy ścianie) zmieniał wysokość pomiędzy 1.6 m, 1.8 m i 2 m. Bazując na wynikach RMS DS (Rys. 1), najlepszą pozycją dla anteny TX jest środek pomieszczenia, gdzie najdłuższa możliwa odległość TX-RX wynosi połowę długości przekątnej pomieszczenia. Wpływa to na zakres wartości uzyskanych dla tłumienia w wolnej przestrzeni co jest szczególnie istotne w warunkach bezpośredniej widoczności (LOS). W przypadku krótszej odległości TX-RX względna moc odbierana jest wyższa, a RMS DS jest mniejsze.



Rysunek 1: CDF średniokwadratowego rozkładu opóźnienia (RMS DS).

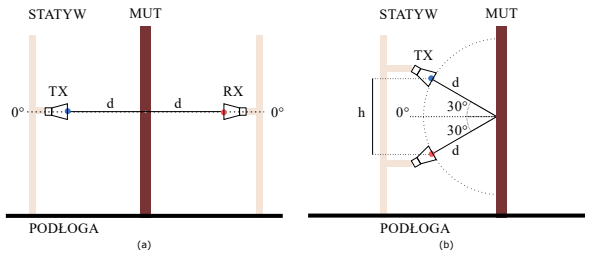
Wysokość anteny i jej pochylenie musi być powiązane z jej charakterystyką promieniowania. W przypadku anten tubowych bez pochylenia, umieszczenie anteny TX zdecydowanie wyżej niż RX zmniejsza maksymalną moc odbieraną. Większa liczba MPC jest zliczana dla zdefiniowanego progu, a RMS DS jest większe niż w przypadku anteny TX umieszczonej niżej. Gdy antena TX zostanie pochylona, a anteny RXs są umieszczone w jej głównym kierunku promieniowania, wyższe umieszczenie anteny TX pomaga uniknąć przeszkód umieszczonych między antenami TX i RX.

Jeśli antena TX jest anteną dookólną to maksymalna względna moc odbierana znacznie spada w porównaniu z anteną tubową. Aby zwiększyć moc odbieraną w RX, można użyć wyższej mocy nadawczej, jednak spowoduje to również wzrost mocy odbieranej ze wszystkich MPC. W związku z tym RMS DS pozostanie wyższy niż w przypadku anteny tubowej.

2.2 Materiały

Każdy materiał używany do budowy środowiska lub mebli ma swoją charakterystykę zależną od częstotliwości propagowanego sygnału. Oznacza to, że w zależności od częstotliwości sygnału, wystąpią różne straty związane z transmisją przez nią (*ang. transmission loss*, TL) lub odbiciem od niej (*ang. reflection loss*, RL). Znając częstotliwość, na której będzie działać sieć, można wybrać odpowiednie materiały w celu zwiększenia jej wydajności.

Wiedzę o cechach materiałów można wykorzystać w dwóch przypadkach: przy projektowaniu sali konferencyjnej lub poprzez dodanie obiektów o znanych właściwościach do istniejących środowisk. Wybór odpowiednich materiałów umożliwia sygnałom przenikanie niektórych obiektów lub blokowanie propagacji w niepożądanych kierunkach.



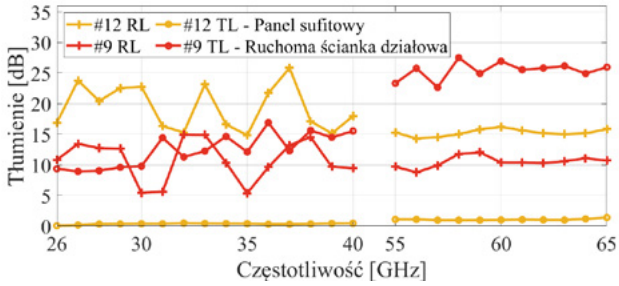
Rysunek 2: Schemat układu pomiarowego dla a) TL, b) RL.

Do uzyskania charakterystyk materiałów przeprowadzono dedykowaną kampanię pomiarową oraz wykorzystano metodę referencyjną. Metoda referencyjna opiera się na substrakcji sygnału odniesienia (RSL) mierzonego bez materiału pomiędzy antenami od sygnału pomiarowego z rozważanym materiałem jak na Rys. 2a) dla TL i Rys. 2b) dla RL. Metodę tę podaje równanie 2:

$$TL/RL\ [\text{dB}] = SL_{MUT}(f)\ [\text{dB}] - RSL(f)\ [\text{dB}], \quad (2)$$

gdzie SL_{MUT} odnosi się do SL otrzymanego z pomiarów w ustawieniu z Rys. 2a) dla TL i Rys. 2b) dla RL. MUT oznacza mierzony materiał (*ang. material under test*).

W sumie uzyskano charakterystyki dla 22 materiałów. Bardzo przezroczystym materiałem w obu mierzonych pasmach (26-40 GHz i 55-65 GHz) jest wełna mineralna, z której zbudowane są panele sufitowe. Jej charakterystyka jest przedstawiona na Rys. 3. Jeśli antena musi być ukryta przed uczestnikami konferencji, co jest częstym wymogiem, obiekt wykonany z wełny mineralnej może posłużyć za osłonę.



Rysunek 3: Charakterystyki tłumienia panelu sufitowego z wełny mineralnej i przegrody pomieszczenia.

Natomiast jeśli zależy nam na zablokowaniu transmisji sygnału w wybranym kierunku, pomoże użycie metalowych przedmiotów. Te jednak wprowadzają dodatkowe mało-stratne odbicia, które znacząco mogą wpłynąć na charakterystykę środowiska propagacyjnego. Istnieją obiekty, takie jak mierzona ruchoma ścianka działowa, które tłumią zarówno transmisję sygnału jak i jego odbicia. Mierzony materiał, którego charakterystyka również jest przedstawiona na Rys. 3, zauważalnie zmienia charakterystykę przy różnych zakresach częstotliwości. Znakomicie nadaje się do tłumienia sygnału w paśmie 55-65 GHz i zauważalnie tłumi sygnał odbity.

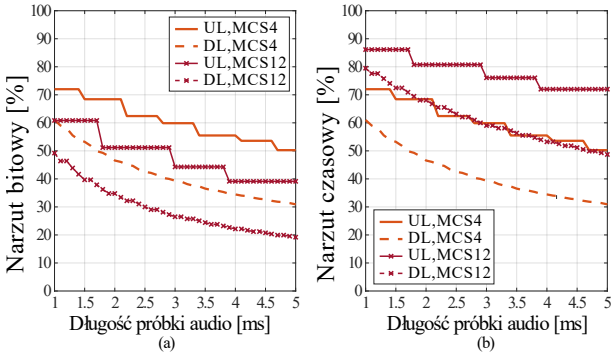
2.3 Sieć

Podstawowym typem danych przesyłanym w sieciach konferencyjnych jest dźwięk w czasie rzeczywistym, co oznacza przesyłanie na bieżąco danych o niewielkiej długości. W standardzie dla systemów konferencyjnych maksymalne opóźnienie dla dźwięku wynosi 20 ms. Z tego powodu zdecydowano, że długość próbki audio może wynosić maksymalnie 5 ms.

Aby określić optymalną długość próbki audio wzięto pod uwagę zużycie energii i, w przypadku transmisji w paśmie 60 GHz, narzut, jaki wprowadza standard IEEE 802.11ay. Do przeanalizowania aspektu związanego z użyciem energii wykorzystano model "per-frame" [5], który zakłada, że na każdą przesłaną ramkę przypada porcja mocy związana z jej przetworzeniem.

Standard 802.11ay dla pasma 60 GHz jest zaprojektowany do pracy z dużymi plikami. Wysyłanie ramek z tak małą liczbą bitów generuje duży narzut. Narzut może być rozpatrywany zarówno w kontekście liczby wykorzystanych bitów jak i czasu koniecznego na przesłanie części ramki niezwiązanej z danymi. Na Rys. 4 przedstawiono narzut dla ramek EDMG PPDU w zależności od długości próbki audio (1ms-5ms). Kolorem czerwonym oznaczono ramki, w których użyto MCS4, a bordowym MCS12. Dodatkowo ramki zostały opisane jako UL (*ang. uplink*) i DL (*ang. downlink*). Zgodnie ze standardem konferencyjnym maksymalnie 8 mikrofonów może być włączonych w tym samym momencie, tak więc ramki UL zawierają in-

formacje dostarczone od jednego użytkownika a ramki DL od ośmiu. Rys. 4a) przedstawia narzut bitowy obliczony zgodnie ze wzorem 3, a Rys. 4b) narzut czasowy zgodnie ze wzorem 4. P, Ng i GI_{pload} oznaczają odpowiednio długości preambuły, nagłówka i GI w bitach. PHY_{total} jest całkowitą długością ramki PPDU, a T_{chip} i T_{PPDU} to czas trwania jednego symbolu i całej ramki PPDU.



Rysunek 4: Narzut w ramkach EDMG PPDU z NGI.

$$Narzut_{bit}\ [\%] = \frac{P + Ng + GI_{pload}}{PHY_{total}} \times 100\ [\%] \qquad (3)$$

$$Narzut_{t}\ [\%] = \frac{(P + Ng + GI_{pload})/T_{chip}}{T_{PPDU}} \times 100\ [\%] \quad (4)$$

Narzut czasowy przyjmuje wyższe wartości: między 60% a 90% dla próbek audio o długości 1 ms (UL i DL) i między 30% a 75% dla próbek audio o długości 5 ms. W każdym z rozpatrywanych przypadków przez przynajmniej 30% czasu transmitowane są fragmenty ramek nieodpowiadające za przesyłanie danych. Zajmują przy tym przynajmniej 20% przesyłanych bitów. Na podstawie uzyskanych wyników zaleca się używanie mniejszej liczby dłuższych próbek zamiast wielu krótszych. W takim przypadku lepiej jest wysłać jedną ramkę zawierającą dane próbki audio o długości 5 ms zamiast pięciu ramek, z których każda zawiera próbkę audio o długości 1 ms. Mimo wszystko użycie standardu IEEE 802.11ay w sieciach konferencyjnych zorientowanych na przesyłanie sygnałów audio to rozwiązane nieefektywne. Użycie dostosowanego protokołu, niezwiązanego z rodziną standardów IEEE 802.11, może zmniejszyć wpływ narzutu wprowadzanego przez nagłówki, preambuły i interwały ochronne.

Jedna z technologii, która mogłaby zostać użyta w sieciach konferencyjnych niezależnie od użytego protokołu, jest MU-MIMO. RX, które są przypisane do jednej grupy, muszą być nieskorelowane. Dla sygnałów pomierzonych w salach przedstawionych w Tabeli 1 obliczono współczynniki korelacji. W małych środowiskach (EA427) korelacja jest wysoka w porównaniu do dużych sal (Aud2, NE308). Korelacja jest również wysoka, gdy komponent LOS jest obecny i odbierany z podobnym opóźnieniem przez różne RX. Tak więc, odległość między różnymi RX jest jednym z czynników odgrywających rolę w grupowaniu stacji w technologii MU-MIMO.

3. PODSUMOWANIE I WNIOSKI

W niniejszej pracy zaproponowano szereg wytycznych, które są związane z inteligentnym projektowaniem i wdrażaniem konferencyjnych sieci bezprzewodowych w dużych pomieszczeniach w częstotliwościach fal milimetrowych. Przedstawiono dane potwierdzające praktyki stosowane podczas wdrażania sieci konferencyjnych.

Wykazano, że kształt i wielkość pomieszczenia powinny być rozpatrywane zarówno w przekroju pionowym jak i poziomym. Symetryczna budowa pomieszczenia i niesymetryczny układ mebli muszą być wzięte pod uwagę. Jeżeli pomieszczenie nie jest zbudowane w układzie amfiteatralnym to AP powinien zostać umieszczony w centrum wdrażanego systemu tak, aby odległości do wszystkich DU była podobna. Jeśli nie jest to możliwe, montując go w rogu pomieszczenia lub na jednej ze ścian należy umieścić go jak najwyżej i pod odpowiednim kątem.

Wełna mineralna została zidentyfikowana jako materiał, który może zostać wykorzystany do ukrycia AP podczas konferencji i nie wprowadzić przy tym znacznego tłumienia. Znajomość charakterystyk tłumiennych materiałów, które znajdują się w salach konferencyjnych jest istotna dla optymalizacji sieci. Istotna może być nie tylko wartość tłumienia jakie wprowadzają, ale też ich zmienność w funkcji częstotliwości.

Biorąc pod uwagę zużycie mocy i narzut wprowadzany w ramkach PPDU w IEEE 802.11ay dla pasma 60 GHz, najlepszym rozwiązaniem jest wysyłanie jak najdłuższych ramek jak najrzadziej. Dla rozwiązań konferencyjnych są to ramki audio o długości 5 ms. Trzeba jednak zaznaczyć, że transmisja w tym standardzie nie jest efektywna i proponuje się stworzenie dedykowanego protokołu transmisji, na tej lub innej częstotliwości w paśmie mmW.

PODZIĘKOWANIA

Praca doktorska komentowana w tym artykule została częściowo ufundowana przez H2020 Marie Skłodowska-Curie Actions Program w ramach Grantu Projektowego nr 766231 WAVECOMBE - ITN - 2017, przy współpracy z Televic Conference, Belgia.

Kampanie pomiarowe na Politechnice Gdańskiej były częściowo finansowane przez Akcję COST CA20120 INTERACT w ramach programu grantowego Short-Term Scientific Mission (STSM).

LITERATURA

[1] 3rd Generation Partnership Project. 2019. "3GPP Release 15 Specifications"

[2] ISO. 2019. “Conference systems — Equipment — Requirements, ISO/FDIS 22259:2018(E)”

[3] IEEE. 2021. "IEEE 802.11ay-2021: IEEE Standard for Information Technology—Telecommunications and Information Exchange between Systems Local and Metropolitan Area Networks—Specific Requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 2: Enhanced Throughput for Operation in License-exempt Bands above 45 GHz"

[4] Drozdowska Monika. 2024. “Intelligent Design of mmWave Conference Networks in Built Environments”, Rozprawa doktorska, Departamento de Comunicaciones, Universitat Politecnica de Valencia

[5] Serrano, Pablo., Garcia-Saavedra, Andres, Bianchi, Giuseppe, Banchs, Albert, and Azcorra, Arturo. 2014. "Per-frame energy consumption in 802.11 devices and its implication on modeling and design". IEEE/ACM Transactions on networking, 23(4), 1243-1256.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

Badania ab initio właściwości elektrooptycznych nanowarstw b-As_xP_{1-x} jako materiałów do nowej generacji fotodetektorów dla pasm telekomunikacyjnych.

Ab initio studies of electro-optical properties of b-As_xP_{1-x} nanolayers as materials for new generation photodetectors for telecommunications bands.

Krzysztof Pyrchla[1]; Robert Bogdanowicz[2]

[1] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Katedra Metrologii i Optoelektroniki, ul. Gabriela Narutowicza 11/12, 80-233 Gdańsk, Polska, krzysztof.pyrchla@pg.edu.pl
[2] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Katedra Metrologii i Optoelektroniki, ul. Gabriela Narutowicza 11/12, 80-233 Gdańsk, Polska, robbogda@pg.edu.pl

DOI: 10.15199/59.2025.4.51

Streszczenie: **Czarny arsenofosforen (b-AsP) jest nowoczesnym nanomateriałem o regulowanej, prostej przerwie energetycznej oraz wysokiej ruchliwości nośników, a jednocześnie charakteryzuje się niską podatnością na degradację w stosunku do innych półprzewodników z czarnego fosforu. W artykule przedstawiono wyniki symulacji *ab initio* dających wgląd w elektronową strukturę pasmową tego materiału i jego parametry fotoniczne. Wskazujemy, że fosforen domieszkowany arsenem przy właściwej koncentracji stanowi materiał o dużym potencjale do zastosowania w optoelektrokomunikacji.**

Abstract: **Black arsenophosphorene (b-AsP) is a modern nanomaterial with a tunable, direct bandgap and high carrier mobility, and at the same time it is characterized by low susceptibility to degradation compared to other black phosphorus semiconductors. The paper presents the results of ab initio simulations providing insight into the electronic band structure of this material and its photonic parameters. We indicate that arsenic-doped phosphorene at the right concentration is a material with great potential for use in optoelectrocommunication.**

Słowa kluczowe: **DFT, Metody ab initio, materiały 2D, fotodetektory, półprzewodniki**

Keywords: **Ab initio Methods, DFT, Photodetectors, Semiconductors, 2D Materials**

1. WSTĘP

W ostatnich latach materiały dwuwymiarowe (2D), takie jak grafen, dichalkogenki metali przejściowych (TMDs), fosforen czy silicen, zyskały ogromne zainteresowanie w środowisku naukowym ze względu na swoje unikalne właściwości fizykochemiczne [1], [7], [13]. Charakteryzujące się atomową grubością, silną anizotropią oraz możliwością precyzyjnej kontroli właściwości elektronicznych i optycznych, materiały te oferują nowe możliwości w projektowaniu zaawansowanych urządzeń elektronicznych i optoelektronicznych [21]. W szczególności ich potencjał w dziedzinie optoelektrokomunikacji, łączącej przesyłanie danych za pomocą sygnałów optycznych z ich konwersją i przetwarzaniem w układach elektronicznych, stanowi obiecujący kierunek rozwoju nowoczesnych technologii komunikacyjnych [19], [23].

Tradycyjne systemy komunikacji optycznej, oparte głównie na komponentach krzemowych, napotykają na ograniczenia związane z miniaturyzacją, efektywnością przełączania oraz integracją z fotonicznymi układami scalonymi [18]. Materiały 2D oferują w tym kontekście szereg przewag, takich jak wysoka mobilność nośników, regulowana przerwa energetyczna, silne sprzężenie światło–materiał, a także możliwość integracji na różnorodnych podłożach [5], [11], [12]. Dzięki temu możliwe jest projektowanie nowej generacji fotodetektorów, modulatorów, źródeł światła czy przełączników optycznych, które cechują się wysoką szybkością działania, niskim poborem mocy i kompatybilnością z elastyczną elektroniką [10], [15].

Wśród materiałów 2D szczególną uwagę zwraca fosforen: jedno lub kilkuwarstwowa forma czarnego fosforu, który dzięki swojej bezpośredniej i regulowalnej przerwie energetycznej (w zakresie od ~0,3 eV dla formy objętościowej BP do ~2 eV dla monowarstwy) oraz wysokiej mobilności nośników (~1000 $cm^2/V{\cdot}s$) stanowi bardzo obiecującą platformę dla fotodetektorów działających w szerokim zakresie widmowym, od światła widzialnego po bliską podczerwień (NIR) [2], [9], [22]. Dodatkowo, silna anizotropia optyczna i elektryczna fosforenu umożliwia projektowanie układów o kierunkowo zależnych właściwościach detekcyjnych, co może być wykorzystane w systemach optycznej polarymetrii lub modulacji kierunkowej [24].

W kontekście optoelektrokomunikacji, gdzie kluczowe znaczenie ma szybka detekcja sygnałów optycznych przy niskim poziomie szumów i minimalnym zużyciu energii, fosforen wyróżnia się korzystną kombinacją parametrów. Jednak jego ograniczona stabilność chemiczna w warunkach atmosferycznych pozostaje istotnym wyzwaniem praktycznym [3]. W odpowiedzi na ten problem, jednym z obiecujących podejść jest domieszkowanie fosforenu atomami arsenu, co prowadzi do utworzenia tzw. arsenofosforenu (AsP). Taka modyfikacja nie tylko zwiększa stabilność materiału poprzez poprawę odporności na utlenianie, ale także pozwala na dostrojenie jego właściwości elektronicznych i optycznych [25].

Badania wykazały, że arsenofosforen może wykazywać poszerzony zakres absorpcji w podczerwieni oraz zwiększoną gęstość stanów elektronowych w rejonie krawędzi przerwy energetycznej, co sprzyja efektywniejszemu pochłanianiu promieniowania NIR i MIR (mid-infrared) [6]. Dzięki temu możliwe jest projektowanie wysokowydajnych, skalowalnych i integracyjnych fotodetektorów opartych na AsP, które odpowiadają wymaganiom nowoczesnych systemów optoelektrokomunikacyjnych, zwłaszcza w zakresie 1,3–1,55 µm — typowym dla transmisji światłowodowej [14]. Jednakże potencjał b-AsP w generowaniu impulsów i wykrywaniu światła w pasmach telekomunikacyjnych nie został jeszcze w pełni zbadany.

W celu zapewnienia bardziej szczegółowego wglądu w właściwości optyczne AsP, w artkule przedstawiono wyniki badania symulacyjnego *ab initio* właściwości optycznych nanowarstw b-AsP i porównano je do właściwości czystych nanowastw fosforenowych. Przyjęto następujący podział treści artykułu: Sposób w jaki utworzono modele atomowe oraz metody i ich implementacje zastosowane w obliczeniach przedstawiono w sekcji drugiej. Wyniki obliczeń struktury pasmowej oraz gęstość ładunku odpowiadającą maksimum pasma walencyjnego i minimum pasma przewodnictw przedstawiono w sekcji trzeciej. Sekcję czwartą stanowi podsumowanie otrzymanych wyników wraz z wnioskami dotyczącymi potencjału b-AsP w kontekście budowy fotodetektorów telekomunikacyjnych.

2. EKSPERYMENT

W celu uzyskania szczegółowego opisu struktury elektronowej fosforenu zastosowano obliczenia *ab initio* w ramach teorii funkcjonału gęstości (DFT), wykorzystując podejście bazujące na liniowej kombinacji orbitali atomowych (LCAO). Metoda ta umożliwia efektywne modelowanie właściwości materiałów o złożonej strukturze warstwowej, zapewniając dobrą równowagę pomiędzy dokładnością a kosztami obliczeniowymi [17]. Zarówno do tworzenia modeli atomowych nanostruktur jak i prowadzenia obliczeń wykorzystano oprogramowanie Quantum ATK [16].

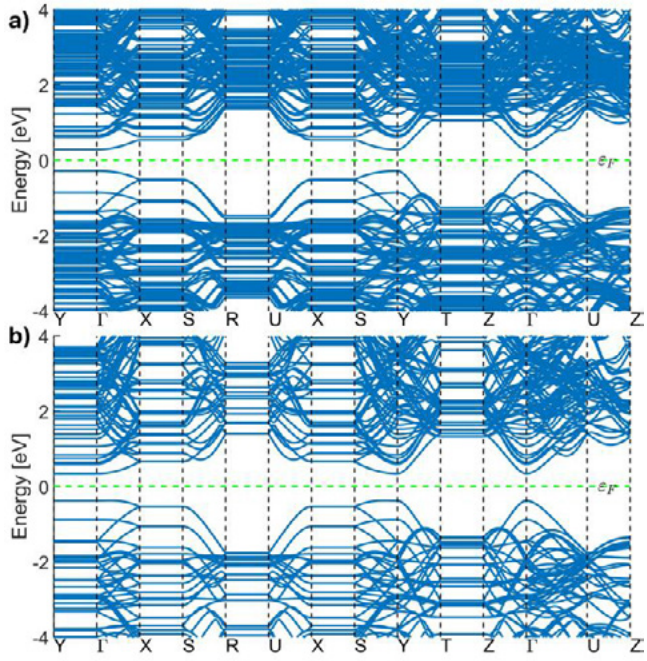
Przygotowano serię modeli atomowych nanostruktury fosforenowej i fosforoarsenowej. Modele reprezentowały struktury nanowarstwy o 1, 2 i 3 warstwach oraz struktury objętościowe dla obydwu struktur. Wymiary każdej z warstw struktury były równe 3 stałym sieci w kierunku zigzag oraz 2 stałym sieci w kierunku armchair, co przekładało się na 24 atomy w warstwie. Ułożenie domieszkowania arsenem zostało wygenerowane w sposób pseudolosowy. Na podstawie danych eksperymentalnych, proporcję atomów określono w każdym przypadku na 40 % As i 60% P.

Przygotowane modele poddano optymalizacji geometrycznej wykorzystując algorytm optymalizacyjny Broyden−Fletcher−Goldfarb−Shanno (LBFGS) [8]. Optymalizacji poddano zarówno położenie atomów jak i wymiary komórki elementarnej. Jako kryterium zakończenia optymalizacji przyjęto spadek modułu największej siły działającej na atom poniżej poziomu 0.01 eV/Å oraz spadek ciśnienia poniżej 10^{-5} eV/Å^3. Siły działające na atomy były obliczane z twierdzenia Hellmann'a–Feynman'a w każdym kroku optymalizacji.

Zarówno do obliczeń sił w czasie optymalizacji oraz elektronowej struktury pasmowej zastosowano metodę LCAO. Aby poprawnie oddać wielkość przerwy energetycznej badanej struktury zastosowano hybrydowy funkcjonał wymiany korelacji HSE06[4]. Dla uzyskania odpowiedniej rozdzielczości w przestrzeni odwrotnej zastosowano siatkę k-punktów generowaną według schematu Monkhorsta-Packa. Ze względu na asymetrię super komórki obliczeniowej, liczba k-punktów była zależna od kierunku i wynosiła 5x5x2, odpowiednio dla kierunków zigzag, armchair i między warstwowego.

3. WYNIKI

W celu zbadania właściwości optycznych b-$As_{0.4}P_{0.6}$, teoretycznie obliczono strukturę pasmową i rozkład przestrzenny gęstości ładunku odpowiadającą maksimum pasma walencyjnego i minimum pasma przewodnictwa. Na rysunku 1 przedstawiono wykres struktury pasmowej trójwarstwowego fosforoarsenu oraz fosforenu.

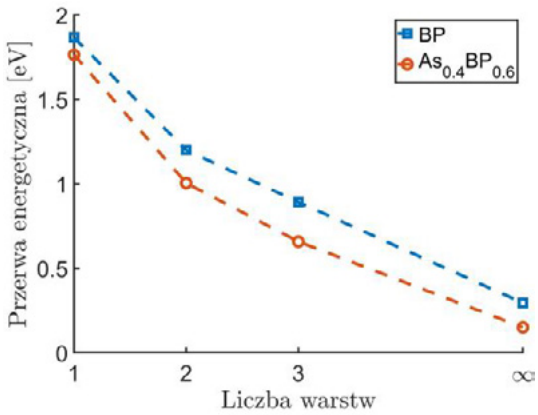


Rys. 1. Struktura pasmowa: a) trójwarstwowego fosforoarsenu oraz b) trójwarstwowego fosforenu obliczona wzdłuż ścieżki wysokiej symetrii przez strefę Brillouina.

Zarówno w przypadku fosforenu domieszkowanego arsenem (rysunek 1a) jak i czystego fosforenu (Rysunek 1b) nie stwierdzono występowania skośnej przerwy energetycznej, co jest korzystne z punktu widzenia zastosowań w sensorach optoelektronicznych. Ponadto, zgodnie z przeprowadzonymi obliczeniami, szerokość przerwy energetycznej dla fosforoarsenu jest niższa o około 0.2 eV.

W celu sprawdzenia zależności przerwy energetycznej od ilości warstw, obliczenia powtórzono dla każdego z przygotowanych modeli. Otrzymane wyniki szerokości

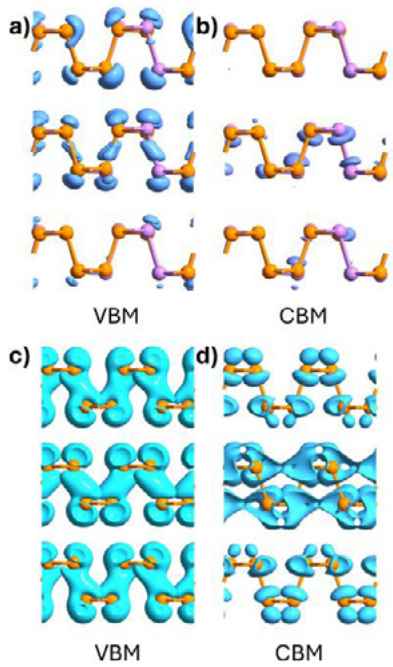
prostej przerwy energetycznej względem liczby warstw struktury przedstawiono na rysunku 2.



Rys. 2. Zależność przerwy energetycznej od liczby warstw dla czystego fosforenu oraz $As_{0.4}P_{0.6}$.

Wraz ze wzrostem liczby warstw, przerwa energetyczna obydwu struktur spada, zachowując obserwowane dla czystego fosforenu prawa skalowania [20]. Dla wszystkich zbadanych modeli, przerwa energetyczna $As_{0.4}P_{0.6}$ była od 0.1 do 0.2 eV niższa niż w przypadku czystego fosforenu. Z tego względu materiał ten, osiąga użyteczną telekomunikacyjnie przerwę energetyczną już dla struktur kilkuwarstwowych (>=3 warstwy) bez potrzeby stosowania materiału objętościowego.

Dla zastosowań optoelektronicznych, poza samą szerokością przerwy energetycznej, istotne są również mechanizmy przewodnictwa materiału. Na rysunku 3 przedstawiono obliczony rozkład przestrzenny wybranych pasm elektronowych.



Rys. 3. Rozkład przestrzenny: a) maksimum pasma walencyjnego (VBM) $As_{0.4}P_{0.6}$ oraz jego b) minimum pasma przewodnictwa (CBM), c) VBM forsofrenu oraz jego CBM.

W przeciwieństwie do BP (Rysunek 3c i d) , pary elektron-dziura w $As_{0.4}P_{0.6}$ są dobrze rozdzielone (Rysunek 3a i b), co sugeruje niższą szybkość rekombinacji i dłuższy czas relaksacji fotonośników.

4. PODSUMOWANIE

Podsumowując, w artykule przedstawiono wyniki badania *ab initio* właściwości optoelektrycznych nanowarstw $As_{0.4}P_{0.6}$ oraz BP. Analiza struktury pasmowej $As_{0.4}P_{0.6}$ wykazała istnienie bezpośredniej przerwy energetycznej w zakresie odpowiadającym podczerwieni bliskiej i środkowej, co jest szczególnie istotne w kontekście detekcji sygnałów w pasmach telekomunikacyjnych (około 1,3–1,55 μm). Wyniki wskazują również, że istnieje możliwość zmiany szerokości przerwy energetycznej badanej nanostruktury. Zapewnia to możliwość wykonania z jednego materiału detektorów pracujących z maksymalną efektywnością w różnych pasmach telekomunikacyjnych co stanie się szczególnie istotne wraz z dalszym rozwojem telekomunikacji optycznej.

Dodatkowo, dobre rozdzielenie par elektron-dziura sugeruje długi czas relaksacji fotonośników. Jest on wymagany dla uzyskania dobrej czułości oraz szybkiego tempa przełączania, które są kluczowe dla wysokowydajnych układów optycznych.

Domieszkowanie fosforenu arsenem pozwala nie tylko na stabilizację struktury i poprawę odporności chemicznej materiału, ale również na korzystne dostrojenie właściwości elektronowych, zwiększając jego funkcjonalność w projektowaniu komponentów aktywnych. Uzyskane wyniki wskazują, że AsBP może pełnić rolę wydajnego i łatwo integrowalnego materiału warstwowego w fotodetektorach przeznaczonych do pracy w elastycznych, zminiaturyzowanych układach fotonicznych nowej generacji, pracujących zarówno w klasycznych pasmach telekomunikacyjnych jak i tych niedostępnych dla sensorów krzemowych.

Obliczenia wykonano z wykorzystaniem komputerów Centrum Informatycznego Trójmiejskiej Akademickiej Sieci Komputerowej" (Computations were carried out using the computers of Centre of Informatics Tricity Academic Supercomputer & Network)

LITERATURA

[1] Chhowalla Manish, Shin Hyeon Suk, Eda Goki, Li Lain-Jong, Loh Kian Ping, Zhang Hua. 2013. 'The chemistry of two-dimensional layered transition metal dichalcogenide nanosheets'. Nature Chemistry 5 (4) : 263–275.]

[2] Engel Michael, Steiner Mathias, Lombardo Antonio, Ferrari Andrea C., Löhneysen Hilbert V., Avouris Phaedon, Krupke Ralph. 2012. 'Light–matter interaction in a microcavity-controlled graphene transistor'. Nature Communications 3 (1) : 906.]

[3] Favron Alexandre, Gaufrès Etienne, Fossard Frédéric, Phaneuf-L'Heureux Anne-Laurence,

Tang Nathalie Y-W., Lévesque Pierre L., Loiseau Annick, Leonelli Richard, Francoeur Sébastien, Martel Richard. 2015. 'Photooxidation and quantum confinement effects in exfoliated black phosphorus'. Nature Materials 14 (8) : 826–832.]
[4] Heyd Jochen, Scuseria Gustavo E., Ernzerhof Matthias. 2003. 'Hybrid functionals based on a screened Coulomb potential'. The Journal of Chemical Physics 118 (18) : 8207–8215.]
[5] Koppens F. H. L., Mueller T., Avouris Ph., Ferrari A. C., Vitiello M. S., Polini M. 2014. 'Photodetectors based on graphene, other two-dimensional materials and hybrid systems'. Nature Nanotechnology 9 (10) : 780–793.]
[6] Kou Liangzhi, Ma Yandong, Sun Ziqi, Heine Thomas, Chen Changfeng. 2017. 'Two-Dimensional Topological Insulators: Progress and Prospects'. The Journal of Physical Chemistry Letters 8 (8) : 1905–1919.]
[7] Li Likai, Yu Yijun, Ye Guo Jun, Ge Qingqin, Ou Xuedong, Wu Hua, Feng Donglai, Chen Xian Hui, Zhang Yuanbo. 2014. 'Black phosphorus field-effect transistors'. Nature Nanotechnology 9 (5) : 372–377.]
[8] Liu Dong C., Nocedal Jorge. 1989. 'On the limited memory BFGS method for large scale optimization'. Mathematical Programming 45 (1–3) : 503–528.]
[9] Liu Han, Neal Adam T., Zhu Zhen, Luo Zhe, Xu Xianfan, Tománek David, Ye Peide D. 2014. 'Phosphorene: An Unexplored 2D Semiconductor with a High Hole Mobility'. ACS Nano 8 (4) : 4033–4041.]
[10] Liu Ming, Yin Xiaobo, Ulin-Avila Erick, Geng Baisong, Zentgraf Thomas, Ju Long, Wang Feng, Zhang Xiang. 2011. 'A graphene-based broadband optical modulator'. Nature 474 (7349) : 64–67.]
[11] Low Tony, Chaves Andrey, Caldwell Joshua D., Kumar Anshuman, Fang Nicholas X., Avouris Phaedon, Heinz Tony F., Guinea Francisco, Martin-Moreno Luis, Koppens Frank. 2017. 'Polaritons in layered two-dimensional materials'. Nature Materials 16 (2) : 182–194.]
[12] Mak Kin Fai, Lee Changgu, Hone James, Shan Jie, Heinz Tony F. 2010. 'Atomically Thin MoS 2 : A New Direct-Gap Semiconductor'. Physical Review Letters 105 (13) : 136805.]
[13] Novoselov K. S., Geim A. K., Morozov S. V., Jiang D., Zhang Y., Dubonos S. V., Grigorieva I. V., Firsov A. A. 2004. 'Electric Field Effect in Atomically Thin Carbon Films'. Science 306 (5696) : 666–669.]
[14] Qiao Jingsi, Kong Xianghua, Hu Zhi-Xin, Yang Feng, Ji Wei. 2014. 'High-mobility transport anisotropy and linear dichroism in few-layer black phosphorus'. Nature Communications 5 (1) : 4475.]
[15] Schall Daniel, Neumaier Daniel, Mohsin Muhammad, Chmielak Bartos, Bolten Jens, Porschatis Caroline, Prinzen Andreas, Matheisen Christopher, Kuebart Wolfgang, Junginger Bernhard, Templ Wolfgang, Giesecke Anna Lena, Kurz Heinrich. 2014. '50 GBit/s Photodetectors Based on Wafer-Scale Graphene for Integrated Silicon Photonic Communication Systems'. ACS Photonics 1 (9) : 781–784.]
[16] Smidstrup Søren, Markussen Troels, Vancraeyveld Pieter, Wellendorff Jess, Schneider Julian, Gunst Tue, Verstichel Brecht, Stradi Daniele, Khomyakov Petr A, Vej-Hansen Ulrik G, Lee Maeng-Eun, Chill Samuel T, Rasmussen Filip, Penazzi Gabriele, Corsetti Fabiano, Ojanperä Ari, Jensen Kristian, Palsgaard Mattias L N, Martinez Umberto, Blom Anders, Brandbyge Mads, Stokbro Kurt. 2020. 'QuantumATK: an integrated platform of electronic and atomic-scale modelling tools'. Journal of Physics: Condensed Matter 32 (1) : 015901.]
[17] Soler José M, Artacho Emilio, Gale Julian D, García Alberto, Junquera Javier, Ordejón Pablo, Sánchez-Portal Daniel. 2002. 'The SIESTA method for *ab initio* order- *N* materials simulation'. Journal of Physics: Condensed Matter 14 (11) : 2745–2779.]
[18] Soref Richard. 2006. 'The Past, Present, and Future of Silicon Photonics'. IEEE Journal of Selected Topics in Quantum Electronics 12 (6) : 1678–1687.]
[19] Sorianello V., Midrio M., Contestabile G., Asselberghs I., Van Campenhout J., Huyghebaert C., Goykhman I., Ott A. K., Ferrari A. C., Romagnoli M. 2018. 'Graphene–silicon phase modulators with gigahertz bandwidth'. Nature Photonics 12 (1) : 40–44.]
[20] Sorkin V., Cai Y., Ong Z., Zhang G., Zhang Y. W. 2017. 'Recent Advances in the Study of Phosphorene and its Nanostructures'. Critical Reviews in Solid State and Materials Sciences 42 (1) : 1–82.]
[21] Wang Qing Hua, Kalantar-Zadeh Kourosh, Kis Andras, Coleman Jonathan N., Strano Michael S. 2012. 'Electronics and optoelectronics of two-dimensional transition metal dichalcogenides'. Nature Nanotechnology 7 (11) : 699–712.]
[22] Xia Fengnian, Wang Han, Jia Yichen. 2014. 'Rediscovering black phosphorus as an anisotropic layered material for optoelectronics and electronics'. Nature Communications 5 (1) : 4458.]
[23] Xia Fengnian, Wang Han, Xiao Di, Dubey Madan, Ramasubramaniam Ashwin. 2014. 'Two-dimensional material nanophotonics'. Nature Photonics 8 (12) : 899–907.]
[24] Yuan Hongtao, Liu Xiaoge, Afshinmanesh Farzaneh, Li Wei, Xu Gang, Sun Jie, Lian Biao, Curto Alberto G., Ye Guojun, Hikita Yasuyuki, Shen Zhixun, Zhang Shou-Cheng, Chen Xianhui, Brongersma Mark, Hwang Harold Y., Cui Yi. 2015. 'Polarization-sensitive broadband photodetector using a black phosphorus vertical p–n junction'. Nature Nanotechnology 10 (8) : 707–713.]
[25] Zhang Shengli, Xie Meiqiu, Li Fengyu, Yan Zhong, Li Yafei, Kan Erjun, Liu Wei, Chen Zhongfang, Zeng Haibo. 2016. 'Semiconducting Group 15 Monolayers: A Broad Range of Band Gaps and High Carrier Mobilities'. Angewandte Chemie 128 (5) : 1698–1701.]

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

KOMPENSACJA NIELINIOWOŚCI DLA KOMUNIKACJI W ŚWIETLE WIDZIALNYM

NONLINEARITY COMPENSATION FOR VISIBLE LIGHT COMMUNICATIONS

Juliusz Bojarczuk[1]; Jakub Kasjanowicz[1]; Michał Marzęcki[1]; Grzegorz Stępniak[1]

[1] Instytut Telekomunikacji i Cyberbezpieczeństwa, Politechnika Warszawska, Warszawa
juliusz.bojarczuk@pw.edu.pl

DOI: 10.15199/59.2025.4.52

Streszczenie: W związku z dynamicznym rozwojem komunikacji w świetle widzialnym, konieczne jest stosowanie coraz to nowszych metod pozwalających na zwiększanie pojemności informacyjnej takich systemów. Jednym z głównych ograniczeń jest nieliniowość diody nadawczej. W referacie przedstawiono wybrane metody pozwalające na kompensację zniekształceń nieliniowych generowanych przez diody LED szeroko stosowane w VLC. Porównano zarówno metody stosowane po stronie nadawczej jak i odbiorczej dla modulacji jedno i wielotonowych pozwalające na zwiększenie pojemności informacyjnej o kilkanaście procent.

Abstract: With the rapid expansion of VLC, ever more sophisticated techniques are needed to push the information throughput of these systems. A key limiting factor is the inherent nonlinearity of the LED. This paper presents selected methods for compensating the nonlinear distortions produced by the LEDs widely used in VLC. Both transmitter-side and receiver-side techniques are compared for single-tone and multi-tone modulation schemes, showing that the proposed solutions can boost the channel capacity by on the order of a dozen percent.

Słowa kluczowe: nieliniowość, Volterra, kompensacja, VLC, LED.

Keywords: nonlinearity, Volterra, compensation, VLC, LED.

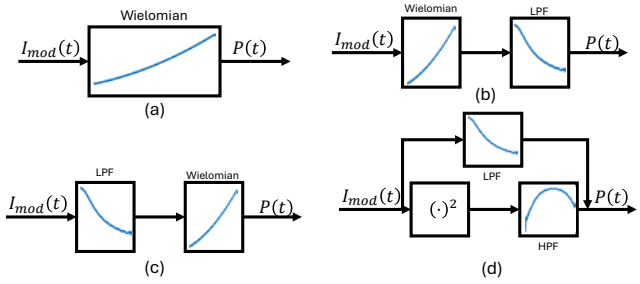
1. WSTĘP

Powszechny dostęp radiowy do sieci Internet przyniósł postęp technologii w wielu różnych dziedzinach życia, np. telemedycynie, pojazdach autonomicznych, konsumpcji multimediów, jak również umożliwia rozwój aplikacji wymagających komunikacji z ogromną liczbą urządzeń, np. IoT. Jednocześnie, widmo radiowe jest w zakresie do ok. 300 GHz już w większości zajęte przez różne standardy komunikacyjne, a sieci 5G (i przyszłe 6G zapewne tym bardziej) wzbudzają niestety szeroki sprzeciw ze strony opinii publicznej z obawy o oddziaływanie na zdrowie. Stąd, do zastosowań sieci 6G poszukuje się alternatywnych sposobów transmisji. Jednym z najbardziej perspektywicznych jest optyczna transmisja krótkiego zasięgu w świetle widzialnym. W szczególności wykorzystanie diod LED (nawet tych, które obecnie służą do oświetlania pomieszczeń) jako źródeł światła dla komunikacji optycznej budzi szczególne nadzieje. Diody te są tanie, powszechne, energooszczędne, odporne na zakłócenia EM, a ograniczenie transmisji do pojedynczego pomieszczenia zapewniają bezpieczeństwo transmisji na poziomie fizycznym [4]. Niestety, jako źródła światła diody LED posiadają pewne ograniczenia. Pierwszym z nich jest pasmo modulacji. Tu, typowe diody oświetleniowe z powłoką fosforescencyjną mają pasmo nie większe niż 10-20 MHz. Prowadzone są natomiast liczne prace nad zwiększeniem tego pasma w innych strukturach – np. RC-LED czy, mikro-LED [12]. Drugim, nieodłącznym ograniczeniem, szczególnie dla zaawansowanych formatów modulacji, bez których nie da się osiągać wielo-gigabitowych przepływności, jest nieliniowość diody. Chociaż dioda półprzewodnikowa jako źródło modulowanego sygnału optycznego znana jest od ponad 50 lat, intensywne prace nad przeciwdziałaniem nieliniowości trwają dopiero od kilkunastu. Istotny problem badawczy stanowi model tej diody. Wykorzystywane są następujące modele: wielomianowy statyczny, oraz dynamiczne: Wienera, Hammersteina, ogólny Volterry i równań różniczkowych (r.r.). Niestety, tylko ostatnie dwa pozwalają na poprawne odwzorowanie sygnału z wyjścia diody. W [2] pokazaliśmy, że możliwe jest wyprowadzenie z r.r. modelu Volterry, którego struktura znacząco się upraszcza i może być przedstawiona jako połączenie bloków filtrów i bloku nieliniowego. Ta struktura ma znaczące implikacje dla systemów komunikacyjnych pracujących z wykorzystaniem LED jako źródeł światła. W tej pracy pokażemy, iż pozwala ona: eliminować całkowicie nieliniowość 2 rzędu przez liniowy predystorter; wyjaśnić dlaczego prekorekcja liniowa typu FFE w nadajniku jest znacząco lepsza niż w odbiorniku (chociaż z teorii komunikacji wiemy, że powinny być równoważne); czy też efektywniej przyporządkowywać bity i moc na podnośną w systemach wielotonowych. Po raz pierwszy zestawiono porównanie metod, co jest istotne z punktu widzenia finalnego wyboru metody pozwalającej na ograniczenie zniekształceń nieliniowych.

2. MODEL DIODY LED

W celu opracowania wydajnych metod kompensacji zniekształceń nieliniowych konieczne jest bazowanie na dokładnych modelach numerycznych opisujących możliwie najlepiej zjawiska nieliniowe degradujące jakość transmisji. W literaturze można spotkać szereg różnych modelów diody LED [2]. Najprostszym modelem jest model statyczny, który opisuje wielomianową relację pomiędzy wyjściową mocą optyczną a przebiegiem prądu modulacji w

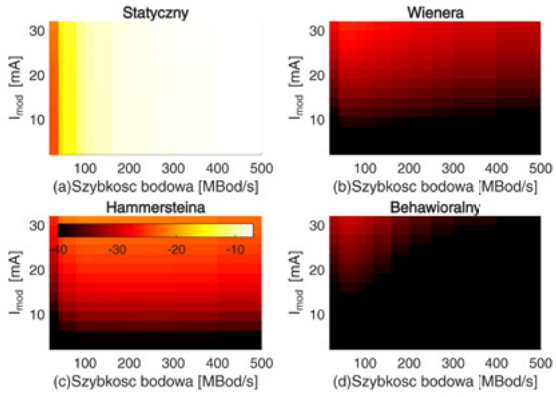
czasie. Najczęściej, współczynniki wielomianu mierzy się w sposób statyczny poprzez pomiar krzywej LIV i jej aproksymację, co jednak znacząco odbiega od charakterystyki diody modulowanej szybkimi sygnałami, co pokazano w [2]. Bardziej złożone modele diody LED uwzględniają jej dynamiczny charakter poprzez konkatenację modelu statycznego z filtrem FIR (zazwyczaj jednobiegunowym). Takie modele nazywane są albo modelem Wienera (filtr FIR jest poprzedzony modelem statycznym) albo Hammersteina (model statyczny jest poprzedzony filtrem FIR) [8]. Najbardziej ugruntowanym modelem diody LED w literaturze jest model ABC wykorzystujący r.r. do określenia zależności pomiędzy zmianą liczby nośników w czasie a wyjściową mocą optyczną [3, 10]. Niestety, zastosowanie r.r. do opisu działania diody LED wiąże się z koniecznością stosowania dużego oversamplingu w celu osiągnięcia odpowiedniej zbieżności rozwiązania. Dlatego też, w pracy [2] wykorzystano szereg Volterry do znalezienia rozwiązania r.r. i zaproponowano uproszczoną strukturę blokową diody LED (zwaną modelem behawioralnym), w której wyjściowa moc optyczna jest złożeniem bloku z filtrem FIR (dolnopasmowym), bloku nieliniowego i filtru górnoprzepustowego. Blokową strukturę modeli przedstawiono na rys. 1.



Rysunek 1: Modele blokowe diody LED: (a) liniowy, (b) Wienera, (c) Hammersteina, (d) Behawioralny.

2.1 Porównanie modelów diody LED

W celu porównania własności transmisyjnych 4 modeli diody LED: statycznego, Wienera, Hammersteina i behawioralnego, jako referencję wybrano fizycznie uzasadniony model diody bazujący na r.r. [3]. Porównanie wykonano w sposób następujący. W pierwszym kroku wygenerowano sygnał PAM-4 z filtrem RCOS o współczynniku poszerzenia pasma równym 1. Następnie, tak przygotowany sygnał wprowadzono do układu r.r., których parametry wskazano w Tab. II z [2] (przy czym bias diody ustalono na 38 mA). Wyjście r.r. opisuje przebieg mocy optycznej w czasie. Następnie, wykorzystując algorytm Levenberga-Marquardta do rozwiązywania nieliniowych problemów optymalizacyjnych, wyestymowano parametry poszczególnych modelów diody LED (dla statycznego są to współczynniki wielomianu, dla Wienera i Hammersteina są to zarówno współczynniki wielomianu jak i filtr $h_1(t)$, zaś dla modelu behawioralnego są to filtry $h_1(t)$ i $h_2(t)$) poprzez minimalizację błędu średniokwadratowego. Na rys. 2 przedstawiono porównanie czterech modeli pod kątem predykcji modelu a faktycznym sygnałem obliczonym z wykorzystaniem r.r. dla różnych prądów modulacji i szybkości symbolowych sygnału PAM-4. Model behawioralny charakteryzuje się najmniejszym błędem średniokwadratowym, dlatego też wykorzystany zostanie on do dalszej analizy kompensacji zniekształceń nieliniowych w diodach LED, gdyż najlepiej wychwytuje charakter zniekształceń. Warto podkreślić, iż we



Rysunek 2: Błąd średniokwadratowy w skali decybelowej dla modelów: (a) statycznego, (b) Wienera, (c) Hammersteina, (d) Behawioralnego dla modulacji PAM-4 o szybkościach symbolowych od 50 do 500 MBod/s i prądów modulacji od 5 do 30 mAp-p.

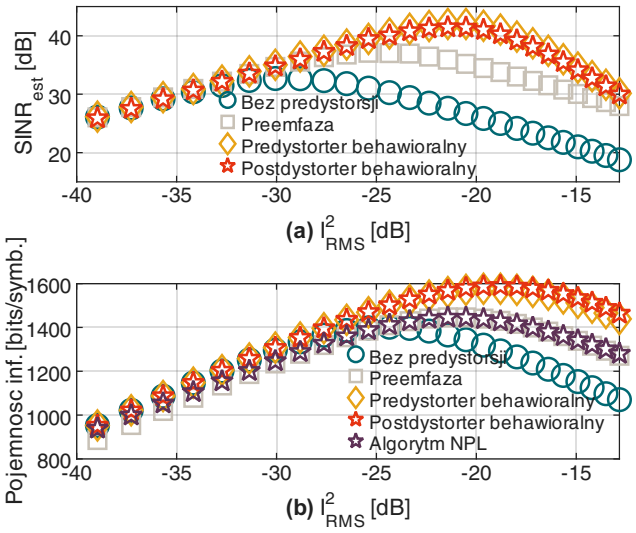
wszystkich modelach zakłada się, że najsilniejszym zniekształceniem jest to drugiego rzędu, zaś wyższe rzędy zostały pominięte, co jest zgodne z założeniami wprowadzanymi dotychczas w literaturze [5].

3. KOMPENSACJA NIELINIOWOŚCI

Metody kompensacji zniekształceń nieliniowych można podzielić na dwie grupy. Pierwsza obejmuje kompensację po stronie nadajnika (tzw. predystortery), zaś druga po stronie odbiornika (tzw. postdystortery). W przypadku predystorsji możliwe jest wykorzystanie modelu behawioralnego w taki sposób, aby odbierany sygnał nie zawierał części nieliniowej [9]. Inną metodą jest zastosowanie liniowej preemfazy (odwracającą charakterystykę częstotliwościową diody) po stronie nadajnika, która spowoduje zmniejszenie mocy zniekształceń nieliniowych [1]. Możliwa jest również kompensacja nieliniowości po stronie nadajnika poprzez wykorzystanie układu postdystortera bazującego na behawioralnej strukturze diody LED [9].

Porównanie dwóch metod dla modulacji PAM-4 (szybkość modulacji równa 350 MBod/s) z wykorzystaniem modelu r.r. przedstawiono na rys. 3(a). Wraz ze zwiększaniem amplitudy modulacji, stosunek mocy sygnału użytecznego do mocy zniekształcenia i szumu (SINR) rośnie liniowo dla małych prądów modulacji. Gdy dioda LED wprowadzona jest w reżim pracy nieliniowej, następuje spadek SINR proporocjonalny do zwiększanej mocy sygnału[1]. Maksimum SINR dla układu bez predystorsji występuje dla około -30 dB I^2_{RMS} i wynosi 32 dB. Preemfaza pozwala na zwiększenie SINR do około 38 dB kosztem zwiększenia mocy sygnału o około 6 dB. Z kolei pre- i postdystortery behawioralne charakteryzują się taką samą wydajnością i osiągają maksymalny SINR rzędu 42 dB przy zwiększeniu mocy o około 5 dB względem maksimum dla preemfazy. Można uznać, że zarówno preemfaza jak i pre/post dystortery behawioralne skutecznie niwelują zniekształcenia nieliniowe diody, gdyż osiągają większe wartości SINR dla większych wartości mocy sygnału modulującego diodę LED. Na uwagę zasługuje znacznie lepsza wydajność preemfazy liniowej względem liniowej korekcji po stronie nadajnika. Dzieje się tak,

[1] W układach nln. drugiego rzędu, moc zniekształceń nln. rośnie z kwadratem mocy wejściowej.



Rysunek 3: (a) SINR dla sygnału PAM-4 w funkcji mocy sygnału modulującego dla różnych metod kompensacji nieliniowości. (b) Teoretyczna pojemność informacyjna dla OFDM w funkcji mocy sygnału modulującego dla różnych metod kompensacji nieliniowości. *Bez predystorsji=FFE.*

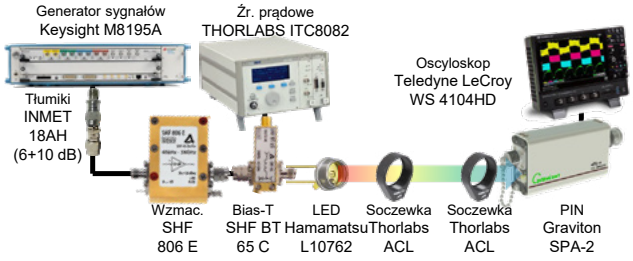
ponieważ preemfaza powoduje obniżenie mocy sygnału na wejściu części nieliniowej diody LED (patrz rys. 1), co z kolei powoduje obniżenie mocy zniekształceń nieliniowych [1].

Na wydajność modulacji OFDM również wpływać będą zniekształcenia nieliniowe diody LED. W przypadku modulacji wielotonowej wyróżnić można kilka metod kompensacji nieliniowości. Po stronie nadajnika możliwe jest stosowanie preemfazy w dziedzinie częstotliwości oraz predystortera behawioralnego [11]. Dodatkowo, możliwe jest zastosowanie adaptacyjnego algorytmu przydziału bitów i mocy NBPL, maksymalizującym pojemność systemu OFDM [7]. Możliwe jest również zastosowanie algorytmu NPL przydzielającego moce podnośnym poprzez maksymalizację pojemności systemu w połączeniu z dobrze znanym algorytmem Levin-Campello [6]. Algorytmy NBPL i NPL opierają się na założeniu, że charakterystyka zniekształcenia nieliniowego nie zmienia się co do kształtu, a tylko jej moc skaluje się ze średnią mocą sygnału nadawczego. 7 Na rys. 3(b) porównano teoretyczną pojemność transmisji OFDM z wykorzystaniem różnych metod kompensacji nieliniowości dla modulacji OFDM z 128 podnośnymi i pasmem 175 MHz.

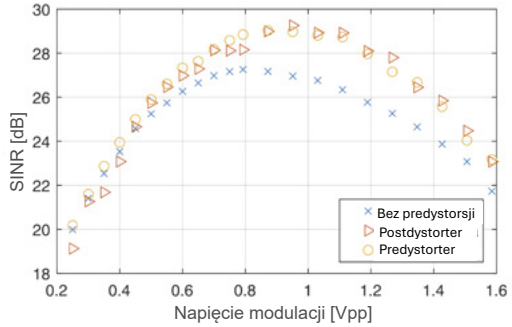
Warto zwrócić uwagę, że dla modulacji PAM-4 dla niższych mocy, preemfaza charakteryzuje się podobnymi wartościami SINR jak w przypadku innych metod kompensacji nieliniowości, zaś dla wyższych mocy jest lepsza niż korektor FFE, ale nadal gorsza niż korektor nieliniowy. Z kolei w przypadku modulacji wielotonowej, dla niższych mocy modulujących preemfaza jest najgorsza, co zostało szczegółowo wyjaśnione w [6].

4. WYNIKI EKSPERYMENTALNE

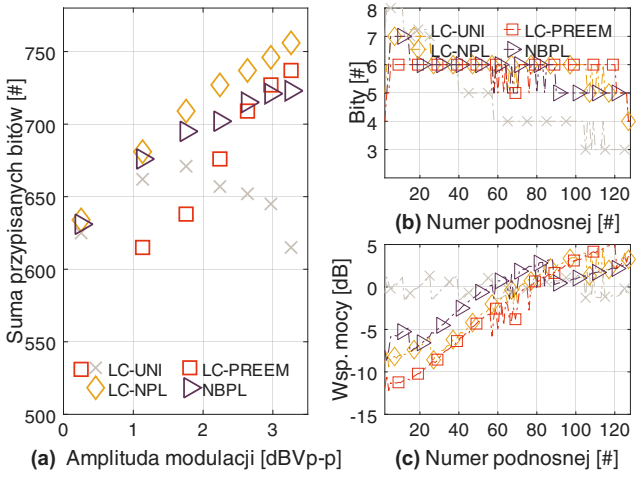
Przeprowadzono również badania mające na celu porównanie wskazanych metod w warunkach eksperymentalnych. Schemat stanowiska testowego przedstawiono na rys. 4. Procedura testowa uwzględnia zarówno modulację PAM-4 jak i OFDM. Proces generowania sygnałów PAM-4 i OFDM opisano w poprzednim rozdziale oraz w [2]. Zależność SINR w funkcji napięcia modulacji dla PAM-4 przedstawiono na rys. 5. Z kolei wyestymowaną pojemność informacyjną dla modulacji OFDM wraz z przypisaniem bitów i mocy pokazano na rys. 6. W eksperymencie z modulacją OFDM wykorzystano 3 różne metody kompensacji nieliniowości tj. algorytm NPL (LC-NPL), preemfazę (LC-PREEM) oraz algorytm NBPL. Dla porównania przedstawiono wyniki dla układu bez predystorsji (LC-UNI).



Rysunek 4: Schemat blokowy układu pomiarowego. Bias diody równy 20 mA.



Rysunek 5: SINR dla sygnału PAM-4 w funkcji mocy sygnału modulującego dla różnych metod kompensacji nieliniowości.



Rysunek 6: (a) Wyestymowana pojemność informacyjna dla OFDM w funkcji mocy sygnału modulującego dla różnych metod kompensacji nieliniowości. (b) Bity i (c) moce przypisane przez wybrane algorytmy w funkcji numeru podnośnej dla 3.26 dBVpp.

W przypadku wyników eksperymentalnych dla modulacji PAM-4 można zauważyć, że zarówno pre- jak i postdystorter pozwalają na zwiększenie wartości SINR o około 2.5 dB kosztem zwiększenia napięcia modulacji o 25%. W przypadku modulacji OFDM, algorytm NPL pozwala uzyskać największą pojemność informacyjną lepszą o około 15% względem przypadku bez metody kompensacji nieliniowości.

5. PODSUMOWANIE

Ciągły i dynamiczny rozwój technik komunikacji w świetle widzialnym stawia konieczność opracowywania coraz to bardziej zaawansowanych metod pozwalających na minimalizację ograniczeń transmisji wynikających z nieliniowego charakteru pracy diod LED. W pracy przedstawiono metody kompensacji nieliniowości zarówno dla modulacji jedno jak i wielotonowych pokazując jednocześnie szerokie zastosowanie behawioralnego modelu diody LED. Krótkie zestawienie stosowanych metod przedstawiono w Tab. 1. Warto podkreślić, że w systemach nieliniowych predystorter po stronie nadajnika charakteryzuje się podobną skutecznością usuwania zniekształceń nieliniowych jak postdystorter (co jest swoistą analogią do pre i post korekcji w systemach liniowych).

Dalsze prace skupiać się mogą na analizie zastosowania rozwiązań bazujących na sztucznych sieciach neuronowych, czy wyprowadzeniu modelu behawioralnego dla wyższych rzędów nieliniowości. Ponadto, znane w literaturze są również modele r.r. opisujących działanie laserów półprzewodnikowych, a więc i możliwe by było również opracowanie, tak jak dla diod LED, modelu behawioralnego dla laserów. Umożliwiłoby to stosowanie podobnych metod znacznie szerzej w telekomunikacji.

Tabela 1: Metody usuwania zniekształceń nieliniowych

Metoda	Lokalizacja	Skuteczność
Modulacje jednotonowe		
Preemfaza	Nadajnik	Częściowa
Predystorter	Nadajnik	Pełna*
Korektor FFE	Odbiornik	Brak
Postdystorter	Odbiorcza	Pełna*
Modulacje wielotonowe		
Preemfaza	Nadajnik	Częściowa
Predystorter	Nadajnik	Pełna*
Algorytm NPL	Nadajnik	Prawie pełna
Algorytm NBPL	Nadajnik	Częściowa+
Korektor FFE	Odbiornik	Brak
Postdystorter	Odbiornik	Pełna*

*do 2. rzędu; +z alokacją bitów

Niniejsza praca powstała w ramach projektu USVEGAN finansowanego ze środków Narodowego Centrum Badań i Rozwoju (nr umowy WPC2/20/USVEGAN/2019).

BIBLIOGRAFIA

[1] J. Bojarczuk, M. Marzecki i G. Stepniak. "Can linear pre-emphasis mitigate the nonlinear distortion of light-emitting diodes?" W: *Optics Letters* 49.9 (maja 2024), s. 2257–2260.

[2] J. Bojarczuk, M. Marzecki, D. Sun i in. "A Behavioral Model of the Light Emitting Diode Nonlinearity". W: *IEEE Transactions on Communications* (2025), s. 1–1. ISSN: 1558-0857. DOI: 10.1109/tcomm.2025.3569749.

[3] X. Deng, S. Mardanikorani, Y. Wu i in. "Mitigating LED nonlinearity to enhance visible light communications". W: *IEEE Transactions on Communications* 66.11 (2018), s. 5593–5607.

[4] M. W. Eltokhey, M. S. Bashir, O. I. Younus i in. "Optical Wireless Communications: Illuminating the Path for Cooper's Law". W: (stycznia 2024). DOI: 10.36227/techrxiv.170620142.21301431/v1.

[5] T. Kamalakis, J. W. Walewski, G. Ntogari i in. "Empirical Volterra-series modeling of commercial light-emitting diodes". W: *Journal of Lightwave Technology* 29.14 (2011), s. 2146–2155.

[6] J. Kasjanowicz, J. Bojarczuk, P. Araszkiewicz i in. "Balancing noise and nonlinear distortion by power loading in FBMC OWC". W: *Applied Optics* 64.16 (maja 2025), s. 4589. ISSN: 2155-3165. DOI: 10.1364/ao.562794.

[7] J. Kasjanowicz, J. Bojarczuk i G. Stepniak. "Bit and Power Loading Algorithms for Nonlinear Optical Wireless Communication Channels". W: *IEEE Communications Letters* 27.12 (2023), s. 3270–3274.

[8] X. Ling, X. Zhang, J. Sun i in. "When Hammerstein meets Wiener: Nonlinearity Modeling for End-to-End Visible Light Communication Links". W: *IEEE Transactions on Communications* 71.1 (2023), s. 310–323. DOI: 10.1109/TCOMM.2022.3222482.

[9] G. Stepniak, M. Marzecki i J. Bojarczuk. "Volterra predistorter for the dynamic nonlinearity of LED". W: *Optics Letters* 47.5 (2022), s. 1161–1164.

[10] G. Stepniak i J. Siuzdak. "Influence of Lighting LED Design Parameters on the Dynamic Nonlinear Response". W: *Journal of Lightwave Technology* 40.4 (2022), s. 954–960. DOI: 10.1109/JLT.2021.3129586.

[11] D. Sun, J. Bojarczuk, Z. Jin i in. "6 Gbps Micro-LED Transmission Using OFDM With Predistortion and Single-Tap Nonlinearity Compensation". W: *IEEE Photonics Technology Letters* 35.14 (2023), s. 781–784.

[12] Z. Wang, Z. Jin, R. Lin i in. "Vertical stack integration of blue and yellow InGaN micro-LED arrays for display and wavelength division multiplexing visible light communication applications". W: *Optics Express* 30 (24 listopada 2022), s. 44260. ISSN: 1094-4087. DOI: 10.1364/OE.475548.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

KONCEPCJA SIECI RADIOWO-FOTONICZNEJ OPARTEJ NA OTWARTEJ ARCHITEKTURZE URZĄDZEŃ WĘZŁOWYCH

THE CONCEPT OF A RADIO-PHOTONIC NETWORK BASED ON AN OPEN ARCHITECTURE OF NODE DEVICES

Zbigniew Zakrzewski

Politechnika Bydgoska im. Jana i Jędrzeja Śniadeckich, Bydgoszcz, zbizak@pbs.edu.pl

DOI: 10.15199/59.2025.4.53

Streszczenie: **W artykule przedstawiono koncepcję otwartej sieci radiowo-fotonicznej, która winna współpracować z systemami xPON na poziomie optycznych zasobów ODN. Zaproponowano, że część radiowa bazuje na rozwiązaniu O-RAN, zaś część fotoniczna na węzłowym rozwiązaniu O-ROADM. Podano przykładowe konfiguracje architektury otwartych węzłów oraz topologii otwartych sieci R-F współpracujących z systemami xPON. Zaprezentowano również główne aspekty dotyczące przenoszenia w ścieżce optycznej sygnałów radiowych z rodziny D-RoF oraz A-RoF.**

Abstract: **The paper presents the concept of an open radio-photonic network that should cooperate with xPON systems at the level of ODN resources. It is proposed that the radio part is based on the O-RAN solution, while the photonic part is based on the O-ROADM node solution. Example configurations of the open node architecture and the topology of open R-F networks cooperating with xPON systems are given. The main aspects related to the transmission of radio signals from the D-RoF and A-RoF families in the optical path are also presented.**

Słowa kluczowe: **A-RoF, D-RoF, O-RAN, O-ROADM, otwarta sieć radiowo-fotoniczna, pasywna sieć optyczna.**

Keywords: **A-RoF, D-RoF, open radio-photonic network, O-RAN, O-ROADM, passive optical network.**

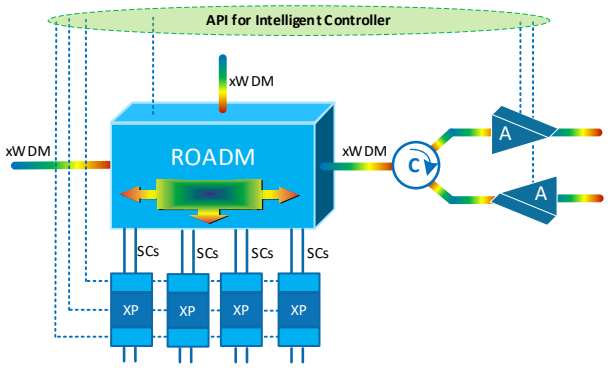
1. WSTĘP

Współczesne sieci światłowodowe i bezprzewodowe integrują się w coraz większym stopniu. Zacieśnienie granicy między domenami optyczną i radiową powoduje, że sygnały radiowe coraz bardziej przenikają sieci światłowodowe. Cyfryzacja sygnałów radiowych powoduje znaczący wzrost zapotrzebowania na ich transport w dosyłowych sieciach szerokopasmowych. W związku z tym pojawia się wiele nowych rozwiązań wskazujących na znacznie efektywniejsze wykorzystanie już istniejących sieci światłowodowych. Do takich zasobów można zaliczyć sieci ODN (Optical Distribution Network) budowane pod instalacje xPON (Passive Optical Network). Rosnąca liczba lokalnych operatorów sieciowych powoduje, że coraz częściej pojawia się potrzeba tworzenia rozwiązań pozwalających na efektywne integrowanie oferowanych przez nich rozwiązań radiowych i optycznych. Realizacja tego zadania jest możliwa, ale przy założeniu, że poszczególne części sieci będą mogły być w łatwy i dostępny sposób integrowane. W ostatnich latach budowanie sieci bazujących na wirtualizacji oraz otwartej architekturze i programowalnych urządzeniach węzłowych staje się standardem [3,5]. Dzięki takiemu podejściu mali operatorzy są w stanie kreować własne usługi bazując na sprzęcie należącym niejednokrotnie do różnych operatorów [4]. Tworzenie i upowszechnianie interfejsów API (Application Programming Interface), które stanowią styk między domenami lub między oprogramowaniem i sprzętem może przynieść wiele korzyści, ale często spotyka się z oporem dostawców.

W rozdziale 2 opracowania przedstawiono cechy dotyczące otwartości architektury urządzeń wchodzących w skład sieci radiowych oraz fotonicznych. Rozdział 3 został poświęcony wskazaniu technik łączenia otwartej sieci radiowo-fotonicznej do sieci ODN przy wykorzystaniu optycznego pasma E. W rozdziale 4 nakreślono kluczowe zagadnienia dotyczące przenoszenia sygnałów radiowych formatów A-RoF oraz D-RoF w łączach z koegzystencją sygnałów optycznych w formacie xWDM. Piąty rozdział zawiera wnioski oraz wskazanie badań, które będą prowadzone w kierunku określenia optymalizacji rozwiązań dotyczących automatyzacji pracy sieci radiowo-fotonicznej pod kątem doboru węzłów oraz ścieżek optycznych przenoszących jednocześnie skrajnie różne sygnały optyczne.

2. OTWARTA ARCHITETKURA URZĄDZEŃ SIECIOWYCH

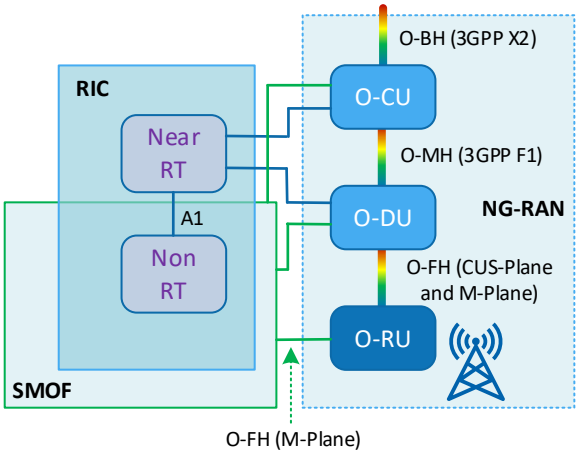
W tak zwanych sieciach radiowo-fotonicznych zauważyć można prawie całkowitą eliminację systemów transmisyjnych bazujących na mediach miedzianych. W skutek tego tylko układy elektroniczne wykorzystują komunikację poprzez miedziane ścieżki nieprzekraczające kilku centymetrów. Do transmisji na większe odległości są wykorzystywane media radiowe lub światłowodowe.



Rys. 1. Otwarta architektura węzła optycznego z portami klienckimi – OpenROADM [6]

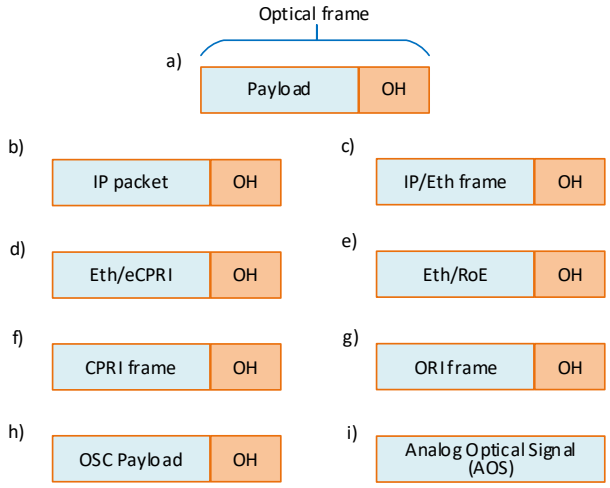
Taka sytuacja powoduje, że szerokopasmowe sygnały radiowe rzadziej doświadczają tzw. wąskich gardeł.

Na rysunku 1 przedstawiono przykład otwartej architektury węzła optycznego opartego na ROADM (Reconfigurable Optical Add-Drop Multiplexer). Zostało to osiągnięte dzięki dekompozycji urządzenia. Poszczególne podzespoły węzła mogą być niezależnie sterowane poprzez interfejs API, co gwarantuje możliwość wykreowania usługi świadczonej przez węzeł [4], który może brać udział w transmisji różnorodnych sygnałów pochodzących z różnych ścieżek optycznych. Dzięki takiemu podejściu pojawiający się nowy format optycznego sygnału nie wymusza potrzeby zbyt częstej aktualizacji sprzętu. Warunkiem w tym miejscu jest stosowanie inteligentnych xponderów (XP), które mogą pełnić funkcję prostych transponderów, muxponderów lub switchponderów, zależnie od sygnału klienckiego docierającego do wkładki klienckiej. Kolejna zaletą dekompozycji jest możliwość wyniesienia xpondera do lokalizacji sieci dostępowej, co umożliwia znacznie sprawniejsze wdrożenie rozwiązania IPoWDM lub RoWDM (Radio-over-Wavelength Division Multiplexing). Bardzo istotnym składnikiem jest wzmacniacz optyczny A (Amplifier), którego zadaniem jest kompensacja tłumień wnoszonych przez ścieżkę optyczną w pasywnej strukturze płaszczyzny użytkownika urządzeń ROAMD oraz światłowodowych połączeń między węzłami. Zarządzany szerokopasmowy wzmacniacz optyczny daje możliwość równomiernej kompensacji zwłaszcza tłumienia, którego doznają optyczne sygnały przelotowe nienależące do sieci podłączanych poprzez XP w ramach otwartej części sieci optycznej.



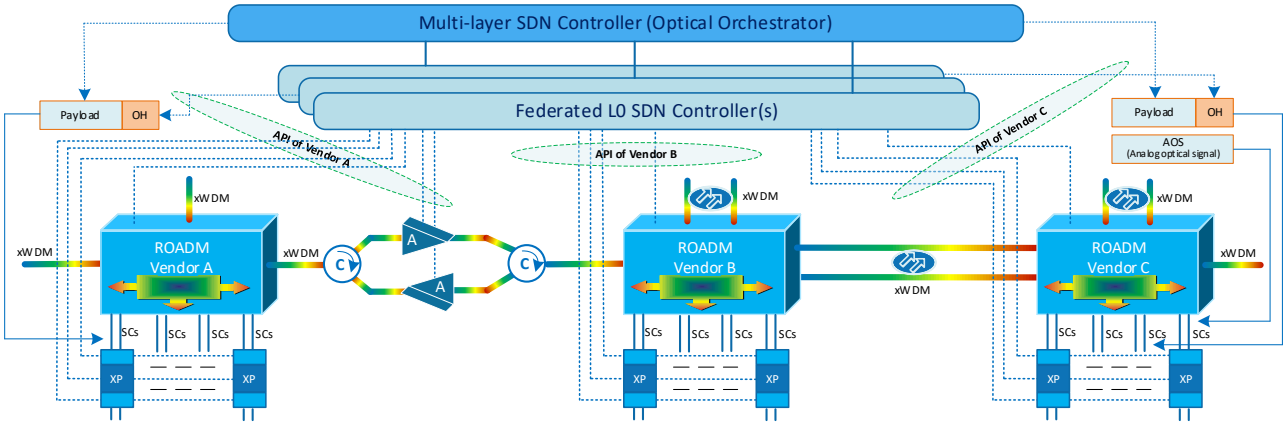
Rys. 2. Otwarta architektura systemu mobilnego w zakresie domeny radiowej – OpenRAN [3]

Podobnie wygląda sytuacja w sieciach dostępu radiowego. Tutaj zasoby fizyczne na poziomie medium są znacznie mniejsze, zatem potrzeby związane z efektywnym wykorzystaniem sprzętu w tym zakresie są jeszcze wyższe. Za dobry przykład może posłużyć rozwiązanie OpenRAN (rysunek 2) [3], które zostało wprowadzone w systemach mobilnych 5G i na pewno będzie rozbudowywane i upowszechniane w kolejnych generacjach, zwłaszcza dostępowych systemów mobilnych. Gdy podzespoły sieciowe O-CU (Open Central Unit), O-DU (Open Distributed Unit) oraz O-RU (Open Radio Unit) są łączone przy pomocy sieci optycznej, wówczas znikają przeszkody polegające na buforowaniu pakietów lub sygnałów. W przypadku sieci RAN takie newralgiczne odcinki leżą na poziomie otwartej sieci fronthaul (O-FH) oraz otwartej sieci midhaul (O-MH). Na odcinku sieci O-FH mogą być świadczone tzw. usługi czasu rzeczywistego (płaszczyzna CUS – control, user and synchronization), zaś w przypadku sieci O-MH istnieje możliwość zapewnienia usług bliskich czasu rzeczywistego. Według O-RAN Alliance [3] jest to możliwe przy zastosowaniu odpowiedniego oprogramowania w formie otwartych inteligentnych sterowników RIC (RAN Intelligent Controller). Jeśli wprowadzimy efektywną integrację domen fotonicznej i radiowej, wówczas zaistnieje możliwość zintegrowanego doboru zasobów na warstwach L0 oraz L1, gdzie L0 oznacza warstwę optyczną, zaś L1 definiuje warstwę radiową.



Rys. 3. Przykłady propozycji optycznych ramek służących do synchronicznego transportowania danych pochodzących ze źródeł klienckich: a) ogólny format ramki (OH – nagłówek), b) transport IP poprzez ścieżkę optyczną, c) transport IP w ramce Ethernet, d) transport informacji z sieci RAN w formacie eCPRI poprzez ramki ethernet, e) transport informacji z sieci RAN w formacie RoE (Radio-over-Ethernet), f) transport ramek w formacie CPRI (3GPP Option 8), g) transport ramek radiowych w formacie ORI (Open Radio equipment Interface), h) ramkowanie w nadzorczym kanale usługowym (OSC – Optical Service Channel), i) sygnał analogowy pozbawiony ramkowania i nagłówka (format A-RoF)

Na rysunku 3 przedstawiono przykładowe formaty ramek optycznych służących do synchronicznego transportowania danych pochodzących z portów leżących po stronie XP, które dopasowują zasoby ścieżki optycznej do strumienia danych lub sygnału pochodzącego z wkładki klienckiej. Ramki przedstawione na rysunkach 3b) i 3c) służą do transportowana klasycznych ramek ethernetowych oraz pakietów IP, ramki z rysunków od 3d) do 3g) przedstawiają transport scyfryzowanych sygnałów radiowych z pasma podstawowego D-RoF (Digitized Radio-over-Fiber), zaś rysunek 3h) przedstawia ramkę transportującą dane z kanału sterującego określonym kanałem optycznym lub grupą kanałów. Na rysunku 3i) zauważamy różnicę, gdyż jest to sygnał analogowy, który jest pozbawiony nagłówka (OH).

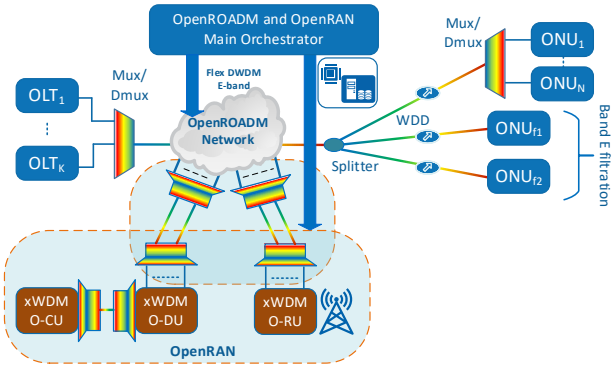


Rys. 4. Przykład przelotowej sieci optycznej bazującej na otwartej architekturze multiplekserów ROADM oraz pozostałych układów i urządzeń aktywnych należących do różnych lokalnych operatorów sieciowych [5,6]

Tego rodzaju kanał optyczny będzie potrzebował osobnego wydzielonego kanału sterującego (rysunek 4). Kanał sterujący może pełnić funkcje kanału współdzielonego dla większej grupy kanałów przenoszących sygnały analogowe z rodziny A-RoF (Analog Rafio-over-Fiber). Tego rodzaju format może być wykorzystany do transportu wyżej wymienionych ramek lub pakietów w przypadku, gdy nie jest wymagana synchronizacja na poziomie łącza optycznego. Na rysunku 4 przedstawiono przykładowy wycinek sieci ze sterowaniem optycznymi komponentami składowymi węzłów pochodzących od różnych operatorów. Poprawna praca wydzielonej optycznej sieci transportowej w całości jest podporządkowana nadrzędnemu sterownikowi SDN (Software Defined Network). W przypadku otwartej architektury węzłów sieci optycznej w zasadzie każdy podzespół może być sterowany z poziomu osobnego interfejsu API, co końcowo daje szerokie możliwości kreowania ścieżki optycznej. Ograniczenia fizyczne poszczególnych składników muszą być dostępne na poziomie sfederalizowanych sterowników warstwy optycznej L0.

3. PODŁĄCZANIE OTWARTEJ SIECI RADIOWO-FOTONICZNEJ DO ODN

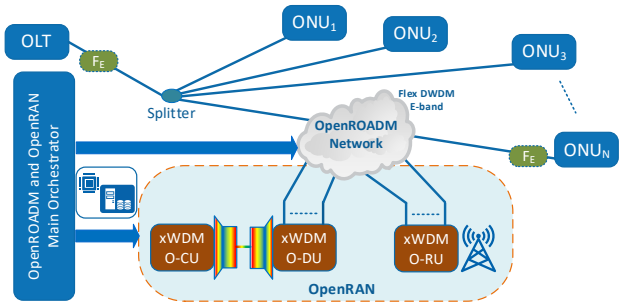
Dystrybucyjne sieci ODN przenikają obszary zurbanizowane coraz skuteczniej. Dzięki temu w wielu przypadkach nie ma potrzeby instalowania nowych sieci światłowodowych, gdyż systemy z rodziny xPON nie zajmują wszystkich zasobów światłowodów SMF (Single Mode Fiber), co zostało przedstawione w [9].



Rys. 5. Koegzystencja otwartej sieci radiowo-fotonicznej i sieci ODN na poziomie łącza dosyłowego

Analiza wykazała, że bezpieczne jest stosowanie optycznego pasma E, jednak obecne ograniczenia wynikają z faktu, że na rynku jest mało urządzeń węzłowych i przelotowych pracujących w tym paśmie.

Elastyczny system DWDM jest przygotowany do pracy w optycznych pasmach S, C oraz L, zatem uruchomienie dopełnienia w paśmie E wymaga utworzenia standardu, aby mogły być produkowane rekonfigurowalne urządzenia węzłowe ROADM. W tej sytuacji wskazane jest, aby już były to rozwiązania oparte na architekturze otwartej [6]. Jest to możliwe do zrealizowania, jeśli sygnały są przenoszone przez chmurę sieciową (rysunek 3) w sposób prawie przezroczysty. Prawie, gdyż każda fizyczna sieć wnosi na optycznej ścieżce tłumienie, co skutkuje potrzebą stosowania wzmacniaczy optycznych obejmujących optyczne pasmo sygnałów xPON. Kontrola dyspersji chromatycznej ścieżki nie powinna być problemem w przypadku, gdy ta ścieżka nie wydłuża się.

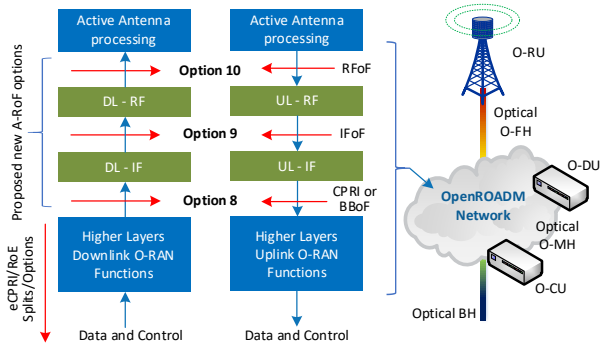


Rys. 6. Koegzystencja radiowej sieci O-RAN i optycznej sieci bazującej na O-ROADM z siecią ODN na poziomie łącza dystrybucyjnego

Zintegrowane otwarte sieci radiowo-fotoniczne mogą być podłączane w dosyłowej części ODN (rysunek 5) lub dystrybucyjnej (rysunek 6). W obydwu przypadkach ograniczenia wynikające z zasięgu systemów xPON pozwalają na swobodną realizację połączeń O-RAN, zwłaszcza w zakresie sieci O-FH (interfejs między O-DU i O-RU), gdzie odległość nie może być większa niż 20km. Cel wprowadzania sieci opartej na otwartych urządzeniach ROADM wydaje się być najbardziej istotny w masowych zastosowaniach O-RAN, gdzie potrzeba dostarczania przetworzonych sygnałów radiowych do aktywnych anten jest realizowana przez kilku operatorów.

4. TRANSPORT SYGNAŁÓW RADIOWYCH D-ROF I A-ROF W OTWARTEJ SIECI R-F

Sygnały radiowe, które są przygotowane do transportowania z miejsca przetworzenia do miejsca promieniowania cechują się określonymi wymaganiami. Dekompozycja funkcji radiowych w obszarze sieci dosyłowej spowodowała pojawienie się koncepcji Open RAN. Tym samym pojawiły się poddomeny nacechowane określonymi wymaganiami czasowymi. Wymagania te z powodzeniem mogą być spełnione przez sieci optyczne, zwłaszcza całkowicie optyczne.



Rys. 7. Realizacja podziału rozszerzonych funkcji sieciowych O-RAN ze względu na transportowe możliwości optycznej sieci opartej na O-ROADM

Kolejnym aspektem jest dostępność zasobów fizycznych i potrzeba realizacji dedykowanej usługi [4]. Na poziomie sieci radiowej otwartość w zarządzaniu zasobami radiowymi i sprzętowymi jest oczywista, gdyż jest ich mało. W przypadku sieci optycznej obciążonej obcym ruchem już nie jest to tak oczywiste, stąd decyzja o wyborze większości wolnego optycznego pasma E [9]. Zaproponowane formaty ramek optycznych (rysunek 3) umożliwiają przetransportowanie istniejących i przyszłych sygnałów. W pracy [8] zaproponowano nowe opcje podziału funkcji o numerach 9 i 10, jak też rozszerzenie funkcji opcji 8 (CPRI) na transport analogowych sygnałów BBoF – rysunek 7. Otwarta sieć oparta na multiplekserach OADM wpisuje się bardzo dobrze w możliwość reakcji na potrzeby transportowania sygnałów z rodziny A-RoF ścieżką optyczną, która współdzieli zasoby fizyczne z sieciami xPON. Przy tym należy zaznaczyć, że dekompozycja sieci Wi-Fi przy pomocy takiego zasobu optycznego także może być zrealizowana [10].

Pozostaje określenie otwartości w komunikacji sterowników ze sprzętem poprzez odpowiednie API. Na poziomie części radiowej już jest to możliwe w przypadku sygnałów D-RoF, natomiast przy wprowadzeniu transportu A-RoF węzły O-DU oraz O-RU wymagają sprzętowego rozszerzenia [8]. Część optyczna O-ROADM jest w większości do tego przygotowana, tzn. układ ROADM w podstawach działania przenosi sygnały na warstwie L0 w sposób przezroczysty. Nacisk należy położyć na część XP oraz wkładki przetworników, które będą rozpoznawały sygnały z rodziny A-RoF. W drodze komunikacji sterowników z podzespołami otwartego systemu O-ROADM ma zastosowanie język YANG [1,6], którego funkcjonalność jest przygotowana do modelowania danych konfiguracyjnych, danych stanu, zdalnych wywołań procedur i powiadomień dla protokołów zarządzania siecią. Zestaw instrukcji tego języka wskazuje na możliwość sprawnego dodania rozszerzeń w kierunku interpretacji sygnałów A-RoF, które na warstwie L0 przyjmują postać sygnałów podobnych do stosowanych w koherentnych interfejsach sieci EON (Elastic Optical Network).

5. WNIOSKI I DALSZE BADANIA

Wyniki przeprowadzonej szybkiej analizy zaproponowanej koncepcji pokazują, że fizycznie istnieje możliwość wprowadzenia rozwiązania otwartego na sterowanie siecią radiowo-fotoniczną, w której węzły należą do różnych lokalnych operatorów. Takie otwarte łączne podejście do sieci radiowo-fotonicznych nie było prezentowane w literaturze, chociażby ze względu na wysoki koszt wdrażania takich technik w sieciach o zasięgu regionalnym lub lokalnym. Istnieje jednak duże prawdopodobieństwo, że pojawi się taka możliwość, gdy systemy mobilne następnych generacji oraz sieci Wi-Fi zintegrują się na poziomie zasobów radiowych, zaś sieci optyczne z węzłami całkowicie optycznymi będą często stosowane, na podobieństwo współczesnych ethernetowych sieci LAN.

Do poszukiwania optymalnej optycznej ścieżki może być zastosowany protokół OpenFlow (OF) wspomagany poprzez GMPLS/RSVP-TE [7] oraz OSPF-TE [2], ale interfejs API musi być opracowany w nowej wersji przy wykorzystaniu języka YANG [1], aby współbieżnie mogły być w otwartej sieci transportowane sygnały D-RoF oraz A-RoF [8]. Będzie to kolejnym etapem badań w połączeniu z ustalaniem parametrów przenoszonych sygnałów optycznych, które nie wpłyną na sygnały xPON.

LITERATURA

[1] IETF 7950. 2016. “The YANG 1.1 Data Modeling Language”.

[2] IETF 8363. 2018. “GMPLS OSPF-TE Extensions in Support of Flexi-Grid Dense Wavelength Division Multiplexing (DWDM) Networks”.

[3] O-RAN Alliance. 2025. “O-RAN Architecture Description”. WG1.TS.OAD-R004-v13.00.

[4] The OpenROADM MSA Group, 2024. “A Description of the OpenROADM Service Model”, ver. 13.1.

[5] The OpenROADM MSA Group, 2024. “Network Model White Paper”, ver. 13.1.

[6] The OpenROADM MSA Group. 2024. “Device Model White Paper”, ver. 13.1.

[7] Zbigniew Zakrzewski. 2016. „Koncepcja dystrybucyjnej mikrofalowo-fotonicznej sieci SDR-SDN pracującej pod kontrolą OpenFlow/GMPLS”. *PTiWT* 8-9/2016 : 1025-1035.

[8] Zbigniew Zakrzewski. 2020. “D-RoF and A-RoF Interfaces in an All-Optical Fronthaul of 5G Mobile Systems”. *Appl. Sci*. 10 (1212).

[9] Zbigniew Zakrzewski. 2023. „Integracja O-RAN oraz xPON na warstwie optycznych zasobów ODN”. *PTiWT*, 4/2023 : 277-280.

[10] Zbigniew Zakrzewski. 2024. „Zastosowanie interfejsów D-RoF/A-RoF w sieciach 802.11 następnych generacji”. *PTiWT* 4/2024 : 233-236.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

SPEKTROSKOPIA RAMANOWSKA Z SONDAMI ŚWIATŁOWODOWYMI – WYBRANE ZASTOSOWANIA

RAMAN SPECTROSCOPY WITH FIBER OPTIC PROBES - SELECTED APPLICATIONS

Marcin Gnyba[1]

[1] Politechnika Gdańska, Gdańsk, margnyba@pg.edu.pl

DOI: 10.15199/59.2025.4.54

Streszczenie: **Spektroskopia ramanowska to metoda badania struktury molekularno-krystalograficznej materiałów za pomocą nieelastycznego rozpraszania światła. Wykorzystywanie pomiarów optycznych w paśmie VIS-NIR pozwala na szeroką implementację światłowodów. W komunikacie przedstawiono wybrane możliwości zastosowań i przykłady badań w kierunku realizacji lokalnych światłowodowych sieci transmisji danych, pozwalających na diagnostykę procesów technologicznych lub detekcję śladowych ilości substancji.**

Abstract: **Raman spectroscopy is a method of investigation of the molecular and crystallographic structure of materials with the help of inelastic light scattering. The use of optical measurements in the VIS-NIR band allows for extensive implementation of optical fibers. The paper presents the examples of applications and of research on implementation of local fiber optic data transmission networks, allowing to diagnose technological processes or detection of trace amounts of chemicals.**

Słowa kluczowe: **diagnostyka *in-situ*, analiza struktury molekularnej, światłowody, spektroskopia ramanowska.**

Keywords: ***in-situ* diagnostics, molecular composition analysis, optical fibers, Raman spectroscopy.**

1. WSTĘP

Sieci pomiarowe mogą wykorzystywać światłowody do transmisji sygnałów, umożliwiając precyzyjne monitorowanie różnych parametrów, takich jak temperatura, ciśnienie czy odkształcenia. Sieci te mogą być wykorzystywane w różnorodnych zastosowaniach, od kontroli infrastruktury i procesów po badania naukowe. Jednym z zastosowań jest diagnozowanie procesów technologicznych w celu poprawy jakości produktu oraz zwiększenia uzysku. Efektywnym sposobem diagnostyki procesu jest monitorowanie *in-situ* jego przebiegu, w miejscu jego przeprowadzania i w czasie rzeczywistym jego trwania. Możliwość wykorzystania światłowodów do transmisji sygnałów oraz możliwość monitorowania przebiegu procesów w czasie rzeczywistym oraz daje spektroskopii ramanowskiej potencjalną przewagę nad innymi metodami badania struktury molekularnej-krystalograficznej materiałów. W komunikacie przedstawiono wybrane zagadnienia związane z dostosowywaniem światłowodowych systemów ramanowskich do diagnostyki *in-situ* oraz przykłady zastosowań. Przedstawiony opis nie wyczerpuje wszystkich możliwych zastosowań światłowodowych systemów ramanowskich, które znajdują między innymi rosnący potencjał zastosowań w światłowodowych sensorycznych sieciach rozłożonych, wykorzystujących sprzężenie z optyczną reflektometrią w dziedzinie czasu i pozwalających na detekcję zmian temperatury z odległości rzędu kilkudziesięciu kilometrów [1]. Alternatywne rozwiązania do detekcji śladowych ilości substancji, stosowane w spektroskopii ramanowskiej ze wzmocnieniem powierzchniowym SERS (ang.: Surface Enhanced Raman Spectroscopy), opierają się na wykorzystaniu intensyfikacji efektu ramanowskiego obserwowanego na granicy pomiędzy badanym ośrodkiem a cienką warstwą (lub nanocząsteczkami) metalu szlachetnego. Wiąże się to jednak z koniecznością przygotowania specjalnych próbek, które nie są reprezentatywne dla wielu rodzajów procesów, w których istotny jest typ zastosowanego podłoża technologicznego.

2. UKŁADY RAMANOWSKIE DO MONITORINGU IN-SITU

2.1. Sposób działania

Spektroskopia ramanowska polega na rejestracji i analizie widma promieniowania rozproszonego nieelastycznie w wyniku oddziaływania wiązki promieniowania monochromatycznego (najczęściej laserowego) z dipolami indukowanymi w drgającym układzie atomowym (molekule). W wyniku tego oddziaływania materia jest wzbudzana do wyższego poziomu energetycznego, wirtualnego lub rzeczywistego. W wyniku oddziaływania z promieniowaniem wzbudzającym, układ atomowy może w szczególności zostać przeniesiony do oscylacyjnego poziomu energetycznego innego niż początkowy. Występuje to dla rozpraszania Ramana. W promieniowaniu rozpraszanym pojawiają się fotony o częstotliwości mniejszej od promieniowania wzbudzającego (pasmo stokesowskie) lub większej (pasmo antystokesowskie). Różnice częstotliwości, a ściślej - liczby falowej, pomiędzy promieniowaniem rozproszonym i wzbudzającym, wyrażane za pomocą przesunięcia Ramana, odpowiadają częstotliwościom drgań własnych układu atomów i są jego cechą charakterystyczną – „odciskiem palca”. Mogą pochodzić od różnych ruchów oscylacyjnych występujących

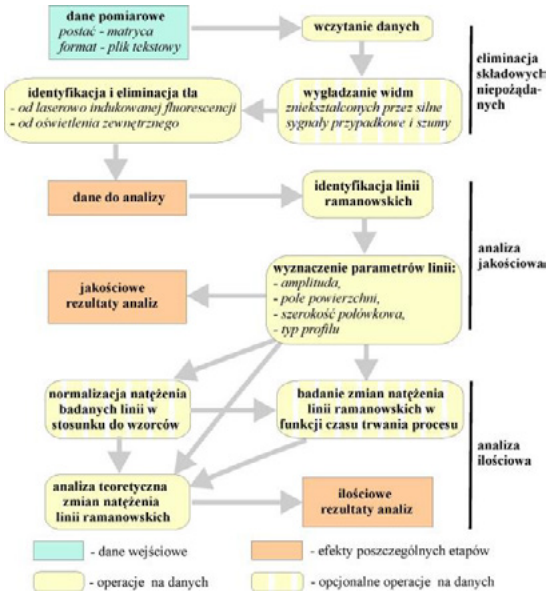
w materiale, zarówno drgań poszczególnych wiązań molekuł, jak i większych fragmentów np. sieci polimeru.

2.2. Wybór długości fali

Istotnym problemem w spektroskopii ramanowskiej jest wybór optymalnej długości fali wiązki wzbudzającej. Natężenie sygnału rozpraszania ramanowskiego jest odwrotnie proporcjonalne do czwartej potęgi długości fali promieniowania wzbudzającego. Jeżeli więc rozpatrywać jako kryterium wyboru natężenie sygnału ramanowskiego, efektywniejsze wydaje się zastosowanie laserów z zakresu promieniowania widzialnego VIS (a jeszcze lepiej z zakresu ultrafioletu UV) niż źródeł z zakresu podczerwieni IR. Z drugiej jednak strony, dla wielu grup materiałów, zwłaszcza organicznych lub zawierających zanieczyszczenia, trzeba rozważyć również wpływ na pomiar bardzo słabego sygnału ramanowskiego sygnału zakłócającego od fluorescencji indukowanej wiązką laserową. Jej wpływ zależy od rodzaju badanego materiału, toteż nie można go z góry przewidzieć, chociaż jest najsilniejszy w zakresie długości fali promieniowania wzbudzającego od 270 do 700 nm. Ponadto, przy doborze długości fali lasera należy uwzględnić zakres widmowy czułości fotodetektora oraz jego skorelowanie z mierzonym zakresem widmowym sygnału Ramana. Przykładowo, przy zastosowaniu długości fali 830 nm oraz krzemowej matrycy CCD praktyczny zakres pomiarowy w paśmie stokesowskim jest ograniczony do ok. 2000 cm^{-1}.

2.3. Przetwarzanie danych

Z uwagi na niskie natężenie sygnału Ramana oraz obecność szumów i zewnętrznych sygnałów zakłocających, dla uzyskania jakościowej lub ilościowej informacji o strukturze molekularno-krystalograficznje materiału konieczne jest przetworzenie danych pomiarowych zgodnie z procedurami przedstawionymi na Rys.1



Rys. 1 Zastosowany algorytm analizy danych pomiarowych.

3. Przykłady zastosowań systemów ramanowskich z sondami światłowodowymi

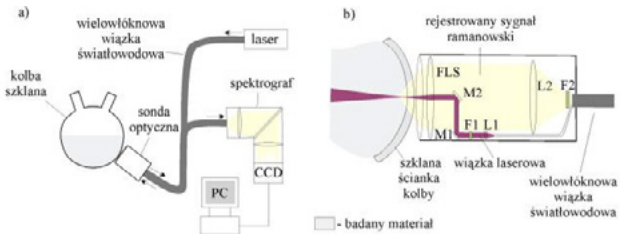
3.1. Założenia

Ponieważ, zarówno długość fali promieniowania wzbudzającego, jak i pasmo rejestrowanego sygnału ramanowskiego, znajdują się w zakresie widmowym od ultrafioletu powietrznego do bliskiej podczerwieni, ułatwione staje się wprowadzanie sygnałów do komory technologicznej i ich wyprowadzanie. W odróżnieniu np. od spektroskopii w podczerwieni, można do tego celu wykorzystać szklane elementy obudów reaktorów lub wzierniki. W przypadku braku takich elementów w konstrukcji komory technologicznej, możliwe jest wprowadzanie i wyprowadzanie sygnałów za pomocą standardowych światłowodów szklanych lub plastikowych. Monitoring in-situ procesów umożliwia kontrolę przebiegu procesu w trakcie ich trwania, co pozwala na lepsze poznanie mechanizmów reakcji, a także określenie jej poprawności i dokonanie niezbędnych modyfikacji parametrów procesu lub jego przerwanie w razie konieczności. W porównaniu z trybem *ex-situ*, w pomiarach *in-situ* znacznie trudniejsze jest pozycjonowanie badanego obiektu. Istotnym problemem jest zaprojektowanie efektywnego sprzężenia optycznego pomiędzy komorą technologiczną i systemem pomiarowym w taki sposób, aby zapewnić wzbudzenie wybranego miejsca badanego obiektu oraz transmisję sygnału ramanowskiego z wnętrza komory, a jednocześnie zapobiec zakłóceniom przebiegu procesu syntezy i pracy aparatury pomiarowej, co można osiągnąć. Istotą problemu jest takie zaprojektowanie torów optycznych, aby elementy układu pomiarowego nie musiały być umieszczane wewnątrz komory technologicznej i nie zakłócały przebiegu procesu technologicznego. Wiązka wzbudzająca może być wprowadzana do wnętrza poprzez transparentne elementy komory (ścianki, wzierniki) i w taki sam sposób może być transmitowany sygnał rozpraszania ramanowskiego. Umożliwia to zachowanie zwartości konstrukcji i szczelności komory, co jest warunkiem prawidłowego stosowania monitoringu ramanowskiego do diagnostyki procesów próżniowych i procesów wymagających wysokiej czystości. W aspekcie bezpieczeństwa oraz ograniczenia możliwości uszkodzenia czułych układów elektronicznych systemu ramanowskiego przez pola elektromagnetyczne generowane przez stanowisko technologiczne zastosowanie światłowodów do transmisji sygnałów optycznych pozwala też na umieszczenie głównych bloków systemu ramanowskiego w odpowiedniej odległości od stanowiska technologicznego.

3.2. Układ do monitorowania procesów zol-żel w ciekłych materiałach organiczno-nieorganicznych

System ramanowski wykorzystujący promieniowanie laserowe o długości fali 830 nm został skonfigurowany w laboratorium VTT Electronics (Oulu, Finlandia) na bazie spektrometru Ramstas [2] i przedstawiony schematycznie na Rys.2a. Wykorzystuje on laser półprzewodnikowy fali ciągłej, zapewniający moc promieniowania wzbudzającego do 200 mW. Wykonany w VTT spektrograf typu *axial transmissive*, wyposażono w holograficzną siatkę transmisyjną (zakres długości fali od 850 do

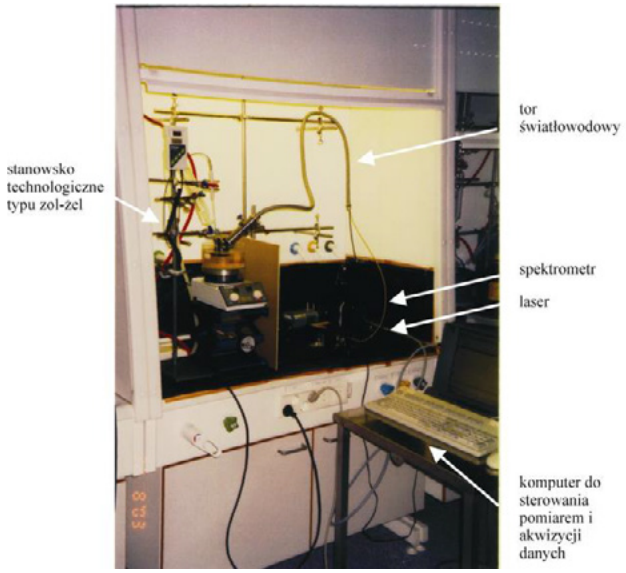
1000 nm) oraz kamerę CCD Oriel InstaSpec IV, chłodzoną termoelektrycznie do -30°C, co jest niezbędne z uwagi na niski poziom sygnału Ramana, na poziomie szumów cieplnych kamery CCD. Pomiary z laserem 830 nm przeprowadzono w zakresie stokesowskim, w przedziale od 200 do 2000 cm^{-1}, z rozdzielczością 8 cm^{-1}. Do pomiarów w paśmie poszerzonym od 2000 do do 3500 cm^{-1} wykorzystywano laser Ti:Sp o długości fali 720 nm i mocy sygnału 50 mW. Do kalibracji spektralnej systemu wykorzystano lampę rtęciową. Transmisja wiązek optycznych pomiędzy aparaturą ramanowską i stanowiskiem technologicznym odbywała się za pomocą wiązek światłowodowych zakończonych sondami, których konstrukcja została przedstawiona na Rys.2b. Ramanowskie sondy światłowodowe, oprócz soczewek skupiających wiązkę laserową na badanym obiekcie i zapewniających akwizycję istotnej części sygnału rozproszonego, muszą zawierać filtry: linii laserowej w torze nadawczym oraz pasmowo-zaporowy, usuwający sygnał Rayleigha, który może nasycać detektor. Z uwagi na konieczność: (1) skupienia wiązki laserowej wewnątrz kolby, (2) uniknięcia wzbudzenia rozpraszania ramanowskiego w ściance (jej linie ramanowskie pokryłyby się z liniami sieci nieorganicznej SiO_2) oraz (3) zapewnienia dużej efektywności transferu sygnału ramanowskiego z wnętrza komory, sondy pomiarowe zostały wyposażone w skupiające układy optyczne o aperturze równej 2,5 cm i długości ogniskowej od 1,5 do 2,5 cm.



Rys. 2. Schemat systemu ramanowskiego do badania ciekłych próbek objętościowych: a) główne bloki funkcjonalne, b) przekrój sondy ramanowskiej.

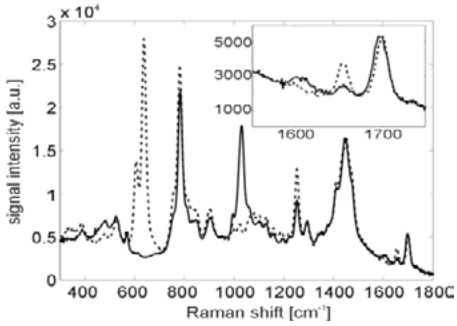
Dedykowany układ mechaniczny wykorzystywał mikrometryczny posuw XYZ do pozycjonowania sond w stosunku do kolb, w których zachodził proces. Sondy umieszczone były ukośne z dołu, bocznie lub ukośnie z góry w stosunku do kolb, co nie zakłócało pracy mieszacza magnetycznego, a skutecznie uniezależniało pomiary od poziomu cieczy w kolbie. Połączenie stanowiska technologicznego ze światłowodowym systemem do diagnostyki *in-situ* procesu zol-żel przedstawiono na Rys. 3.

Przykładem wykorzystania układu jest były procesu żelowania hybrydowych układów organiczno-nieorganicznych zawierających 3-aminopropyl (trimetoksy) - i 3-glicydoksypropyl (trimetoksy)silany zawierających różne ketony. Wykorzystano następujące prekursory: 3-Aminopropyl(trimethoxy)silane (APTES, ≥97%). 3-glycidoxypropyl(trimethoxy)silane (GPTS, 98%), alkohol izopropylowy (IPA, ≥99.8%), aceton (Ac≥99.8%), keton metylowoizobutylowy (Mibk, ≥99%) i woda dejonizowana, na bazie których dokonywano żelowania i tworzono produkty iminowe Wykazano wpływ wody na proces otwierania pierścienia epoksydowego w GPTS, jak również potwierdzono wpływ rozpuszczalników na ścieżkę reakcji.

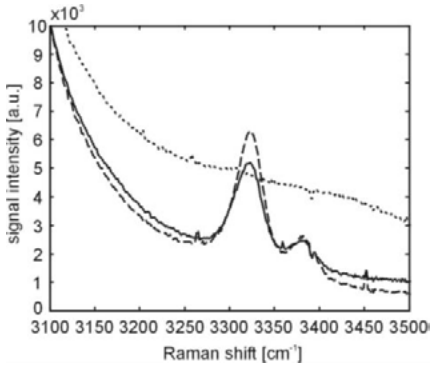


Rys.3. System do monitorowania w czasie rzeczywistym procesu technologicznego żelowania sprzęgnięty światłowodami ze stanowiskiem technologicznym i połączony w aparaturą do akwizycji danych.

Przykłady pozyskiwania informacji o procesie żelowania z widm pomiarowych przedstawiono na Rys. 4 i 5.



Rys. 4. Porównanie widma Ramana roztworu po inicjalizacji żelowania (linia kropkowana) i widma produktu końcowego zarejestrowanego po około 22 godzinach mieszania (linia ciągła).

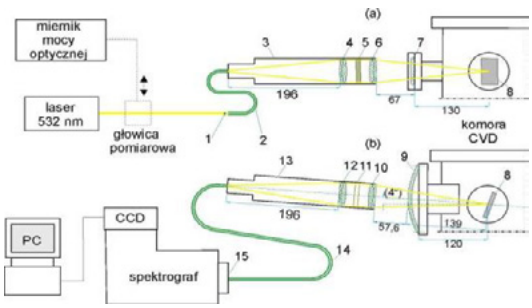


Rys.5 Porównanie widma Ramana po inicjalizacji żelowania Linia przerywana przedstawia widmo silanów, linia stała - rozwiązanie tuż po dodaniu rozpuszczalników, krzywa przerywana - widmo zarejestrowane po 2 min. mieszania [3].

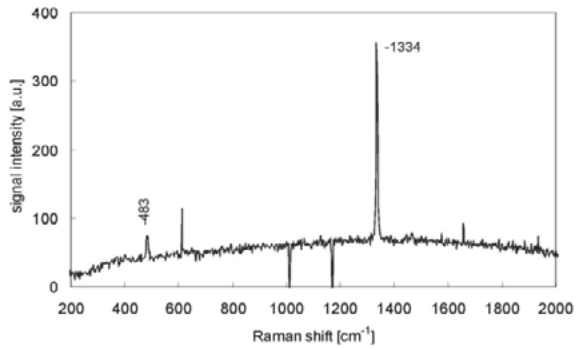
Przykładowo, zanik linii 1256 cm^{-1} na Rys. 4, przypisanej do pierścienia epoksydowego, potwierdza jednoznacznie otwarcie pierścienia epoksydowego. Znikający dublet przy 641 cm^{-1} i 607–611 cm^{-1} wynika z symetrycznego i asymetrycznego rozciągania ν(Si(-O-CH_3)), co oznacza dekompozycję części trimetoksysilanowej. Nowe pasma od 1000 do 1100 cm^{-1}, które pojawiają się po dodaniu wody, można przypisać sieci organicznej opartej na węglu i metanolowi, który jest produktem ubocznym polimeryzacji nieorganicznej. Zmniejszenie natężenia dubletu Ramana przy 3310 cm^{-1} i 3370 cm^{-1} (Rys. 5) jest przypisane wiązaniom NH_2 i N-H z APTES, które ulegają otwarciu. Potwierdzono w ten sposób kluczowe reakcje dla żelowania bez ingerencji w proces.

3.3. Układ do diagnostyki in-situ procesu CVD

Optoelektroniczny system Ramana dedykowany do monitorowania procesów przeprowadzanych w układzie CVD wspomaganym plazmą mikrofalową przedstawiono na Rys. 6. System CVD wykorzystuje chłodzoną wodą komorę próżniową o średnicy 160 mm, wyposażoną w szklane wzierniki w ścianach, jak również stanowisko pompowe, sekcją mikrofalową, sekcję pola magnetycznego prądu stałego, grzejnik podłoży i przepływomierze gazu. Z uwagi na warunki pracy komory założono dostęp zdalny przez wzierniki. System ramanowski składa on z lasera o długości fali 532 nm i mocy 200 mW, spektrografu Kaiser Holospec f/1.8i oraz termoelektrycznie chłodzonej kamery CCD Andor DVP401-BV. Zestaw dwóch dedykowanych sond światłowodowych w konfiguracji teleskopowej, umożliwia sprzęgnięcie systemu ramanowskiego z komorą. Zastosowane filtry optyczne eliminują sygnały zakłócające: emisję spontaniczną lasera i rozpraszanie Rayleigha. Dzięki użyciu w sondach skupiających soczewek klejonych o dużej średnicy (ok. 50 mm), uzyskano możliwość rejestracji sygnału pochodzącego od cienkich warstw o grubości submikronowej z dystansu roboczego 200 mm. Zastosowanie światłowodów do transmisji sygnałów optycznych pozwala prowadzić badania zdalnie.



Rys. 6. System ramanowski o długim dystansie roboczym do monitorowania procesu CVD: (a) część wzbudzenia, (b) część akwizycji; 1 - złącze, 2 - włókno optyczne, 3 - sonda w torze wzbudzenia, 4, 6, 10, 12 - soczewki, 5 - filtr pasmowy linii laserowej 532 nm, 7, 9 – w zierniki komory CVD, 8 – rosnąca warstwa 11 – filtr pasmowo-zaporowy typu Notch, 13 - sonda zbierająca, 14 - wiązka wielowłóknowa, 15 - adapter; wymiary w mm. [4]



Rys. 7. Widmo testowe warstwy diamentowej umieszczonej w komorze CVD.

Przykładowe widmo cienkiej warstwy diamentowej położonej na szkle kwarcowym zarejestrowanej przez prototyp systemu Ramana pracującego na dużej odległości (20 cm) z zewnątrz komory CVD, potwierdzające poprawną pracę systemu przedstawiono na Rys. 7. Linia 1334 cm^{-1} pochodzi od fazy diamentowej o hybrydyzacji sp^3, a linia 483 cm^{-1} - od podłoża technologicznego – szkła kwarcowego [4].

4. PODSUMOWANIE

W artykule przedstawiono wybrane przykłady wykorzystania światłowodów w systemach ramanowskich. Dzięki przesunięciu widma mierzonych sygnałów z zakresu średniej podczerwieni do pasma widzialnego i bliskiej podczerwieni możliwe jest sprzęganie badanych obiektów z aparaturą pomiarową za pomocą torów światłowodowych, zarówno ze światłowodów specjalnych (włókna o średnicy rdzenia 100 i 200 µm, wiązki wielowłóknowe, itp.) jak i potencjalnie ze standardowych światłowodów telekomunikacyjnych. Podczas prac uzyskiwano możliwość stosowania włókien o długościach rzędu metrów, co umożliwia odseparowanie stanowisk technologicznych od aparatury pomiarowej i tworzenie lokalnych sieci światłowodowych.

LITERATURA

[1] Li, J., Zhang, M. Physics and applications of Raman distributed optical fiber sensing. *Light Sci Appl* 11, 128 (2022). https://doi.org/10.1038/s41377-022-00811-x

[2] P. Niemelä i wsp., „A rugged raman spectrometer for process measurements"; *Proceedings of Finnish Optics Days - Oulu*, 1998, s. 35.

[3] M. Keranen, A. Maaninen, M. Gnyba, T. Maaninen, "Spectroscopic studies of a ring opening process between epoxy-and aminosilanes and imine formation reactions in aqueous solutions", *Journal of Sol-Gel Science and Technology*,87,3, s. 725-733,2018

[4] M. Gnyba, M. Kozanecki, P. Wroczyński, B.B. Kosmowski, R. Bogdanowicz, "Long-working- distance Raman system for monitoring of µPA ECR CVD process of thin diamond/DLC layers growth", *Photonics Letters of Poland*,1,2,76-78,2009

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

Ramy statystycznej walidacji modeli uczenia maszynowego w telekomunikacji

An approach for statistical validation of machine learning models in telecommunication

Sebastian Zarębski [1,2]; Krzysztof Rusek [3]; Piotr Chołda [4]

[1] NOKIA, Kraków, sebastian.zarebski@nokia.com
[2,3,4]AGH University of Krakow, Institute of Telecommunications, Kraków,
[2] zarebski@agh.edu.pl, [3] krzysztof.rusek@agh.edu.pl [4] piotr.cholda@agh.edu.pl

DOI: 10.15199/59.2025.4.55

Streszczenie: Modele uczenia maszynowego (ML) rewolucjonizują podejmowanie decyzji i optymalizację w różnych branżach. W krytycznych obszarach, takich jak telekomunikacja, zwłaszcza w przypadku stacji bazowych 4G i 5G, wczesne wykrywanie błędów w testach regresyjnych jest niezbędne do zapewnienia wysokiej jakości usług (QoS) i niezawodności sieci. Kluczowe jest wykazanie, że modele ML zapewniają statystycznie istotne korzyści w porównaniu z przypadkowymi wynikami. Niniejszy artykuł przedstawia ramy walidacji statystycznej dla oceny wytrenowanych modeli ML, zapewniające ich skuteczność w wymagających środowiskach oprogramowania. Dzięki proponowanej metodzie wykazujemy zdolność modelu ML do uczenia się trendów z danych, odróżniając jego wydajność od losowej linii bazowej. Stosując różnorodne testy statystyczne dostarczamy dowodów na przewagę klasyfikatora ML. Ramy te oferują praktyczną ścieżkę budowania zaufania do komponentów ML, zapewniając ich pomyślną integrację w infrastrukturach.

Abstract: Machine Learning (ML) models are revolutionizing decision-making and optimization across industries. In critical areas, such as telecommunications, especially for 4G and 5G base stations, early detection of faults in software regression testing is essential to ensure high QoS and network reliability. It is essential to show that ML models deliver statistically significant advantages over randomness. This paper presents a resilient statistical validation framework for evaluating post-training ML models, ensuring their effectiveness in challenging software settings. Our method distinctly proves an ML model's capacity to learn trends from data, separating its performance from a random baseline. By employing a variety of statistical tests, we provide evidence for the ML classifier's consistent and significant outperformance. This framework offers a practical path to building trust in ML components, ensuring their successful integration into infrastructures.

Słowa kluczowe: ocena modeli, uczenie maszynowe, testowanie hipotez, walidacja statystyczna

Keywords: hypothesis testing, machine learning, model evaluation, statistical validation

1. WSTĘP

Modele uczenia maszynowego (ML) fundamentalnie zmieniają sposób, w jaki organizacje podejmują decyzje, optymalizują procesy operacyjne i napędzają innowacje. Jednak wdrożenie modelu ML bez rygorystycznej oceny ryzyka może prowadzić do błędnych decyzji opartych na niewiarygodnych prognozach. Kluczowe jest nie tylko zbudowanie modelu, ale – co ważniejsze – przeprowadzenie jego rygorystycznej walidacji statystycznej. Niniejszy artykuł przedstawia metodyczne podejście do oceny modeli ML w ramach biznesowych, koncentrując się na systematycznej weryfikacji, czy model jest znacząco lepszy niż przypadkowe zgadywanie lub ustalone wartości bazowe, a tym samym wnosi realną wartość.

Naszym nowatorskim wkładem jest zastosowanie kompleksowego ramowego podejścia do walidacji statystycznej w celu wczesnego wykrywania błędów w testach regresyjnych stacji bazowych 4G i 5G. Podczas gdy indywidualne testy statystyczne i modele ML są szeroko stosowane, ich systematyczne połączenie i interpretacja zapewniają **statystycznie solidną pewność dla wdrożeń ML o znaczeniu krytycznym w telekomunikacji**, unikalnie wypełniając lukę w praktykach przemysłowych, które często wykorzystują ograniczone metryki. To zwiększa poziom pewności wymagany przed integracją ML w wrażliwej infrastrukturze, gdzie niezawodność i QoS są kluczowe.

Niniejszy artykuł jest zorganizowany w następujący sposób: Sekcja 2 omawia znaczenie wykrywania błędów oprogramowania w telekomunikacji, podkreślając potrzebę zapobiegania awariom i stratom ekonomicznym w systemach 4G i 5G w celu utrzymania QoS i niezawodności sieci. Sekcja 3 przedstawia naszą metodologię, w tym model uczenia maszynowego, metryki wydajności i testy walidacyjne. Sekcja 4 pokazuje, w jaki sposób te testy potwierdzają skuteczność naszej metody w ocenie istotności statystycznej prognoz. Na koniec, Sekcja 5 podkreśla nasz wkład oraz wpływ proponowanego podejścia do walidacji na zapewnienie niezawodnego uczenia maszynowego w infrastrukturze krytycznej.

2. OPIS PROBLEMU

Głównym wyzwaniem jest wczesne wykrywanie błędów w testowaniu oprogramowania, kluczowe dla IT i telekomunikacji. Wczesne rozpoznanie błędów zmniejsza

koszty i poprawia jakość oprogramowania, co jest istotne dla QoS i niezawodności w sieciach 5G. Nawet drobne usterki oprogramowania w stacjach bazowych 4G i 5G mogą mieć poważne konsekwencje, w tym **przerwy w świadczeniu usług, zerwane połączenia, pogorszenie wydajności sieci (np. zwiększone opóźnienia lub zmniejszona przepustowość), a nawet wpływ na usługi ratunkowe**, prowadząc do strat ekonomicznych. Kluczowym problemem jest zwykle **nierównowaga klas**: w naszym przypadku w NOKIA jedynie 2,16% przypadków testowych faktycznie wykazuje błędy ($y_{\text{true}} = 1$), podczas gdy większość kończy się bez problemów. Rozważmy teraz codzienny test regresji, w którym przypadek testowy jest identyfikowany przez swoją nazwę i powiązane metadane. Na przykład:

- **Nazwa testu:**
  `NR_5G_SA_Handover_Latency_UL_Perf_Test_10ms`
- **Powiązana zmiana w kodzie (opis programisty):**
  `Fix: Buffer overflow in signaling_module.h, changed base_station_core.c for 4x4 MIMO. See JIRA-1234.`
  `Code snippet:`
  `‘if (buffer_size > MAX_BUFFER_SIZE) { /* handle error */ }‘`

Wyzwanie polega na jak najwcześniejszym przewidywaniu, które konkretne testy prawdopodobnie zakończą się niepowodzeniem w związku z codziennymi aktualizacjami oprogramowania (wiadomości z commitów i fragmenty kodu).

Surowe opisy tekstowe, składające się z połączonych wiadomości z tzw. commitów (fragmentów zmienionego kodu) i nazw testów, są przekształcane w spójne dane wejściowe tekstowe, które następnie są tokenizowane i przetwarzane przez **duży model językowy (LLM) oparty wyłącznie na części kodującej (enkoder)**, konkretnie model **RoBERTa-large** [1]. Model RoBERTa-large został wybrany ze względu na swoją zdolność do uchwycenia bogatych osadzeń kontekstowych pochodzących z obszernego zbioru danych tekstowych. Enkoder modelu językowego wyodrębnia te osadzenia, które następnie są przekazywane przez **pojedynczą głęboką warstwę liniową** (o wymiarze wejściowym 1024 i wyjściowym 1) z **funkcją aktywacji ReLU**. Proces ten kończy się **jednym neuronem wyjściowym** z **funkcją aktywacji sigmoid** w celu klasyfikacji binarnej, rozróżniającej przypadki wadliwe i niewadliwe. Model był trenowany przez okres **5 epok** z wykorzystaniem **stopy uczenia wynoszącej** 2×10^{-5}. Należy zauważyć, że ogólna dokładność w wykrywaniu błędów pozostaje niska z powodu krytycznej i nieuchwyconej wiedzy specyficznej dla domeny (takiej jak specjalistyczne skróty), co stanowi wyzwanie wykraczające poza zakres przedstawionego tutaj badania.

Aby uzasadnić zastosowanie tego klasyfikatora ML, należy wykazać jego wyższość nad trywialnymi podejściami. Ustanawiamy **ograniczoną losową linię bazową** jako główny punkt odniesienia. Ta linia bazowa jest starannie zaprojektowana: generuje losowe binarne predykcje, które w każdej iteracji cyklu rozwoju oprogramowania dokładnie odpowiadają liczbie pozytywnych predykcji (błędów) dokonanych przez model ML. Zapewnia to, że losowa linia bazowa ma tę samą „skłonność” do identyfikowania błędów co model ML, co pozwala ocenić, czy model ML umieszcza te predykcje w statystycznie lepszej lokalizacji. Badanie koncentruje się na statystycznej walidacji tego klasyfikatora opartego na RoBERTa w porównaniu z tą ograniczoną losową linią bazową, aby potwierdzić jego podstawową zdolność uczenia się i znaczącą poprawę. Eksperyment porównuje predykcje klasyfikatora ML z pełnym zestawem $N_{\text{RANDOM_DRAWS}} = 19\,683$ niezależnych uruchomień losowej linii bazowej. Liczba tych porównań wynika z trzeciej potęgi liczby 27 tygodni danych z wynikami testów, konkretnie $27^3 = 19\,683$. Ten wybór zapewnia solidną i niearbitralną podstawę do porównań statystycznych, gwarantując:

- **Wysoka precyzja wartości** p: Liczba permutacji bezpośrednio wpływa na rozdzielczość dostępnych wartości p. Przy $N_{\text{RANDOM_DRAWS}} = 19\,683$ najmniejsza niezerowa wartość p, którą można oszacować, wynosi około $1/19\,683 \approx 5.08 \times 10^{-5}$. Wysoka rozdzielczość jest kluczowa dla rozróżniania krytycznych wyników, szczególnie przy ocenie modeli dla systemów wrażliwych, które wymagają wysokiego poziomu zaufania do wydajności [2, 3].
- **Wystarczająca siła statystyczna**: Duża liczba losowań zapewnia wystarczającą moc statystyczną do wykrywania subtelnych różnic między modelem ML a losową bazą odniesienia, zmniejszając ryzyko błędów drugiego rodzaju, w których skuteczny model mógłby zostać błędnie uznany za nieistotny.
- **Odporność dzięki centralnemu twierdzeniu granicznemu**: Indywidualne metryki wydajności mogą nie podlegać rozkładowi normalnemu, ale centralne twierdzenie graniczne sugeruje, że przy wielu niezależnych losowaniach ($N_{\text{RANDOM_DRAWS}}$) rozkład próbkowy metryk oceny zbliża się do normalności. To wspiera wiarygodne wnioskowanie statystyczne na podstawie zagregowanych metryk, umożliwiając porównania modelu ML z losową bazą odniesienia [4, 5].
- **Efektywność obliczeniowa**: Chociaż większa liczba losowań zapewniłaby jeszcze wyższą rozdzielczość wartości p, wybór $27^3 = 19\,683$ został dokonany ze względu na praktyczne ograniczenia pamięci obliczeniowej.

3. METODYKA

Zbiór danych obejmuje wykonania testów zebrane w 27 różnych punktach czasowych (tygodniach), co odzwierciedla około rok testowania. Każda partia zawiera kilka tysięcy indywidualnych przypadków testowych, co daje łącznie ponad 100 000 wyników testów. Zbiór odzwierciedla rzeczywiste środowisko testowania regresji. Następnie y_{true} (prawdziwe etykiety błędów) i y_{pred} (przewidywania modelu ML) są binarne (0 lub 1), zgodnie z rozkładem Bernoulliego dla każdego przypadku testowego i rozkładem dwumianowym w agregacji tygodniowej. $y_{\text{pred_proba}}$ (przewidywane prawdopodobieństwa modelu ML) mają wartości ciągłe od 0 do 1. Ich rozkład, analizowany za pomocą histogramów i wykresów gęstości, wykazuje skośność i brak normalności, co jest typowe dla wyników klasyfikacji na niezrównoważonych zbiorach danych.

Rozpoznając ograniczenia dokładności w przypadku niezrównoważonych danych, wybrano dwa kluczowe wskaźniki. **Współczynnik korelacji Matthewsa (MCC)** jest niezawodny w przypadku niezrównoważonej klasyfikacji binarnej [6]. **Średni procent wykrytych błędów (APFD)** ocenia *efektywność* wykrywania błędów [7], co jest kluczowe dla QoS [8], mierząc

wczesne wykrywanie błędów w uporządkowanych przypadkach testowych na podstawie przewidywanego prawdopodobieństwa awarii. APFD, obliczany jako $APFD = 1 - (\sum_{i=1}^{m} TF_i)/(nm) + 1/(2n)$, wykorzystuje m błędnych przypadków testowych, n wszystkich przypadków testowych oraz pozycję TF_i i-tego błędnego przypadku na posortowanej liście. Model ML porządkuje na podstawie prawdopodobieństwa awarii, traktując zestaw testów jako równie ważny, ale priorytetowo traktując krótsze testy wśród przewidywanych awarii. To priorytetyzowanie przyspiesza rozwiązywanie problemów, skraca czas wprowadzenia aktualizacji oprogramowania na rynek i maksymalizuje wpływ wczesnego wykrywania błędów.

Aby określić istotność statystyczną, zastosowano testowanie hipotez ze ścisłym poziomem istotności $\alpha = 0.01$, aby zminimalizować **błędy pierwszego rodzaju**, aby zapobiec zakłóceniom spowodowanym przez modele o słabej wydajności. Przewidując różnice w rozkładzie danych, zastosowano testy nieparametryczne zamiast testu T: **test permutacyjny** [9], **test McNemara** [10], **test chi-kwadrat i dokładny test Fishera** [11] oraz **test U Manna-Whitneya** [12]. Zapewnia to solidny widok statystyczny, wspierając wydajność modelu i umożliwiając *statystyczną pewność* dla wdrożenia produktu. Testy permutacyjne i U Manna-Whitneya oceniały rozkłady MCC i APFD z 19 683 niezależnych uruchomień w celu zmierzenia ogólnych różnic w wydajności.

4. WYNIKI

Eksperyment dotyczył poważnej nierównowagi klas, gdzie tylko 2.16% przypadków stanowiły rzeczywiste błędy. Nasz klasyfikator ML, zoptymalizowany pod kątem wykrywania błędów, oznaczył 51.12% przypadków jako potencjalne błędy. Celem było statystyczne wykazanie wartości tej strategii w porównaniu z przypadkowym podejściem. Tabela 1 przedstawia średnie i odchylenia standardowe wyników MCC i APFD dla wszystkich uruchomień eksperymentalnych.

Tabela 1: Podsumowanie wyników MCC i APFD

Metryka	Model	Średnia ± Odch. std.
MCC	Klasyfikator ML	$0.0176 \pm 3.47 \times 10^{-18}$
	Losowy	-0.0016 ± 0.0030
APFD	Klasyfikator ML	$0.5112 \pm 1.11 \times 10^{-16}$
	Losowy	0.4835 ± 0.0149

Jak pokazano w tabeli, nawet wstępnie wytrenowany klasyfikator ML konsekwentnie przewyższa ograniczoną losową bazę porównawczą, osiągając wyższe średnie wyniki MCC i APFD. Minimalne odchylenia standardowe wskazują na bardzo spójną wydajność modelu we wszystkich symulacjach, w przeciwieństwie modelu losującego.

4.1 Testy statystyczne dla pojedynczych uruchomień

Tabela 2 przedstawia zbiorcze wyniki testów McNemara, Chi-kwadrat i dokładnego testu Fishera, przeprowadzonych indywidualnie dla każdego z 19 683 losowych uruchomień.

Tabela 2: Zagregowane wyniki testów statystycznych dla pojedynczych uruchomień ($\alpha = 0.01$)

Test	$p < \alpha$	Średnia wartość p
McNemar	0	0.197
Chi-kwadrat	19 683	3.84×10^{-102}
Dokładny test Fishera	19 683	3.89×10^{-102}

Wyniki można podsumować następująco:

- **Test McNemara:** Średnia wartość p wynosząca 0.197 nie jest statystycznie istotna przy $\alpha = 0.01$, a wszystkie 19 683 testy wykazały brak istotności. Oznacza to brak istotnych różnic w wzorcach błędów między klasyfikatorem ML a losową bazą porównawczą, ponieważ oba zostały ustawione na przewidywanie tej samej liczby pozytywnych przypadków na grupę danych.
- **Test Chi-kwadrat:** Bardzo mała wartość p (3.84×10^{-102}) dla wszystkich 19 683 testów wskazuje na silne powiązanie wyników klasyfikatora ML z rzeczywistymi wzorcami, a nie przypadkiem.
- **Dokładny test Fishera:** Mała średnia wartość p (3.89×10^{-102}) potwierdza wyniki testu Chi-kwadrat, wskazując, że obserwowane rozkłady są mało prawdopodobne, aby były losowe.

4.2 Zagregowane testy statystyczne na rozkładach wyników

Te testy skupiają się na porównaniu rozkładów wyników MCC i APFD między klasyfikatorem ML a losową bazą porównawczą w 19 683 uruchomieniach. Tabela 3 przedstawia wyniki testów permutacyjnych i testów U Manna-Whitneya, potwierdzając przewagę klasyfikatora ML.

- **Test permutacyjny dla MCC**: Dodatnia różnica średniej MCC (0.019 186) oraz wartość p równa 0.0 wskazują, że różnica nie jest przypadkowa.
- **Test permutacyjny dla APFD**: Podobnie, różnica (0.027 740) i niska wartość p wskazują na istotność różnicy.
- **Test U Manna-Whitneya dla MCC**: Bardzo niska wartość p potwierdza wyższość klasyfikatora ML w MCC.
- **Test U Manna-Whitneya dla APFD**: Wyniki wskazują na wyższość klasyfikatora ML w APFD.

Podsumowując, te testy statystyczne dostarczają jednoznacznych dowodów na skuteczność proponowanego klasyfikatora ML w porównaniu z przypadkowym podejściem.

5. PODSUMOWANIE

Główne osiągnięcie tego artykułu to **promowanie rygorystycznej metodologii walidacji statystycznej** jako fundamentu integracji uczenia maszynowego w systemach krytycznych, szczególnie w telekomunikacji. Nasze wyniki przedstawiają proponowane podejście do systematycznej walidacji modeli ML w środowiskach krytycznych. Poprzez zastosowanie różnorodnych testów statystycznych i ocenę kluczowych metryk, dostarczamy

Tabela 3: Zagregowane wyniki testów statystycznych na rozkładach MCC i APFD

Test / metryka	**Zaobserwowana różnica**	**p-value**	**CI (rozkład zerowy)**
Testy permutacyjne			
Permutacyjny MCC	0.019 186	0.0	−0.000 194 to 0.000 195
Permutacyjny APFD	0.027 740	0.0	−0.000 344 to 0.000 343
Testy U Manna-Whitneya			
Mann-Whitney MCC	387 420 489.0	0.0	–
Mann-Whitney APFD	377 638 038.0	0.0	–

Uwaga: Dla testów permutacyjnych i Manna-Whitneya, $p < 1/100\,000$ oznacza $p < 1 \times 10^{-5}$.

dowodów na bezpieczne i niezawodne wdrażanie klasyfikatorów ML w złożonych środowiskach programistycznych. Takie ramy walidacyjne są kluczowe dla budowania zaufania i pewności wobec rozwiązań AI w infrastrukturze krytycznej.

Rozwój zaawansowanych sieci, takich jak 5G, wymaga rozsądnej i przemyślanej automatyzacji, co zwiększa znaczenie jakości i niezawodności oprogramowania. Wdrażanie modeli ML bez dokładnej oceny jest zbyt ryzykowne. Nasze ramy, z solidnymi punktami odniesienia, metrykami uwzględniającymi kontekst oraz różnorodnymi testami statystycznymi, oferują **jasną i praktyczną ścieżkę do budowania weryfikowalnego zaufania i pewności** wobec komponentów ML. Uosabiają one zasady odpowiedzialnej stucznej inteligencji (AI), promują **przejrzystość** w ocenie, ustanawiają jasną **odpowiedzialność** za wydajność oraz zapewniają **niezawodność** takich systemów. Te elementy są kluczowe dla wdrażania rozwiązań AI/ML w przedsiębiorstwach i zgodności z normami, takimi jak ETSI ZSM [13].

Proponujemy to podejście walidacyjne jako podejście inżynierskie **odpowiednie do szerszego zastosowania**. Przyjęcie kompleksowych ram oceny statystycznej, takich jak nasze, nie jest tylko opcją; jest to **warunek konieczny do budowy niezawodnych, wydajnych i godnych zaufania inteligentnych sieci przyszłości.** Zdecydowanie zachęcamy do rozważenia go jako standardowej, niepodlegającej negocjacjom procedury w cyklu życia produktowego każdego systemu ML przeznaczonego dla infrastruktury krytycznej.

BIBLIOGRAFIA

[1] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach*. arXiv. http://arxiv.org/abs/1907.11692

[2] Good, P. (2005). *Permutation, Parametric and Bootstrap Tests of Hypotheses* (3rd ed.). Springer-Verlag.

[3] Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press.

[4] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

[5] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.

[6] Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

[7] Rothermel, G., Untch, R. H., Chu, C., & Harrold, M. J. (2001). Prioritizing test cases for regression testing. *IEEE Transactions on Software Engineering*, *27*(10), 929–948. https://doi.org/10.1109/32.962562

[8] Rothermel, G., & Harrold, M. J. (1999). Analyzing regression test selection techniques. *IEEE Transactions on Software Engineering*, *25*(6), 529-551. https://doi.org/10.1109/32.815323

[9] Ojala, M., & Garriga, G. C. (2010). Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, *11*, 1835-1863.

[10] Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*(7), 1895-1923.

[11] Zhai, Y., Song, W., Liu, X., Liu, L., & Zhao, X. (2018). A Chi-Square statistics based feature selection method in text classification. In *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)* (pp. 160-163). IEEE. https://doi.org/10.1109/ICSESS.2018.8663882

[12] Nachar, N. (2008). The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution. *Tutorials in Quantitative Methods for Psychology*, *4*(1), 13-20.

[13] ETSI. (2019). *Zero-touch network and Service Management (ZSM); Reference Architecture* (ETSI GR ZSM 002 V1.1.1). ETSI.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

KLASYFIKACJA RUCHU INTERNETOWEGO ZA POMOCĄ PRZETWARZANIA NEUROMORFICZNEGO

INTERNET TRAFFIC CLASSIFICATION USING NEUROMORPHIC PROCESSING

Szymon Betka[1]; Sylwester Kaczmarek[2]; Marek M. Landowski[3]

[1] Adtran, Networks sp. z o.o.,Gdynia, szymon.betka@adtran.com
[2] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, kasyl@eti.pg.edu.pl
[3] Intel, Gdańsk, marek.m.landowski@intel.com

DOI: 10.15199/59.2025.4.56

Streszczenie: **W ostatnich latach na fali popularności sieci neuronowych, zaczęły pojawiać się prace, w których autorzy publikują modele wykazujące się dużą dokładnością w klasyfikacji zaszyfrowanego ruchu internetowego. W szczególności, interesującym kierunkiem jest użycie sieci impulsowych, inspirowanych sposobem w jaki działa ludzki mózg, wykazując że takie modele także nadają się do tego zadania. Bazując na tych wynikach, w niniejszej pracy zaimplementowano model impulsowych sieci neuronowych w oparciu o oprogramowaniu Intel® Lava, który następnie przetestowano na symulatorze neuromorficznego procesora Intel® Loihi 2 w celach porównawczych.**

Abstract: **In recent years, with the rise in popularity of neural networks, studies have begun to emerge in which authors publish models demonstrating high accuracy in classifying encrypted internet traffic. Particularly interesting is the use of spiking neural networks, biologically inspired by the way the human brain works, showing that such models are also suitable for this task. Based on these results, this work implemented model of spiking neural network in the Intel® Lava framework and tested it on the Intel® Loihi 2 neuromorphic processor simulator for accuracy comparisons.**

Słowa kluczowe: **impulsowe sieci neuronowe, klasyfikacja ruchu internetowego, przetwarzanie neuromorficzne**

Keywords: **spiking neural networks, internet traffic classification, neuromorphic computing**

1. WSTĘP

W wyniku zwiększania bezpieczeństwa w Internecie o szyfrowanie przesyłanych danych, klasyczne metody identyfikacji źródeł ruchu oparte o informację z nagłówka pakietu czy jego zawartości w celu przypisania klasy i dalszej obsługi w systemie stają się mało dokładne. W takich sytuacjach dodatkowo poddaje się analizie czas przybycia czy rozmiar pakietu lub próbuje się wywnioskować inne cechy jakimi objawia się ruch w celu poprawnej klasyfikacji. Na fali popularności sieci neuronowych, na przykład w [11] zbadano możliwość wykorzystania splotowych sieci neuronowych (*convolutional neural networks* – CNNs), osiągając przy tym zadowalające wyniki klasyfikacji ruchu internetowego, także gdy był on przesyłany poprzez tunel prywatnej sieci virtualnej (*virtual private network* - VPN), czy przy użyciu rozproszonej komunikacji TOR (*the onion router*). Zbliżone podejście do rozwiązań w oparciu o sieci splotowe zastosowali autorzy rozwiązania [10], z tą różnicą, że wykorzystali do tego celu impulsową sieć neuronową (*spiking neural network* – SNN). Zaletą impulsowych sieci neuronowych jest ich możliwość rozpoznawania zależności czasowych zawartych w danych [7] (czas pomiędzy impulsami także niesie informację), co stanowi znaczącą różnicę z pozostałymi sieciami neuronowymi. Kolejną zaletą jest możliwość uruchamiania ich na procesorach neuromorficznych, które naśladują sposób działania ludzkiego mózgu i charakteryzują się wysoką równoległością oraz niskim zużyciem energii. Przykładowe rozwiązania to między innymi Intel Loihi 2 [5], IBM TrueNorth [1] czy SpiNNaker [3] – każde z nich reprezentuje inne podejście do architektury procesora neuromorficznego.

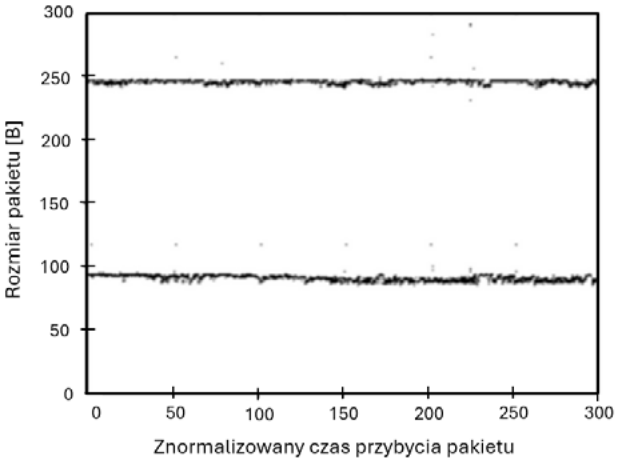
W poniższej pracy przedstawiono wstępne badania dokładności klasyfikacji zaszyfrowanego ruchu przy użyciu sieci impulsowych zaimplementowanych za pomocą oprogramowania Intel Lava [9] oraz symulatora procesora neuromorficznego Intel Loihi 2. W dalszej części artykułu przedstawiono architekturę rozwiązania wraz z metodą przeprowadzonego treningu oraz testowania. Wyniki badań porównawczych zamieszczono w rozdziale trzecim. A podsumowanie całości zawiera rozdział czwarty.

2. ARCHITEKTURA ROZWIĄZANIA

W celu przeprowadzenia badań z zastosowaniem impulsowych sieci neuronowych zastosowano symulator procesora Loihi 2. Symulator Loihi to narzędzie programowe opracowane przez firmę Intel jako część otwartego oprogramowania Intel Lava, które umożliwia testowanie funkcjonalne i rozwijanie aplikacji neuromorficznych bez potrzeby fizycznego dostępu do układów Loihi. Symulator działa na klasycznych procesorach CPU, emulując zachowanie sieci neuronowych opartych na impulsach, co umożliwia szybkie prototypowanie, debugowanie oraz testowanie ich dokładności. Symulator wspiera programowalne dynamiki neuronów i synaps, zgodne z architekturą Loihi 2, oraz umożliwia tworzenie złożonych, modułowych sieci neuronowych. Dzięki integracji z językiem Python i otwartym oprogramowaniem Intel Lava, użytkownicy mogą programowalnie testować funkcjonalność sieci impulsowych, a następnie uruchamiać je w środowisku

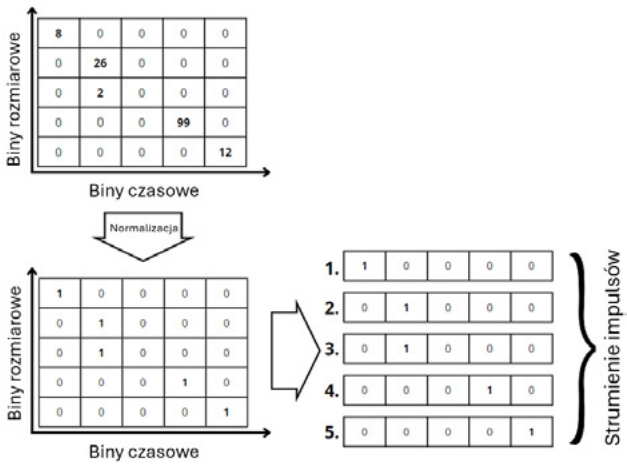
symulacyjnym.

Proponowana impulsowa sieć neuronowa do klasyfikacji ruchu internetowego bazuje na implementacji sieci klasyfikującej zbiór danych MNIST [4] w bibliotece Lava, a sposób przetwarzania pakietów opiera się na metodach użytych w [10]. Dane wejściowe sieci neuronowej stanowią histogramy reprezentujące przepływ pakietów pogrupowanych według ich rozmiarów (rys. 1).



Rys. 1. Przykładowy histogram dla VoIP przesyłanego tunelem VPN [10]

Jak podkreślają sami autorzy rozwiązania [10], tego typu podejście nie jest niczym nowym i jest bardzo podobne do rozwiązania z [11]. W naszych badaniach natomiast, klasyfikowane histogramy przed podaniem na wejście sieci impulsowej zostają przetworzone na impulsy binarne, ponieważ eksperymenty wykazały, że sieć impulsowa osiąga lepsze wyniki w klasyfikacji dla impulsów binarnych niż dla impulsów stopniowanych. Proces zamiany histogramu na strumienie binarnych impulsów wejściowych jest pokazany na rys. 2.



Rys. 2. Proces zamiany histogramu na strumienie impulsów binarnych (zobrazowanie dla histogramu 5x5)

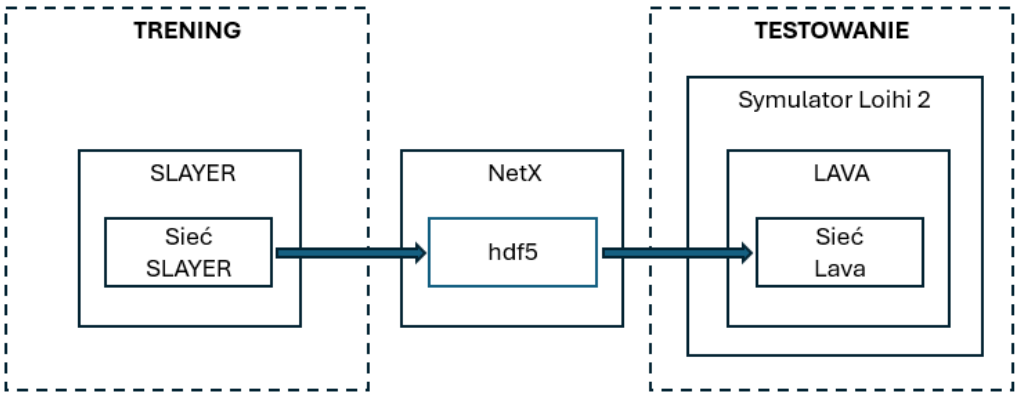
Wartości histogramów poddaje się na początku normalizacji, w której wartości we wszystkich niezerowych pojemnikach zostają zamienione na wartość równą 1, co powoduje, że histogram opisuje jedynie fakt wystąpienia jakiegoś pakietu w określonym binie czasowym i rozmiarowym, a nie konkretną liczbę pakietów znajdujących się w danych slotach lub liczbę zajmowanych bajtów. W ten sposób dla każdego histogramu zarejestrowanego ruchu internetowego otrzymano 300 strumieni binarnych impulsów o długości 300, które następnie są podawane na odpowiednie neurony wejściowe sieci w celu ich klasyfikacji.

W stosunku do rozwiązania [10] znacząco zmniejszono liczbę neuronów w warstwie ukrytej ze 100 do 50 neuronów, a wynika to z faktu, że model sieci nie wykazywał korzyści z posiadania większej liczby neuronów w tej warstwie dla każdego z badanych problemów klasyfikacji. Opisywana implementacja klasyfikatora w języku Lava różni się od [10,11] również pod względem zestawu hiperparametrów. Oprócz standardowych parametrów treningu sieci neuronowej, do których zaliczamy m.in. liczbę neuronów, liczbę epok treningowych i rozmiar serii danych wejściowych, w oprogramowaniu Lava konieczne jest także odpowiednie dobranie wartości parametrów neuronów dla każdej z warstw sieci, takich jak: wartość natężenia progowego, spadek natężenia oraz spadek napięcia. Na bazie wcześniejszych obserwacji treningu sieci ustalono, że wartości progowe neuronów muszą znajdować się w przedziale <0.1; 2> oraz wartość progowa neuronów drugiej warstwy musi być mniejsza niż wartość pierwszej warstwy. Natomiast wartości pozostałych parametrów neuronów ustalono w zakresie <0.01; 0.5>.

Proces poszukiwania najlepszego zestawu parametrów polegał na przeszukiwaniu przestrzeni kilku tysięcy kombinacji wartości parametrów. Następnie sieć trenowano dla każdego wylosowanego zestawu parametrów, a w trakcie tego procesu zapisywano model wykazujący się najlepszymi wynikami. W ten sposób otrzymano zestaw hiperparametrów neuronów poszczególnych warstw, dla której sieć najlepiej klasyfikowała nowe dane pochodzące ze zbioru testowego.

W bibliotece Lava istnieją dwie możliwości treningu impulsowych sieci neuronowych: konwersja ANN do SNN wykorzystując bibliotekę Bootstrap albo trening bezpośredni sieci poprzez bibliotekę SLAYER. W opisywanym rozwiązaniu zdecydowano się na wykorzystanie drugiego rozwiązania, w którym trening poszczególnych modeli sieci został zaimplementowany w SLAYER (rys. 3), a następnie wytrenowane modele wyeksportowano za pomocą API NetX (używając do tego formatu hdf5) do środowiska symulacyjnego Intel Loihi 2, w celu uruchamiania i testowania klasyfikatora.

W celu uruchomienia wytrenowanego modelu sieci impulsowej na symulatorze procesora neuromorficznego, należało także zdefiniować odpowiednie komponenty Lava (procesy, modele procesów), niezbędne do ich poprawnego działania. W związku z tym dodano procesy, odpowiadające za kodowanie histogramów na odpowiadające im impulsy na wejście sieci neuronowej, czy procesy zawierające funkcje do odczytywania predykcji sieci na podstawie impulsów na jej wyjściu. Pozostałe procesy i odpowiadające im modele procesów odpowiadają za adaptacje impulsów wejściowych i wyjściowych z CPU na sieć impulsową w symulatorze.



Rys. 3. Schemat treningu i testowania proponowanej sieci impulsowej

3. WYNIKI BADAŃ

W celu oceny dokładności zaproponowanego modelu impulsowej sieci neuronowej do klasyfikacji ruchu internetowego takiego jak transmisja wideo, VoIP, transfer plików, czat oraz przeglądanie stron WWW użyto dwóch zbiorów danych: ISCX [2,8] w celu porównania wyników z publikacją [10] oraz VPN/Non-VPN Network Application Traffic Dataset (VNAT) [6] w celu dokonania dodatkowej ewaluacji dokładności modelu. Zbiory te następnie podzielono na trzy podzbiory histogramów: 65% – podzbiór treningowy, 15% – podzbiór testowy oraz 20% – podzbiór do badań dokładności modelu. Ze względu na nierównomierną ilość danych w poszczególnych kategoriach ruchu, zadbano o proporcjonalne rozłożenie danych, dzięki czemu procentowy udział danego ruchu w poszczególnych podzbiorach był jednakowy. W tab. 1 przedstawiono wyniki dokładności modelu z publikacji [10] dla zbioru danych ISCX porównane z zaproponowanym modelem sieci zaimplementowanej w Intel Lava. Proponowany klasyfikator osiągnął średnią dokładność na poziomie 97%, w porównaniu do 99,36% uzyskanych przez autorów z pracy [10].

Tab. 1. Porównanie dokładności pomiędzy modelem zaprezentowanym w [10] a implementacją w Intel Lava

Kategoria ruchu	Dokładność [%] modelu z [10]	Dokladność [%] modelu w Lava
Video	99.9	97.4
VoIP	97.8	99.4
Transfer plików	99.9	98.0
Czat	99.7	95.5
WWW	99.5	94.8
średnia	99.4	97.0

Ogólnie rzecz biorąc, model w Lava radził sobie nieco słabiej od tamtego modelu, choć w przypadku ruchu VoIP osiągnął lepszą dokładność, bo na poziomie 99,4%. Klasyfikator najgorzej rozpoznawał ruch związany z przeglądaniem stron internetowych myląc go z czatem, co wpłynęło na dokładność wyników dla tych kategorii.

W kolejnym eksperymencie sprawdzono klasyfikator pod kątem rozpoznawania kategorii ruchu, gdy jest on szyfrowany poprzez wysyłanie danych w tunelu VPN lub poprzez sieć TOR (tab. 2). W tym przypadku model sieci impulsowej w Lava, również radził sobie słabiej. Z eksperymentów wynikało, że sieć impulsowa bardzo często błędnie klasyfikowała tylko metodę zastosowanego szyfrowania, a nie do jakiej kategorii ruch należał. Model sieci najsłabiej poradził sobie z klasyfikacją ruchu związanego z przeglądaniem stron internetowych (70.1%) oraz nieszyfrowanego ruchu VoIP (73.5%). Najlepszą dokładność osiągnął dla video przesyłanego w tunelu VPN (97,5%) oraz niezaszyfrowanego czatu (97.1%).

Tab. 2. Dokładność klasyfikacji dla zbioru testowego ISCX z uwzględnieniem szyfrowania

Kategoria ruchu	Dokładność [%] modelu z [10]	Dokładność [%] modelu w Lava
Video	99,9	82,6
Video (TOR)	100	95,4
Video (VPN)	100	97,5
VoIP	96,7	73,5
VoIP (TOR)	99,9	90,0
VoIP (VPN)	96,7	86,1
Transfer plików	99,9	92,0
Transfer plików (TOR)	99,9	96,2
Transfer plików (VPN)	100	95,8
Czat	99,4	97,1
Czat (TOR)	99,6	96,9
Czat (VPN)	100	96,3
WWW	99,4	70,1
WWW (TOR)	99,7	93,4
średnia	99,4	90,2

Słabe wyniki dla VoIP, gdy klasyfikator miał dodatkowo rozpoznawać czym ruch był szyfrowany, chociaż wcześniej bez tego rozróżnienia średnie wyniki były lepsze skłoniły nas do sprawdzenia, z jaką dokładnością model w Lava będzie klasyfikować ruch internetowy ze zbioru danych VNAT, w którym co prawda nie ma danych dla sieci TOR, czy ruchu związanego z przeglądaniem stron WWW, natomiast sam zbiór jest lepszej jakości, gdyż występują w nim tylko źródła ruchu związane z jedną z pięciu aplikacji: video, VoIP, transfer plików, czat oraz zdalny dostęp (tab. 3). W tym przypadku, gdy klasyfikator miał tylko określić kategorię niezależnie czy ruch byl szyfrowany czy nie, jego średnia dokładność wzrosła do 98.6%, poprawiając wyniki klasyfikacji dla każdej z kategorii (video, VoIP, transfer plików, czat) ze zbioru ISCX. Klasyfikator, również odznaczył się 98,9% dokładnością w rozpoznawaniu ruchu związanego ze zdalnym dostępem.

W kolejnym eksperymencie sprawdzono jak poradzi sobie klasyfikator w rozpoznawaniu kategorii ruchu ze zbioru VNAT, gdy jest on zaszyfrowany (tab. 4). W tym

przypadku, model w Lava ponownie poradził sobie dużo lepiej bo z dokładnością na poziomie 98.9%, względem zbioru danych ISCX (90.2%). Klasyfikator najlepiej rozpoznawał ruch VoIP (99.9%), a najsłabiej niezaszyfrowany strumień video (97%). Model wykazał się także dobrą dokładnością w rozpoznawaniu zdalnego dostępu w tunelu VPN.

Tab. 3. Dokładność klasyfikacji dla zbioru danych VNAT

Kategoria ruchu	Dokładność [%] modelu w Lava
Video	97.9
VoIP	99.9
Transfer plików	98.3
Czat	97.9
Zdalny dostęp	98.9
średnia	98.6

Tab. 4. Dokładność klasyfikacji dla zbioru danych VNAT z uwzględnieniem szyfrowania

Kategoria ruchu	Dokładność [%] modelu w Lava
Video	97.0
Video (VPN)	99.8
VoIP	99.9
VoIP (VPN)	99.9
Transfer plików	98.8
Transfer plików (VPN)	99.1
Czat	98.3
Czat (VPN)	98.6
Zdalny dostęp	99.1
Zdalny dostęp (VPN)	99.4
średnia	98.9

4. PODSUMOWANIE

Wyniki przedstawionych badań potwierdzają, że impulsowe sieci neuronowe stanowią interesującą alternatywę dla metod klasyfikacji źródeł ruchu oraz dobrze sobie radzą w rozpoznawaniu gdy ruch jest zaszyfrowany. Opracowany model w Intel Lava jest wczesną implementacją osiągającą słabsze wyniki dla zbioru ISCX, podczas gdy ze zbiorem VNAT, dokładność klasyfikatora wzrasta. Zbiór ISCX, posiada w swoich plikach ruch niezwiązany z żadną z klasyfikowanych aplikacji, podczas gdy w zbiorze VNAT występują tylko strumienie pakietów generowane przez daną aplikację, co jest jednym z powodów tak dużych różnic w dokładności klasyfikacji pomiędzy zbiorami danych. Dodatkowo, w modelu sieci impulsowej w Lava, nie było możliwości zaimplementowania wszystkich parametrów z [10] oraz same architektury różnią się w modelach synaps oraz neuronów.

Dużym ułatwieniem w przeprowadzeniu powyższych badań było samo środowisko otwartego oprogramowania Intel Lava, które dzięki modularnej architekturze i wysokopoziomowym API w Pythonie, umożliwia elastyczne prototypowanie oraz testowanie modeli sieci impulsowych, które można wygodnie uruchomić w symulatorze procesora Loihi 2.

W przypadku kontynuacji badań, należałoby przetestować działanie zaproponowanego klasyfikatora na innych zbiorach danych, w celu dalszej ewaluacji dokładności modelu oraz podjąć próbę uruchomienia go na rzeczywistym procesorze Loihi 2. Interesującym kierunkiem byłoby także dopracowanie metody dobierania wartości parametrów specyficznych dla architektury Intel Loihi 2, czy ustalenie optymalnych zestawów hiperparametrów w celu poprawy dokładności klasyfikacji ruchu internetowego.

LITERATURA

[1] Akopyan Filipp, i in. (2015). "TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip". *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 34 (10): 1537 – 1557.

[2] Draper-Gil G. i in. 2016. „Characterization of encrypted and VPN traffic using time-related features*". ICISSP 2016 - Proceedings of the 2nd International Conference on Information Systems Security and Privacy*, 407–414.

[3] Furber Steve B., i in. (2014). "The SpiNNaker Project". *Proceedings of the IEEE*, 102 (5): 652 – 665.

[4] Guarav R. "Lava Tutorial: MNIST training on GPU and evaluation on Loihi2". https://r-gaurav.github.io/2024/04/13/Lava-Tutorial-MNIST-Training-on-GPU-and-Evaluation-on-Loihi2.html (data dostępu 26.05.2025).

[5] Intel. „Next-Level Neuromorphic Computing: Intel Lab's Loihi 2 Chip". https://open-neuromorphic.org/neuromorphic-computing/hardware/loihi-2-intel/ (data dostępu 26.05.2025).

[6] Steven J. "VPN/Non-VPN Network Application Traffic Dataset (VNAT)". https://www.ll.mit.edu/r-d/datasets/vpnnonvpn-network-application-traffic-dataset-vnat (data dostępu 26.05.2025).

[7] Kasabov N., i in. 2016. "Evolving spatio-temporal data machines based on the NeuCube neuromorphic framework: Design methodology and selected applications". *Neural Networks* (78): 1-17.

[8] Lashkari A. i in. 2017. „Characterization of tor traffic using time based features". *ICISSP 2017 - Proceedings of the 3nd International Conference on Information Systems Security and Privacy*, 253–262.

[9] Lava architecture. https://lava-nc.org/lava_architecture_overview.html (data dostępu 26.05.2025).

[10] Rasteh A. i in. 2022. „Encrypted internet traffic classification using a supervised Spiking Neural Network". *Neurocomputing,* (503): 272-282.

[11] Shapira T. i in. 2019. „FlowPic: Encrypted Internet Traffic Classification is as Easy as Image Recognition*". INFOCOM 2019 - IEEE Conference on Computer Communications Workshops*, 680–687.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

OPTYMALIZACJA SIECI DOSTĘPOWYCH PRZY UŻYCIU ALGORYTMÓW EWOLUCYJNYCH

ACCESS NETWORK OPTIMIZATION WITH EVOLUTIONARY ALGORITHMS

Mateusz Burczaniuk[1]; Tomasz Kuryłowicz[1]; Michał Ostapowicz[1]; Hoang Nghia Le[1]

[1] Suntech S. A. Warszawa

DOI: 10.15199/59.2025.4.57

Streszczenie: **Optymalizacja sieci dostępowych to ważne zagadnienie dla firm telekomunikacyjnych. Często jest wykonywana ręcznie przez ekspertów. Nieliczne mechanizmy automatyczne potrafią zaplanować wszystkie wymagane zasoby. Artykuł prezentuje rozwiązanie problemu przy użyciu algorytmów: genetycznego (GA), optymalizacji rojem cząstek (PSO), kukułki (CS) i pszczelego (ABC). Umożliwiają one zaprojektowanie całej sieci, w tym: kabli, muf, studzienek, duktów, połączeń kabli, itd. Jakość wyników potwierdzono na realistycznych przypadkach testowych.**

Abstract: **Access network optimization is a crucial task for telecoms. It is often performed manually by teams of professionals, and only a few automated mechanisms can plan all necessary resources. This paper presents novel mechanisms to address the problem using genetic algorithm (GA), particle swarm optimization (PSO), cuckoo search (CS) and artificial bee colony (ABC). They enable designing an entire telecommunications network with all necessary resources: cables, splices, manholes, ducts, cable joints, etc. Their quality and validity were confirmed through realistic test cases.**

Słowa kluczowe: **algorytm genetyczny, algorytmy stadne, optymalizacja sieci, planowanie sieci dostępowych, PSO.**

Keywords: **access network planning, genetic algorithm, network optimization, PSO, swarm algorithms.**

1. WSTĘP

Przedmiotem niniejszego artykułu jest analiza zastosowania algorytmów ewolucyjnych w optymalizacji sieci dostępowych. Badanie obejmuje zarówno segment dostępowy sieci stacjonarnej w technologii FTTH, jak i podsieci pomiędzy centralami a antenami w sieci 5G.

Planowanie sieci dostępowej jest złożonym procesem, bardzo istotnym dla operatorów telekomunikacyjnych. Zazwyczaj przebiega w kilku etapach. Pierwszym krokiem jest dekompozycja sieci, która dzieli dużą sieć na mniejsze podsieci łatwiejsze do analizy. Następnie, przeprowadzana jest właściwa optymalizacja, która ma na celu zaplanowanie zasobów sieci telekomunikacyjnej dla każdej z tych podsieci. Celem zadania optymalizacji jest podłączenie punktów docelowych, które są ONU (FTTH) lub antenami (5G), z punktami źródłowymi które są OLT (FTTH) lub centralami (5G). Planowane sieci powinny być optymalne kosztowo, bezpieczne i niezawodne.

Problem można by uprościć do znajdowania najkrótszych ścieżek w grafie. Jednak rzeczywiste przypadki są bardziej złożone. Rozwiązanie problemu powinno uwzględniać różne typy planowanych elementów sieciowych, pojemności duktów. Preferowane jest korzystanie z istniejącej infrastruktury, jednak często trzeba budować nową lub dzierżawić ją od innych operatorów. W związku ze złożonością problemu, zazwyczaj rozwiązywany jest on ręcznie przez ekspertów.

Artykuł prezentuje nowe podejście do konstrukcji ścieżek kablowych oparte na wektorach kierunkowych ustalanych indywidualnie dla każdego punktu docelowego. Analizowane są cztery algorytmy. Oprócz algorytmów genetycznego, kukułki i optymalizacji rojem cząstek badany jest nowy w tym kontekście algorytm pszczeli. Należy podkreślić, że używany model danych zawiera wprost obiekty warstwy fizycznej: wielopoziomowe dukty/mikrodukty, pojemności kabli, studzienki, mufy, połączenia kabli i pozwala na tworzenie nowych duktów. Ten szczegółowy model idzie dalej niż powszechne w literaturze proste, abstrakcyjne rozważania, które zazwyczaj nie uwzględniają hierarchii duktów czy łączenia kabli.

Praca jest zorganizowana następująco. Sekcja 2 wprowadza czytelnika do tematu optymalizacji sieci. Sekcja 3 zawiera przegląd literatury na badany temat. Sekcja 4 przedstawia zaproponowane rozwiązanie problemu. Sekcja 5 przedstawia natomiast, jak dostosowane zostały używane algorytmy ewolucyjne i stadne. W Sekcji 6 znajduje się przedstawienie uzyskanych wyników, a Sekcja 7 podsumowuje badania.

2. OPTYMALIZACJA SIECI

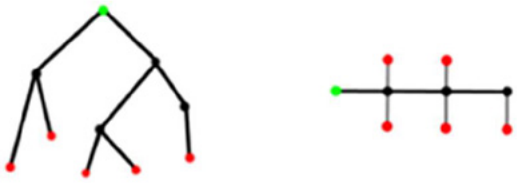
Zadanie optymalizacji polega na łączeniu punktów źródłowych i docelowych za pomocą ścieżek kablowych. Każdy punkt docelowy ma własne zapotrzebowanie określające, ile włókien należy do niego doprowadzić.

Kable są prowadzone w duktach. Dukty mogą istnieć już wcześniej albo być utworzone w ramach zadania. Każdy dukt ma własną pojemność i może pomieścić ilość kabli zależną od konkretnych modeli: kabla i duktu. Wielu dostawców stosuje kilkupoziomową strukturę duktów, gdzie dukty zawierają mikrodukty, a dopiero do nich wkładane są kable. Prezentowane podejście obsługuje oba przypadki w podobny sposób, więc dukty i mikrodukty będą określane tutaj łącznie jako dukty. Wewnątrz muf czy studzienek kable mogą być dzielone, a ich włókna dzielone lub łączone zgodnie z zapotrzebowaniem. Sposób łączenia włókien tych kabli należy dokładnie określić.

Celem zadania jest zaprojektowanie takich kabli, aby wszystkie zapotrzebowania punktów docelowych zo-

stały spełnione. W razie potrzeby możliwe jest dotworzenie nowych duktów czy studzienek. Pełne rozwiązanie problemu powinno zawierać informacje o wszystkich zasobach planowanych do utworzenia: kablach, duktach, mufach, studzienkach, ich położeniu, wzorcach, a także zawierać opis, którymi duktami prowadzone są kable i jak włókna kabli są łączone ze sobą.

W ramach prezentowanego rozwiązania kable są planowane zgodnie z jedną z rozważanych topologii: drzewa lub szyny. W topologii drzewa kable wynikowe tworzą las w rozumieniu teorii grafów, czyli mają formę jednego lub wielu drzew. W topologii szyny najważniejszymi kablami są szyny, które pobierają sygnał od punktów źródłowych i do nich podłączane są kable dostarczające sygnał do konkretnych punktów docelowych. Topologie te nie zapewniają odporności na awarie wewnątrz sieci. Do zapewnienia niezawodności w tym zakresie potrzebne byłoby użycie innych topologii, np. pierścienia.



Rys. 1. Przykłady topologii drzewa (z lewej) i szyny (z prawej). Kolorem zielonym oznaczono punkty źródłowe, a czerwonym docelowe.

3. PRZEGLĄD LITERATURY

Temat planowania i optymalizacji sieci telekomunikacyjnych jest szeroko znany w literaturze. Penttinen [16] podsumowuje sposoby planowania sieci 4G i podkreśla nowe aspekty specyficzne dla sieci 5G, jak elastyczność, niezawodność i bezpieczeństwo. Celli et al. [3] przedstawili działanie algorytmu genetycznego (GA) w rozwiązywaniu problemów NP-trudnych związanych z topologią sieci, planowaniem pojemności i minimalizacją kosztu.

Jahromi i Rad [8] zaproponowali bazujące na GA podejście do optymalizacji topologicznej sieci komunikacyjnych systemu elektroenergetycznego, spełniające rygorystyczne kryteria jakości usług. Hurley [7] wprowadził framework oparty na metodzie symulowanego wyżarzania uwzględniający przy optymalizacji realistyczne dane o tłumieniu sygnału i ukształtowaniu terenu. Luna et al. [12] dokonali przeglądu zastosowań algorytmów ewolucyjnych w automatycznym planowaniu sieci komórkowych.

Dreifuerst et al. [5] pokazali, że głębokie uczenie ze wzmocnieniem i optymalizacja Bayesowska przewyższają tradycyjne metody w pokrywaniu sieci i planowaniu pojemności. Ostatnie badania [2, 14, 20] zastosowały optymalizację rojem cząstek (PSO) do optymalnego rozmieszczania punktów dostępowych, minimalizując straty propagacyjne. Tychogiorgos i Leung [21] rozszerzyli metodę Network Utility Maximizaton (NUM) do problemów niewypukłych, proponując sprawiedliwą i efektywną alokację zasobów w sieciach komunikacyjnych. Wu i Zeng [23] użyli PSO do bazującego na SDN-ach (sieci definiowane programowo) dynamicznego zarządzania zasobami sieci 5G, poprawiając ich przepustowość i redukując przeciążenia.

Wiele prac badało optymalizowanie topologii sieci, aby efektywnie alokować zasoby. Przykładowo, Bauknecht [1] użył GA do projektowania wirtualnej topologii w wielowarstwowych sieciach, a Hildmann et al. [6] stosowali PSO do optymalizacji topologii sieci bezprzewodowej. Efektywne użycie zasobów przekłada się nie wprost na redukcję kosztów. Na przykład, Wnuk i Kozdrowski [22] skupili się na minimalizowaniu zasobów optycznych takich jak transpondery i wzmacniacze. Shakya et al. [18] stosowali GA do optymalizacji struktur drzewiastych w systemach wewnątrzbudynkowych (IBDAS).

Singh [19] pokazał, że algorytm kukułki jest lepszy od PSO w optymalizacji sieci komórkowej, szczególnie pod kątem pobranej mocy i straty/osłabienia sygnału. Kozdrowski et al. [11] pokazali przewagę algorytmów inspirowanych naturą (algorytmy ewolucyjny, pszczeli i przeszukiwania harmonicznego (ang. Harmony search) nad podejściami deterministycznymi przy projektowaniu dużych sieci optycznych. De Oliveira et al. [15] zaprezentowali Hybrydowy Świetlikowo-Genetyczny Algorytm pokonujący zarówno algorytm genetyczny, jak i świetlika przy projektowaniu sieci optycznych.

4. ZAPROPONOWANE PODEJŚCIE

Artykuł prezentuje cztery algorytmy: genetyczny (GA), optymalizacji rojem cząstek (PSO), kukułki (CS) i pszczeli (ABC) dopasowane do rozważanego problemu. Głównymi optymalizowanymi elementami są: pozycje nowo tworzonych studzienek m_i oraz wektory kierunkowe optymalizacji dla punktów docelowych d(v).

Optymalizacja przebiega w zbliżony sposób dla wszystkich tych algorytmów. W pierwszym kroku położenia m_i i wektory d(v) są inicjalizowane losowo i tworzone są początkowe rozwiązania problemu. Następnie, przez określoną liczbę iteracji konstruowane są nowe rozwiązania (więcej w następnym akapicie). Za wynik algorytmu przyjmuje się najlepsze rozwiązanie znalezione podczas całego procesu.

Konstruowanie pojedynczego rozwiązania zaczyna się od modyfikacji lokalizacji m_i oraz wektorów kierunkowych d(v) zgodnie z wybranym algorytmem ewolucyjnym. Następnie tworzone są ścieżki składające się na rozwiązanie. Odbywa się to przyrostowo przez podłączanie kolejnych punktów do projektowanej sieci. Punkty są podłączane w losowej kolejności do już utworzonego drzewa lub szyny, które są wtedy odpowiednio rozszerzane. W razie, gdy nie powstało jeszcze drzewo lub odpowiednia szyna, lub podłączenie do nich jest nieoptymalne, punkt może być podłączony do nowego drzewa lub szyny. Każdy punkt v jest podłączany do sieci najkrótszą ścieżką zgodnie z opisaną poniżej metryką. Ścieżka ta jest konstruowana od punktu v. Kolejne krawędzie są do niej dodawane przy użyciu metryki długości d'(e):

$$d'(e) = c_e\, k\, (1 + \cos \alpha) \qquad (1)$$

gdzie c_e to koszt krawędzi e, k to współczynnik większy od 0, natomiast α to kąt między wektorem kierunkowym a krawędzią. Jeśli k = ½, a $\alpha = 0$, to to d'(e) przyjmuje wartość c_e. Jeśli natomiast $\alpha = \pi$, to d'(e) przyjmuje

wartość 0. Zatem metryka d'(e) powoduje wytłumienie kosztu, gdy wektor kierunkowy jest przeciwny do kierunku krawędzi i względne wzmocnienie, gdy są one zgodne. Metryka ta jest używana wyłącznie podczas konstruowania rozwiązania i nie ma wpływu na finalną wartość kosztu rozwiązania. Podczas eksperymentów współczynnik k przyjmował wartość ¼.

Podczas wyznaczania ścieżek kabli uwzględniane są dukty, do których będą wkładane i ich pojemności. Rozważane są dukty podłączające nowe studzienki do sieci, nowe dukty łączące wybrane punkty ze sobą, istniejące dukty i potencjalne nowe dukty naśladujące te istniejące.

5. UŻYTE ALGORYTMY EWOLUCYJNE

W rozwiązaniu wykorzystywane są algorytmy: genetyczny (GA), kukułki (CS), optymalizacji rojem cząstek (PSO) i pszczeli (ABC). Są to algorytmy populacyjne, przetwarzające pewną populację rozwiązań przez wiele iteracji. W tym punkcie zostaną omówione ich implementacje. Na potrzeby tego rozdziału lokalizacje nowo tworzonych studzienek m_i oraz wektory kierunkowe dla punktów docelowych d(v) będą łącznie określane jako współrzędne rozwiązania.

5.1. Algorytm genetyczny

Algorytm genetyczny składa się z czterech głównych elementów: selekcji, mutacji, krzyżowania i sukcesji [17]. W używanej implementacji stosowana jest selekcja turniejowa i sukcesja z pełnym zastępowaniem starej populacji przez nową. Mutacja z pewnym prawdopodobieństwem modyfikuje współrzędne rozwiązania. Używa do tego funkcji częściowo wzorowanej na Δ(t,y) przedstawionej przez Michalewicza [13]. Stosowane krzyżowanie polega na losowej wymianie pomiędzy rozwiązaniami współrzędnych rozwiązań.

5.2. Optymalizacja rojem cząstek

W optymalizacja rojem cząstek (PSO) populacja składa się z cząstek. Każda cząstka ma swoje położenie i prędkość [10]. Położenie cząstki określa pojedyncze rozwiązanie problemu. W omawianej implementacji PSO rolę położenia cząstki pełnią współrzędne rozwiązania. Są one w każdej iteracji przesuwane zgodnie z prędkością w każdym z tych kierunków. Prędkość jest modyfikowana w oparciu o wzór:

$$v' = \omega v + c_1 r_1 (x_{best} - x) + c_2 r_2 (x_{lbest} - x) \quad (2)$$

gdzie v to poprzednia prędkość cząstki, ω to parametr inercji, r_1,r_2 to liczby losowe z przedziału [0,1], c_1 i c_2 to odpowiednio parametry wpływu globalnego określającego wpływ najlepszego ogólnie znanego rozwiązania oraz lokalnego określającego wpływ najlepszego rozwiązania uzyskanego w przeszłości przez cząstkę. x_{best} to współrzędne najlepszego znalezionego do tej pory rozwiązania, x_{lbest} – najlepszego rozwiązania znalezionego przez daną cząstkę, a x – współrzędne aktualnego rozwiązania.

5.3. Algorytm kukułki

W algorytmie kukułki [24] mamy zbiór analizowanych rozwiązań. W każdej iteracji dla każdego rozwiązania wykonywane są loty Lévy'ego, które modyfikują współrzędne rozwiązania. Loty Lévy'ego polegają na błądzeniu losowym, którego krok jest losowany na podstawie rozkładu Lévy'ego. W niniejszej pracy jest to dokonywane na podstawie wzoru zaczerpniętego od Gutowskiego [5]:

$$\Delta x = r^{-\frac{1}{c}} - 1 \quad (3)$$

gdzie r to liczba losowa z przedziału [0,1], a c to współczynnik lotów Lévy'ego.

Powstaje w ten sposób nowe rozwiązanie, które zastępuje poprzednie pod warunkiem, że jest od niego lepsze. Ponadto, ustalony parametrem procent najgorszych rozwiązań jest odrzucany i zastępowany przez nowe losowe rozwiązania.

5.4. Algorytm pszczeli

W algorytmie pszczelim (ABC) [9] w każdej iteracji dla każdego osobnika populacji a losowany jest inny osobnik b i współrzędne rozwiązania a są modyfikowane losowo w stronę wartości osobnika b lub w stronę przeciwną. Tak powstaje nowe rozwiązanie, które zastępuje starsze, jeśli jest od niego lepsze. Jeśli przez k kolejnych iteracji nie doszło do poprawy danego rozwiązania, to jest ono zastępowane przez nowe losowe rozwiązanie.

6. WYNIKI

Analiza rezultatów została podzielona na kilka etapów. W pierwszej kolejności zbadano hiperparametry dla poszczególnych algorytmów, następnie zestawiono i porównano ich wyniki ze sobą, a na samym końcu zaprezentowane zostało jedno z uzyskanych rozwiązań.

Wyniki porównywano na podstawie kosztu urządzeń, kabli i duktów. Koszt ten był szacowany jako suma sumy długości planowanych kabli oraz sumy długości planowanych duktów pomnożonej przez wybrany mnożnik. Założono, że koszt obiektów punktowych typu mufy czy studzienki zawiera się w koszcie planowanych obiektów liniowych. Pominięte zostały koszty stałe, np. obsługi prawnej. Koszt rozwiązań jest wyrażony w sztucznej jednostce kosztu [j].

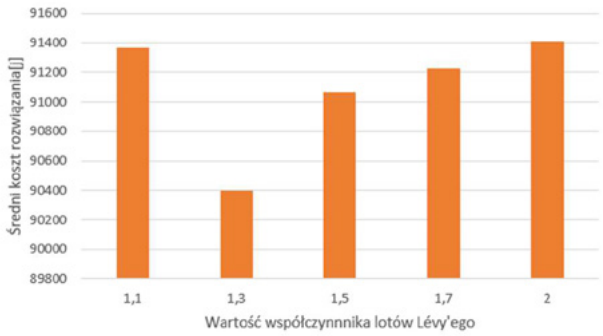
6.1. Dobór hiperparametrów

Wartości większości hiperparametrów ustalono na podstawie literatury przedmiotu oraz własnych doświadczeń autorów. Część z nich dobrano na podstawie eksperymentów. Zbadano 3 różne hiperparametry dla GA oraz PSO, 2 dla CS i 1 dla ABC. Dla każdej kombinacji hiperparametrów przeprowadzono po jednym uruchomieniu dla każdego z 3 analizowanych przykładów testowych przy 2 wariantach topologii, jedynie dla ABC przeprowadzono więcej badań ze względu na badanie tylko jednego parametru. Przykłady testowe wygenerowano na podstawie układu dróg prawdziwych polskich miejscowości: miasta Bielsk Podlaski, dzielnicy Piekar Śląskich Kozłowa Góra, wsi Wyględów (oznaczane kolejno jako BP, KG i W), by odwzorowywały sytuacje rzeczywiste. Miały one odpowiednio 318, 545 i 242 punkty docelowe, przy czym mogły się różnić kształtem czy gęstością, co miało przełożenie na wyniki. W tej fazie testów dla GA, PSO i CS liczba iteracji i liczebność populacji wynosiły po 5 dla przykładów BP oraz W i po 10 dla przykładu .KG. Z kolei

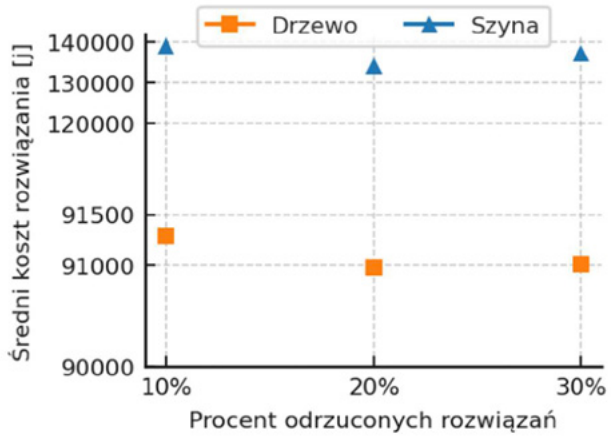
dla ABC wykorzystano liczby iteracji: 18 i 20 oraz populacje liczebności 12, 15, 20. Za lepsze uznano te hiperparametry, dla których uzyskano mniejsze koszty, przy czym każdy hiperparametr analizowano osobno.

Dla GA przetestowano prawdopodobieństwa mutacji i krzyżowania oraz rozmiar turnieju. Najlepszymi wartościami dla drzewa były odpowiednio (0.1, 0.5, 2), a dla szyny (0.005, 0.1, 2). Dla PSO badano parametry inercji, wpływu lokalnego i globalnego. Wybrano wartości odpowiednio: dla drzewa (0.8, 2.2, 2) i dla szyny (0.8, 1.5, 1.8). Dla CS sprawdzono wartości współczynnika lotów Lévy'ego oraz procent odrzucanych najgorszych rozwiązań. Najlepszy wyniki uzyskano dla wartości odpowiednio: (1.3, 0.2) dla drzewa i (2, 0.2) dla szyny. Dla ABC zweryfikowano jedynie maksymalną liczby iteracji bez poprawy rozwiązania. Najlepszą jego wartością dla szyny było 4, a dla drzewa 8.

W trakcie analizy wyników dokonano kilku interesujących obserwacji. Prawdopodobnie znaleziono ekstremum współczynnika lotów Lévy'ego dla CS z topologią drzewa. Chociaż brakuje gwarancji, że jest to minimum globalne, to *Rys. 2* pokazuje spadek kosztu rozwiązania wokół pewnej wartości. Zauważono także, że w przypadku tego algorytmu najlepsze wyniki są uzyskiwane, gdy procent odrzucanych najgorszych rozwiązań jest w okolicy 20%. Zachowanie to wyglądało bardzo podobnie dla topologii szyny i drzewa, co przedstawiono na Rys. 3.



Rys. 2. Średni koszt rozwiązania w zależności od współczynnika lotów Lévy'ego dla CS z topologią drzewa



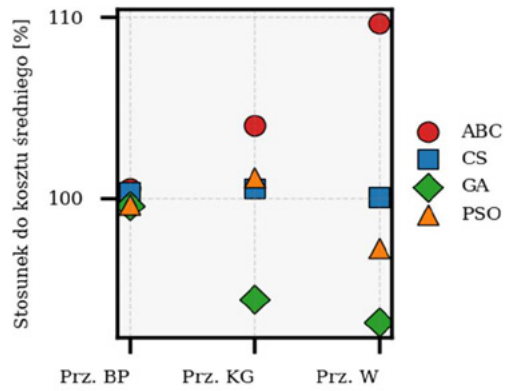
Rys. 3. Średni koszt rozwiązania w zależności od procenta odrzucanych rozwiązań dla CS

6.2. Porównanie algorytmów

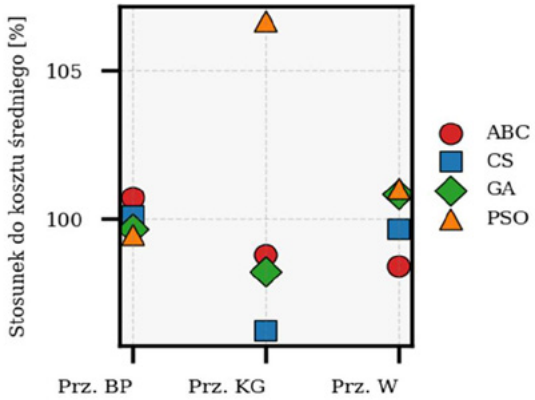
Następnie, przeprowadzono testy porównawcze dla badanych algorytmów z ustalonymi w poprzednim punkcie hiperparametrami. Wykonano je z różnymi liczbami iteracji (18, 20) i liczebnościami populacji (12, 15, 20). Średnie koszty uzyskanych w ten sposób wyników wraz z odchyleniami standardowymi przedstawiono w Tab. 1. Stosunek tych kosztów do ich średniej arytmetycznej dla danego przykładu przedstawiono na Rys. 4. i Rys. 5.

Tab. 1. Średnie wyniki oraz odchylenia standardowe dla topologii szyny i drzewa.

Algorytm	*Szyna*		*Drzewo*	
	Śr. wynik [j]	Odch. st. [j]	Śr. wynik [j]	Odch. st. [j]
		Przykład BP		
ABC	*224 636,0*	10701	*159 908,8*	2264
CS	*223 220,9*	9159	*159 619,3*	1716
GA	*222 210,2*	4495	*158 445,2*	2300
PSO	*221 810,0*	14910	*158 483,4*	1223
		Przykład KG		
ABC	*132 783,4*	12187	*79 472,4*	888
CS	*129 366,0*	10693	*76 824,3*	6998
GA	*132 027,9*	14732	*72 138,3*	7425
PSO	*143 295,8*	6157	*77 293,9*	988
		Przykład W		
ABC	*28 492,7*	556	*30 611,5*	203
CS	*28 855,5*	1135	*27 946,5*	676
GA	*29 191,8*	1189	*26 016,2*	2544
PSO	*29 242,0*	1135	*27 147,9*	553



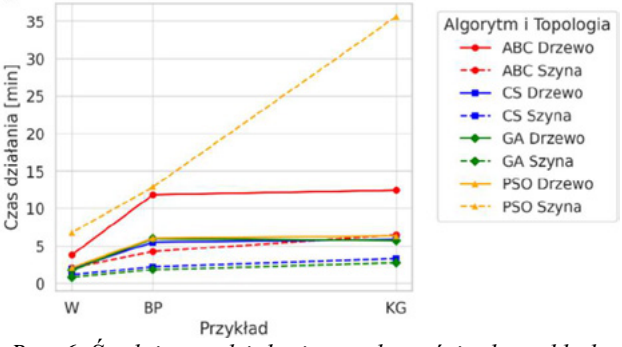
*Rys. 4. Średni koszt rozwiązania dla danego algorytmu względem średniej dla topologii **drzewa***



*Rys. 5. Średni koszt rozwiązania dla danego algorytmu względem średniej dla topologii **szyny***

W przypadku topologii szyny trudno jest jednoznacznie wybrać najlepszy algorytm. W zależności od przykładu wygrywały 3 różne algorytmy, a GA również prezentował się dobrze. Jedyny gorszy wynik uzyskało na Przykładzie KG PSO. W przypadku topologii drzewa wyniki są bardziej klarowne. Najlepszy jest algorytm genetyczny, a najgorszy pszczeli, który dla jednego przykładu jako jedyny znalazł dla topologii drzewa gorsze rozwiązanie niż dla topologii szyny.

Rys. 6 przedstawia zależność czasu wykonywania zadania od liczby punktów docelowych w przykładzie. Wykres ten pokazuje, że PSO działa zauważalnie wolniej w przypadku topologii szyny. Nie licząc tego przypadku, algorytm ABC jest zauważalnie wolniejszy od pozostałych.



Rys. 6. Średni czas działania w zależności od przykładu i topologii

6.3. Przykładowe rozwiązanie

Ze względu na stopień złożoności zadania, najlepszą metodą potwierdzenia wysokiej jakości wyników jest zeprezentowanie konkretnego przypadku. Aby lepiej przedstawić rezultat, została wykorzystana warstwa podkładu mapowego dostarczona przez OpenStreetMap [25].

Rys. 7. przedstawia przykładowe rozwiązanie dla Przykładu W z poprzedniego rozdziału, otrzymane za pomocą GA, przy wybranej topologii drzewa z limitem 100 nowo utworzonych studzienek. Punkty docelowe są niebieskie, a punkty źródłowe są zielone. Projektowane kable układają się w cztery drzewa oznaczone różnymi kolorami. Rozwiązanie to jest prawidłowo skonstruowane, określony jest ponadto dla niego szczegółowo sposób łączenia kabli i umieszczania ich w duktach.

7. WNIOSKI

Prezentowane podejście pozwala na zaprojektowanie poprawnej i efektywnej kosztowo sieci. Umożliwia zaprojektowanie wszystkich potrzebnych zasobów. Badane metody skracają czas potrzebny na rozwiązanie analizowanych problemów i zwalniają projektantów sieci z potrzeby wykonywania ręcznie czasochłonnych operacji. Szczególnie ciekawe wydają się algorytmy: genetyczny oraz kukułki.

W przyszłości praca może być kontynuowana. Optymalizacja sieci może rozpatrywać kolejne topologie sieci. Można pogłębić dopasowanie metody do problemu, wykorzystać dodatkowe optymalizacje lokalne.

Rys. 7. Jedno z rozwiązań dla Przykładu W, uzyskane przy użyciu GA dla topologii drzewa

LITERATURA

[1] Bauknecht, Uwe. 2021. „A Genetic Algorithm Approach to Virtual Topology Design for Multi-Layer Communication Networks". GECCO '21, 928–936. Association for Computing Machinery.

[2] Brahma, Debolina, Arindam Deb, i Jibendu Sekhar Roy. 2021. „Optimal Antenna Placement in a Distributed Antenna System using RSSI Optimization for Indoor-to-Outdoor Propagation". 2021 IEEE 2nd International Conference on Applied Electromagnetics, Signal Processing, & Communication (AESPC), 1–4.

[3] Celli, G., E. Costamagna, i A. Fanni. 1995. „Genetic Algorithms for Telecommunication Network Optimization". 1995 IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century, 2:1227–1232.

[4] Dreifuerst, Ryan M., Samuel Daulton, Yuchen Qian, Paul Varkey, Maximilian Balandat, Sanjay Kasturia, Anoop Tomar, Ali Yazdan, Vish Ponnampalam, i Robert W. Heath. 2021. „Optimizing Coverage and Capacity in Cellular Networks Using Machine Learning". ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings 2021 (June): 8138–42.

[5] Gutowski, Marek, 2001. „Lévy flights as an underlying mechanism for global optimization algorithms", https://doi.org/10.48550/arXiv.math-ph/0106003

[6] Hildmann, Hanno, Dina Y. Atia, Dymitr Ruta, S. S. Khrais, i Abdel F. Isakovic. 2018. „A Model for Wireless-Access Network Topology and a PSO-Based Approach for Its Optimization". WCO@FedCSIS.

[7] Hurley, Stephen. 2002. „Planning Effective Cellular Mobile Radio Networks". Vehicular Technology, IEEE Transactions on 51 (April): 243–53.

[8] Jahromi, Abdolhamid Eshraghniaye, i Zohreh Besharati Rad. 2013. „Optimal Topological Design of Power

Communication Networks Using Genetic Algorithm". Scientia Iranica 20 (3): 945–57.
[9] Karaboga, Dervis, i Bahriye Basturk. 2007. „Artificial Bee Colony (ABC) Optimization Algorithm for Solving Constrained Optimization Problems". W Foundations of Fuzzy Logic and Soft Computing, red. Patricia Melin et al., 789–98. Springer.
[10] Kennedy, J., i R. Eberhart. 1995. „Particle Swarm Optimization". Proceedings of ICNN'95 - International Conference on Neural Networks 4:1942–48.
[11] Kozdrowski, S., Mateusz Żotkiewicz, Kacper Wnuk, Arkadiusz Sikorski, i Sławomir Sujecki. 2020. „A Comparative Evaluation of Nature Inspired Algorithms for Telecommunication Network Design". Applied Sciences 10 (19): 6840.
[12] Luna, Francisco, Juan Durillo, Antonio Nebro, i Enrique Alba. 2010. „Evolutionary Algorithms for Solving the Automatic Cell Planning Problem: A Survey". Engineering Optimization 42 (7): 671–90.
[13] Michalewicz, Zbigniew. 1996. Genetic Algorithms + Data Structures = Evolution Programs. Springer Berlin Heidelberg
[14] Nascimento, Antonio I. S., i Carmelo J. A. Bastos-Filho. 2010. „A Particle Swarm Optimization Based Approach for the Maximum Coverage Problem in Cellular Base Stations Positioning". 2010 10th International Conference on Hybrid Intelligent Systems, 91–96.
[15] Oliveira, Bruno Quirino de, Marcos Sousa, i Flávio Henrique Vieira. 2020. „A Hybrid Firefly-Genetic Algorithm for Planning of Optical Transport Networks". JCIS 35 (November): 243–55.
[16] Penttinen, Jyrki T. J. 2019. „9 5G Network Planning and Optimization 9.1 Overview".
[17] Sastry, Kumara, David Goldberg, i Graham Kendall. 2005. „Genetic Algorithms". W Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques, red. Edmund K. Burke i Graham Kendall, 97–125. Springer US.
[18] Shakya, Siddhartha, Kin Poon, i Anis Ouali. 2018. „A GA Based Network Optimization Tool for Passive In-Building Distributed Antenna Systems". Proceedings of the Genetic and Evolutionary Computation Conference, 1371–78.
[19] Singh, Sarabjeet. 2017. „Base Station Localization and Optimized Placement Using Cuckoo Search (CSA) Algorithm". https://api.semanticscholar.org/CorpusID:214753528.
[20] Tsilimantos, D., D. Kaklamani, i G. Tsoulos. 2008. „Particle Swarm Optimization for UMTS WCDMA Network Planning". 2008 3rd International Symposium on Wireless Pervasive Computing, 283–87.
[21] Tychogiorgos, G., i K. K. Leung. 2014. „Optimization-Based Resource Allocation in Communication Networks". Computer Networks 66: 32–45.
[22] Wnuk, Kacper i Stanislaw Kozdrowski. 2021. „Multi-Objective Optimization in Optical Networks". SoftCOM 2021, 1–3.
[23] Wu, Guanqaun, i Desheng Zeng. 2022. „Optimization of Network Resource Management Based on Software-Defined Networking in the 5G Environment". Journal of Supercomputing 78 (15): 16721–44.
[24] Yang, Xin-She, i Suash Deb. 2009. „Cuckoo Search via Lévy Flights". 2009 World Congress on Nature & Biologically Inspired Computing (NaBIC), 210–14.
[25] OpenStreetMap contributors. "Planet dump retrieved from https://planet.osm.org", https://www.openstreetmap.org, 2025.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

UCZENIE REPREZENTACJI SIECI IEEE 802.11 Z WYKORZYSTANIEM GRAFOWYCH SIECI NEURONOWYCH

REPRESENTATION LEARNING OF IEEE 802.11 NETWORKS USING GRAPH NEURAL NETWORKS

Maksymilian Wojnar

AGH Akademia Górniczo-Hutnicza, Kraków, maksymilian.wojnar@agh.edu.pl

DOI: 10.15199/59.2025.4.58

Streszczenie: Wraz ze wzrostem gęstości sieci Wi-Fi oraz rozwojem zaawansowanych mechanizmów koegzystencji, takich jak Multi-Access Point Coordination (MAPC), rośnie zapotrzebowanie na skuteczne metody modelowania topologii sieci. W niniejszej pracy zaproponowano wykorzystanie grafowych sieci neuronowych do uczenia reprezentacji sieci IEEE 802.11, umożliwiając ich dalszą analizę, wizualizację oraz zastosowanie w planowaniu transmisji MAPC.

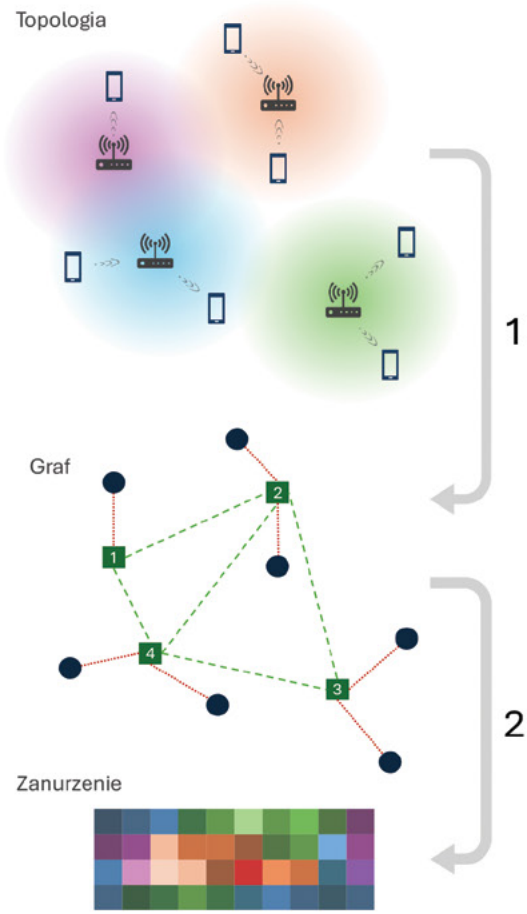
Abstract: Increasing the density of Wi-Fi networks and developing advanced coexistence mechanisms, such as Multi-Access Point Coordination (MAPC), requires effective network topology modeling methods. This paper proposes the use of graph neural networks to learn representations of IEEE 802.11 networks, enabling their further analysis, visualization and application in MAPC transmission planning.

Słowa kluczowe: grafowe sieci neuronowe, IEEE 802.11, uczenie reprezentacji, wizualizacja

Keywords: graph neural networks, IEEE 802.11, representation learning, visualization

1. WSTĘP

Wraz ze wzrostem gęstości wdrożeń sieci Wi-Fi, kolejne wersje standardu IEEE 802.11 wprowadzają mechanizmy mające na celu poprawę koegzystencji wielu urządzeń w ograniczonej przestrzeni. Istotnym kierunkiem rozwoju jest optymalizacja pracy wielu stacji w ramach jednej sieci, czego przykładami są: kształtowanie wiązki (ang. *beamforming*) w standardzie 802.11n, *multi-user multiple-input and multiple-output* (MU-MIMO) w 802.11ac czy *orthogonal frequency-division multiple access* (OFDMA) w 802.11ax. Nowsze standardy koncentrują się także na współdzieleniu kanału między sąsiadującymi sieciami, m.in. przez kolorowanie ramek w 802.11ax lub planowane w 802.11bn multi-access point coordination (MAPC) [8].



Rysunek 1: Konwersja topologii sieci na graf (1) oraz wygenerowanie zanurzenia sieci o danym rozmiarze z grafu (2). Wierzchołki grafu to stacje (koła) i AP (kwadraty), krawędzie to asocjacje (czerwone) oraz relacje między AP (zielone). Krawędzie zawierają atrybuty, przykładowo moc odebranego sygnału.

MAPC to nowy mechanizm umożliwiający koordynację działań wielu punktów dostępowych (ang. *access points*, AP), przy zachowaniu rozproszonego dostępu do kanału. Celem rozwiązania jest umożliwienie równoległych transmisji w sąsiadujących sieciach z minimalną interferencją. MAPC oferuje istotny potencjał poprawy wydajności przez szerokie możliwości

konfiguracji, jednak wiąże się z trudnym problemem optymalizacyjnym: oprócz wyboru par stacja-AP, konieczne jest równoczesne ustalenie modulacji i schemat kodowania (ang. *modulation and coding scheme*, MCS) oraz mocy nadawczej, aby zapewnić niezawodność połączeń i ograniczyć interferencje.

Aby skutecznie rozwiązywać problemy o wysokiej złożoności, często stosuje się techniki uczenia maszynowego (ang. *machine learning*, ML), które zyskały istotne znaczenie w literaturze dotyczącej optymalizacji sieci bezprzewodowych [5]. Zainteresowanie to znajduje także odzwierciedlenie w działaniach grupy standaryzacyjnej IEEE 802.11, czego dowodem jest powołanie AIML Topic Interest Group. Jednym z omawianych scenariuszy zastosowania ML w nadchodzących standardach jest planowanie transmisji w ramach mechanizmu MAPC [6]. Wśród proponowanych podejść pojawia się zastosowanie hierarchicznych algorytmów wielorękich bandytów (ang. hierarchical multi-armed bandits, H-MAB) [6, 7]. Możliwe jest jednak wykorzystanie bardziej zaawansowanych modeli, takich jak generatywne sieci neuronowe, które umożliwiają próbkowanie złożonych rozkładów wielowymiarowych z wysoką precyzją.

Efektywnie zastosowanie tego typu modeli wymaga uzyskania odpowiednich reprezentacji analizowanych obiektów. W kontekście sieci 802.11 oznacza to konieczność wyznaczenia numerycznych reprezentacji topologii, które następnie mogą posłużyć jako wejście do modelu ML. W niniejszej pracy zaproponowano metodę uczenia reprezentacji sieci IEEE 802.11 z wykorzystaniem grafowych sieci neuronowych (ang. *graph neural networks*, GNN), podejścia sprawdzonego w literaturze [4]. Uzyskane w ten sposób zanurzenia (ang. *embeddings*) mogą być wykorzystane m.in. do analizy, wizualizacji oraz generowania nowych konfiguracji MAPC. Według wiedzy autora, jest to pierwsza praca, w której GNN zostały zastosowane do uczenia reprezentacji konfiguracji MAPC.

2. GRAFOWE SIECI NEURONOWE

Topologię sieci IEEE 802.11 można modelować jako graf $G = (V, E)$, gdzie V to zbiór wierzchołków, a E krawędzi. Szczegóły konstrukcji grafu opisano w Sekcji 3. GNN, który jest sercem proponowanej metody, to iteracyjne przekształcenie grafu, które zachowuje jego strukturę i opiera się na dwóch głównych operacjach: agregacji i aktualizacji [1]. Niech $\mathbb{X}$ będzie przestrzenią atrybutów wierzchołków, krawędzi lub globalnych wartości, natomiast $\mathbb{Y}$ przestrzenią zagregowanych atrybutów z sąsiedztwa. Agregacja $\rho : \mathbb{X}^n \to \mathbb{X}$ łączy informacje z sąsiedztwa (np. przez sumę lub średnią), a aktualizacja $\phi : \mathbb{X} \times \mathbb{Y} \to \mathbb{X}$ przekształca atrybuty mając na wejściu poprzednie wartości oraz informacje z sąsiedztwa.

W typowej architekturze GNN stosuje się dwie lub trzy sieci neuronowe wykorzystywane do aktualizacji wierzchołków, krawędzi i wartości globalnych (jeśli istnieją). Powtarzane kilkukrotnie operacje agregacji i aktualizacji tworzą mechanizm *message passing*, który umożliwia efektywną propagację informacji w grafie, dzięki czemu sieci GNN są uniwersalne, dobrze skalują się do dużych topologii i radzą sobie z grafami o zmiennej strukturze. Zgodnie z modelem GraphNetwork według Algorytmu 1 z [1], pojedynczy krok *message passing* składa się z trzech głównych etapów:

1. aktualizacja krawędzi – nowe atrybuty krawędzi obliczane są na podstawie poprzednich wartości oraz atrybutów wierzchołków końcowych, z użyciem współdzielonej sieci neuronowej,
2. agregacja na wierzchołkach – każdy wierzchołek agreguje atrybuty krawędzi do niego przylegających,
3. aktualizacja wierzchołków – nowe atrybuty wierzchołków obliczane są na podstawie poprzednich wartości oraz zagregowanych atrybutów krawędzi, również z wykorzystaniem współdzielonej sieci neuronowej.

W pełnym wariancie modelu, w ostatnim kroku następuje także agregacja informacji z całego grafu (wszystkich krawędzi i wierzchołków) oraz aktualizacja atrybutów globalnych.

3. METODOLOGIA

Graf topologii sieci konstruowany jest z dwóch typów wierzchołków (reprezentujących stacje oraz AP) i dwóch typów krawędzi (stacja-AP i AP-AP), co ilustruje Rysunek 1. Wierzchołki początkowo nie zawierają atrybutów. Krawędzie stacja-AP tworzone są dla każdej aktywnej asocjacji, natomiast krawędzie AP–AP łączą pary znajdujące się w zasięgu słyszalności (na podstawie ramek *beacon*), gdy siła sygnału przekracza ustalony próg (przykładowo para AP 1-3 na Rysunku 1 nie jest połączona z powodu zbyt słabego sygnału).

Każda krawędź zawiera zestaw atrybutów. Dla krawędzi stacja-AP są to: informacja o udziale w rywalizacji o kanał, udziale w ostatniej możliwość transmisji (ang. *transmission opportunity*, TXOP), użytym MCS, mocy transmisji (w dBm) oraz stosunku poprawnie odebranych do wszystkich wysłanych podramek. Krawędzie AP-AP opisują interferencję między węzłami i zawierają informację o sile sygnału (w dBm). W przypadku braku danego parametru, przypisywana jest wartość specjalna $\emptyset$.

Model GNN został zaimplementowany w architekturze wariacyjnego autoenkodera (ang. *variational autoencoder*, VAE) [2], którego celem jest uczenie reprezentacji wejścia x i jego rekonstrukcja $\hat{x}$. Aby umożliwić rekonstrukcję atrybutów z brakującymi wartościami, zastosowano dedykowaną funkcję kosztu:

$$\begin{aligned}\mathcal{L}_k = \mathbb{E}_{x\sim D_k,\hat{x}=\mathrm{GNN}(x)}\,[&\mathcal{L}_{\mathrm{CE}}(\mathbb{1}_{x=\emptyset}, p(\hat{x}=\emptyset)) \\ &+(1-\mathbb{1}_{x=\emptyset})\cdot\mathcal{L}_{\not x}(x,\hat{x})]\,,\end{aligned} \tag{1}$$

gdzie D_k oznacza zbiór danych treningowych dla atrybutu k, $\mathcal{L}_{\mathrm{CE}}$ to entropia krzyżowa, $\mathcal{L}_{\not x}$ jest funkcją

kosztu dopasowaną do typu atrybutu (np. entropia krzyżowa lub błąd średniokwadratowy), natomiast $\mathbb{1}_{x=\emptyset}$ to funkcja wskaźnikowa, przyjmująca wartość 1 dla brakującego atrybutu. Dodatkowo przestrzeń reprezentacji jest regularyzowana przez dywergencję Kullbacka-Leiblera (D_{KL}) względem standardowego rozkładu normalnego [2]. Ostateczna funkcja kosztu uwzględnia średnią po wszystkich atrybutach oraz regularyzację: $\mathcal{L} = \frac{1}{n}\sum_{k=1}^{n}\mathcal{L}_k + \beta D_{\mathrm{KL}}$, gdzie n to liczba atrybutów, a β oznacza czynnik balansujący rekonstrukcję i regularyzację. Szczegóły architektury i użytych hiperparametrów dostępne są w repozytorium implementacji[1].

Zbiór treningowy obejmuje 1000 zróżnicowanych topologii złożonych z 1 do 10 AP, z których każdy obsługuje od 1 do 10 stacji. Pozycje węzłów, w zależności od scenariusza, były ustalane deterministycznie lub losowo, z uwzględnieniem przeszkód (ścian i pomieszczeń) lub w otwartym środowisku. Szczegóły parametrów symulacyjnych dostępne są w repozytorium. Dla każdej topologii wygenerowano po 50 konfiguracji MAPC przy użyciu czterech metod: (1) losowego agenta (każdy AP transmituje z prawdopodobieństwem 50% do losowej stacji) oraz trzech podejść z [7], a więc (2) H-MAB oraz dwóch modeli analitycznych: (3) T-Optimal (maksymalizacja przepływności) i (4) F-Optimal (maksymalizacja sprawiedliwości). Dane te posłużyły do treningu GNN. Do symulacji H-MAB oraz obliczania przepływności wygenerowanych konfiguracji MAPC wykorzystano symulator *mapc-sim*[2]. Dodatkowo utworzono zbiór walidacyjny obejmujący 200 topologii wygenerowanych z tymi samymi założeniami, również po 50 konfiguracji na topologię i metodę. Zbiór ten wykorzystano do oceny i wizualizacji nauczonych reprezentacji.

4. WYNIKI

Wytrenowano trzy warianty modelu GNN: duży (L), średni (M) i mały (S). Wszystkie wykorzystują tę samą architekturę z trzema sieciami neuronowymi (dla wierzchołków, krawędzi i globalnych atrybutów), różniąc się jedynie szerokością warstw, rozmiarem zanurzeń i siłą regularyzacji. Wyjściem każdej sieci jest macierz $|V| \times m$, gdzie $|V|$ to liczba wierzchołków w grafie, a m to rozmiar zanurzenia. Szczegóły architektury można znaleźć w Tabeli 1 oraz repozytorium.

W drugiej części Tabeli 1 przedstawiono dokładność rekonstrukcji wybranych atrybutów. Pozostałe atrybuty były rekonstruowane z dokładnością bliską 100%, zatem pominięto te wyniki dla czytelności. W przypadku zmiennych ciągłych, takich jak siła sygnału, zastosowano normalizację do standardowego rozkładu normalnego. Dokładność rekonstrukcji oceniano na podstawie zgodności przyporządkowania do tych samych przedziałów po dyskretyzacji (krok 0,1). Widoczny jest istotny spadek jakości rekonstrukcji zmiennych ciągłych w porównaniu do zmiennych kategorycznych, co sugeruje, że są one trudniejsze do zakodowania przez GNN. Może to negatywnie wpływać na skuteczność dalszych modeli wykorzystujących generowane zanurzenia. Co ciekawe, mały model osiąga najlepsze wyniki w rekonstrukcji MCS oraz siły sygnału, z kolei duży model lepiej radzi sobie z pozostałymi atrybutami. Warto zaznaczyć, że mały model nie uwzględnia regularyzacji D_{KL} ($\beta = 0$), co znacznie wpływa na jego zachowanie – reprezentacje są precyzyjne, ale trudne do użycia w modelach generatywnych z powodu skomplikowanego rozkładu danych.

Tabela 1: Charakterystyka i dokładność modeli GNN

Model	*L*	*M*	*S*
Opis			
Liczba parametrów	776k	77k	21k
Siła regularyzacji β	0,7	0,7	0
Rozmiar zanurzenia	64	16	2
Dokładność			
MCS	87%	83%	97%
Siła sygnału	18%	15%	50%
Prawd. sukcesu	98%	98%	96%
Moc nadawcza	100%	96%	98%

Wizualizacje potwierdzają, że nauczone reprezentacje są dobrze pogrupowane i zorganizowane semantycznie. Na Rysunku 2 przedstawiono zanurzenia krawędzi z dużego modelu po zastosowaniu analizy głównych składowych (ang. *principal component analysis*, PCA), pokolorowane według szybkości transmisji. Widoczne są klastry odpowiadające podobnym parametrom transmisji. Z kolei reprezentacje z małego modelu (Rysunek 3) oferują dobrą separację względem mocy transmisji, jednak nie przypominają rozkładu normalnego z powodu braku regularyzacji.

Rysunek 4 przedstawia reprezentacje całych grafów (jako średnia zanurzeń aktywnych krawędzi), zredukowane przez *t-distributed stochastic neighbor embedding* (t-SNE) [3]. Kolory wskazują metodę generowania konfiguracji MAPC. Można zauważyć dobrą separację metod, przykładowo losowy agent (żółty) tworzy wyraźny odrębny klaster (konfiguracje są losowe i znacznie różnią się od tych inteligentnie dobranych), namiast T-Optimal (fioletowy) grupuje się w zbiory małych i szybkich, ale niesprawiedliwych konfiguracji. F-Optimal i H-MAB równoważą przepływność i sprawiedliwość, zatem odpowiadające im punkty są wymieszane i bardziej rozproszone.

5. WNIOSKI

W niniejszej pracy zaprezentowano wykorzystanie GNN do uczenia reprezentacji konfiguracji MAPC w sieciach IEEE 802.11. Przedstawiono modele o różnej złożoności, trenowane z regularyzacją przestrzeni ukrytej oraz bez niej. Uzyskane wyniki potwierdzają,

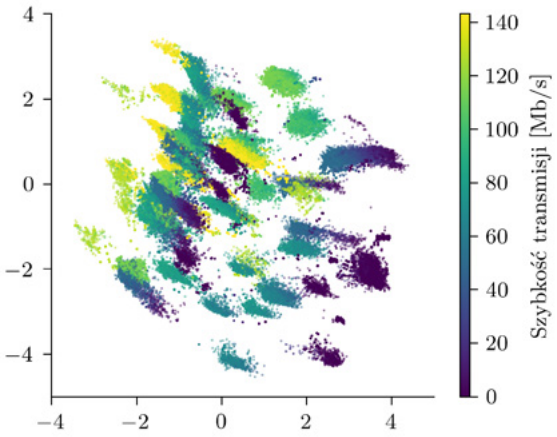
[1] https://github.com/ml4wifi-devs/mapc-gnn (dostęp: 02.06.2025).
[2] https://github.com/ml4wifi-devs/mapc-sim (dostęp: 02.06.2025).

że GNN skutecznie rekonstruują atrybuty transmisji, a reprezentacje odzwierciedlają semantyczne własności połączeń, tworząc klastry podobnych konfiguracji.

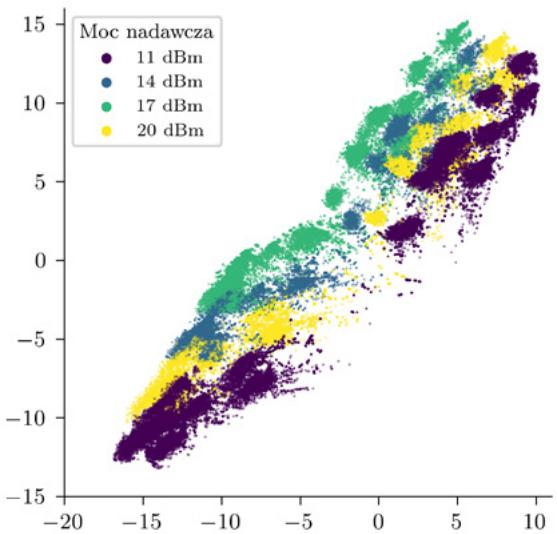
Nauczone reprezentacje mogą być wykorzystane jako wejście do modeli generatywnych, umożliwiając trening metod ML o pożądanych właściwościach. W ramach przyszłych prac, planowane jest wykorzystanie nauczonych reprezentacji w modelach generatywnych, a także dodanie historii zachowania węzłów.

PODZIĘKOWANIA

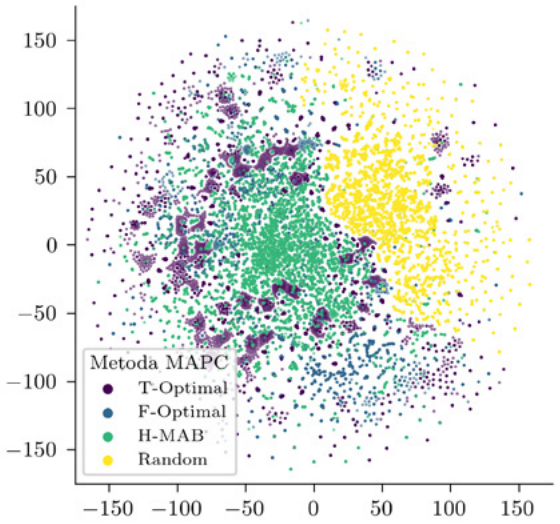
Praca powstała w wyniku realizacji projektu badawczego o nr DEC-2020/39/I/ST7/01457 finansowanego ze środków Narodowego Centrum Nauki. Dziękujemy za wsparcie Państwowemu Instytutowi Badawczemu, numer grantu POIR.04.02.00-00-D008/20-01 „Krajowe laboratorium sieci i usług 5G wraz z otoczeniem" (akronim PL-5G), za udostępnienie sprzętu obliczeniowego.



Rysunek 2: Reprezentacje aktywnych krawędzi z dużego modelu zwizualizowane przy pomocy PCA. Kolory oznaczają uzyskaną szybkość transmisji.



Rysunek 3: Reprezentacje aktywnych krawędzi z małego modelu. Kolory oznaczają użytą moc transmisji.



Rysunek 4: Reprezentacje grafów z dużego modelu zwizualizowane przy pomocy t-SNE. Kolory oznaczają metodę użytą do wygenerowania konfiguracji.

LITERATURA

[1] Battaglia Peter et al. 2018. „Relational inductive biases, deep learning, and graph networks". arXiv 1806.01261.

[2] Kingma Diederik, Welling Max. 2013. „Auto-Encoding Variational Bayes". arXiv 1312.6114.

[3] van der Maaten Laurens, Hinton Geoffrey. 2008. „Visualizing Data using t-SNE". Journal of Machine Learning Research 9 (86) : 2579 – 2605.

[4] Mohsenivatani Maryam et al. 2024. „Graph Representation Learning for Wireless Communications". IEEE Communications Magazine 62 (1) : 141 – 147.

[5] Szott Szymon et al. 2022. „Wi-Fi Meets ML: A Survey on Improving IEEE 802.11 Performance With Machine Learning". IEEE Communications Surveys & Tutorials 24 (3) : 1843 – 1893.

[6] Szott Szymon et al. 2025. „MAPC C-SR Scheduling with Multi-Armed Bandits Revisited". URL: https://mentor.ieee.org/802.11/dcn/25/11-25-0710-01-aiml-mapc-c-sr-scheduling-with-multi-armed-bandits-revisited.pptx (dostęp: 29.05.2025).

[7] Wojnar Maksymilian et al. 2025. „Coordinated Spatial Reuse Scheduling With Machine Learning in IEEE 802.11 MAPC Networks". arXiv 2505.07278. Przyjęty do publikacji w IEEE Journal on Selected Areas in Communications.

[8] Yu Ross Jian. 2025. „Specification Framework for TGbn". URL: https://mentor.ieee.org/802.11/dcn/24/11-24-0209-15-00bn-specification-framework-for-tgbn.docx (dostęp: 29.05.2025).

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

Wykorzystanie multimodalnego uczenia maszynowego oraz wyjaśnialnego AI do automatyzacji monitorowania i zarządzania stanem sieci komórkowej 5G+

The use of multimodal machine learning and explainable AI for automating the monitoring and management of 5G+ cellular network conditions

Michał Błaszczak[1,2]; Ireneusz Jabłoński[3,4]; Aleksandra Knapińska[2]; Krzysztof Walkowiak[2];

[1] Nokia, Lotnicza 12, 54-155 Wrocław, michal.blaszczak@nokia.com

[2] Wroclaw University of Science and Technology, Wyb. Wyspiańskiego 27, 50-370 Wrocław

[3] Brandenburg University of Technology Cottbus-Senftenberg, Conrad-Zuse Str. 1, 03046 Cottbus

[4] Fraunhofer Institute for Photonic Microsystems, Maria-Reiche Str. 2, 01109 Dresden

DOI: 10.15199/59.2025.4.59

Streszczenie: Sieci mobilne 5G+ (5G i kolejne generacje) stają się coraz bardziej złożone, co utrudnia ich analizę oraz zarządzanie nimi. Ilość oraz heterogeniczność generowanych danych sieciowych sprawiają, że tradycyjne metody unimodalne stosowane do ich analizy stają się niewystarczające. Niniejsza praca dokonuje przeglądu danych sieciowych oraz ich modalności i omawia możliwości zastosowania metod multimodalnych do ich analizy. Przedstawiono również związane z tym wyzwania oraz ogólną motywację do podejmowania takich działań.

Abstract: 5G+ (5G and beyond) mobile networks are becoming more and more complex which makes their analysis and management more difficult. The amount and heterogeneity of produced network data is infeasible by traditional unimodal analysis methods. This work reviews both network data and its modalities and discusses possible use of multimodal methods for the analysis. Associated challenges are discussed and high level motivation is presented.

Słowa kluczowe: sieci bezprzewodowe, pomiar wydajności, wspomaganie decyzji, dane multimodalne, wyjaśnialna sztuczna inteligencja

Keywords: wireless networks, performance measurement, decision support, multimodal data, explainable AI

1. INTRODUCTION

Traditional approaches to performance modeling of complex systems, like the one observed in 5G+ network infrastructure environment, face limitations due to data sparsity and heterogeneity. A significant portion of data regarding system performance is captured through numerical measurements, such as counters, which are then used to calculate key performance indicators (KPIs). Technical documentation, i.e. regarding feature activation, further contributes essential contextual knowledge. In addition, measurement activities conducted in laboratory and field conditions throughout the development life cycle serve as valuable sources of information for the automation of monitoring and management processes in the system.

The 5G+ network system and its processes are complex, which coincides with the emerging need to automate state monitoring and system resources management recently observed in both academia and industry. There are studies that have shown that observing complexity using a single data modality — and the models appropriate for it — is subjected to certain limitations [1, 2, 3]. It is common for complex systems and the processes occurring within them to be observed through multiple data sources, which may be of different nature (modality), as it is the case with 5G+ cellular networks. Considering the above, contributions to algorithm development and the application of multimodal modeling, including explainable approaches, become justified, particularly for tasks such as automated monitoring of the real-time state of cellular networks and the management of their resources. There are works in domains different from telecommunication that show the benefits of multimodal approaches over the unimodal ones.

The goal of this paper is to present the current state of work regarding automation of monitoring and management of 5G+ cellular network conditions, its challenges, and propose a novel approach through the use of the multimodal machine learning (MML).

2. CHALLENGES IN MODELING 5G+ NETWORK PERFORMANCE

Modeling the network performance (NP) faces several challenges, both intrinsic and extrinsic. The complexity of networks is increasing due to the growing technological and technical sophistication of network infrastructures, the expanding portfolio of provided services, and the increasingly diverse mix of services tailored to specific user groups. Moreover, user profiles and mobility patterns are evolving, and service distribution is taking place over increasingly vast geographical areas. All these factors con-

tribute to making the monitoring and management of such systems highly non-trivial, particularly when automation is involved, and this is an example of the latter challenge. Methods relying solely on one data type, which are unable to capture the complex nature of the phenomenon, are examples of the former challenge.

2.1 Evolution of network complexity in 5G+ systems

Mobile networks experience growth of network traffic as a result of several factors. Network complexity increases due to emerging new services and the needs of the customers. 5G+ networks have been evolving to accommodate these challenges. In this context, service assurance (SA) and network KPI monitoring are indispensable. However, since network is virtualized and the number of introduced features increases, human troubleshooting and resources monitoring and management become infeasible [4].

2.2 Limitations of unimodal modeling

Unimodal methods are the ones that rely solely on one data type, i.e. time series data. There are methods and tools available for numerical data that enable deterministic, statistical, and machine learning-based approaches, including those grounded in artificial intelligence. These approaches support tasks involving descriptive, predictive, and even prescriptive inference, with a focus on both regression and classification activities. Analogously, such methods and tools can be applied to other data modalities but their capabilities are currently limited. Nevertheless, these capabilities are steadily expanding because of the growing interest of researchers and engineers in these domains. While all these models have proven effective in narrow contexts, their effectiveness in context rich 5G+ network environment is utmost questionable. It remains true even if overspreaded on all unimodal models. Although they may be relatively simple and fast, a main drawback is that they can miss contextual information [5]. For example, there is a shift in the numerical data which can be attributed to feature activation or change in the configuration, but since the model relays on numerical data only, the shift remains unexplained. Besides explainability, in complex scenarios, the accuracy of predictions made by the unimodal models may be too inaccurate to be useful for decision support.

2.3 Lack of explainability

In critical systems, of which a 5G+ network is a perfect example, decisions must be characterized by several factors like the accuracy of the predictions made, and explainability. It is a well known fact that, although capable of achieving high accuracy, deep learning (DL) models lack explainability [6]. Because of that, accountability and trustworthiness of a solution based on DL model is limited, so its usage in 5G+ network is also limited [7].

2.4 Automation of network operation, monitoring, and management

The automation of network operation, monitoring, and management represents one of the key cognitive and engineering challenges nowadays. Self-organizing network (SON) is a concept of self-adaptation capabilities of the network to offer users services on-demand [8]. However, there are several challenges that must be solved in order to enable autonomous intelligent 5G+ network: training issues, lack of bounding performance, lack of explainability, uncertainty in generalization and lack of interoperability [9]. Autonomous network (AN) is another concept, which aims to ensure both trustworthy and efficient operation and management at lower cost. However, because the network operates as distributed dynamic systems governed by various technical and economic constraints, its structural and functional complexity imposes a challenge [10].

3. TYPES OF MULTIMODAL DATA IN 5G+ NETWORK ENVIRONMENT

The 5G+ network environment generates an enormous amount of diverse multimodal data. These include both numerical and non-numerical data. Example of the former are KPIs. Example of the latter is documentation (text, diagrams, images) or reports after feature introduction (text, diagrams, tables). Data can be also classified as structured (i.e. KPIs) and unstructured (i.e. documentation).

3.1 Numerical data

Network Performance should be measured in a way that is quantifiable and understandable. Counters and KPIs represent metrics that satisfies both criteria. Due to multiplicity of KPIs, they are organized into categories and subcategories, based on the area they describe, i.e. throughput. Yet another organization is placing them in two categories based on the area they characterize: traffic conditions and environmental conditions. Example KPIs belonging to the aforementioned groups are as follows. The first group includes indicators such as #user equipment (UE), data volume (DV) or physical resource block (PRB) utilization. Second group is represented by channel quality indicator (CQI) and block error rate (BLER).

3.2 Non-numerical data

Besides NP measurement, its expected and observed behaviour, architecture specifications or any other network-related information, are stored in a documentation which consists of different data modalities. For example, every new feature that is introduced, in addition to describing what functionality is being introduced, discusses several different topics. One of them is the evaluation of the impact that the operation of the feature has on performance. Another topic discusses interdependencies of the new feature with other features. Yet another topic discusses configuration settings and procedures necessary to activate, configure, and optimize the feature. The aforementioned documentation consists of, but is not limited to, natural language texts, images, tables and diagrams. Some of data is static i.e. text, but some is interactive, i.e. a dashboard showing feature dependencies.

3.3 Use-case

There are several use-cases in 5G+ network performance modeling that encompass multimodal data. One of the examples is assessment of the performance changes due to the deployment of new functionality. The impact

of a functionality deployment depends on the parameter settings, which may cause KPIs to change in a vast range. The description of the impact of the parameters is in the multimodal documentation. The classical approach is to assess the changes by a domain expert whose knowledge is supported by the aforementioned documentation.

4. MULTIMODAL ML FOR TELECOM: PRINCIPLES AND CROSS-DOMAIN APPLICATIONS

Sparsity and heterogeneity of multimodal data combined with the potential of ML methods create an opportunity to revolutionize the modeling of a phenomenon of interest. There is a plethora of MML solutions that have demonstrated improved accuracy over unimodal models in domains like computer vision or medicine. However, in context of 5G+ networks, the usage of MML is highly unexplored which creates a perfect research gap to fulfill.

4.1 Definition of MML

MML is a type of machine learning that aims to utilize data of different modalities to improve modeling of a phenomenon of interest [11]. Using data of different modalities can improve model accuracy and enables broader applicability of the solution. It is possible due to the leveraging of complementary information coming from different modalities, as well as facilitating knowledge transfer if there is scarcity of one data type [12].

4.2 Handling heterogeneous data

Data heterogeneity is the core of MML, but, at the same time, brings challenges unseen in unimodal ML. Representation tries to address the challenge of how to represent multimodal data in a manner that exploits the complementarity and redundancy of multiple modalities. Translation aims to map the data from one modality to another. Alignment tries to find relations between entities of different modalities. Fusion, apparently the most important one, aims to join entities of different modalities. Co-learning addresses the challenge of transferring knowledge between different modalities [11].

4.3 Related applications of multimodal fusion

Multimodal fusion has proven records of improved model performance in different domains. There is a multimodal fusion detection system for autonomous vehicles that achieves 3.7% improvement in accuracy compared to a unimodal CNN detection system [13]. Another example is a multimodal social media video classification system that surpassed an unimodal neural network by 11.6% reaching 88% accuracy [14]. There is also a medical survey reviewing methods that combine medical imaging with clinical data. Finding is that multimodal fusion models showed an increased accuracy over traditional single modality models, where improvement ranged from 1.2% up to 27.7% [15]. There is also an example from predictive maintenance field, in which the model combines time-series sensor data with event logs, with an improvement in fault prediction accuracy ranging from 2.6% up to 16.9% [16]. Last but not least, there is a network related example. A model that combines textual descriptions with packet behaviour data achieves 95% mean area under the curve (AUC) and outperforms baseline models by 11.6% [17].

5. PROPOSED MULTIMODAL FRAMEWORK FOR NETWORK PERFORMANCE MODELING

Based on the literature review and conceptual studies, it is advisable to develop intelligent analytics module, of multimodal 5G+ cellular network data – e.g., in a proposed architecture presented in Figure 1 – to improve the deployment of new features through automation of monitoring and management of network resources. Proposed architecture schema shows input data, possible analytical paths and expected result. Analysis can be done using multimodal graph learning (MGL) (upper path), or MML or foundation models adapted for time series data analysis (TSFMs) (lower path) utilizing different fusion techniques. Results can be both of predicitve and prescriptive nature.

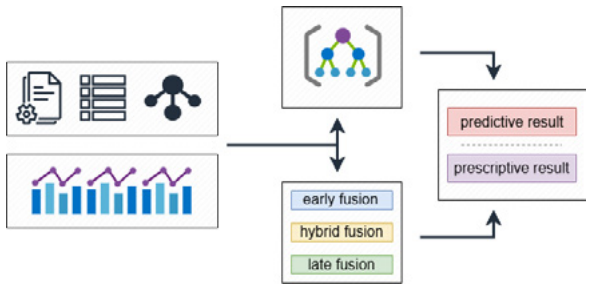


Fig. 1: Architecture of the proposed module.

5.1 Use of multimodal machine learning

Although the domain of MML algorithms is well established, there is scarcity of papers discussing the usage of MML in the telecommunication domain, making it the perfect research gap to fulfill. An example of such solution in another domain is a multimodal time series analysis model based on the spiking neural networks. Classification results on a biological time-series dataset shows accuracy of 98.75% which outperforms baseline models [18].

5.2 Use of multimodal graph learning

Another approach is the usage of the MGL. High complexity and diversity of data in the considered case makes the usage of the aforementioned group of algorithms a perfect choice. Moreover, there are no works discussing the usage of MGL in the telecommunication domain, thus, it is a perfect research gap to fulfill. An example of such algorithm is TimeGNN, which learns dynamic temporal graph representations. These representations are capable of capturing the evolution of inter-series patterns along with the correlations of multiple series. Although the performance of TimeGNN is comparable with other state-of-the-art methods, its inference times are 4 to 80 times faster [19].

5.3 Use of foundation models adapted for time series data analysis

Yet another approach is the use of TSFMs. One of the most prevalent aspects of the current TSFMs is the

development based on a single modality. Developing multimodal TSFM for telecommunication domain offers an ideal opportunity to fulfill this research gap. For example, the text-guided time series forecasting (TGTSF) model proves that integrating textual cues with time series data mitigate the information insufficiency. This approach achieves an 80% improvement in the mean squared error (MSE) over the best performing transformer-based baseline model and a 96% increase over DLinear [20]. Another example, modality-aware transformer, makes use of the modality-aware structure to effectively exploit information within each modality as well as foster cross-modal understanding in the financial domain. Proposed model achieves lowest mean squared error (MSE) and mean absolute error (MAE) when compared with several transformer-based time series forecasting baseline methods [21].

6. CONCLUSIONS

There are unresolved issues related to: a) reliable and automated monitoring and management of 5G+ network performance, b) the use of only numerically encoded data in existing research focused on (a). The literature review conducted in the article indicates that it is possible to enrich the input information for network performance modeling by adding multimodal data streams. This, in turn, based on the review of MML methods, may lead to improved reliability of the models used in decision-making related to cellular network performance characteristics, making the network resource management process increasingly automated. The article also highlights the need for collecting diverse data and telecommunications knowledge and integrating (fusing) them in the inference process, which constitutes a research gap to be addressed.

7. ACKNOWLEDGEMENT

This work was financed by the Polish Ministry of Education and Science via "Implementation doctorate" program (Agreement no. DWD/8/0032/2024).

REFERENCES

[1] Karchmer, Ari. 2024. „On Stronger Computational Separations Between Multimodal and Unimodal Machine Learning". arXiv:2404.02254.

[2] Zubić, Nikola et al. 2025. „Limits of Deep Learning: Sequence Modeling through the Lens of Complexity Theory". arXiv:2405.16674.

[3] Esteghamati, Mohsen Zaker et al. 2024. „Beyond development: Challenges in deploying machine learning models for structural engineering applications". arXiv:2404.12544.

[4] Tran, Nguyen Phuc et al. 2023. "ML KPI Prediction in 5G and B5G Networks". In: *2023 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*. IEEE, (2023).

[5] Summaira, Jabeen et al. 2022. „A Review on Methods and Applications in Multimodal Deep Learning". arXiv:2202.09195.

[6] Usama, Muhammad et al. 2020. „Examining Machine Learning for 5G and Beyond through an Adversarial Lens". arXiv:2009.02473.

[7] Senevirathna, Thulitha et al. 2025. „A Survey on XAI for 5G and Beyond Security: Technical Aspects, Challenges and Research Directions". IEEE Communications Surveys & Tutorials 27 (2): 941–973.

[8] Fourati, Hasna et al. 2021. „Comprehensive survey on self-organizing cellular network approaches applied to 5G networks". Computer Networks 199: 108435.

[9] Shafin, Rubayet et al. 2020. „Artificial Intelligence-Enabled Cellular Networks: A Critical Path to Beyond-5G and 6G". IEEE Wireless Communications 27 (2): 212–217.

[10] Sifakis, Joseph et al. 2025. „A Reference Architecture for Autonomous Networks: An Agent-Based Approach". arXiv:2503.12871.

[11] Baltrušaitis, Tadas et al. 2018. „Multimodal Machine Learning: A Survey and Taxonomy". IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (2): 423–443.

[12] Li, Songtao et al. 2024. „Multimodal Alignment and Fusion: A Survey". arXiv:2411.17040.

[13] Person, Michael et al. 2019. „Multimodal Fusion Object Detection System for Autonomous Vehicles". Journal of Dynamic Systems, Measurement, and Control 141 (7): 071017.

[14] Trzcinski, T. 2018. „Multimodal social media video classification with deep neural networks". 10808: 108082U.

[15] Huang, Shih-Cheng et al. 2020. „Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines". npj Digital Medicine 3 (1): 136.

[16] Liu, Zengkun et al. 2024. „Advancing predictive maintenance: a deep learning approach to sensor and event-log data fusion". Sensor Review 44 (5): 563–574.

[17] Masukawa, Ryozo et al. 2025. „PacketCLIP: Multi-Modal Embedding of Network Traffic and Language for Cybersecurity Reasoning". arXiv:2503.03747.

[18] Liu, Chengzhi et al. 2024. „MTSA-SNN: A Multi-modal Time Series Analysis Model Based on Spiking Neural Network". arXiv:2402.05423.

[19] Xu, Nancy et al. 2023. „TimeGNN: Temporal Dynamic Graph Learning for Time Series Forecasting". arXiv:2307.14680.

[20] Xu, Zhijian et al. 2024. „Beyond Trend and Periodicity: Guiding Time Series Forecasting with Textual Cues". arXiv:2405.13522.

[21] Emami, Hajar et al. 2024. „Modality-aware Transformer for Financial Time series Forecasting". arXiv:2310.01232.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

SAMOLUBNE ZACHOWANIA AGENTÓW UCZENIA ZE WZMOCNIENIEM W SIECIACH IEEE 802.11

SELFISH BEHAVIOUR OF REINFORCEMENT LEARNING AGENTS IN IEEE 802.11 NETWORKS

Kamil Szczęch

AGH Akademia Górniczo-Hutnicza, Kraków, kamil.szczech@agh.edu.pl

DOI: 10.15199/59.2025.4.60

Streszczenie: Rozwój technologii uczenia maszynowego wprowadza nowe możliwości ale także stwarza nowe wyzwania. Niepoprawne wdrożenie technik wspomaganych agentami korzystającymi z uczenia ze wzmocnieniem może doprowadzić do zmniejszenia wydajności systemu. Celem pracy jest analiza egoistycznych zachowań agentów uczenia ze wzmocnieniem stosowanych w sieciach standardu IEEE 802.11 do modyfikacji parametrów dostępu do kanału. Sprawdzony został wpływ działania takich agentów na wydajność stacji stosujących się do reguł standardu IEEE 802.11.

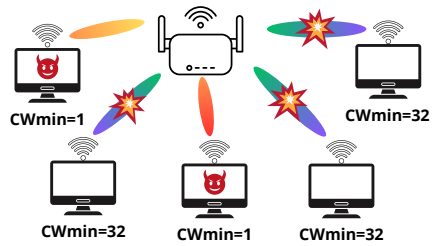
Abstract: The development of machine learning methods offers new solutions, but also causes new challenges. The incorrect implementation of techniques supported by reinforcement learning agents can lead to reduced system performance. The purpose of this study is to analyze the selfish behavior of reinforcement learning agents used in IEEE 802.11 networks to modify channel access parameters. We assess the impact of such selfish agents on stations that adhere to the IEEE 802.11 channel access rules.

Słowa kluczowe: lokalne sieci bezprzewodowe, optymalizacja, uczenie maszynowe, zarządzanie zasobami

Keywords: machine learning, optimization, resource management, wireless local area networks

1. WSTĘP

Rozwój sztucznej inteligencji daje możliwość zastosowania nowych narzędzi w celu optymalizacji sieci bezprzewodowych. Niestety nieskoordynowane zarządzanie lokalną siecią bezprzewodową może doprowadzić do nierównego podziału zasobów wśród jej użytkowników. Rozproszony system, w którym każda stacja samodzielnie dobiera parametry dostępu do kanału według swojego upodobania może doprowadzić do wystąpienia zachowań samolubnych. Przykładem systemu rozproszonego może być sieć standardu IEEE 802.11. Każda ze stacji samodzielnie dobiera parametry dostępu do kanału, z których kluczowym jest okno współzawodnictwa (ang. *contention window*, CW). Dobór tego parametru może zostać przekazany agentowi uczenia maszynowego ze wzmocnieniem (ang. *reinforcement learning*, RL). Jeśli stacje nie komunikują się między sobą, a nagroda agenta zostanie źle zaprojektowania, to może dojść do zdominowania sieci przez grupę samolubnych stacji. Schemat takiego scenariusza jest przedstawiony na rysunku 1, gdzie agenci sterujący stacją wybierają wartość parametru CW niższą od domyślnej. Prowadzi to do zwiększonej liczby kolizji i może uniemożliwić transmisję stacjom działających według standardu.



Rysunek 1: Przykład zachowania samolubnego w sieciach IEEE 802.11.

Celem niniejszej pracy jest analiza zachowania sieci standardu 802.11ax, w której część użytkowników optymalizuje swoje połączenie za pomocą agentów RL. Agenci działają w rozproszeniu, a więc nie dochodzi do komunikacji między nimi. Badania zakładają także minimalną komunikację z punktem dostępowym. Dodatkowo przetestowano tylko techniki RL oparte na tak zwanych wielorękich bandytach (ang. *multi-armed bandits*, MAB). MABy znane są z szybkiego znajdowania optimum i mniejszych wymagań sprzętowych od rozwiązań głębokiego uczenia. Niestety nie zawsze są wystarczające ze względu na brak użycia kontekstu w procesie uczenia. Podczas badania przeanalizowano parametry sieci takie jak przepustowość i współczynnik sprawiedliwości Jaina zależnie od liczby stacji sterowanych przez agentów RL. W rozdziale 2 opisano literaturę dotyczącą zarówno zachowań samolubnych oraz sieci zarządzanych przez agenta RL. Następnie opisano scenariusze badawcze oraz implementację agentów RL (rozdział 3). Rozdział 4 przedstawia możliwe strategie stosowane przez agentów RL. W rozdziale 5 opisano wyniki oraz wnioski z nich płynące. Na koniec w rozdziale 6 podsumowano pracę i przedstawiono dalsze kierunki badań.

2. PRZEGLĄD LITERATURY

Znane są badania dotyczące ataków samolubnych w sieciach 802.11 [4]. Ataki te polegają na ustawianiu przez agresora niższej wartości CW dla danej klasy ruchu niż zakłada to standard. Taka zmiana w sieci przy małej liczbie stacji agresywnych skutkuje większą przepustowością stacji agresywnych kosztem pozostałych stacji. Zwiększająca się liczba agresorów w sieci doprowadza jednak do spadku przepustowości całej sieci w tym i samych stacji oszukujących. Problemy te spowodowane są większą liczbą kolizji. Gdy cała sieć składa się wyłącznie z oszustów przepustowość sieci jest znacząco mniejsza niż gdyby wszystkie stacje działały według standardu.

Użycie uczenia maszynowego w sieciach 802.11 zyskuje na popularności – rośnie liczba publikacji w tym temacie [3]. Popularnym podejściem są scentralizowane rozwiązania gdzie cała sieć zarządzana jest przez punkt dostępowy (ang. *access point*, AP), a parametry takie jak CW, agregacja ramek czy uruchomienie lub wyłączenie RTS/CTS dobierane są przez jednego głównego agenta, który połączony jest z punktem dostępowym [5]. Wyklucza to problem zachowania egoistycznego jednej stacji, ponieważ agent posiada wiedzę na temat całej sieci.

Istnieją próby poprawy działania sieci poprzez systemy rozproszone. W pracy [1] badacze sprawdzają możliwość poprawy wydajności mechanizmu CW właśnie przez budowę rozproszonego systemu wykorzystującego agentów RL. Każda stacja używa agenta RL, który musi jednak pozyskać informacje na temat liczby aktywnych stacji, a także o czasie przez który stacje transmitowały. Sprawia to, że nagroda jest liczona na podstawie danych, które nie zależą bezpośrednio tylko od jednej stacji, a co za tym idzie są trudniejsze do zagregowania i wymagają dodatkowej komunikacji. Użyto także w tej pracy głębokiego uczenia ze wzmocnieniem. Metody głębokie mają większy kontekst podczas procesu uczenia ale wymagają większych zasobów sprzętowych co często utrudnia ich implementację w sprzęcie. Przy podobnym sposobie liczenia nagrody badacze sprawdzali także sposoby obrony przed atakami opartymi na zmniejszaniu wartości CW [7]. Jednak i tu wykorzystano dodatkową wymianę informacji między AP, a stacją, a także użycie głębokich metod uczenia. Podsumowując, dotychczas w literaturze nie badano w pełni rozproszonych systemów agentowych dla sieci 802.11, które nie komunikowałyby się bezpośrednio ze sobą.

3. ŚRODOWISKO BADAWCZE

W tym rozdziale omówiono scenariusz badawczy oraz komunikację agentów ze środowiskiem symulacyjnym.

3.1 Scenariusz symulacyjny

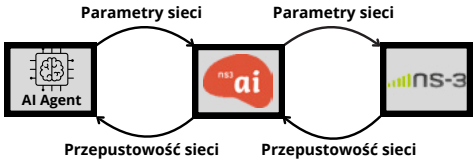
Analizujemy sieć działającą w standardzie 802.11ax z jednym AP i 10 stacjami. Każda stacja znajduje się w losowej odległości od AP, lecz nie dalej niż 10 m. Wszystkie stacje są stacjonarne i generują jednorodny ruch do AP o wartości 100 Mb/s co powoduje nasycenie sieci. Pozostałe parametry są podane w tabeli 1.

W sieci znajduje się od zera do dziesięciu 'stacji RL', tj. stacji sterowanych przez agenta RL, który dobiera tym stacjom wartość CW. Dobór CW przez agenta wykonywany jest co 0.5 s. Czas symulacji wynosi 200 sekund, a więc podczas symulacji agent przeprowadza 400 takich zmian. Agent przez pierwsze 100 sekund jest w fazie eksploracji gdzie poszukuje najlepszej konfiguracji. Po 100 sekundach agent przechodzi w fazę eksploatacji gdzie jego wybory są bardziej stałe i z tej fazy były zbierane wszystkie wyniki. Z 10 niezależnych przebiegów wyliczono średnie wartości, a także 95% przedziały ufności.

Tabela 1: Parametry symulacyjne.

Parametr	**Wartość**
Pasmo	5 GHz
PHY/MAC	IEEE 802.11ax
Strumienie przestrzenne	1, SISO
Model kanału	Log-distance
Szerokość kanału	20 MHz
Modulacja i kodowanie	1024 QAM, 5/6
CWmin/CWmax	15/1023
Wielkość A-MPDU	64 kB
RTS/CTS	Wyłączone

3.2 Architektura



Rysunek 2: Schemat wymiany informacji między ns-3 a agentem RL

Symulacje wykonano w ns-3, który pozwala na dynamiczną zmianę parametrów CW podczas symulacji co jest kluczowe przy użyciu agenta RL. Symulator połączony jest z agentem poprzez moduł ns3-ai, który pozwala na przekazywanie informacji między agentem RL, a symulatorem [8]. Symulator do agenta przekazuje informację o przepustowości sieci, a agent na podstawie tych danych wylicza nagrodę i wyznacza nową wartość CW (rysunek 2).

Agenci RL są napisani w języku Python korzystając z biblioteki reinforced-lib [6]. Użyty rodzaj agenta to MAB typu ε-greedy. Każdy agent ustawia wartość minimalnego (CWmin) i maksymalnego (CWmax) okna współzawodnictwa na tę samą wartość. W celu uproszczenia problemu, a także by przyśpieszyć proces uczenia agenta, wybór wartości CW została ograniczony do dziewięciu wartości. Agent wybiera wartość X, która jest wykładnikiem we wzorze służącym do wyliczenia wartości CW:

$$CW_{\min} = CW_{\max} = 2^{X+1} - 1, X \in [0, 9]. \quad (1)$$

4. STRATEGIE AGENTÓW

Rozważamy agentów, którzy korzystają z czterech różnych strategii, a każda z nich różni się definicją nagrody.

4.1 Strategia zachłanna

W strategii zachłannej, zadaniem agenta jest maksymalizowanie przepustowości stacji. Nagroda ma postać

ilorazu przepustowości stacji (T) przez ruch generowany przez stację (D), dzięki czemu jest z przedziału $[0, 1]$. To podejście wydaje się najbardziej intuicyjne ale także najbardziej narażone na wystąpienie zachowań samolubnych.

4.2 Strategia ostrożna

W strategii ostrożnej zadaniem agenta jest minimalizowanie prawdopodobieństwa wystąpienia kolizji podczas transmisji. Kolizje były wyliczane jako różnica ramek wysłanych przez stacje, a ramek otrzymanych przez AP:

$$R_i = \frac{\text{TX-RX}}{\text{TX}} \tag{2}$$

gdzie TX oznacza liczbę wysłanych ramek przez stację, a RX liczbę dostarczonych ramek do AP.

4.3 Ślepo-sprawiedliwa strategia

W ślepo-sprawiedliwej strategii agent dąży do wyidealizowanego, sprawiedliwego dostępu do kanału. Założono, że czas nadawania przez stacje ($czasTX$) przemnożony przez liczbę stacji w sieci ($liczbaSTA$) powinien się równać czasowi obserwacji sieci przez agenta ($czasObs$):

$$R_i = \frac{\text{RX}}{\text{TX}} \times e^{-5 \times |(czasTX \times liczbaSTA) - czasObs|^3} \tag{3}$$

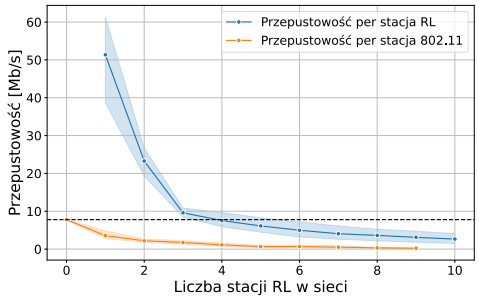
Gdy w sieci znajduje się 10 stacji, a czas obserwacji wynosi 0.5 sekundy to czas nadawania powinien dążyć do wartości 0.05 sekundy. Takie podejście pozwala nam na minimalna dodatkową komunikację z punktem dostępowym. Jedyną informacją, która musi być przesłana przez AP to liczba stacji obecnie nadających w sieci. Nagrodę ostatecznie wymnożono jeszcze przez prawdopodobieństwo sukcesu transmisji. Daje to agentowi lepszy kontekst o działaniu sieci, a także zapobiega nauczeniu się , że najlepszym sposobem na osiągniecie sprawiedliwej komunikacji jest brak transmisji.

4.4 Całkowicie-sprawiedliwa strategia

Ostatnia strategia jest wzięta z literatury [1], gdzie zastosowano techniki głębokiego uczenia ze wzmocnieniem. Naszym celem jest sprawdzenie czy agenci RL używający MAB są w stanie działać w systemach rozproszonych. Nagroda w całkowicie-sprawiedliwej strategii jest podobna do poprzednio omawianej nagrody z tą różnicą, że iloraz $czasTX$ i $liczbaSTA$ nie zmierza do $czasObs$ a do realnego czasu nadawania wszystkich stacji w sieci. Wadą tej metody jest wymóg wymiany dodatkowych informacji: AP musi przekazać informację o liczbie stacji w sieci oraz o ich sumarycznym czasie transmisji.

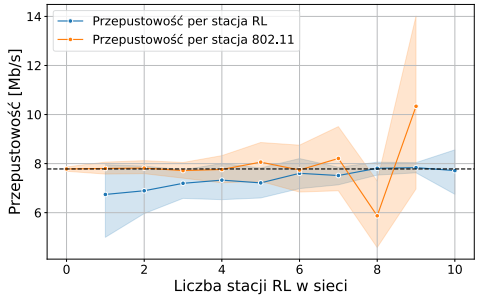
5. WYNIKI

Przebadano wpływ liczby stacji RL na przepustowość sieci, analizując każdą strategię osobno. Scenariusz badawczy był jednorodny, co oznacza, że w każdej symulacji wszyscy agenci RL w sieci stosowali tę samą, z góry wybraną strategię. Na Rysunku 3 przedstawiono przepustowość stacji (zarówno tych z domyślnymi ustawieniami 802.11, jak i stacji RL) dla strategii zachłannej. Można zauważyć, że po pojawieniu się choć jednej stacji RL w sieci przepustowość stacji stosujących się do standardu gwałtowanie spada. Z drugiej strony przepustowość stacji RL jest znacznie wyższa. Taka dysproporcja oznacza, że stacje RL dominują kanał, czyli dochodzi do zachowań samolubnych. Agenci RL mają tendencję do wybierania niskich wartości parametru CW, co prowadzi do zwiększonej liczby kolizji w sieci. Wynika to z faktu, że algorytm DCF staje się w takich warunkach mniej efektywny. Gdy liczba stacji wykorzystujących podejście RL przekracza cztery, ich średnia przepustowość może być niższa niż przepustowość stacji działających w oparciu o domyślne ustawienia standardu. Oznacza to, że zastosowanie tej techniki nie tylko zwiększa liczbę kolizji i obniża łączną przepustowość sieci, ale także może skutkować pogorszeniem wydajności samej stacji RL – do poziomu gorszego niż w przypadku przestrzegania standardowych parametrów.



Rysunek 3: Przepustowość stacji gdy stacje RL używają strategii zachłannej (linia przerywana wyznacza domyślną przepustowość)

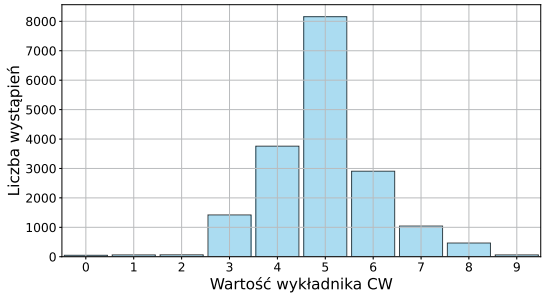
Jeśli poprzednia strategia doprowadzała do zwiększenia liczby kolizji w następnym kroku przetestowano strategię ostrożną, w której agent minimalizuje prawdopodobieństwo kolizji. Niestety ta strategia także okazała się nie przynosić oczekiwanych efektów. Zaobserwowano, że gdy liczba stacji RL była mniejsza niż cztery, agenci wciąż wybierali niskie wartości CW (między 1 a 7), co ponownie prowadziło do zachowań samolubnych. Natomiast w scenariuszach z większą liczbą stacji RL decyzje agentów stawały się chaotyczne, a wybór wartości CW wydawał się wręcz losowy. Skutkowało to spadkiem przepływności sieci, choć nie tak znaczącym, jak miało to miejsce w przypadku strategii zachłannej. Porównanie spadku przepustowości względem innych strategii przedstawiono na rysunku 6.



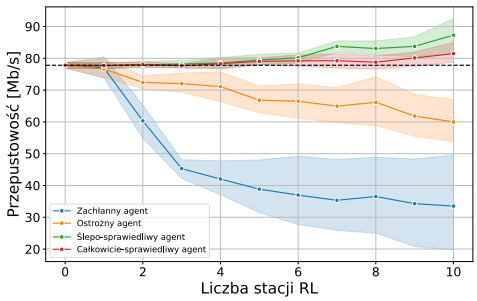
Rysunek 4: Przepustowość stacji gdy stacje RL używają strategii ślepo-sprawiedliwej (linia przerywana wyznacza domyślną przepustowość)

Wyniki dla strategii ślepo-sprawiedliwej przedstawia

rysunek 4. Strategia ta nie wykazaa zachowa samolubnych pomimo wystpienia stacji RL w sieci. Stacje RL uzyskiwały podobne przepustowości do stacji stosujących się do standardu. Jest to wskazane zachowanie według zasady wstecznej kompatybilności urządzeń w sieci. Co więcej jeśli sieć składa się wyłącznie ze stacji RL to przepustowość agregowana sieci zwiększała się.



Rysunek 5: Histogram wartości wykładnika CW (X) wybranych przez agentów RL w fazie eksploatacji. Agenci używali ślepo-sprawiedliwej strategii



Rysunek 6: Sumaryczna przepustowość sieci zależnie od liczby stacji RL w sieci gdy pozostałe z 10 stacji są stacjami 802.11

Po przeprowadzonym teście strategii ślepo-sprawiedliwej wykonano eksperyment, w którym testowano różne wartości parametru CW dla topologii składającej się z 10 stacji. Okazało się, że przepustowość sieci może zostać jeszcze zwiększona. Optymalne ustawienia sieć osiągała przy wartości CW = 255.

Widać więc, że pomimo zwiększenia wydajności, agent RL nie wybrał optymalnych ustawień tego parametru dla przedstawionej topologii. Powód braku znalezienia optimum można zobaczyć na rysunku 5, przedstawiającym wartości CW wybierane przez agentów RL w trakcie eksploatacji sieci. Najczęściej wybierane przez agentów wartości to wartości środkowe zakresu. Agenci RL decydują się na wartości środkowe, ponieważ najłatwiej im było „spotkać się pośrodku" dostępnego zakresu CW — problemem w tym przypadkum może nie być sama funkcja nagrody, lecz zastosowana technika RL. W związku z tym zbadano również inną funkcję nagrody, której skuteczność została wcześniej potwierdzona w literaturze, lecz przy wykorzystaniu technik głębokiego uczenia. Niestety, także w strategii całkowicie sprawiedliwej wystąpiła sytuacja analogiczna do poprzedniej. Wybory agentów RL ponownie skupiały się wokół środkowych wartości przedziału. Zaobserwowane zjawiska prowadzą do wniosku, że techniki MAB są niewystarczające dla systemu rozproszonego o tak rygorystycznych założeniach.

6. PODSUMOWANIE

W pracy przeanalizowano działanie agentów RL do doboru CW w sieciach IEEE 802.11. Wykazano, że proste strategie (maksymalizujące przepustowość, minimalizujące kolizje) prowadzą do zachowań samolubnych, które pogarszają wydajność sieci. Strategie nastawione na sprawiedliwość ograniczają to zjawisko, lecz nie osiągają optymalnych ustawień sieci. Zatem zastosowane podejście okazało się niewystarczające w środowisku rozproszonym bez komunikacji między agentami. W ramach przyszłych prac zbadamy bardziej zaawansowane metody uczenia ze wzmocnieniem.

PODZIĘKOWANIA

Praca powstała w wyniku realizacji projektu badawczego o nr DEC-2023/05/Y/ST7/00004 finansowanego ze środków Narodowego Centrum Nauki.

LITERATURA

[1] A. Kumar, G. Verma, C. Rao, A. Swami and S. Segarra, „Adaptive Contention Window Design Using Deep Q-Learning", ICASSP 2021, Toronto, ON, Canada, 2021, pp. 4950-4954.

[2] Szott, S., Natkaniec, M., & Pach, A. R. 2010. An IEEE 802.11 EDCA Model with Support for Analysing Networks with Misbehaving Nodes. EURASIP Journal on Wireless Communications and Networking, 2010(1), 209895.

[3] Szott, S. et al. 2022. Wi-Fi Meets ML: A Survey on Improving IEEE 802.11 Performance With Machine Learning. IEEE Communications Surveys & Tutorials, 24(3), 1843–1893.

[4] Szott, S., Natkaniec, M., Canonico, R., & Pach, A. R. 2008. Impact of Contention Window Cheating on Single-Hop IEEE 802.11e MANETs. Proc. of IEEE WCNC.

[5] Szczęch, K., Szott, S. 2024. Koncepcja interfejsu radiowego opartego na uczeniu maszynowym w projekcie MLDR, Przegląd Telekomunikacyjny, Wiadomości Telekomunikacyjne,R. 97 nr 4, s. 61–64.

[6] Wojnar M., Szott S., Rusek K., Ciezobka W. 2024. „Reinforced-lib: Rapid prototyping of reinforcement learning solutions." SoftwareX 26: 101706.

[7] Yazdani-Abyaneh, A. H., Hirzallah, M., & Krunz, M. 2024. Countermeasuring Aggressors via Intelligent Adaptation of Contention Window in CSMA/CA Systems. IEEE Access, 12, 88216-88230.

[8] Yin, H., Liu, P., Liu, K., Cao, L., Zhang, L., Gao, Y., & Hei, X. 2020. Ns3-Ai: Fostering Artificial Intelligence Algorithms for Networking Research. Proceedings of the 2020 Workshop on Ns-3, 57–64.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

KONFIGUROWALNY GENERATOR RUCHU SIECIOWEGO W TECHNOLOGII FPGA DO TESTOWANIA SIECI DO 100 GBIT/S

CONFIGURABLE FPGA-BASED NETWORK TRAFFIC GENERATOR
FOR TESTING NETWORKS UP TO 100 GBIT/S

Maciej Barzowski[1]; Sylwester Kaczmarek[2]; Bartosz Matwiejczyk[3]; Magdalena Młynarczuk[4]

[1] Adtran, Gdynia, maciej.barzowski@adtran.com
[2] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, kasyl@eti.pg.edu.pl
[3] Intel, Gdańsk, bartosz.matwiejczyk@intel.com
[4] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, magmlyna@pg.edu.pl

DOI: 10.15199/59.2025.4.61

Streszczenie: **W artykule przedstawiono projekt i implementację konfigurowalnego generatora ruchu sieciowego zbudowanego w technologii FPGA (Intel MAP-200), umożliwiającego generowanie ruchu na poziomie 10G, 25G, 40G oraz 100G. Rozwiązanie umożliwia dynamiczną konfigurację strumieni pakietów z poziomu aplikacji graficznej działającej przez interfejs PCIe. Generator wspiera warstwy L2–L4 modelu ISO/OSI i stanowi elastyczne, nisko kosztowe narzędzie testowe do analizy wydajności i odporności nowoczesnych sieci komputerowych.**

Abstract: **This paper presents the design and implementation of a configurable network traffic generator based on FPGA technology (Intel MAP-200), capable of producing high-speed traffic at 10G, 25G, 40G, and 100G. The solution allows for dynamic stream configuration via GUI communicating through PCIe. The generator supports ISO/OSI layers L2–L4 and offers a flexible, cost-effective testing tool for evaluating performance and resilience of modern computer networks.**

Słowa kluczowe: **FPGA, generator pakietów, PCIe, sieci komputerowe, testowanie wydajności.**

Keywords: **FPGA, packet generator, PCIe, computer networks, performance testing.**

1. WSTĘP

Wraz z rosnącym zapotrzebowaniem na wysokoprzepustowe i stabilne sieci komputerowe, wzrasta znaczenie narzędzi umożliwiających testowanie infrastruktury w warunkach kontrolowanych. Komercyjne rozwiązania, takie jak Spirent [6] czy Keysight Ixia [7], oferują zaawansowane funkcje, lecz ich wysoka cena oraz ograniczona elastyczność są barierą dla wielu użytkowników. W artykule opisano alternatywne rozwiązanie – konfigurowalny generator ruchu sieciowego zbudowany w technologii FPGA [1,5], który umożliwia generację pakietów Ethernet/IPv4/UDP przy przepustowości do 100Gbit/s, zachowując przy tym niski koszt i dużą elastyczność. Urządzenie może być szczególnie przydatne w środowiskach badawczo-rozwojowych oraz w małych i średnich firmach, które nie dysponują budżetem na profesjonalne testery. Dodatkowo, dzięki zastosowaniu otwartej architektury sprzętowej, możliwa jest łatwa adaptacja generatora do indywidualnych potrzeb użytkownika.

2. DOSTĘPNE ROZWIĄZANIA

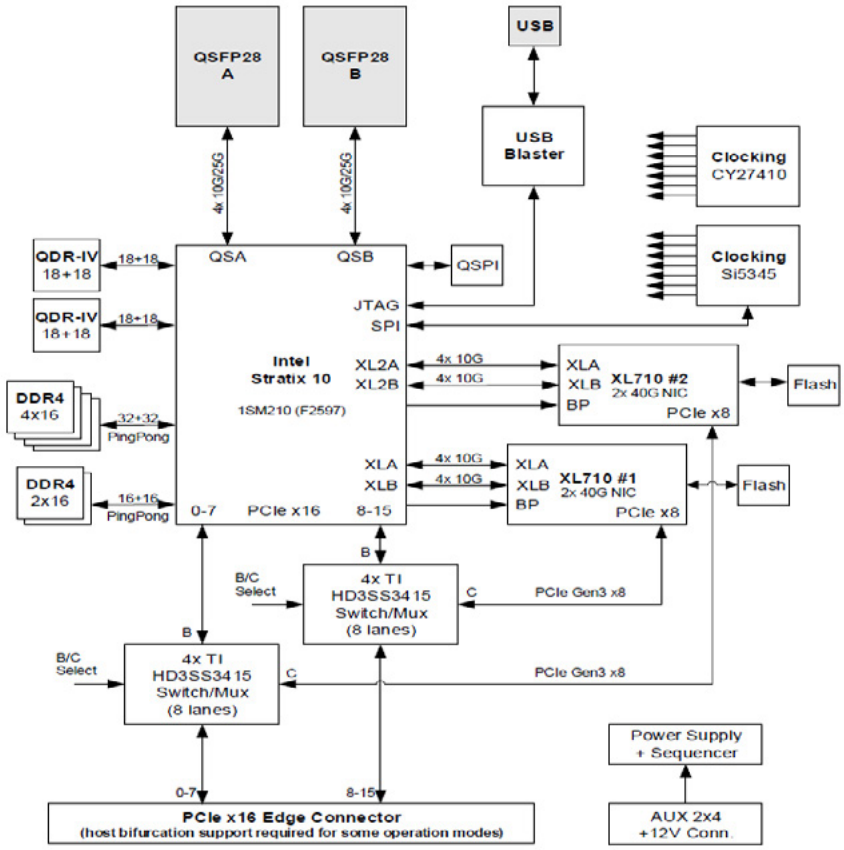
Na rynku dostępnych jest wiele narzędzi przeznaczonych do generowania i testowania ruchu sieciowego, zarówno komercyjnych, jak i open source. Wśród rozwiązań komercyjnych szczególną popularnością cieszą się produkty takich firm jak Keysight Ixia [7], Spirent Communications [6] czy Xena Networks [8], które oferują zaawansowane platformy do symulacji ruchu w sieciach o wysokiej przepustowości. Inne narzędzia, takie jak PacketExpert firmy GL Communications [12] czy Threat Generator od Palo Alto Networks [9], umożliwiają tworzenie realistycznych scenariuszy testowych obejmujących m.in. ruch złośliwy, ataki typu DDoS czy emulację aplikacji. Obok komercyjnych rozwiązań funkcjonują również otwarte platformy, takie jak Ostinato [10] czy TRex [11], które są chętnie wykorzystywane w środowiskach akademickich i przez entuzjastów technologii ze względu na elastyczność, możliwość modyfikacji oraz brak opłat licencyjnych.

3. ZAŁOŻENIA

Projekt zakładał zrealizowanie generatora ruchu sieciowego opartego o kartę prototypową Intel MAP-200 (schemat blokowy karty został przedstawiony na rys. 1). Generator obsługuje prędkości 10Gbit/s, 25Gbit/s, 40Gbit/s oraz 100Gbit/s i może być dowolnie konfigurowany niezależnie na każdym z dwóch portów QSFP. Generator wspiera trzy warstwy modelu ISO/OSI tj. warstwę 2, warstwę 3 (IPv4), warstwę 4 (UDP).

Podawanie danych do generatora odbywa się za pośrednictwem plików konfiguracyjnych w formacie YAML, co zapewnia przejrzystość i łatwość edycji parametrów wejściowych. Aby uruchomić generowanie ruchu sieciowego, wymagane jest przygotowanie dwóch niezależnych plików konfiguracyjnych.

Pierwszy z nich odpowiada za definicję **zawartości strumieni pakietów** — użytkownik może zdefiniować aż **2688 unikalnych strumieni**, z których każdy może mieć



Rys. 1. Schemat blokowy karty MAP-200

odrębne parametry, takie jak adresy MAC i IP, porty źródłowe i docelowe, rozmiar pakietów, nagłówki protokołów. Tak duża liczba konfigurowalnych strumieni umożliwia realistyczne odwzorowanie złożonych scenariuszy testowych, takich jak równoległa transmisja danych od wielu nadawców czy symulacja zróżnicowanego obciążenia sieci.

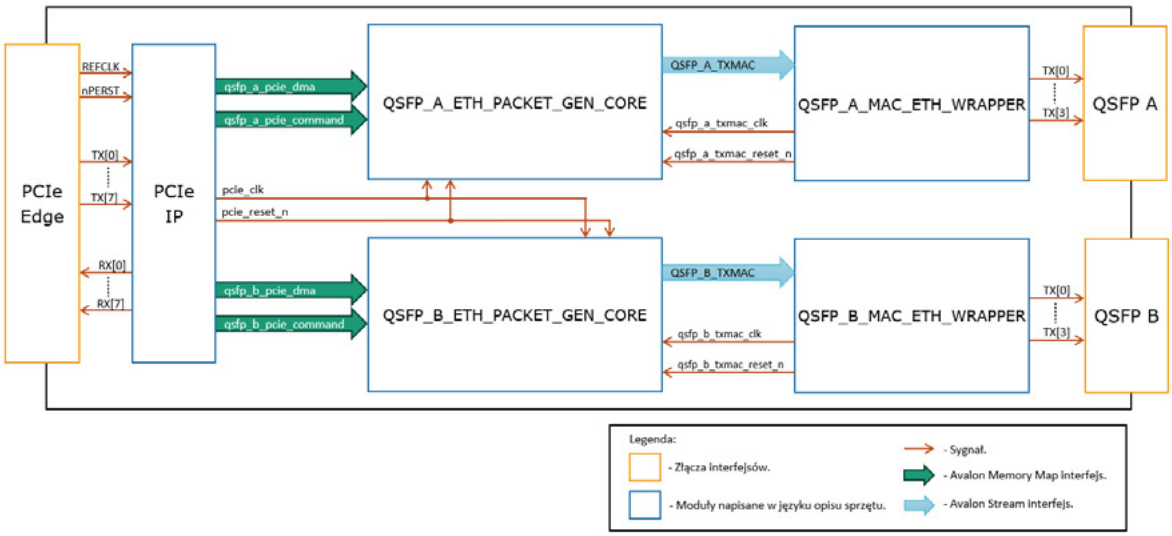
Drugi plik YAML zawiera **listę poleceń sterujących** – tzw. **komendy**, które umożliwiają precyzyjne zarządzanie sekwencją i warunkami wysyłania pakietów. System wspiera do **1024 niezależnych komend**, które mogą określać takie operacje jak rozpoczęcie lub zakończenie transmisji określonych strumieni, ustawienie liczby powtórzeń, długość trwania transmisji.

Taka dwuetapowa struktura konfiguracji zapewnia dużą elastyczność oraz możliwość łatwego skalowania scenariuszy testowych – od prostych przypadków po bardzo złożone środowiska z dużą liczbą zróżnicowanych strumieni i skomplikowaną logiką sterującą.

4. IMPLEMENTACJA

Generator został zaimplementowany przy użyciu języka opisu sprzętu SystemVerilog [2,4], co umożliwiło precyzyjne odwzorowanie architektury sprzętowej oraz elastyczność w definiowaniu parametrów konfiguracji. Do weryfikacji funkcjonalnej projektu na etapie symulacyjnym wykorzystano zaawansowany symulator **Siemens Questa** [15], który zapewnił dokładne i wydajne środowisko do testowania zachowania układu przed implementacją w docelowym układzie FPGA. Schemat blokowy generatora został przedstawiony na rys. 2 i obrazuje strukturę funkcjonalną systemu w podziale na trzy główne moduły logiczne: moduł PCIe (Peripheral Component Interconnect Express), rdzeń Ethernet Packet Generator Core oraz moduł Mac Ethernet Wrapper. Każdy z tych bloków pełni odrębną rolę i odpowiada za inny etap przetwarzania danych od momentu ich odbioru z systemu hosta, aż po transmisję gotowych ramek Ethernet w wybranym standardzie IEEE 802.3 [3].

Pierwszym z tych bloków jest **moduł PCIe**, będący kontrolerem interfejsu komunikacyjnego między systemem hosta a układem FPGA. Odpowiada on za odbieranie danych przesyłanych za pomocą szybkich różnicowych linii szeregowych PCI Express oraz ich konwersję do formatu równoległego, zgodnego z interfejsem **Avalon Memory Mapped**, wykorzystywanym wewnętrznie w architekturze układu FPGA. Dodatkowo w module PCIe został zaimplementowany **mechanizm DMA (Direct Memory Access)**, umożliwiający wydajne, bezpośrednie przesyłanie dużych porcji danych konfiguracyjnych i sterujących bez nadmiernego obciążania procesora hosta. W obecnej wersji projektu zastosowano konfigurację opartą o **PCIe generacji 3**, jednak z ograniczeniem do **ośmiu**



Rys. 2. Schemat blokowy generatora

różnicowych linii transmisyjnych, mimo że karta MAP-200 udostępnia fizycznie do szesnastu takich linii. W efekcie szerokość interfejsu danych dostępnego dla mechanizmu DMA wynosi **256 bitów**. Dla porównania, interfejs komend operuje na szynie o szerokości **32 bitów**, co w zupełności wystarcza do przesyłania prostych instrukcji sterujących. Moduł PCIe pochodzi z biblioteki standardowych komponentów udostępnianych w środowisku projektowym **Intel Quartus** i jest kompatybilny z resztą platformy.

Drugim, kluczowym modułem generatora jest **Ethernet Packet Generator Core**, którego wewnętrzna struktura została zaprezentowana na rys. 3. Jego głównym zadaniem jest przetwarzanie danych konfiguracyjnych dostarczanych przez użytkownika w celu generowania strumieni pakietów Ethernet zgodnych z wymaganiami testowymi. Struktura tego bloku jest podzielona na trzy główne sekcje. Pierwszą z nich jest **PCIe Wrapper**, czyli moduł odbierający dane z interfejsów Avalon MM – osobno dla konfiguracji strumieni i komend sterujących – i zapisujący je do odpowiednich przestrzeni pamięci RAM. Dane są buforowane w pamięciach podręcznych, z których korzystają kolejne komponenty. Drugą sekcję stanowi **Packet Generator Engine**, który na podstawie otrzymanych komend oblicza odpowiednie parametry transmisji i generuje dane odpowiadające zawartości pakietów. W ramach tej sekcji uwzględniane są takie parametry jak liczba pakietów, ich długość, zawartość nagłówków czy kolejność transmisji. Trzecia sekcja odpowiada za **budowę i enkapsulację pakietów sieciowych**. Składa się ona z łańcucha modułów, które kolejno tworzą warstwy pakietu: dane użytkownika, warstwę transportową (UDP), warstwę sieciową (IPv4), a następnie warstwę łącza danych (Ethernet). Ostatnim elementem tej sekcji jest moduł **Avalon ST Resizer**, który dokonuje zmiany szerokości danych z wewnętrznego formatu 64-bajtowego na format zgodny z wymaganiami danego interfejsu MAC, do którego pakiety będą kierowane.

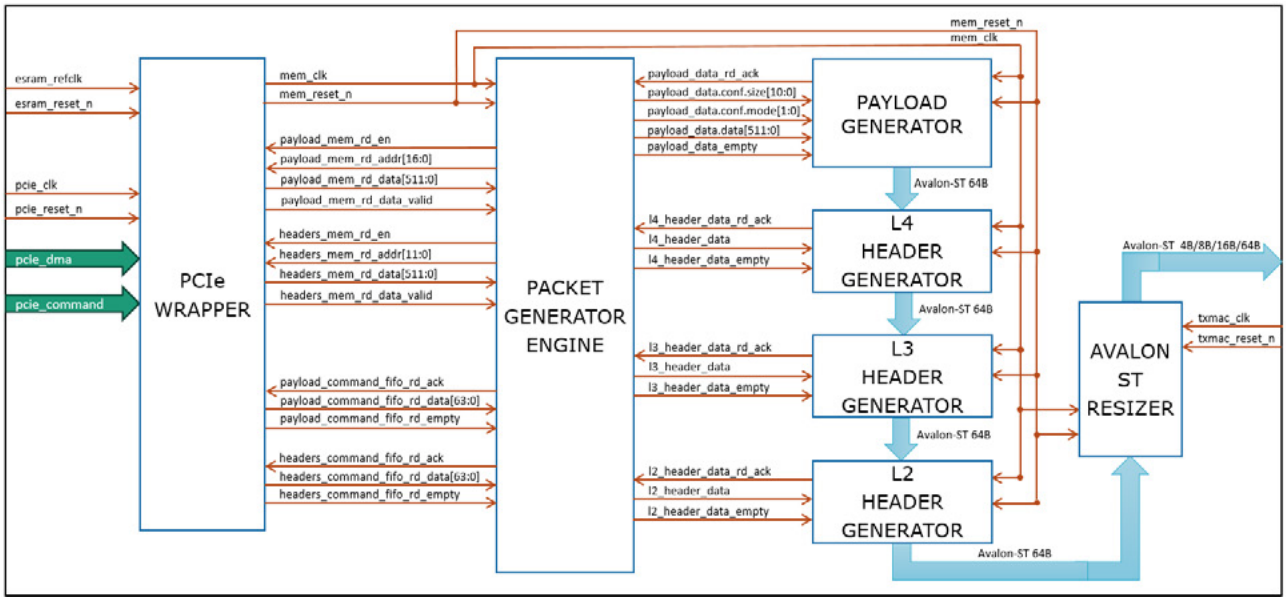
Trzecim i końcowym blokiem w strukturze generatora jest **Mac Ethernet Wrapper**. Moduł ten zawiera zestaw bloków **MAC (Media Access Controller)** przystosowanych do obsługi wszystkich obsługiwanych przepływności: 10Gbit/s, 25Gbit/s, 40Gbit/s oraz 100Gbit/s. Wybór odpowiedniego wariantu MAC odbywa się na etapie syntezy projektu, za pomocą parametru konfiguracyjnego ustawianego w pliku top-level. Moduły MAC pochodzą z biblioteki komponentów oferowanych przez środowisko **Intel Quartus Prime** i są zgodne ze standardami IEEE. Moduł ten odpowiada za końcową fazę transmisji danych – konwersję pakietów na odpowiedni format linii transmisyjnej oraz synchronizację z zegarem fizycznego interfejsu transmisyjnego.

Całościowa struktura generatora została zaprojektowana w sposób modułowy i skalowalny, co umożliwia w przyszłości na łatwą rozbudowę o nowe funkcje, takie jak obsługa dodatkowych protokołów czy wyższych szybkości transmisji.

Projekt został zaimplementowany przy użyciu oprogramowania Quartus firmy Intel. W skład procesów, które składają się na pełny proces implementacji wchodzi synteza logiczna, synteza fizyczna, place&rout, analiza czasowa oraz generowanie pliku binarnego zawierającego konfigurację docelową układu FPGA. Należy dodać, iż współczesne metodologie projektowe uwzględniają również zastosowanie elementów bibliotecznych takich jak moduły MAC (Media Access Controller) dla wszystkich użytych w projekcie prędkości ETH oraz moduł PCIe umożliwiający realizację takich funkcji jak konfiguracja urządzenia poprzez rejestry oraz transfer danych bezpośrednio do pamięci hosta za pomocą mechanizmu DMA (Direct Memory Access). W tabeli 1 zamieszczono ilość wykorzystanych zasobów układu FPGA Stratix 10 MX w zrealizowanym konfigurowalnym generatorze ruchu sieciowego.

Tab. 1. Wykorzystanie zasobów w układzie FPGA

Zasób	*Ilość*	*% zajętości*
ALM	109 528	15.6
Bity pamięci	16.6 Mbit	11.9
Bloki RAM	1102	16.1
PLL	25	13.9

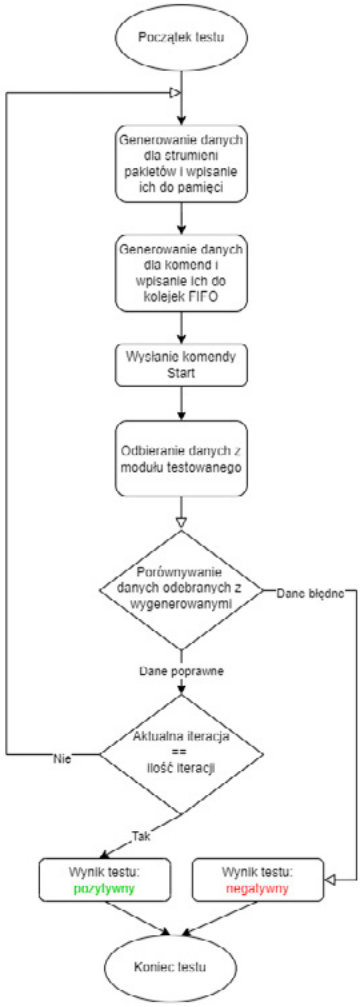


Rys. 3. Schemat blokowy modułu Ethernet Packet Generator Core

5. TESTOWANIE

Generator ruchu sieciowego został przetestowany w symulacji oraz fizycznie na karcie MAP-200. W przypadku testowania metodą symulacji wykorzystano gotowy framework SVUnit od AgileSoC [14], który wspiera automatyczne testowanie. Do każdego modułu został utworzony osobny test (rys. 4), aby sprawdzić jego poprawność działania. Pierwszym etapem testowania było wygenerowanie w sposób losowy danych parametrów dla strumieni pakietów. Generowano całą pamięć zawierającą opis 2688 strumieni. Kolejnym etapem było także losowo wygenerowanie komend w liczbie losowanej z przedziału 1000-1024. Każda komenda miała losowo wygenerowane parametry początkowego strumienia pakietu, liczby inkrementalnych strumieni pakietów następujących po sobie oraz liczbę powtórzeń danej sekwencji. Po wygenerowaniu danych wysyłana była komenda startu, aby uruchomić mechanizm przetwarzania danych tak, aby wygenerować strumienie danych interfejsu Avalon Memory Map. Test odbierał dane od modułu testowego i porównywał je z wygenerowanymi danymi bajt po bajcie w poszukiwaniu różnic. Jeśli różnice zostały wykryte test automatycznie się kończył z wynikiem negatywnym wskazując wykryte błędy. Testowanie w symulacji przeprowadzono dla każdej dostępnej dla generatora szybkości przesyłania danych oraz pełną przepustowością. Jedynym ograniczeniem był czas trwania symulacji ponieważ, aby przesymulować 1 sekundę działania urządzenia potrzeba 3 tygodni. Po uruchomieniu wszystkich testów w regresji i otrzymaniu wyniku pozytywnego można było przejść do testowania generatora w sprzęcie.

Testowanie w sprzęcie polegało na podłączeniu generatora ruchu sieciowego do kart sieciowych o różnych obsługiwanych szybkościach aby sprawdzić poprawność przesyłanych pakietów. Dane odbierane były przy pomocy aplikacji Wireshark. Przetestowano zarówno losowo wygenerowane komendy jak i dane strumieni, a także warunki brzegowe tj. pakiety o długości payloadu poniżej 64B oraz duże pakiety 1500B. Każdy parametr był generowany w każdej możliwej konfiguracji metadanych.



Rys. 4. Diagram sekwencji testowej

6. PODSUMOWANIE

W dobie gwałtownego wzrostu liczby użytkowników oraz zwiększającego się zapotrzebowania na wysokoprzepustowe usługi sieciowe, takie jak strumieniowanie wideo w jakości 4K, komunikacja w czasie rzeczywistym (VoIP) czy obsługa chmur obliczeniowych, testowanie infrastruktury sieciowej staje się kluczowe dla zapewnienia ciągłości i niezawodności pracy sieci.

Technologia FPGA, dzięki swojej elastyczności i możliwości programowania umożliwia tworzenie dedykowanych rozwiązań o wysokiej wydajności, zdolnych do generowania symulowanego ruchu sieciowego z prędkościami od 10Gbit/s, przez 25Gbit/s i 40Gbit/s, aż do 100Gbit/s. Zrealizowany generator umożliwia dokładną kontrolę nad charakterystyką generowanego ruchu, umożliwiając symulację różnych typów obciążeń sieciowych, takich jak ruch typowy dla użytkowników końcowych, jak również ruch intensywny, charakterystyczny dla ataków DDoS (Distributed Denial of Service). Narzędzie to jest również nieocenione w testowaniu wydajności i skalowalności urządzeń sieciowych, takich jak rutery, przełączniki czy firewalle, a także w ocenie systemów zapewniających jakość usług (QoS).

Konfigurowalny generator ruchu sieciowego o wysokich przepustowościach jest kluczowym narzędziem do testowania, monitorowania i oceny wydajności sieci w wielu różnych scenariuszach, od infrastruktury korporacyjnej po usługi dostawców internetowych oraz aplikacje wymagające dużych zasobów sieciowych. Umożliwia symulację rzeczywistych warunków pracy oraz optymalizację systemów przed ich wdrożeniem.

Wyniki testów potwierdzają zdolność generatora do precyzyjnego odwzorowywania ruchu sieciowego przy zachowaniu wysokiej wydajności, nawet w przypadku dużych przepustowości. Generator ruchu umożliwia dokładne testowanie stabilności infrastruktury sieciowej, a także na analizę wpływu intensywnego ruchu na wydajność i niezawodność urządzeń sieciowych oraz aplikacji.

Opracowane narzędzie może znaleźć szerokie zastosowanie w różnych dziedzinach, od centrów danych i usług chmurowych, przez testowanie infrastruktury dostawców usług internetowych (ISP), po rozwój i testowanie nowoczesnych technologii sieciowych, takich jak sieci 5G. Dzięki zastosowaniu FPGA jako podstawy architektury możliwe jest szybkie adaptowanie generatora do zmieniających się wymagań oraz scenariuszy testowych, co czyni go wszechstronnym narzędziem do badań i testów sieciowych.

Aktualnie rozwiązanie to jest przenoszone na produkcyjną kartę **Intel® Stratix 10 MX** [13]. Równocześnie opracowywana jest dedykowana aplikacja webowa, która umożliwi wygodne zarządzanie generatorem przez przeglądarkę internetową, m.in. konfigurację parametrów pakietów (takich jak rozmiar, prędkość wysyłania, typy ramek) oraz monitoring aktualnych statystyk transmisji. System webowy będzie komunikował się z modułem FPGA poprzez warstwę pośrednią działającą na systemie operacyjnym hosta. Możliwa będzie również obsługa wielu kart w ramach jednej instancji aplikacji.

Równolegle prowadzone są prace nad modułem analizy pakietów, który umożliwi odbiór i interpretację przesyłanych danych w czasie rzeczywistym. Dzięki temu rozwiązaniu możliwe będzie nie tylko generowanie, lecz także kontrola i weryfikacja przesyłanych ramek, co znacząco zwiększy funkcjonalność całego systemu i umożliwi jego zastosowanie w bardziej zaawansowanych scenariuszach testowych.

LITERATURA

[1] Ashenden P. J. 2008. "The Designer's Guide to VHDL" (3rd ed.). *Morgan Kaufmann*.
[2] Ciletti M. D. 2004. "Advanced Digital Design with the Verilog HDL". *Prentice Hall*.
[3] IEEE Computer Society. 2018. "IEEE Standard for Ethernet. IEEE Std 802.3-2018". https://standards.ieee.org (data dostępu 14.09.2024).
[4] Stuart Sutherland S., S. Davidmann & P. Flake. 2010. "SystemVerilog for Design: A Guide to Using SystemVerilog for Hardware Design and Modeling" (2nd ed.). *Springer*.
[5] Wolf W. 2004. "FPGA-Based System Design". *Prentice Hall*.
[6] https://www.spirent.com/products/testcenter-hardware (data dostępu 20.05.2025).
[7] https://www.keysight.com/us/en/products/network-test/network-test-hardware.html (data dostępu 20.05.2025).
[8] https://xenanetworks.com/products (data dostępu 20.05.2025).
[9] https://docs.paloaltonetworks.com/advanced-threat-prevention (data dostępu 20.05.2025).
[10] https://userguide.ostinato.org (data dostępu 20.05.2025).
[11] https://trex-tgn.cisco.com (data dostępu 20.05.2025).
[12] https://www.gl.com/voip-ethernet-test-tools.html (data dostępu 20.05.2025).
[13] https://www.intel.com/content/www/us/en/products/details/fpga/development-kits/stratix/10-mx.html (data dostępu 20.05.2025).
[14] https://agilesoc.com/open-source-projects/svunit (data dostępu 20.05.2025).
[15] https://eda.sw.siemens.com/en-US/ic/questa-one/simulation/questa-one-sim/ (data dostępu 20.05.2025).

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

Poprawa dokładności określenia głębokości akwenu za pomocą echosondy interferometrycznej

Improving the Accuracy of Water Depth Determination Using an Interferometric Echosounder

Piotr Grall[1]

[1] Politechnika Gdańska, piograll@pg.edu.pl

DOI: 10.15199/59.2025.4.62

Streszczenie: W artykule przedstawiono metodę poprawy dokładności określania głębokości akwenu za pomocą echosondy interferometrycznej. W pracy zaprezentowano zastosowanie Zmodyfikowanej Metody Prony'ego (ZMP), która umożliwia poprawę dokładności wyznaczenia kierunków sygnałów odbitych, jednak jej skuteczność zależy od znajomości liczby źródeł echa. Aby umożliwić praktyczne wykorzystanie ZMP, zastosowano opracowaną wcześniej oryginalną metodę estymacji liczby sygnałów echa. Bazuje ona na analizie wartości osobliwych macierzy utworzonej z sygnałów odbiorczych oraz oszacowaniu poziomu szumu i współczynnika korelacji przestrzennej sygnału. Skuteczność podejścia potwierdzono na danych rzeczywistych pochodzących z pomiarów batymetrycznych wykonanych w porcie Everglades w USA.

Abstract: The article presents a method for improving the accuracy of water depth determination using an interferometric echosounder. The study introduces the application of the Modified Prony Method (MP), which enhances the accuracy of estimating the directions of reflected signals. However, its effectiveness depends on knowing the number of echo sources.To enable practical application of MP, an original method for estimating the number of echo signals has been developed. It is based on the analysis of singular values of a matrix constructed from the received signals, as well as on the estimation of the noise level and the spatial correlation coefficient of the signal. The effectiveness of the approach was confirmed using real bathymetric measurement data collected in the Port of Everglades, USA. The results obtained using the classical Least Squares (LS) method were compared with those from the MP method supported by the developed echo count estimation technique.

Słowa kluczowe: metoda Prony'ego, metody wysokiej rozdzielczości,określanie kierunku.

Keywords: Direction of Arrival, Model-Based Estimation, Prony Method.

1. WSTĘP

Obecnie do określenia rozkładu głębokości w akwenach stosuje się najczęściej echosondy wielowiązkowe lub interferometryczne. Urządzenia te różnią się zasadniczo sposobem działania. W echosondzie wielowiązkowej stosuje się technikę formowania odbiorczej wiązki akustycznej (tzw. beamforming) i estymowana jest odległość echa dla każdej wiązki (typowo 120-400 wiązek) co pozwala uzyskać punkty estymujące położenie dna. Natomiast, w echosondzie interferometrycznej, dla danego zakresu obserwacji R, estymuje się kierunki ech kilku źródeł odbicia. Operacja ta powtarzana jest dla kolejnych zestawów próbek sygnału odebranego otrzymując w ten sposób estymacje położenia dna (typowo 2000-5000 punktów). Wyznaczenie kierunków echa w echosondzie interferometrycznej wymaga zastosowania metod wysokiej rozdzielczości w celu określenia parametrów przyjętego modelu sygnału odebranego. Podstawową metodą określenia parametrów modelu sygnału odebranego jest metoda najmniejszych kwadratów (ang. *Least Square* - LS)[4]. Inne metody, takie jak Zmodyfikowana Metoda Prony'ego [8], TLS [5], root-MUSIC [6] czy metody z grupy ESPRIT [10], pozwalają na poprawę dokładności określenia kierunku przez echosondę interferometryczną w stosunku do metody najmniejszych kwadratów, wymagają one jednak znajomości liczby sygnałów echa przed wykonaniem obliczeń. Ze względu na zmienne właściwości sygnału echa, określenie liczby sygnałów echa, dla każdego zestawu odebranych sygnałów, jest zagadnieniem złożonym, gdyż liczba ta nie jest znana *a priori* i musi zostać oszacowana w trakcie obliczeń. W niniejszym artykule zaprezentowana zostanie metoda określenia liczby sygnałów echa, która w połączeniu ze Zmodyfikowaną Metodą Prony'ego pozwala na poprawę dokładności określenia kierunku echosondy interferometrycznej, w stosunku do dokładności otrzymywanej za pomocą metody najmniejszych kwadratów.

2. WYZNACZENIE KIERUNKU W ECHOSONDZIE INTERFEROMETRYCZNEJ

W echosondzie interferometrycznej przyjmuje się następujący model sygnału [3, 4, 7]:

1. Liniowa antena odbiorcza składa z N równo oddalonych od siebie elementów i jest wykorzystana do pomiaru sygnałów ech, rozchodzących się w tej samej płaszczyźnie co antena odbiorcza;
2. W danej chwili dokładnie M niezależnych od siebie, płaskich fal akustycznych, dociera do każdego elementu odbiorczego;
3. Sygnały docierające do elementów odbiorczych są wąskopasmowe;

4. Sygnał na wyjściu każdego elementu odbiorczego jest w stanie ustalonym (składowe przejściowe są pomijalnie małe).

Przy spełnieniu powyższych założeń i przyjęciu, że tłumienie sygnału pomiędzy elementami odbiorczymi jest pomijalnie małe sygnały docierające do każdego elementu odbiorczego n, w danej chwili (dla danego zasięgu R) można przedstawić w następującej postaci (rys. 1):

$$s(n) = \sum_{i=1}^{M} a_i \, e^{j\,\omega_i n} + w(n), \tag{1a}$$

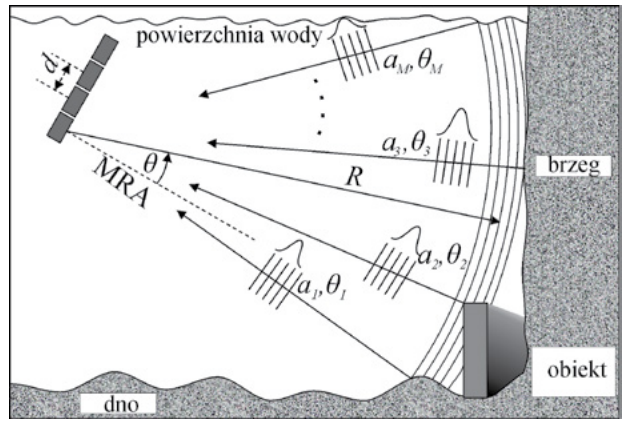
$$\omega_i = u_i \, d, \tag{1b}$$

$$u_i = k \sin\theta_i, \tag{1c}$$

$$n = 0,\, 1,\, 2,\, \cdots,\, N-1;\; k = \frac{2\pi}{\lambda};\; \lambda = \frac{c}{f_c}, \tag{1d}$$

$$\theta_i = \arcsin(\frac{\omega_i}{d\,k}), \tag{1e}$$

gdzie: $s(n)$ – sygnał w n-tym elemencie odbiorczym w danej chwili, n – numer porządkowy elementu odbiorczego, θ_i – kierunek propagacji echa sygnału mierzony względem osi akustycznej anteny (*ang. Main Response Axis - MRA*), ω_i – unormowana pulsacja, a_i – zespolony sygnał echa z kierunku θ_i w referencyjnym elemencie odbiorczym ($n = 0$), u_i – pulsacja przestrzenna sygnału echa z kierunku kierunku θ_i, k – liczba falowa, λ – długość fali akustycznej, f_c – częstotliwość fali nośnej, d – odległość między elementami anteny odbiorczej, c – prędkość dźwięku w wodzie, $w(n)$ szum akustyczny.



Rysunek 1: Model sygnału odebranego w echosondzie interferometrycznej.

Bazując na tym modelu można sformułować następujący układ równań liniowych (jest to metoda najmniejszych kwadratów w sformułowaniu liniowej predykcji w przód i wstecz) [9]:

$$\begin{bmatrix} s(L-1) & s(L-2) & \cdots & s(0) \\ s(L) & s(L-1) & \cdots & s(1) \\ \vdots & \vdots & \ddots & \vdots \\ s(N-2) & s(N-3) & \cdots & s(N-L-1) \\ s^*(1) & s^*(2) & \cdots & s^*(L) \\ s^*(2) & s^*(3) & \cdots & s^*(L+1) \\ \vdots & \vdots & \ddots & \vdots \\ s^*(N-L) & s^*(N-L+1) & \cdots & s^*(N-1) \end{bmatrix} \begin{bmatrix} \hat{g}_1 \\ \hat{g}_2 \\ \vdots \\ \hat{g}_L \end{bmatrix} = \begin{bmatrix} s(L) \\ s(L+1) \\ \vdots \\ s(N-1) \\ s^*(0) \\ s^*(1) \\ \vdots \\ s^*(N-L-1) \end{bmatrix}, \tag{2}$$

lub w skróconej wersji:

$$\mathbf{A}\hat{\mathbf{g}} = \mathbf{b}. \tag{3}$$

Z powyższego układu równań należy wyznaczyć wektor współczynników $\hat{\mathbf{g}}$, a następnie podstawić je do równania [8]:

$$1 - \sum_{i=1}^{L} g_i z^{-i} = 0. \tag{4}$$

Równanie to ma L rozwiązań na płaszczyźnie zespolonej postaci:

$$z_i = A \cdot e^{j\omega_i}, \tag{5}$$

a unormowaną pulsację obliczamy ze wzoru:

$$\omega_i = \arg z_i, \tag{6}$$

i wykorzystujemy ją od obliczenia kierunku echa ze wzoru (1e). W ten sposób otrzymujemy pary w biegunowym układzie współrzędnych (R,θ), gdzie kąt θ obliczany jest względem MRA.

W procedurze tej pojawiają się dodatkowe echa, gdy $L > M$, które należy odrzucić w procesie selekcji ech.

Równanie (3) można rozwiązać metodą najmniejszych kwadratów wykorzystując rozkład SVD macierzy $\mathbf{A}$:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathbf{H}}, \tag{7}$$

gdzie: $\mathbf{\Sigma}$ – macierz diagonalna wartości osobliwych, $\mathbf{\Sigma} = diag(\sigma)$ o wymiarach $(L \times L)$, σ – wektor wartości osobliwych σ_i macierzy $\mathbf{A}$, $i \in \langle 1, L\rangle$, $\mathbf{U}, \mathbf{V}$ – macierze unitarne o wymiarach $2(N-L) \times 2(N-L)$ i $L \times L$.

Rozwiązanie równania (3) przyjmuje wówczas postać [5]:

$$\hat{\mathbf{g}} = \sum_{i=1}^{L} \frac{1}{\sigma_i} \left(\mathbf{u_i^H b}\, \mathbf{v_i}\right), \tag{8}$$

gdzie: $\mathbf{u_i}$ i $\mathbf{v_i}$ są kolejnymi wektorami macierzy $\mathbf{U}$ oraz $\mathbf{V}$, a σ_i są uporządkowanymi malejąco, dodatnimi wartościami osobliwymi i znajdują się na głównej

przekątnej macierzy $\mathbf{\Sigma}$.

W Zmodyfikowanej Metodzie Prony'ego odrzuca się najmniejsze wartości osobliwe z równania (8), przez co usuwa się częściowo wpływ szumu na wartości elementów macierzy $\mathbf{A}$ i tym samym poprawia się dokładność estymacji kierunku θ [8].

$$\tilde{\mathbf{g}} = \sum_{i=1}^{M} \frac{1}{\sigma_i} \left(\mathbf{u_i^H b\, v_i} \right). \tag{9}$$

Amplitudy poszukiwanych sygnałów echa wyznacza się na podstawie obliczonych kierunków w osobnej procedurze [4].

3. METODA OKREŚLENIA LICZBY SYGNAŁÓW ECHA

Najtrudniejszym zagadnieniem dla zastosowania Zmodyfikowanej Metody Prony'ego jest określenie liczby sygnałów M echa dla każdego zasięgu obserwacji R. W tym celu opracowana została metoda, która na podstawie amplitudy szumu, wymiarów macierzy $\mathbf{A}$ oraz współczynnika korelacji sygnałów odebranych, pozwala na obliczenie minimalnej wartości osobliwej σ_l macierzy $\mathbf{A}$, którą uznajemy za sygnał echa [1]. W metodzie tej wartość graniczną σ_l oblicza się z wzoru:

$$\sigma_l = r\, n_g\, \sigma_n, \tag{10a}$$

$$r = \sqrt{4(N-L)L}, \tag{10b}$$

$$n_g = \sqrt{\frac{d_{SNR}}{d_d + 1}}, \tag{10c}$$

$$d_d = \frac{\gamma_d}{1-\gamma_d},\ d_{SNR} = \frac{\sigma_s^2}{\sigma_n^2}, \tag{10d}$$

,

gdzie: r – mnożnik zależny od wymiarów macierzy $\mathbf{A}$. n_g – współczynnik wzmocnienia szumu, σ_n – odchylenie standardowe szumu, σ_s – odchylenie standardowe sygnału odebranego, d_d – stosunek sygnału do szumu dekorelacji,d_{SNR} – stosunek sygnału do szumu, γ_d – współczynnik korelacji sygnału odebranego. Liczba wartości σ_i macierzy $\mathbf{A}$ większych od wyznaczonej wartości σ_l jest poszukiwaną liczba sygnałów M. Współczynnik r jest mnożnikiem, który stosuje się w celu oszacowania maksymalnej wartości osobliwej macierzy o danych wymiarach, jeżeli jej elementy stanowią jedynie próbki szumu. W opracowanej metodzie zakłada się, że szum występujący w macierzy $\mathbf{A}$ jest sumą szumu tła oraz tzw. szumu dekorelacji, który wynika z przestrzennych zależności pomiędzy źródłem sygnału echa i anteną odbiorczą. Zjawisko to pogarsza rzeczywisty stosunek sygnału do szumu w stosunku do nominalnego d_{SNR}. Można powiedzieć, że poziom szumu tła zostaje wzmocniony, na skutek występowania zjawiska dekorelacji. Sumaryczny wpływ dekorelacji przestrzennej szacowany jest za pomocą współczynnika korelacji sygnału γ_d.

W opracowanej metodzie wartość tego współczynnika szacuje się dwuetapowo [2]. Najpierw określa się przebieg dna za pomocą metody LS, a następnie dokonuje się podziału anteny odbiorczej na dwie podanteny. Formuje się następnie dwie wiązki odbiorcze, poprzez odpowiednie opóźnienie sygnałów w elementach odbiorczych, tak aby osie wiązek w każdej chwili skierowane były w kierunku echa od dna. Następnie określa się lokalny współczynnik korelacji (tj. odpowiadający rozmiarom fragmentu dna, który w danym momencie jest źródłem echa) pomiędzy tak wyznaczonymi sygnałami z podanten, wykorzystując odpowiednią liczbę kolejnych próbek K.

$$\gamma_d = \left| \frac{cov\left(s_1(K), s_2(K)\right)}{\sqrt{var\left(s_1(K)\right) var\left(s_2(K)\right)}} \right|. \tag{11}$$

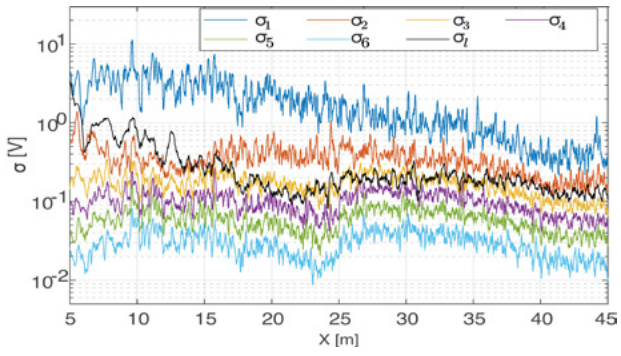
4. WERYFIKACJA METODY Z WYKORZYSTANIEM DANYCH Z POMIARÓW W WARUNKACH RZECZYWISTYCH

W celu zbadania dokładności zaproponowanej metody wykorzystano dane zarejestrowane z pomocą echosondy interferometrycznej EdgeThech 6205. Urządzenie składa się z dwóch głowic pracujących przemiennie na częstotliwości 550 kHz. Każda głowica zbudowana jest z 10 elementów odbiorczych. Do badań wykorzystano dane pomiarowe zarejestrowane w dniu 13.12.2016 r. w Porcie Everglades, w stanie Miami, USA. Dane uzyskano dzięki uprzejmości producenta urządzenia firmy EgdeTech. Rejestrację przeprowadzono w basenie portowym, o głębokości 14m. Głowica urządzenia umieszczona była na głębokości 1m. Odległość między profilami pomiarowymi wynosiła 20 m. Rejestrację przeprowadzono z zakresem obserwacji 50 m. Do wyznaczenia dokładności określania głębokości wybrano fragment dna – obszar o wymiarach 60 × 60 m, o płaskim ukształtowaniu dna. Wybrany fragment został przesondowany ośmiokrotnie – cztery raz za pomocą prawej głowicy i cztery razy za pomocą lewej głowicy, na kierunkach profili 000°, 090°, 180°, 270°. Z każdego profilu wybrano 200 zarejestrowanych sygnałów echa obejmujących swoim zakresem wybrany obszar (rys.2). Otrzymano w ten sposób ponad 5,3 mln punktów pomiarowych głębokości dna. Na podstawie punktów pomiarowych oszacowano dokładności otrzymanego pomiaru głębokości za pomocą metod LS oraz ZMP, obszarów dna o szerokości 1 m, a następnie uśredniono wyniki dla wszystkich 200 impulsów, kolejnych profili pomiarowych dla poszczególnych głowic. Dla metody ZMP dokonano oszacowania liczby sygnałów echa za pomocą opracowanej metody. Przykład wyznaczenia wartości granicznej σ_l zaprezentowano na rys. 3.

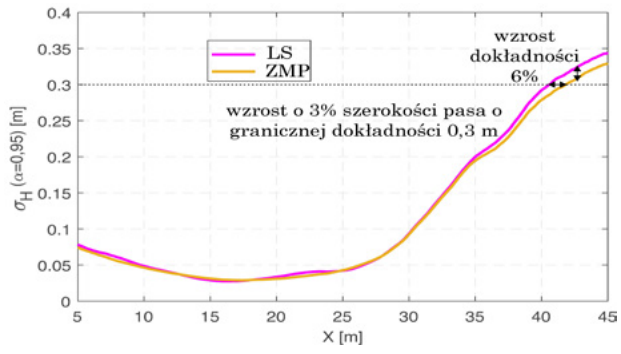
Zarówno dla prawej, jak i lewej głowicy uzyskano podobną poprawę dokładności określenia głębokości na skraju zasięgu obserwacji (wyniki zaprezentowano jedynie dla prawej głowicy rys. 4). Zastosowanie

zaproponowanej metody pozwoliło na wzrost dokładności określania pozycji na skraju zakresu obserwacji o 5-6%, a co za tym idzie zwiększenie szerokości pasa dna o akceptowalnej dokładności o 3-4%. W rezultacie zastosowana metoda może prowadzić do wzrostu wydajności czasu prac hydrograficznych, pozwalając na zwiększenie odległości pomiędzy profilami pomiarowymi tak, aby uzyskane dane pomiarowe nadal spełniały przyjęte kryteria dokładności.

Rysunek 2: Rejon pomiarów oraz zrealizowane przejścia sondażowe. Czerwony prostokąt oznacza rejon wybrany do badań.



Rysunek 3: Przykład określenia wartości granicznej σ_l dla jednego cyklu odbiorczego w funkcji odległości poprzecznej X.



Rysunek 4: Dokładność określenia głębokości H w funkcji odległości poprzecznej X dla prawej głowicy.

5. PODSUMOWANIE

Oryginalność zaproponowanej metody polega na wykorzystaniu informacji o rzeczywistej liczbie sygnałów echa (w metodzie LS informacja ta nie jest wykorzystywana). Opracowana metoda pozwala na oszacowanie liczby sygnałów echa, dzięki czemu możliwe jest częściowe usunięcie wpływu zakłóceń na dokładność określenia kierunku echa poprzez zastosowanie Zmodyfikowanej Metody Prony'ego. Zwiększenie dokładności określenia kierunku przekłada się, z kolei, na wzrost dokładności określenia głębokości i w rezultacie na zwiększenie szerokości akceptowalnego pasa przeszukiwania. Zastosowanie opracowanej metody może tym samym prowadzić do wzrostu wydajności pomiarów głębokości. Należy również podkreślić, że poprawa dokładności została uzyskana jedynie przez zmianę algorytmu przetwarzania sygnału, bez zmiany parametrów pracy anteny nadawczo-odbiorczej echosondy.

LITERATURA

[1] Grall Piotr. 2023. Rozprawa doktorska „Określanie liczby sygnałów echa w zmodyfikowanej metodzie Prony'ego w zastosowaniu do echosondy interferometrycznej". Politechnika Gdańska.

[2] Grall Piotr, Marszal Jacek. 2021. „Method for the correlation coefficient estimation of the bottom echo signal in the shallow water application using interferometric echo sounder". Vibrations in Phisical Systems, 32 (1) :1–6.

[3] Kraeutner P. H. 1998. Ph. D. Thesis „Small aperture acoustic imaging using model based array signal processing". Simon Fraser University.

[4] Kraeutner P. H., Bird J. S. 1997. „Principal components array processing for swath acoustic mapping". Proceedings of Oceans'97 : 1246–1254.

[5] Rahman M. A. ,Yu K. B. 1987. „Total least squares approach for frequency estimation using linear prediction". IEEE Trans. Acoustics, Speech, Signal Processing, ASSP-35 (10) :1440– 1454.

[6] Rao B. D. , Arun K. S. 1992. „Model based processing of signals:a state space approach". Proceedings of the IEEE, 80 (2) :283–309.

[7] Ribeiro M. P., Ewins D. J., Robb D. A. 2003. „Non-stationary analysis and noise filtering using a technique extended from the original prony method". Mechanical Systems and Signal Processing, 17 (3) :533–549.

[8] Tufts D. W., Kumaresan R. 1982. „Estimation of frequencies of multiple sinusoids: making linear prediction perform likemaximum likelihood". Proceedings of the IEEE, 70(9) :975–989.

[9] Ulrych T. J., Clayton R. W. 1976. „Time series modelling and maximum entropy". Phys. Earth Planet. Interiors,12 :188–200.

[10] van der Veen A.-J. , Swindlehurst A. L. 1993. „Subspace-based signal analysis using singular value decomposition". Proceedings of the IEEE, 81(9) :1277–1308.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

ALGORYTM ZNAJDOWANIA ZER I BIEGUNÓW FUNKCJI ZESPOLONEJ W ANALIZIE PROPAGACJI FAL ELEKTROMAGNETYCZNYCH

THE ALGORITHM FOR FINDING ZEROS AND POLES OF A COMPLEX FUNCTION IN ELECTROMAGNETIC WAVE PROPAGATION ANALYSIS

Sebastian Dziedziewicz[1]

[1] Politechnika Gdańska, Gdańsk, sebastian.dziedziewicz@pg.edu.pl

DOI: 10.15199/59.2025.4.63

Streszczenie: Niniejsza praca jest omówieniem fragmentów rozprawy doktorskiej autora. Przedstawia ona algorytm służący do znajdowania zer i biegunów funkcji zespolonej w przestrzeni $\mathbb{C} \times \mathbb{R}$ tj. zależnej od dwóch argumentów (jednego zespolonego i jednego rzeczywistego). Opracowana technika opiera się na uogólnionej zasadzie argumentu Cauchy'ego i trójwymiarowej triangulacji Delaunaya. Skuteczność metody została potwierdzona na kilku przykładach numerycznych obejmujących analizę propagacji fal w strukturach elektromagnetycznych.

Abstract: This paper is a discussion of excerpts from the author's doctoral dissertation. It presents an algorithm for finding zeros and poles of a complex function in the space $\mathbb{C} \times \mathbb{R}$, i.e. dependent on two arguments (one complex and one real). The developed technique is based on the generalization of Cauchy's argument principle and three-dimensional Delaunay triangulation. The effectiveness of the method has been confirmed in several examples involving the analysis of wave propagation in electromagnetic structures.

Słowa kluczowe: algorytm poszukiwania pierwiastków, zera i bieguny funkcji zespolonych, propagacja fal elektromagnetycznych, metody numeryczne

Keywords: root finding algorithm, zeros and poles of comples function, electromagnetic waves propagation, numerical methods

1. WSTĘP

Postępująca ewolucja technologii oraz coraz wyższe pasma częstotliwości w komunikacji bezprzewodowej wymagają złożonych i wydajnych narzędzi projektowania urządzeń elektronicznych. Zrozumienie działania nowoczesnych układów mikrofalowych wymusza badanie zjawisk elektromagnetycznych zarówno na poziomie makroskopowym, jak i kwantowym w materiałach. W efekcie konieczna jest analiza elektromagnetyczna polegająca na bezpośrednim rozwiązaniu równań Maxwella, co prowadzi do sformułowania problemu własnego w postaci macierzowej $\mathbf{M}(\lambda)\mathbf{u} = 0$ z odpowiednimi warunkami brzegowymi. Wartości własne, definiowane w dziedzinie zespolonej, opisują właściwości propagacyjne i radiacyjne struktur elektromagnetycznych oraz ich reakcję na pola elektromagnetyczne. Na przykład współczynnik propagacji γ jest zazwyczaj zespolony (mody prowadzone, wyciekające lub hybrydowe) i określa własności propagacyjne fal w danym ośrodku, natomiast wektor własny związany jest z rozkładem pola w strukturze. Rozwiązanie układu sprowadza się do znalezienia pierwiastków wyznacznika macierzy $\mathbf{M}$ i może być przedstawione jako:

$$F(z) = \det(\mathbf{M}(z)) = 0 \tag{1}$$

gdzie $z \in \mathbb{C}$ to miejsce zerowe funkcji $F(z)$ odpowiadające wartości własnej λ.

Istnieje wiele metod znajdowania zer i biegunów (osobliwości) funkcji zespolonych w zadanym obszarze poszukiwań. Można wyróżnić metody lokalne [1], bazujące na próbkowaniu funkcji jedynie w niewielkim fragmencie dziedziny oraz globalne, analizujące cały obszar obliczeniowy w celu znalezienia wszystkich zer. Przykładem podejścia globalnego jest algorytm GRPF [2] oraz jego pierwotne odmiany [3, 4], opierające się na analizie znaku części rzeczywistej oraz urojonej funkcji. Weryfikacja miejsc zerowych bazuje zwykle na zasadzie argumentu Cauchy'ego (CAP) [5], realizowanej poprzez całkowanie odpowiedniego obszaru.

Często interesująca jest analiza zmian pierwiastków względem dodatkowego parametru rzeczywistego t. Problem ten opisuje równanie:

$$F(z, t) = 0, \tag{2}$$

gdzie $z = \xi(t)$ jest miejscem zerowym reprezentującym krzywą w przestrzeni $\mathbb{C} \times \mathbb{R}$. W zagadnieniach propagacji fal elektromagnetycznych z może oznaczać współczynnik propagacji γ, a parametr t – np. częstotliwość (charakterystyka dyspersyjna).

Dotychczasowe algorytmy globalne nie zostały bezpośrednio uogólnione na przypadek wielu zmiennych (np. przestrzeń $\mathbb{C} \times \mathbb{R}$). Standardowe rozwiązanie tego problemu opiera się na "plastrowaniu" – wielokrotnych analizach funkcji dla różnych wartości parametru t. Pomimo poprawnych wyników metoda ta jest mało efektywna, zwłaszcza przy złożonych krzywych $\xi(t)$, które mogą się rozwidlać, przecinać lub przebiegać równolegle do płaszczyzny $t = \text{const}$. Wymaga to odpowiedniego doboru sposobu dyskretyzacji, który jest trudny do automatyzacji. Dodatkowo metoda nie pozwala na bezpośrednie śledzenie zależności pomiędzy kolejnymi pierwiastkami $z_n(t_n)$.

Alternatywnym podejściem są algorytmy śledzenia pierwiastków w funkcji parametru rzeczywistego lub zespolonego [3, 6, 7]. Są one bardziej efektywne niż metoda "plastrowania", ale wymagają precyzyjnych punktów startowych i mogą ominąć część rozwiązań w przypadku rozwidlenia lub przecięcia krzywych. Ograniczenia te można minimalizować [8], co jednak zwiększa koszt obliczeniowy.

2. SFORMUŁOWANIE PROBLEMU

Niniejsza praca stanowi omówienie fragmentu rozprawy doktorskiej pt. "Analiza nieliniowych problemów własnych dla prowadnic i struktur rezonansowych w optyce i technice mikrofalowej", w ramach której opracowana została technika zwiększająca efektywność znajdowania pierwiastków dla funkcji z dodatkowym parametrem [9]. Wykorzystuje ona cechy algorytmów globalnych oraz eliminuje problemy występujące w śledzeniu. Celem nowego algorytmu jest otrzymanie pełnego rozkładu zer i biegunów danej funkcji $F(z,t)$ w całej analizowanej przestrzeni $\mathbb{C}\times\mathbb{R}$. Odbywa się to poprzez iteracyjną dyskretyzację obszaru pooszukiwań Ω oraz uogólnieniu dyskretnej zasady Cauchy'ego na przestrzeń $\mathbb{C}\times\mathbb{R}$ w celu wyznaczenia krzywych $\xi(t)$ reprezentujących zera i bieguny funkcji zespolonych zmieniające się w funkcji dodatkowego parametru [7, 10], które w jest stosowane również w śledzeniu.

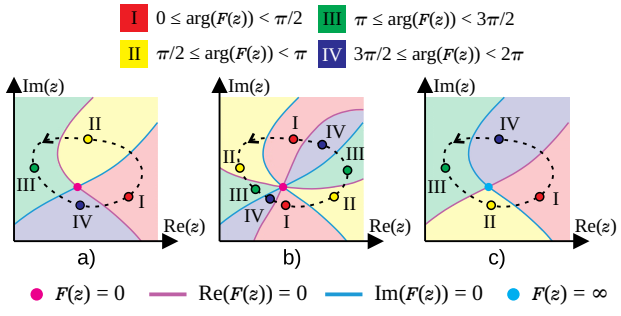
Zgodnie z CAP, liczba zer i biegunów funkcji wewnątrz dowolnego konturu C może być reprezentowana przez całkę:

$$q = \frac{1}{j2\pi}\oint_C \frac{F'(z)}{F(z)}dz \tag{3}$$

gdzie parametr $q \in \mathbb{Z}$ jest różnicą sumy zer oraz biegunów funkcji wraz z ich krotnościami. W praktyce wykorzystuje się dyskretną zasadę argumentu Cauchy'ego (DCAP), która nie wymaga obliczania pochodnej ani całkowania funkcji. Parametr q można oszacować poprzez próbkowanie funkcji wzdłuż zdyskretyzowanego konturu C [11], a przyrost fazy funkcji $F(z)$ na każdym odcinku spełnia warunek:

$$\left|[\arg F(z)]_{z\in(z_p,z_{p+1})}\right| \leq \pi \tag{4}$$

gdzie punkty $z_1, z_2, ..., z_P, z_{P+1} = z_1$ są wynikiem dyskretyzacji konturu C (patrz Rysunek 1)

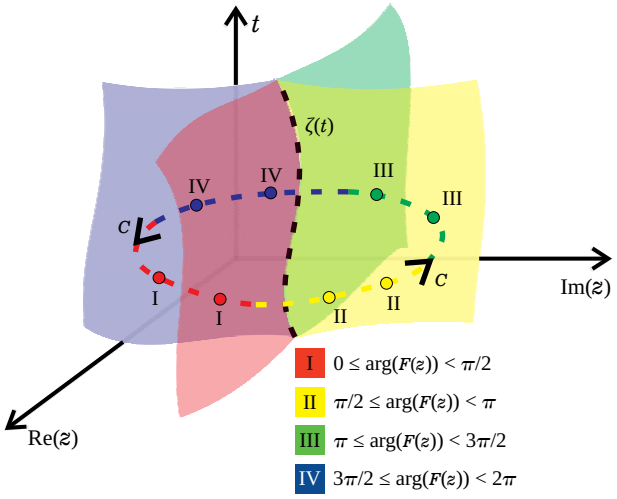


Rysunek 1: Przykłady dyskretnej zasady argumentu Cauchy'ego: a) pojedynczy pierwiastek q = 1; b) podwójny pierwiastek q = 2; c) osobliwość q = -1.

Potwierdzenie istnienia pojedynczego pierwiastka wewnątrz odpowiednio małego obszaru wymaga jedynie czterech punktów rozmieszczonych w różnych ćwiartkach układu współrzędnych. Dzięki temu DCAP umożliwia rozszerzenie klasycznej metody bisekcji na płaszczyznę zespoloną.

2.1 Uogólnienie dyskretnej zasady argumentu Cauchy'ego w przestrzeni $\mathbb{C}\times\mathbb{R}$

Rozważmy funkcję ciągłą i holomorficzną $F(z,t)$ w obszarze $\Omega = z \in \mathbb{C}, t \in \mathbb{R}$. Jeśli zero/biegun funkcji zależy od parametru t, to $z = \xi(t)$ opisuje krzywą w przestrzeni $\mathbb{C}\times\mathbb{R}$ (zgodnie z równaniem (2)). Wówczas CAP oraz jej dyskretną wersję można zastosować do dowolnej krzywej zamkniętej $C \subset \Omega$, niekoniecznie leżącej na płaszczyźnie $t = \text{const}$. Prawdziwe są więc wzory (3), (4) oraz ich interpretacja. Parametr q można oszacować poprzez próbkowanie funkcji i obserwację zmiany fazy po obwodzie konturu C, co ilustruje Rysunek 2.



Rysunek 2: Krzywa $\xi(t)$ w przestrzeni $\mathbb{C}\times\mathbb{R}$ oraz kontur C potwierdzający istnienie pierwiastka wewnątrz jego obszaru.

2.2 Założenia algorytmu

Algorytm rozpoczyna się od zdefiniowania funkcji $F(z,t)$ oraz obszaru poszukiwań $\Omega \subset \mathbb{C}\times\mathbb{R}$ o dowolnej geometrii. W przeciwieństwie do metody GRPF [2], tutaj obszar jest dyskretyzowany trójwymiarową siatką złożoną z czworościanów w przestrzeni $\mathbb{R}^3$ $(\mathrm{Re}(z), \mathrm{Im}(z), t)$, tworzoną metodą triangulacji Delaunaya. W węzłach siatki obliczana jest wartość funkcji oraz analizowane są różnice faz pomiędzy węzłami. Krawędzie, na których występują znaczące zmiany fazy (przekraczające $\pi/2$)[1], określane są jako "podejrzane" ("kandydujące"), a ich środki są traktowane jako potencjalne zera/bieguny funkcji. Podsumowując algorytm można przedstawić w następujących krokach:

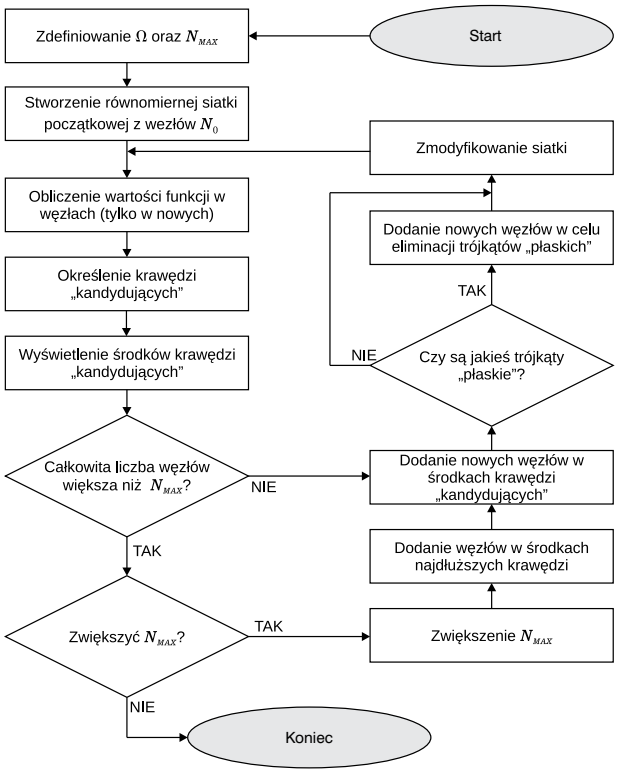
1. Dyskretyzacja obszaru Ω przy użyciu trójwymiarowej siatki czworościanów.
2. Obliczenie wartości funkcji $F(z,t)$ we wszystkich węzłach siatki.
3. Analiza przyrostu fazy funkcji wzdłuż każdej krawędzi na podstawie DCAP. Na tej podstawie wyznaczana jest zmiana fazy $\Delta\Theta_p$ lub w dyskretnej wersji odpowiadająca temu różnica kwadrantów $|\Delta Q_p|$.

[1]Zapis ten jest w pewnym sensie odpowiednikiem zmiany argumentu o dwa kwadranty, ale działa bardziej ogólnie. Nie narzuca się dyskretyzacji argumentu funkcji względem kątów: $0, \pi/2, \pi, 3\pi/2$.

4. Wyznaczenie zbioru krawędzi "kandydujących", tj. spełniających warunek istotnej zmiany fazy ($\Delta\Theta_k > \pi/2$).
5. Dodawanie nowych węzłów w środkach krawędzi "kandydujących", iteracyjne zagęszczanie siatki wokół poszukiwanych krzywych.
6. Wyznaczenie zbioru rozwiązań w postaci krzywych $\xi(t)$ na podstawie środków krawędzi "kandydujących".

Schemat działania algorytmu przedstawiono na Rysunku 3. Użytkownik może zwiększyć dokładność wyników poprzez podwyższenie maksymalnej liczby węzłów N_{max}, co powoduje dodatkowe zagęszczenie siatki. Oto kilka dodatkowych uwag, które mogą pomóc jeszcze lepiej zrozumieć detale opisanego algorytmu:

- Jeśli wartość funkcji w węźle jest zerem lub nieskończonością, wszystkie przylegające krawędzie są traktowane jako "podejrzane".
- Po zwiększeniu N_{max} część nowych węzłów jest dodawana równomiernie, na środkach najdłuższych krawędzi całego obszaru (eliminacja trójkątów "płaskich")[2].

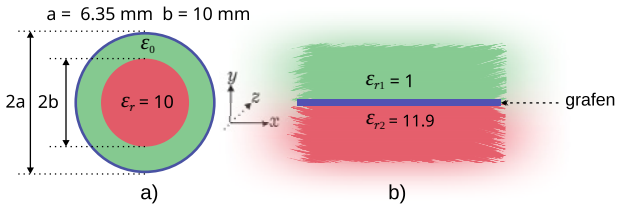


Rysunek 3: Schemat blokowy działania proponowanego algorytmu.

Algorytm, choć wydaje się kosztowny numerycznie, w praktyce ograniczony jest głównie przez złożoność obliczeniową samej funkcji oraz charakterystykę badanego obszaru. Podobnie jak klasyczne algorytmy globalne, nie gwarantuje on znalezienia wszystkich rozwiązań dla dowolnie skomplikowanych funkcji przy skończonej liczbie iteracji, ale ryzyko pominięcia rozwiązania jest ograniczone dzięki trójwymiarowemu podejściu i możliwości zagęszczania siatki.

3. WYNIKI NUMERYCZNE

W celu potwierdzenia stabilności i wydajności proponowanego podejścia przeanalizowano dwa przykładowe problemy propagacyjne. W obu przypadkach wyznaczono współczynnik propagacji w funkcji częstotliwości dla struktur przedstawionych na Rysunku 4, wraz z określonymi tam parametrami. Wyniki porównano z metodą "plastrową", stosując algorytm GRPF w wersji Self-Adaptive Mesh Generator [12] dla różnych częstotliwości.



Rysunek 4: Przekroje poprzeczne analizowanych struktur i ich parametry: a) falowód cylindryczny wypełniony dielektrykiem; b) linia transmisyjna z grafenem.

3.1 Falowód cylindryczny wypełniony dielektrykiem

Pierwszy przykład obejmuje współosiowy falowód dielektryczny, w którym występują złożone mody opisane funkcjami Bessela [13]. Rozkład pola można wyznaczyć analitycznie metodą dopasowania rodzajów. Obszar analizy w przestrzeni (γ, f) przyjęto jako $\Omega = \{\gamma \in \mathbb{C}, f \in \mathbb{R} : -4.5 < \mathrm{Re}(\gamma) < 4.5 \wedge -0.25 < \mathrm{Im}(\gamma) < 2.25 \wedge 2 < f < 7\}$ przy parametryzacji $N_{max} = 100\,000$ utworzono 1 000 węzłów początkowych.

3.2 Linia transmisyjna z grafenem

Drugi przykład dotyczy propagacji fal powierzchniowych (mody TM/TE) w linii transmisyjnej z cienką warstwą grafenu na podłożu krzemowym; model przewodzenia opisano w [14]. Analizowany obszar znormalizowanego współczynnika propagacji wynosi $\Omega = \{\gamma \in \mathbb{C}, f \in \mathbb{R} : -100 < \mathrm{Re}(\gamma) < 350 \wedge 25 < \mathrm{Im}(\gamma) < 300 \wedge 1 < f < 7\}$. Również tutaj przy $N_{max} = 100\,000$ wygenerowano 1 000 węzłów startowych.

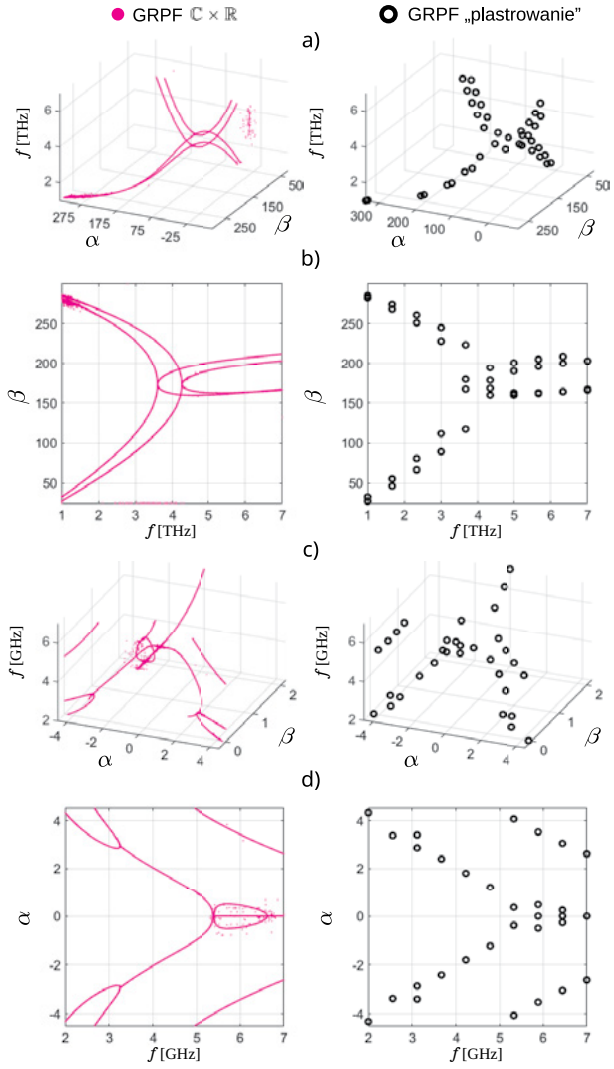
3.3 Charakterystyki dyspersyjne

Obliczone charakterystyki dyspersyjne pokazano na Rysunku 5, gdzie a) i c) to widoki trójwymiarowe $F(\gamma, f) = 0$, natomiast b) i d) przedstawiają odpowiednio rzuty na osie urojoną $F(\beta, f) = 0$ i rzeczywistą $F(\alpha, f) = 0$. Proponowany algorytm (lewa kolumna) zapewnia pełniejszy obraz zer/biegunów w badanym obszarze. Prawa kolumna przedstawia wyniki otrzymane "plastrowaniem" wykonanym w 10 przekrojach. Niezależna analiza na każdej z tych częstotliwości wymagała około 10 000 wywołań funkcji, co łącznie dało $\approx 100\,000$ punktów, czyli tyle samo, ile proponowany algorytm wykorzystał w całym obszarze Ω. Wyniki obu metod są zgodne, jednak metoda "plastrowa" wymaga znacznie gęstszej dyskretyzacji w funkcji częstotliwości, by uzyskać równie kompletną charakterystykę.

[2]Niestety, w kolejnych iteracjach zagęszczanie siatki może prowadzić do źle uwarunkowanej geometrii. Wskutek czego powstają trójkąty „płaskie" o ogromnej dysproporcji długości boków.

4. PODSUMOWANIE

Przedstawiony algorytm GRPF w przestrzeni $\mathbb{C} \times \mathbb{R}$ skutecznie rozszerza wyszukiwanie miejsc zerowych funkcji zespolonych o dodatkowy parametr. Zapewnia stabilne, globalne rozwiązanie i kompletny przegląd charakterystyk przy umiarkowanej liczbie wywołań funkcji. Rezultaty mają charakter przybliżony; dla dokładnego położenia pierwiastków można je doprecyzować metodami śledzenia lub lokalnym "plastrowaniem". Kod źródłowy algorytmu jest dostępny na: github.com/PioKow/GRPF3D



Rysunek 5: Uzyskane pierwiastki w postaci charakterystyki dyspersji znormalizowanego współczynnika propagacji γ jako funkcji częstotliwości f dla struktur: a), b) - falowód cylindryczny wypełniony dielektrykiem; c), d) - linia transmisyjna z grafenem. Lewa kolumna: wyniki proponowanego algorytmu; prawa kolumna: metoda "plastrowania".

LITERATURA

[1] William H. Press et al. *Numerical Recipes in Fortran 77: The Art of Scientific Computing*. 2nd ed. Cambridge University Press, England, 1992. ISBN: 052143064X.

[2] Piotr Kowalczyk. "Global Complex Roots and Poles Finding Algorithm Based on Phase Analysis for Propagation and Radiation Problems". In: *IEEE Transactions on Antennas and Propagation* 66.12 (Sept. 2018), pp. 7198–7205. DOI: 10.1109/TAP.2018.2869213.

[3] Jerzy Julian Michalski and Piotr Kowalczyk. "Efficient and Systematic Solution of Real and Complex Eigenvalue Problems Employing Simplex Chain Vertices Searching Procedure". In: *IEEE Transactions on Microwave Theory and Techniques* 59.9 (2011), pp. 2197–2205. DOI: 10.1109/TMTT.2011.2160277.

[4] Piotr Kowalczyk. "Complex Root Finding Algorithm Based on Delaunay Triangulation". In: *ACM Transactions on Mathematical Software (TOMS)* 41 (3 June 2015), pp. 1–13. DOI: 10.1145/2699457.

[5] James Ward Brown and Ruel Vance Churchill. *Complex Variables and Applications*. McGraw-Hill, New York, NY, USA, 2009.

[6] Michal Mrozowski. "An efficient algorithm for finding zeros of a real function of two variables". In: *IEEE Trans. Microw. Theory Tech.* 36.3 (Mar. 1988), pp. 601–604.

[7] Sebastian Dziedziewicz, Rafal Lech, and Piotr Kowalczyk. "A Self-Adaptive Complex Root Tracing Algorithm for the Analysis of Propagation and Radiation Problem". In: *IEEE Transactions on Antennas and Propagation* 69.8 (2021), pp. 5171–5174. DOI: 10.1109/TAP.2021.3060047.

[8] Sebastian Dziedziewicz et al. "Multipath Complex Root Tracing". In: *2022 24th International Microwave and Radar Conference (MIKON)*. 2022, pp. 1–4. DOI: 10.23919/MIKON54314.2022.9924911.

[9] Sebastian Dziedziewicz et al. "Global Complex Roots and Poles Finding Algorithm in C × R Domain". In: *IEEE Access* 11 (2023), pp. 68809–68817. DOI: 10.1109/ACCESS.2023.3292961.

[10] Piotr Kowalczyk and Wojciech Marynowski. "Efficient Complex Root Tracing Algorithm for Propagation and Radiation Problems". In: *IEEE Transactions on Antennas and Propagation* 65.5 (June 2017), pp. 2540–2546. DOI: 10.1109/TAP.2017.2684198.

[11] Peter Kravanja and Marc Van Barel. "A Derivative-Free Algorithm for Computing Zeros of Analytic Functions". In: *Computing* 63.1 (July 1999), pp. 69–91. DOI: 10.1007/s006070050051.

[12] Sebastian Dziedziewicz et al. "Self-Adaptive Mesh Generator for Global Complex Roots and Poles Finding Algorithm". In: *IEEE Transactions on Microwave Theory and Techniques* 71.7 (2023), pp. 2854–2863. DOI: 10.1109/TMTT.2023.3238014.

[13] Michal Mrozowski. *Guided electromagnetic waves: properties and analysis*. Research Studies Press, Wiley, New York, 1997.

[14] Juan Sebastian Gomez-Diaz, Juan R Mosig, and Julien Perruisseau-Carrier. "Effect of spatial dispersion on surface waves propagating along graphene sheets". In: *IEEE Trans. Antennas Propag.* 61.7 (Mar. 2013), pp. 3589–3596.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

DETEKCJA LOGICZNO-KORELACYJNA EFEKTU LUKSEMBURSKIEGO

LOGIC-CORRELATIVE DETECTION OF THE LUXEMBOURG EFFECT

Tomasz A. Miś[1]

[1] Centrum Badań Kosmicznych PAN, Zakład Fizyki Plazmy, Warszawa, tmis@cbk.waw.pl

DOI: 10.15199/59.2025.4.64

Streszczenie: **Referat prezentuje proponowaną metodykę zautomatyzowanej detekcji efektu luksemburskiego (wywołanej w absorpcyjnej warstwie jonosfery, wzbudzonej przez stację zakłócającą, wtórnej modulacji słabszej stacji pożądanej) opartej o zależność logiczną, uruchamiającą zapis danych, i zależność analityczną o wyższym koszcie obliczeniowym, ostatecznie potwierdzającą zaistnienie efektu w zebranych danych. Proponowane algorytmy umożliwią automatyzowanie zbierania danych zawierających efekt luksemburski w celu dalszych analiz jonosferycznych.**

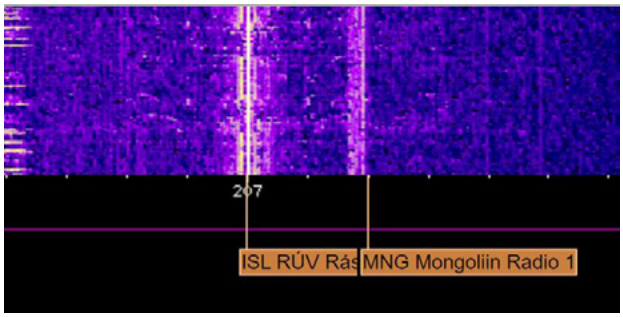
Abstract: **The paper presents the proposed methodology for automated detection of the Luxembourg effect (induced in the absorptive layer of the ionosphere excited by the interfering station, secondary modulation of the weaker wanted station) based on a logical detection, which triggers the data recording, and an analytical detection with a higher computational cost, which finally confirms the existence of the effect in the recorded data. The proposed algorithms shall enable the automation of data collection containing the Luxembourg effect for further ionospheric analyses.**

Słowa kluczowe: **broadcasting, częstość zderzeń, efekt luksemburski, jonosfera, warstwa E**

Keywords: **broadcasting, collision frequency, Luxembourg effect, ionosphere, E layer**

1. WSTĘP

Znany i badany od niemalże wieku efekt luksemburski (efekt jonosferycznej modulacji skrośnej) występuje powszechnie w zakresie częstotliwości niskich - fal długich (30-300 kHz) [1], średnich (300 kHz – 3 MHz) oraz krótkich (3-30 MHz). Powszechnie uznawany jest za efekt zakłócający, zniekształcający odbiór stacji radiodyfuzyjnych niższej mocy, emitujących sygnały docierające na dalekie dystanse dzięki odbiciu od warstwy D jonosfery Ziemi. Objawia się on występującą dodatkową modulacją słabszego sygnału (pożądanego), będącą efektem wzbudzenia jonosfery przez sygnał stacji silniejszej (zakłócającej). Wzbudzenie to ma charakter dwuetapowy: początkowe przekazanie energii jonosferycznym elektronom i jonom przez sygnał zakłócający, objawiające się modyfikacją ich częstości zderzeń, i późniejsze zmodulowanie odbijającego się od tego samego obszaru jonosfery sygnału słabszego [7]. Efekt luksemburski, ze względu na fizyczne własności jonosfery, posiada maksymalną głębokość modulacji rzędu kilku procent [6] i typową szerokość widma do kilkuset herców, przy czym modulacja ta może mieć charakter asymetryczny, w zależności od wzajemnego położenia stacji zakłócającej i pożądanej [3]. Przykładowy efekt luksemburski asymetryczny pokazano na Rys. 1.



Rys. 1. Asymetryczny efekt luksemburski w zakresie fal długich, na częstotliwości 209 kHz emisji A3E (utworzona wąska dolna wstęga boczna – LSB – dla stacji odbieranej w Enschede, Niderlandy, z prawie zupełnie wytłumionymi wstęgami bocznymi stacji pożądanej; stacja pożądana mongolska, poza rastrem CCIR/ITU). Obserwacja Autora, 3 stycznia 2022 r.

Jak wykazano już na przełomie lat 50. i 60., przy pomiarze głębokości modulacji efektu luksemburskiego możliwe jest wyznaczenie podstawowych parametrów jonosfery, takich, jak gęstość elektronowa lub częstość zderzeń [2]. Czyni to to zjawisko, zwyczajowo niepożądane, potencjalnie niezwykle użytecznym narzędziem przy określaniu częstości zderzeń w danej warstwie jonosfery. Warstwa ta określona jest jednoznacznie przez dobór częstotliwości, na jakiej zachodzi badany efekt (np. dla fal długich i średnich będzie to warstwa E), przy czym dla częstotliwości niższych sąsiadujących (zakłócająca blisko pożądanej) podejście teoretyczne prezentowane np. przez ITU [7] będzie cechować się większą dokładnością ze względu na silniejsze zawężenie obszaru występowania efektu w jonosferze.

Dawne kampanie testowe w zakresie fal krótkich [2] i średnich [4] pozwalały w sposób elementarny scharakteryzować własności efektu luksemburskiego. Wykorzystanie – przy założeniu dostarczenia niektórych danych, np. aktualnej gęstości elektronowej, z innych źródeł, np. okolicznych jonosond – tego zjawiska do wyznaczania częstości zderzeń w niższej warstwie jonosfery jest niezwykle atrakcyjne dla prowadzonych obserwacji i badań jonosferycznych np. za pomocą radioteleskopów w sieci LOFAR. W ten sposób możliwe będzie uszczegółowienie lokalizacji rejestrowanych zjawisk

jonosferycznych, ze wskazaniem konkretnych warstw jonosfery, wewnątrz których mogą one zachodzić.

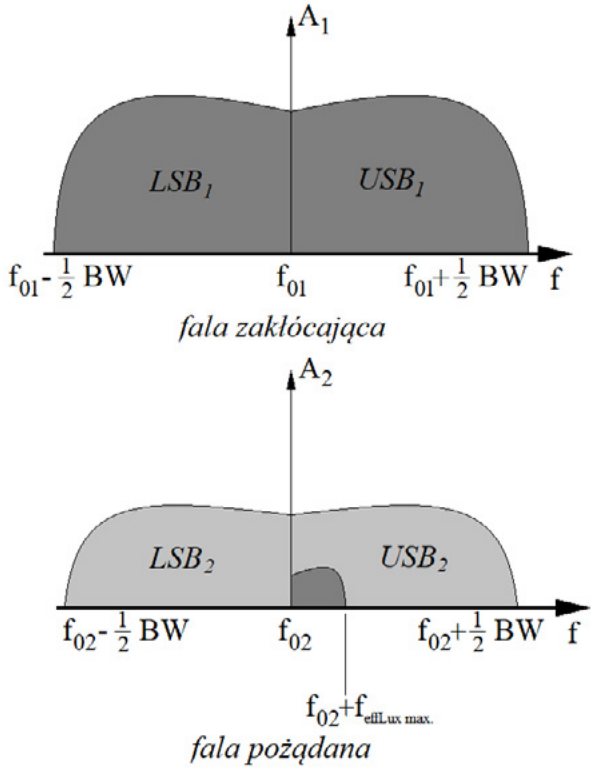
Warto zauważyć, iż do celów wsparcia jonosferycznych badań radioteleskopowych nie istnieje konieczność zestawiania dedykowanych łączy radiowych, wrażliwych na efekt luksemburski – możliwe jest wykorzystanie z powodzeniem istniejących stacji radiodyfuzyjnych w zakresie fal długich i średnich, nadających ze znanymi mocami promieniowanymi i o znanych parametrach anten nadawczych.

Monitorowanie automatyczne wybranych (dla danego obszaru) stacji radiowych –zakłócających i pożądanych, wyselekcjonowanych wcześniej w taki sposób, by wyeliminować możliwe nieefemeryczne sygnały zakłócające – jest jednak zadaniem wymagającym użycia dużej ilości pamięci zdolnej zachować zarejestrowane dane, by później przetworzyć je w poszukiwaniu pożądanego efektu. Niezbędne staje się zaprojektowanie nieskomplikowanego, łatwego w implementacji systemu logicznego, który wskaże jednoznacznie obecność efektu luksemburskiego, umożliwiając rozpoczęcie zapisu tylko użytecznych danych. Ostateczna weryfikacja istnienia efektu luksemburskiego w zapisanych sygnałach może odbyć się, przykładowo, przy wykorzystaniu operacji korelacji wzajemnych zdemodulowanych sygnałów zakłócającego i pożądanego.

2. PODSTAWY WIDMOWE

Na rys. zaprezentowano schematycznie struktury widmowe sygnału silniejszego-zakłócającego o częstotliwości środkowej f_{01} i sygnału słabszego-pożądanego o częstotliwości środkowej f_{02}. Oba sygnały należą do stosowanych powszechnie w radiodyfuzji typów transmisji A3E (analogowej dwuwstęgowej telefonii z niewytłumioną falą nośną). Założono bezpośrednią bliskość częstotliwości pracy obu stacji (sąsiadowanie w rastrze) – różnica między f_{01} i f_{02} będzie równa szerokości pasma (BW), wynoszącej dla fal średnich i krótkich typowo 9 kHz. Stacja słabsza-pożądana ulega efektowi luksemburskiemu asymetrycznemu o zakresie widma ograniczonym częstotliwością $f_{02}+f_{effLux\ max.}$, ukazanemu jako dodatkowe, słabsze pasmo częstotliwości w obszarze wstęgi górnej USB_2. Częstotliwość $f_{effLux\ max.}$ posiada wartości rzędu kilkuset herców, daleko niższe od $f_{02}+½BW$ równe 4.5 kHz – stanowi odfiltrowany fragment pasma USB_1 (nie przeskalowane całe pasmo USB_1).

Na schemacie uwzględniono brak sumowania się widm efektu luksemburskiego i USB_2, uwidaczniając w ten sposób brak nakładania się poszczególnych częstotliwości tych widm. Dodatkowe częstotliwości we wstędze USB_2 spowodują jednak zwiększenie mocy sygnału w zakresie $f_{02}+½BW$, które – dla efektu luksemburskiego asymetrycznego – staną się podstawą do wykonania logicznej detekcji efektu celem uruchomienia automatycznego zapisu.



Rys. 2. Schemat widm sygnałów zakłócającego i pożądanego przy zaistnieniu asymetrycznego efektu luksemburskiego.

3. DETEKCJA LOGICZNA

Symultaniczna analiza dwóch branych pod uwagę sygnałów – pożądanego i zakłócającego – na wstępnym etapie warunkującym zapis danych może zostać zredukowana do analizy pojedynczego sygnału, podzielonego na dwie części posiadające stosunek sygnału do szumu >1 (będących nawet na granicy wykrywalności) – wstęgę boczną dolną (LSB) i wstęgę boczną górną (USB). Przy założeniu występowania asymetrycznego efektu luksemburskiego, naturalną będzie różnica w poziomach sygnałów dla USB i LSB, przy czym, wg schematu z rys. 2, różnica jednego i drugiego przy wystąpieniu efektu luksemburskiego będzie różna od zera.

Forma logiczna (bez istotnych działań arytmetycznych) analizowanej różnicy może przyjąć formę analogiczną do zaprezentowanej przez dr. Bartłomieja Jarmolińskiego [5] w 2024 r. koncepcji prawdziwości przetwarzania, głoszącej, iż logicznym fałszem będzie przedstawienie obrazu obrazu, logiczną prawdą zaś przedstawienie zmodyfikowanego obrazu obrazu. Zależność można ująć w formę bramki logicznej (możliwej do dalszego przedstawienia jako układu składającego się z kombinacji kilku bramek negacji, koniunkcji i alternatywy) o tabeli prawdy ukazanej w Tab. 1. „Modyfikacją" warunkującą otrzymanie prawdziwej wartości operacji w ostatniej kolumnie będzie tu wystąpienie efektu luksemburskiego.

W niniejszym ujęciu wartości „rzeczywistości" i jej „obrazu" odpowiadają detekcji sygnałów odpowiednio wstęg LSB i USB. W przypadku braku sygnału LSB i

detekcji sygnału USB, efekt luksemburski zaistnieje – jedna wstęga nie jest obrazem (symetrycznym odbiciem) drugiej. Gdy zaistnieje jedynie wstęga LSB, bez USB, wówczas efekt nie jest wykrywalny (efekt lewostronny!). W przypadku zaistnienia obu wstęg konieczne jest porównanie poziomu obu sygnałów – jeśli poziom jest identyczny (wykazuje takie samo przesunięcie względem zera), wówczas szukany efekt nie występuje; gdy poziomy różnią się – wówczas efekt jest wykryty.

Tab. 1. Tabela prawdy dla bramki Jarmolińskiego.

Rzeczywistość ***(LSB)***	*Obraz* ***(USB)***	*Obraz obrazu*	*Obraz obrazu zmod.*
0	1	0	1
0	0	1	0
1	1	0	1
1	0	1	0

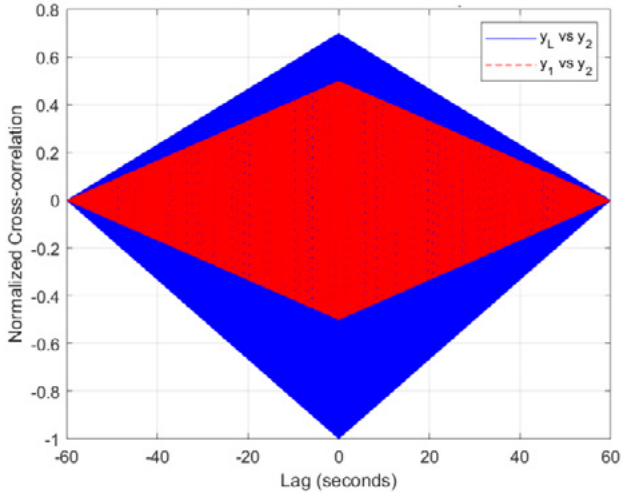
Fizyczna aplikacja tego rodzaju monitorowania wymaga więc podziału na minimum dwa symultanicznie rozważane przypadki – przypadek dla zerowych poziomów wstęg bocznych sygnału pożądanego i przypadek dla identycznych poziomów wstęg bocznych sygnału pożądanego, analizowanych przy bieżącym dostosowywaniu przesunięcia zera (wraz z nasilaniem się lub osłabianiem sygnału transmisji). Nie jest wymagana żadna dodatkowa operacja arytmetyczna – detekcja funkcjonuje jedynie w oparciu o związki logiczne plus bieżące dostosowywanie przesunięcia poziomu zerowego. Analiza jest uproszczona w przypadku określenia występowania jedynie asymetrycznego (w obszarze danej wstęgi bocznej) efektu luksemburskiego – jeśli efekt ten nie jest przewidywalny, wówczas ilość niezbędnych do zrealizowania przypadków wzrośnie do czterech: dwa przypadki dodatkowe uwzględnią w analizie logicznej zamianę stronami LSB i USB.

Przedstawiona metoda logiczna detekcji efektu luksemburskiego wydatnie ogranicza czynności obliczeniowe przy bieżąco rejestrowanych sygnałach i ostatecznie umożliwia istotną redukcję wymagających zapisu ilości danych koniecznych do późniejszego przetworzenia. Bardziej złożone operacje matematyczne finalnie potwierdzające rejestrację efektu luksemburskiego – np. wspomniana korelacja wzajemna sygnałów zakłócającego i pożądanego – będą możliwe do przeprowadzenia nie w czasie rzeczywistym i na zredukowanych ilościach danych, istotnie przyspieszając proces całej analizy.

4. DETEKCJA KORELACYJNA

Dalsze określenie zaistnienia efektu luksemburskiego w zapisanym sygnale można przeprowadzić na podstawie korelacji wzajemnej zarejestrowanych sygnałów stacji pożądanej i stacji zakłócającej. Podstawą do rozróżnienia – wykrycia modyfikacji-efektu luksemburskiego – będą wtedy widoczne zmiany w kształtach obwiedni wykreślonych korelacji wzajemnych dwóch próbek sygnału o jednakowej długości jako funkcji przesunięcia czasowego między nimi.

Metodę tę można wstępnie sprawdzić w prostej symulacji uwzględniającej trzy sygnały: y1 (sygnał pożądany), y2 (sygnał zakłócający) i yL (sygnał z symulowanym efektem luksemburskim: suma y1 oraz składników widmowych y2 poniżej 200 Hz). Składniki widmowe sygnału y1 to funkcje harmoniczne o arbitralnych (przykładowych) częstotliwościach 100, 900, 2000 i 2900 Hz; y2 posiada składowe o częstotliwościach 70, 110, 190, 1000 i 2100 Hz. Różnice widmowe między tymi sygnałami mają za zadanie uproszczone oddanie różnic w strukturach rzeczywistych sygnałów modulujących dwóch różnych stacji radiodyfuzyjnych (mimo pozostawania w niniejszej symulacji widmowo stałymi w czasie). Wszystkie składniki posiadają amplitudę unormowaną równą 1 i czas trwania 60 s. Wyznaczone za pomocą oprogramowania MATLAB unormowane funkcje korelacji wzajemnych pokazano na rys. 3; każda funkcja składowa została podana próbkowaniu z częstością 1 kHz.



Rys. 3. Korelacje wzajemne sygnału symulującego efekt luksemburski i sygnału zakłócającego oraz sygnału pożądanego („czystego”) i sygnału zakłócającego.

Wykreślona korelacja wzajemna dwóch różnych widmowo sygnałów, y1 i y2, posiada obwiednię symetryczną względem osi przesunięć czasowych – osiąga ona tę samą wartość bezwzględną. Sygnał y1 wzbogacony o fragment widma z sygnału y2 – wywołany symulowany efekt luksemburski – posiada obwiednię o kształcie wyraźnie asymetrycznym - widoczna różnica świadczy o zredukowanym podobieństwie między dwoma sygnałami.

Na podstawie przeprowadzonej symulacji można stwierdzić podstawową własność wykreślanej funkcji korelacji wzajemnej badanych sygnałów, wskazującej na istnienie różnicy atrybuowanych efektowi luksemburskiemu. Metoda ta zostanie następnie sprawdzona na rzeczywistych danych zawierających zarejestrowany efekt luksemburski. W przypadku uzyskania zbliżonych wyników, umożliwiających nieskomplikowane potwierdzenie wystąpienia poszukiwanego efektu, w połączeniu z zaprezentowaną uprzednio detekcją logiczną z wykorzystaniem bramki Jarmolińskiego, opisana metodyka będzie mogła być wdrożona do zastosowania przy zau-

tomatyzowanym monitorowaniu sygnałów wrażliwych na jonosferyczną modulację skrośną celem dalszego uzyskania istotnych parametrów jonosferycznych, ze szczególnym naciskiem na częstość zderzeń w warstwie E.

W dalszych pracach, oprócz wdrożenia algorytmu do pracy testowej na sygnałach rzeczywistych (zarejestrowanych uprzednio w terenie), przeprowadzone zostanie również badanie odporności zaproponowanej metodyki na zakłócenia efemeryczne, tj. niewykryte i niewyeliminowane na etapie definicji stacji zakłócającej i pożądanej, wraz z określeniem stopnia/ilości detekcji fałszywych (niezwiązanych z efektem luksemburskim) oraz detekcji w sytuacji występowania szybkozmiennych zaników (fadingu) sygnału stacji pożądanej.

Praca została sfinansowana dzięki środkom przyznanym przez Ministerstwo Nauki i Szkolnictwa Wyższego (MNiSW) na polski wkład do Międzynarodowego Teleskopu LOFAR (decyzja MNiSW nr 30/530252/SPUB/SP/2022).

LITERATURA

[1] Bailey V. A., Martyn D. F. 1934. “Interaction of Radio Waves”. Nature, February 10th: 218.
[2] Benson Robert F. 1962. “Cross modulation of cosmic noise”. Journal of Geophysical Research 67(6): 2569-2572.
[3] de Boer P. T., Vester M. 2018. „Sideband assymetry in ionospheric cross modulation”. Radio Science 53: 640-655.
[4] Hibberd F. H. 1955.”An experimental study of gyro interaction in the ionosphere, at oblique incidence”. Radio Science 67D(1): 25-31.
[5] Bartłomiej Jarmoliński: http://www.artysci-lodzkie.pl/pl/artysta/j/bartlomiej-jarmolinski/ [dostęp: 20.05.2025]
[6] Newton C. C., Hyde F. J., Foster H. G. 1950. „Ionospheric cross-modulation: techniques of measurement”. Proceedings of the Physical Society B 63: 616-623.
[7] Recommendation ITU-R P.532-1. Ionospheric effects and operational considerations associated with artificial modification of the ionosphere and the radio-wave channel.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORISTYKI
KRiT 2025

GENEROWANIE I ZASTOSOWANIE OGRANICZONEJ KSIĄŻKI KODOWEJ DLA RIS W SCENARIUSZU WEWNĄTRZBUDYKOWYM

CREATION AND APPLICATION OF CONSTRAINED CODEBOOK FOR RIS IN INDOOR PROPAGATION SCENARIOS

Tomasz Jaworski[1]; Cyryl Prentki[1]; Paweł Kryszkiewicz[2]; Adrian Kliks[2];

[1] Politechnika Poznańska, Poznań, imie.nazwisko@student.put.poznan.pl
[2] Politechnika Poznańska, Poznań, imie.nazwisko@put.poznan.pl

DOI: 10.15199/59.2025.4.65

Streszczenie: Artykuł omawia zastosowanie Rekonfigurowalnych Matryc Antenowych (ang. Reconfigurable Inteligent Surfaces – RIS) w systemach bezprzewodowych jako narzędzia poprawy jakości sygnału poprzez kształtowanie propagacji fal. Przedstawiono model systemu zawierający RIS, nadajnik i odbiornik w środowisku bez widoczności bezpośredniej, oraz sposób obliczania współczynników odbicia w celu maksymalizacji mocy odbieranej. Zaproponowano metodę tworzenia ograniczonej książki kodowej, przy zachowaniu efektywności. Przeprowadzono testy eksperymentalne, wykazując wzrost mocy sygnału i przepływności przy użyciu RIS z ograniczoną książką kodową.

Abstract: This paper discusses the use of Reconfigurable Intelligent Surfaces (RIS) in wireless systems as a tool to improve signal quality by shaping wave propagation. A system model incorporating an RIS, transmitter and receiver in a non-line-of-sight (NLoS) environment, and how to calculate reflection coefficients to maximise received power is presented. A method for creating a constrained, while maintaining efficiency, codebook was proposed. Experimental tests were carried out, showing an increase in signal power and channel bit rate using RIS with a limited codebook.

Słowa kluczowe: RIS, książka kodowa, wzorce

Keywords: RIS, codebook, patterns

1. WSTĘP

Jednym z rozwiązań rozważanych w kontekście rozwoju sieci bezprzewodowych jest wykorzystanie Rekonfigurowalnych Matryc Antenowych (ang. Reconfigurable Inteligent Surfaces – RIS) [1]. RIS to elementy umieszczane w środowisku propagacyjnym (pomiędzy nadajnikiem a odbiornikiem), których zdalne sterowanie umożliwia modyfikację charakterystyk kanału radiowego w celu uzyskania pożądanych efektów po stronie odbiornika [2]. Dzięki temu możliwa jest, np. zmiana kąta odbicia sygnału, niezależnie od reguł optyki geometrycznej, skutkując poprawą jakości odbieranego sygnału, np. zwiększeniem jego mocy lub redukcją zakłóceń [3]. Liczba elementów antenowych w matrycy wpływa na jej zdolność do kształtowania propagacji. Jeśli każdy z L elementów może przyjmować jeden z K stanów (np. fazowych), to możliwych jest K^L różnych wzorców odbicia.

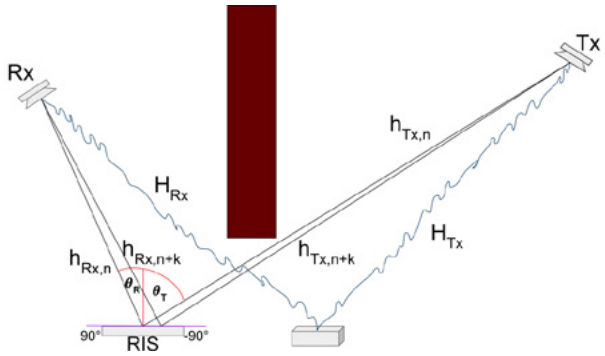
Ze względu na ogromną przestrzeń możliwości, praktyczne rozwiązania skupiają się na stosowaniu książek kodowych, bądź algorytmów korzystających z informacji kontekstowych. Książka kodowa to zbiór wzorców do stosowania z matrycą, gdzie wzorzec stanowi zbiór stanów wszystkich elementów RIS. W literaturze opisano m.in. książki kodowe oparte na: losowości [4], odległościach Euklidesowych wzorców [5] oraz optymalizacji alternatywnej (AO) [6]. Drugą grupą opisanych metod wyznaczania wzorców do stosowania z matrycami RIS są metody algorytmiczne, do których można zaliczyć: modele matematyczne [7], metody uczenia maszynowego [8], czy też heurystyki dopasowujące kanał do odbiornika o nieznanej lokalizacji [9]. Rozwiązania te mają jednak szereg różnych wad, do których można zaliczyć: długi czas wyboru wzorca, wysoką złożoność obliczeniową, duże zużycie energii lub konieczność częstego przełączania matrycy RIS.

Ze względu na specyfikę matryc RIS, ich pomiar wymaga szczególnego podejścia. Metody pomiarowe zależą od oczekiwanych wyników i nie zawsze są wymienne. W [9] analizowano zmiany zysku mocy odebranej w komorze bezechowej. Z kolei autorzy [10] zastosowali okno czasowe na podstawie wcześniej wyznaczonego momentu odebrania fali, w prowizorycznej komorze bezechowej, osiągając dużą precyzję pomiaru. Podejścia te jednak pomijają wielodrogowość sygnałów w rzeczywistych systemach komunikacji radiowej. Systemy o wiele bliższe warunkom rzeczywistym przedstawione są w [7], [11]. Używane tam układy dobrze przedstawiały prawdziwe warunki, ale metodyka tworzenia książek kodowych i interpretacja wyników nie brały pod uwagę wielodrogowości lub pomiary wykonywano w specyficznych warunkach.

W niniejszej pracy, przedstawiono metodę tworzenia książki kodowej, która stara się jak najlepiej uwzględniać powyższe ograniczenia istniejących rozwiązań. Poprzez model matematyczny stworzono zbiór rozwiązań, którego rozmiar następnie zminimalizowano prostymi metodami, w celu otrzymania niewielkiej, lecz skutecznej książki kodowej, jednocześnie uwzględniającej wielodrogową naturę propagacji fal elektromagnetycznych.

2. MODEL SYSTEMU

W ramach tej publikacji zaprojektowano system transmisyjny, zawierający jedną matrycę RIS, statyczny nadajnik (Tx) –pełniący rolę stacji bazowej oraz odbiornik (Rx) – pełniący rolę użytkownika (UE), mogącego znajdować się w dowolnym punkcie wcześniej określonej przestrzeni. Między Rx a Tx postawiono ścianę silnie tłumiącą sygnały radiowe, aby do Tx dochodził tylko sygnał z dróg bez widoczności bezpośredniej anten (NLoS) (Rys. 1).



Rysunek 1: Diagram przedstawiający układ pomiarowy i występujące w nim zjawisko wielodrogowości.

Rx sprawdza na częstotliwości nośnej poziom mocy, który można opisać jako:

$$P_i = \left| H_i + \sum_n h_{n,i} r_n \right|^2 |x|^2 , \tag{1}$$

gdzie x to przesłany, zespolony symbol, H_i to zespolona odpowiedź kanału wszystkich dróg poza drogą przez RIS, $h_{n,i}$ to odpowiedź impulsowa n-tego elementu RIS dla i-tej lokalizacji Rx, a r_n to współczynnik odbicia n-tego elementu RIS.

Jeśli RIS znajduje się w polu dalekim względem Tx i Rx, to amplituda odpowiedzi impulsowych kanałów odbitych od RIS jest równa dla każdego elementu, co można zapisać jako: $\forall_{n\neq\tilde{n}} |h_{n,i}| = |h_{\tilde{n},i}|$. W takim przypadku maksymalizacja odbieranej mocy może zostać osiągnięta ustawiając współczynniki odbicia na:

$$r_n = \exp\left(-j \arg(h_{n,i}) + j \arg(H_i)\right), \tag{2}$$

gdzie arg( ) oznacza argument liczby zespolonej. Znalezienie właściwych r_n jest jednak bardzo trudne, jeśli znana jest tylko odpowiedź połączonego kanału lub moc przy Rx. Najprostszą metodą było by wykonanie N pomiarów, każdy dla zmienionego jednego elementu RIS [9]. Takie podejście może okazać się bardzo czasochłonne, szczególnie dla matryc z dużą ilością elementów lub długimi czasami przełączania elementów. Dodatkowo jeśli H_i jest zależne od czasu koherencji kanału, na przykład w przypadku ruchomego lub nomadycznego UE, niemożliwym może okazać się sprawdzenie dostatecznie dużej ilości elementów RIS do wiarygodnego wyznaczenia r_n. Na przykład dla używanego w tej pracy RIS (OpenSourceRIS [12]) według wykonanych przez autorów pomiarów czas przełączania wzorca wynosił $T_{\mathrm{RIS}} = 22$ms, a czas koherencji kanału dla używanej przez nas częstotliwości (5, 59 GHz) może sięgać 30 ms. Nawet zakładając, że T_{RIS} dla nowoczesnych RIS może być 10-krotnie mniejszy, to nadal możliwe będzie przebadanie tylko 13 konfiguracji, przed tym jak kanał radiowy ulegnie zmianie. Z tych powodów, aby efektywnie używać RIS, do określenia optymalnych wzorców RIS dla danego środowiska, potrzeba jak najmocniej ograniczyć ilość badanych konfiguracji.

3. GENEROWANIE OGRANICZONEJ KSIĄŻKI KODOWEJ

Jedną z opcji jest zbudowanie książki kodowej wykorzystującej wiedzę na temat informacji kontekstowych (np. na temat lokalizacji RIS i Tx) oraz właściwości kanałów widoczności bezpośredniej (LoS): Tx-RIS oraz RIS-Rx.

Dla płaskiego RIS, współczynniki odbicia dla n-tego elementu RIS równają się [7]:

$$\begin{aligned} & r_n^{(LoS)}(\phi_{\mathrm{T}}, \theta_{\mathrm{T}}, \phi_{\mathrm{R}}, \theta_{\mathrm{R}}) = \\ & \exp\Big(j\frac{2\pi}{\lambda}\left(x_n \sin\theta_{\mathrm{T}} \cos\phi_{\mathrm{T}} + y_n \sin\theta_{\mathrm{T}} \sin\phi_{\mathrm{T}}\right) \\ & + j\frac{2\pi}{\lambda}\left(x_n \sin\theta_{\mathrm{R}} \cos\phi_{\mathrm{R}} + y_n \sin\theta_{\mathrm{R}} \sin\phi_{\mathrm{R}}\right)\Big), \end{aligned} \tag{3}$$

gdzie: Tx znajduje się pod kątem elewacji ϕ_{T} i kątem azymutu θ_{T}, Rx znajduje się pod kątem elewacji ϕ_{R} i kątem azymutu θ_{R} (z kątami równymi 0, jeśli dana pozycja znajduje się na normalnej do centrum powierzchni RIS), a x_n i y_n to współrzędne geometryczne N-tego elementu matrycy RIS.

Jednak r_n obliczony w powyższy sposób może nie być zgodny z kanałem H_i. Dlatego nawet dla znanych kątów azymutu i elewacji Tx–RIS–Rx, wielokrotne przesunięcia fazy prekodera φ_s, które są równoważne $\arg(H_i)$ w (2). Co więcej, ponieważ wiele praktycznych realizacji RIS ma tylko kilka możliwych kątów odbicia na element, które można modelować za pomocą funkcji kwantyzacji fazy $Q(\ )$, co daje prekoder:

$$\tilde{r}_n(\phi_{\mathrm{T}}, \theta_{\mathrm{T}}, \phi_{\mathrm{R}}, \theta_{\mathrm{R}}, \varphi_s) = Q\left(r_n^{(LoS)}(\phi_{\mathrm{T}}, \theta_{\mathrm{T}}, \phi_{\mathrm{R}}, \theta_{\mathrm{R}})\, exp(j\varphi_s)\right). \tag{4}$$

Jeśli rozważany jest idealny binarny RIS, funkcja kwantyzacji może być modelowana jako

$$Q(x) = 2H\left(\mathrm{Re}(x)\right), \tag{5}$$

gdzie $H(\ )$ oznacza funkcję skokową Heaviside'a. Należy zauważyć, że w przypadku praktycznej realizacji RIS to odwzorowanie może być nieco inne i może zmieniać się wraz z częstotliwością roboczą [12].

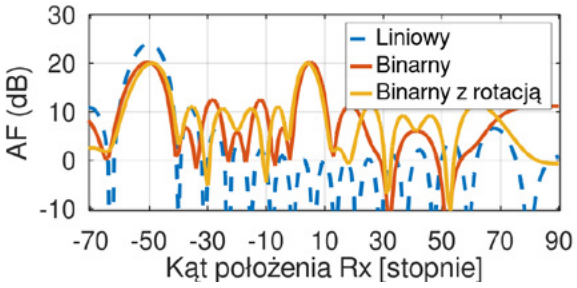
Książkę kodową można zaprojektować przy użyciu informacji kontekstowych o geometrii Tx-RIS-Rx. Bez utraty ogólności można założyć, że wszystkie elementy znajdują się na tej samej wysokości, co daje $\phi_{\mathrm{T}} = \phi_{\mathrm{R}} = 0$. Ponadto, jeżeli położenie Tx w stosunku do RIS nie zmienia się, można założyć, że θ_{T} jest stałą. Należy jednak wziąć pod uwagę wiele azymutów Rx, oznaczonych jako zbiór $\mathbb{R}$, o mocy $|\mathbb{R}|$, z $\theta_{\mathrm{T}} \in \mathbb{R}$. Dla każdej lokalizacji geometrycznej Rx powinniśmy rozważyć wiele obrotów prekodera, oznaczonych jako zbiór $\mathbb{S}$ ($\varphi_s \in \mathbb{S}$). Wówczas całkowita liczba elementów w książce kodowej (zbiór $\mathcal{P}$) wyniesie $|\mathbb{S}||\mathbb{R}|$. Zakładając, że Rx są rozmieszczone w azymutach od 0° do 90°, z rozdzielczością 1° i $\mathbb{S} = \{0, 1, ..., 359\}$, całkowita liczba elementów w książce kodowej wyniesie 32400.

Zaobserwowano jednak, że taka książka kodowa może zostać zmniejszona bez utraty jej efektywności. Możemy tę własność przypisać funkcji kwantyzacji. Aby to uzasadnić, zacznijmy od zdefiniowania charakterystyki odbiciowej matrycy RIS (Array Factor – AF), dla kątów obserwacji: azymutu θ i elewacji ϕ, jako:

$$\mathrm{AF}(\theta,\phi)=\sum_n r_n \exp\Big(-j\frac{2\pi}{\lambda}(x_n \sin\theta_\mathrm{T}\cos\phi_\mathrm{T}+y_n\sin\theta_\mathrm{T}\sin\phi_\mathrm{T}) - j\frac{2\pi}{\lambda}(x_n\sin\theta\cos\phi+y_n\sin\theta\sin\phi)\Big), \tag{6}$$

gdzie θ_T i ϕ_T to rzeczywiste kąty azymutu i elewacji z perspektywy RIS, a r_n to wykorzystany współczynnik odbicia RIS w n-tym jej elemencie.

Biorąc pod uwagę matrycę RIS z 16 rzędami i 16 kolumnami, co daje $N = 256$ elementów, Tx umieszczonym w $\theta_\mathrm{T} = 20°$, Rx pod $\theta_\mathrm{R} = -50°$ i odpowiednio zaprojektowanymi r_n, uzyskamy AF, przedstawiony na Rys. 2. Kąty elewacji są równe 0°. Przedstawiony *Liniowy* AF, zakłada, że RIS pozwala na ciągłe dostrajanie r_n w zakresie $[0; 2\pi]$. Ma on pojedyncze maksimum wynoszące 24 dB. Jednakże, jeśli używana jest binarna matryca RIS, maksymalny zysk wynikający z jej zastosowania, zmniejsza się do około 20 dB. Interesujące jest drugie otrzymywane maksimum dla kąta Rx $\theta \approx 5°$. W kontekście projektowania książki kodowej oznacza to, że jeden wzór odbicia może obsługiwać UE znajdujący się na znacznie różnych azymutach. Co więcej, szerokość wiązek obejmuje kilka stopni, co oznacza, że UE znajdujące się blisko siebie, mogą być obsługiwane przez tę samą wiązkę. Dodatkowo należy zauważyć, że jeśli współczynniki odbicia zostaną przesunięte o $\varphi_s = 90°$, w celu wyrównania odbicia RIS z fazą H_i, otrzymamy AF o bardzo podobnych maksimach.



Rysunek 2: Przykładowy teoretyczny AF wzorca dla: $\theta_T = 20$, $\theta_R = -50$; Liniowy – zakłada możliwość liniowego dostrajania r_n w zakresie $[0; 2\pi]$; Binarny – AF wzorca skwantowanego binarnie z $\varphi_s = 0$; Binarny z rotacją – AF wzorca skwantowanego binarnie z $\varphi_s = 90$.

Analizując zawartość wygenerowanej książki kodowej, zauważono, co wynika z procesu binarnej kwantyzacji RIS, że wiele wzorców się powtarza. Eliminując powtórzenia powstanie podzbiór wzorców $\tilde{\mathcal{P}}$:

$$\forall \mathcal{P}_j \in \{\mathcal{P}\} : (P_j \notin \tilde{\mathcal{P}}) \implies \tilde{\mathcal{P}} \leftarrow (\tilde{\mathcal{P}} \cup \{P_j\}). \tag{7}$$

Rozmiar zbioru $\tilde{\mathcal{P}}$ to w analizowanym przypadku zaledwie 495. Zatem sama ta operacja ogranicza rozpatrywaną przestrzeń rozwiązań o ponad 98%. Z tak przygotowaną książką kodową można podejść do wykonania pomiarów, ponieważ ilość wzorców w tak przygotowanej książce kodowej bez problemu można zmierzyć w wymiernym czasie. Po wykonaniu pomiarów, można dodatkowo zredukować rozmiar książki kodowej, wybierając wzorce dające najwyższy poziom odebranej mocy w zadanych lokalizacjach. W niniejszej pracy w części doświadczalnej, przyjęto metodę, polegającą na prostym wybieraniu jednego wzorca dla jednej lokalizacji. Tak wybrane wzorce utworzą ograniczoną książkę kodową.

4. TEST KSIĄŻKI KODOWEJ

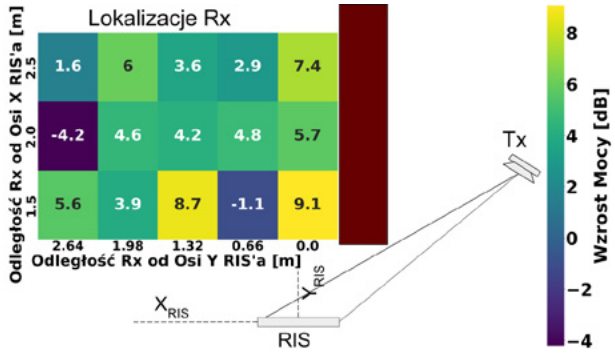
W niniejszej pracy jako miarę parametrów radiowych, przyjęto przepływność kanału, obliczaną na postawie twierdzenia Shannona:

$$C = W log_2(1 + SNR), \tag{8}$$

gdzie, C to całkowita przepływność kanału, W to szerokość pasma, a SNR to stosunek sygnału do szumu. Założono, że szerokość pasma to 80 MHz, a poziom szumu to -50 dBm.

Pomiary poziomów mocy wykonano w pomieszczeniu, dla częstotliwości nośnej 5,53GHz, korzystając z analizatora widma R&S FSV3000 (Rx), generatora sygnałów R&S SMM100A (Tx), anten kierunkowych METIS 5.1-5.8 GHz, matrycy RIS [12] oraz modułów UWB DecaWave Trek1000. Anteny i RIS umieszczono na tej samej wysokości, a moduły UWB zastosowano w celu pomiarów lokalizacji elementów systemu pomiarowego.

W celu wykonania odniesienia, dla efektywności proponowanej przez nas metody, zastosowano zmodyfikowaną metodę element po elemencie [9], do metody kolumna po kolumnie [11].



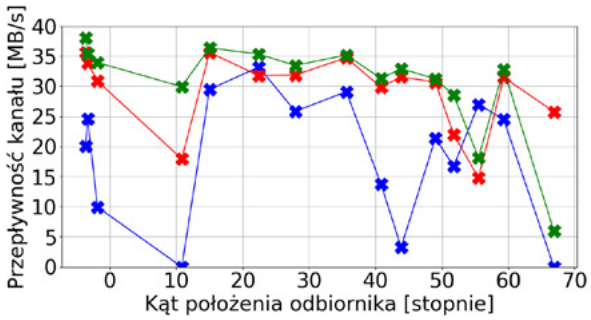
Rysunek 3: Heatmap'a przedstawiająca rozkład mocy odbieranej w badanych ustawieniach Rx, stosując ograniczoną książkę kodową. Moc przedstawiona jest w formie wzrostu względem scenariusza bez RIS. Oś X przedstawia odległość w metrach od ściany blokującej sygnał, a oś Y odległość od płaszczyzny, w której znajduje się RIS (również w metrach).

Pierwszy pomiar wykonano z pełną książką kodową. Następnie przygotowano ograniczoną wersję do drugiej serii pomiarów, w której wykonano pomiary poziomów mocy: referencyjnej (bez matrycy RIS), dla optymalizacji matrycy metodą kolumna po kolumnie i ograniczonej książki kodowej.

Podczas pomiarów ustawiano Rx, w 15 ustalonych położeniach (Rys. 3). Celem wykonywania pomiarów z

matrycą RIS w oddzielnych seriach, było nieidealne ponowne rozmieszczenie Rx, co pozwoliło sprawdzić zaproponowane rozwiązanie w bardziej praktycznym zastosowaniu.

Geometryczny rozkład wyników pomiarów poziomów mocy przedstawiono częściowo na rysunku 3. Na podstawie otrzymanych poziomów mocy dokonano obliczeń przepływności kanału, których wyniki przedstawiono na rysunku 4. Zastosowanie matrycy RIS podniosło przepływność kanału, z punktu widzenia użytkownika, szczególnie wtedy, kiedy przepływność bez RIS byłaby bliska zeru.



Rysunek 4: Wyniki obliczeń przepływności kanału w zmierzonych scenariuszach: *niebieski* – bez matrycy RIS, *czerwony* – z ograniczoną książką kodową, *zielony* – stosując metodę optymalizacji wzorca RIS kolumna po kolumnie.

5. PODSUMOWANIE

Rozwiązanie zaproponowane w niniejszej pracy zauważalnie poprawia przepływność kanału, szczególnie w sytuacjach skrajnych – kiedy kanał bez RIS zapewnia niewielką lub żadną przepływność, umieszczenie matrycy RIS z odpowiednim wzorcem zapewnia transmisję o zbliżonej przepustowości do pozostałych lokalizacji. Metoda ta daje bardzo zbliżone wyniki, w stosunku do optymalizacji elementów RIS metodą kolumna po kolumnie, zapewniając tylko jedno przełączenie matrycy, a więc nie przekraczając czasu koherencji kanału radiowego dla testowanej częstotliwości (co ma miejsce w przypadku optymalizacji). Kosztem tego rozwiązania jest potrzeba znajomości lokalizacji UE. Zastosowana w tej pracy metoda doboru wzorców, do ograniczonej książki, pozostawia też dalsze pole do jej rozwoju, nad czym autorzy aktualnie pracują. Odpowiednie dobranie wzorców, które podczas minimalizowania rozmiarów książki kodowej się w niej znajdą, jest zadaniem trudnym. Rozważyć należy podejścia oparte na: programowaniu liniowym (trudność stanowi maksymalizacja funkcji wypukłej), uczeniu maszynowym (jednak tutaj znów wraca problem zapotrzebowania na moc obliczeniową), klastrowaniu (metoda ta prawdopodobnie dała by dobre rezultaty, gdyby zagęszczono sieć pomiarów) lub algorytmach genetycznych (tutaj wyzwaniem jest odpowiednie zdefiniowanie problemu).

6. PODZIĘKOWANIA

Praca powstała w ramach projektu badawczego OPUS 2021/43/B/ST7/01365 finansowanego przez Narodowe Centrum Nauki w Polsce.

REFERENCES

[1] ETSI, "Reconfigurable Intelligent Surfaces (RIS); Use Cases, Deployment Scenarios and Requirements," ETSI, Tech. Rep. ETSI GR RIS 001 V1.1.1, 2023. [Online]. Available: https://www.etsi.org/committee/1966-ris.

[2] M. Renzo, M. Debbah, and D. Phan-Huy, "Smart radio environments empowered by reconfigurable ai meta-surfaces: An idea whose time has come," *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, no. 129, 2019. DOI: 10.1186/s13638-019-1438-9.

[3] R. Liu, J. Dou, P. Li, J. Wu, and Y. Cui, "Simulation and field trial results of reconfigurable intelligent surfaces in 5g networks," *IEEE Access*, vol. 10, pp. 122 786–122 795, 2022. DOI: 10.1109/ACCESS.2022.3223447.

[4] J. An and L. Gan, "The low-complexity design and optimal training overhead for irs-assisted miso systems," *IEEE Wireless Communications Letters*, vol. PP, pp. 1–1, May 2021. DOI: 10.1109/LWC.2021.3082773.

[5] J. An, C. Xu, L. Wang, Y. Liu, L. Gan, and L. Hanzo, "Joint training of the superimposed direct and reflected links in reconfigurable intelligent surface assisted multiuser communications," *IEEE Transactions on Green Communications and Networking*, vol. 6, no. 2, pp. 739–754, 2022. DOI: 10.1109/TGCN.2022.3143226.

[6] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5394–5409, 2019. DOI: 10.1109/TWC.2019.2936025.

[7] G. C. Trichopoulos, P. Theofanopoulos, B. Kashyap, *et al.*, "Design and evaluation of reconfigurable intelligent surfaces in real-world environment," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 462–474, 2022. DOI: 10.1109/OJCOMS.2022.3158310.

[8] H. Yang, R. Xiong, Y. Xiao, *et al.*, "Codebook configuration for ris-aided systems via implicit neural representations," in *ICC 2024 - IEEE International Conference on Communications*, 2024, pp. 1449–1454. DOI: 10.1109/ICC51166.2024.10623051.

[9] S. Tewes, M. Heinrichs, K. Weinberger, R. Kronberger, and A. Sezgin, "A comprehensive dataset of ris-based channel measurements in the 5ghz band," in *2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring)*, 2023, pp. 1–5. DOI: 10 . 1109 / VTC2023 - Spring57618.2023.10200973.

[10] H.-D. Lang, M. A. Nyffenegger, S. Keller, *et al.*, "Validating convex optimization of reconfigurable intelligent surfaces via measurements," in *2024 18th European Conference on Antennas and Propagation (EuCAP)*, 2024, pp. 1–5. DOI: 10.23919/EuCAP60739.2024.10501015.

[11] X. Pei, H. Yin, L. Tan, *et al.*, "Ris-aided wireless communications: Prototyping, adaptive beamforming, and indoor/outdoor field trials," *IEEE Transactions on Communications*, vol. 69, no. 12, pp. 8627–8640, 2021. DOI: 10.1109/TCOMM.2021.3116151.

[12] M. Heinrichs, A. Sezgin, and R. Kronberger, "Open source reconfigurable intelligent surface for the frequency range of 5 ghz wifi," in *2023 IEEE International Symposium On Antennas And Propagation (ISAP)*, 2023, pp. 1–2. DOI: 10.1109/ISAP57493.2023.10389095.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

PARAMETRY TRANSMISYJNE STANOWISKA DO SONDOWANIA KANAŁÓW RADIOWYCH W ZAKRESIE FAL MILIMETROWYCH

TRANSMISSION PARAMETERS OF THE MILLIMETER-WAVE CHANNEL SOUNDER

Tomasz Grasza[1]; Piotr Zdankowski[1]; Jarosław Wojtuń[1]; Rafał Przesmycki[1]; Jan Kelner[1]; Cezary Ziółkowski[1]

[1] Wojskowa Akademia Techniczna, Warszawa, {tomasz.grasza, piotr.zdankowski, jaroslaw.wojtun, rafał.przesmycki, jan.kelner, cezary.ziolkowski}@wat.edu.pl

DOI: 10.15199/59.2025.4.66

Streszczenie: **Artykuł przedstawia analizę parametrów transmisyjnych elementów składowych stanowiska do sondowania kanałów radiowych w zakresie fal milimetrowych. Głównym celem badań było określenie zysku systemu dla różnych scenariuszach pomiarowych. W tym celu przeprowadzono szczegółowe pomiary parametrów macierzy S przy użyciu wektorowego analizatora obwodów (VNA) w pięciu konfiguracjach antenowych, obejmujących zarówno anteny dookólne, jak i kierunkowe różnego typu. Badania objęły analizę wpływu odległości między antenami oraz zastosowania wzmacniaczy mocy i niskoszumnych na charakterystykę zysku systemowego.**

Abstract: **The article presents an analysis of the transmission parameters of the elements of a station probing radio channels in the millimeter wave range. The main objective of the research was to determine the gain of the system for various measurement scenarios. For this purpose, detailed measurements of the S-matrix parameters were carried out using a vector network analyzer (VNA) in five antenna configurations, including both omnidirectional and directional antennas of different types. The research included an analysis of the effect of the distance between the antennas and the use of power and low-noise amplifiers on the gain characteristics of the system.**

Słowa kluczowe: **fale milimetrowe, kanał radiowy, sondowanie kanałów, tłumienie, zysk systemowy.**

Keywords: **millimeter waves, radio channel, channel probing, attenuation, system gain.**

1. WSTĘP

Fale milimetrowe (mmWave) są obecnie jednym z kluczowych obszarów badań w kontekście nowoczesnych systemów komunikacji bezprzewodowej, w tym sieci 5G oraz przyszłych generacji (6G). Dzięki szerokiemu zakresowi częstotliwości (30–300 GHz) oraz wysokiej przepustowości, fale te oferują znaczący potencjał dla realizacji szybkich połączeń mobilnych oraz transmisji danych na dużą skalę. Wykorzystanie fal milimetrowych w praktyce komunikacyjnej napotyka jednak na istotne wyzwania związane z ich specyficznymi właściwościami propagacyjnymi. W szczególności fale milimetrowe są podatne na zakłócenia wynikające z tłumienia, odbić oraz rozproszeń sygnału, co powoduje konieczność szczegółowego badania ich charakterystyk propagacyjnych [1-3, 9, 10].

Sondowanie kanałów radiowych jest zatem istotnym procesem, który umożliwia szczegółową analizę warunków propagacyjnych fal radiowych w zróżnicowanych środowiskach, takich jak obszary miejskie, przemysłowe oraz wnętrza budynków. W kontekście fal milimetrowych szczególnie istotne staje się badanie parametrów takich jak strata propagacyjna (path loss), opóźnienie rozproszenia (delay spread), szerokość pasma koherencji oraz charakterystyki odbić i rozproszenia sygnału [4,5,7,8]. Dzięki zastosowaniu zaawansowanych metod sondowania możliwe jest dokładne modelowanie warunków propagacyjnych oraz optymalizacja parametrów transmisji w środowiskach o zróżnicowanej charakterystyce.

W literaturze można znaleźć liczne prace badające propagację fal milimetrowych w różnych środowiskach. Przykładem jest praca [3], która przedstawia dynamiczne sondowanie szerokopasmowe w scenariuszu makrokomórkowym przy częstotliwości 15 GHz, wskazując na znaczące straty sygnału w warunkach NLOS (Non-Line-Of-Sight) spowodowane efektem cienia radiowego przez rogi budynków oraz absorpcją przez roślinność Z kolei praca [6] analizuje wpływ środowiska przemysłowego na propagację fal milimetrowych, wskazując na rozproszenia i odbicia w warunkach dużego zagęszczenia przeszkód. Praca [9] koncentruje się na kanałach zwrotnych w zastosowaniach przemysłowych, analizując wpływ lokalnych warunków na tłumienie sygnału. Natomiast praca [8] przedstawia zastosowanie metod uczenia maszynowego do optymalizacji skanowanych fazowo układów antenowych, co może wspierać efektywność pomiarów w zmiennych warunkach propagacyjnych.

Analiza literatury wskazuje, że dobór metody sondowania kanału radiowego jest silnie uzależniony od charakterystyki środowiska oraz typu analizowanego sygnału [7, 8, 12].

2. STANOWISKO DO SONDOWANIA KANAŁÓW RADIOWYCH

2.1. Metody sondowania kanałów radiowych

Ze względu na specyfikę propagacji fal milimetrowych, w badaniach nad kanałami radiowymi stosuje się różnorodne metody sondowania, które umożliwiają precyzyjne odwzorowanie parametrów propagacyjnych w zróżnicowanych warunkach środowiskowych [9,10]. Odpowiedni dobór techniki badawczej zależy zarówno od

charakterystyki środowiska, jak i od rodzaju analizowanego sygnału oraz dynamiki propagacji.

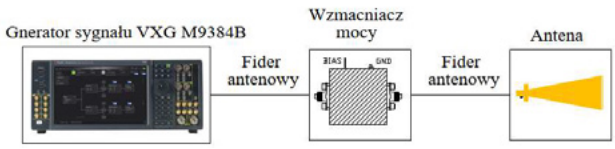
Jedną z metod sondowania kanałów radiowych jest technika impulsowa. Polega ona na generowaniu krótkich impulsów sygnałowych oraz analizie ich odbić i tłumienia w kanale. Dzięki krótkiemu czasowi trwania impulsu uzyskuje się wysoką rozdzielczość czasową i przestrzenną, co jest szczególnie przydatne w środowiskach o dużej dynamice, takich jak miejskie obszary zabudowane.

Innym podejściem jest sondowanie metodą korelacyjną, wykorzystujące sekwencje pseudolosowe (PN) do wyznaczenia funkcji odpowiedzi kanału. Dzięki zastosowaniu długich ciągów pseudolosowych możliwe jest uzyskanie wysokiej rozdzielczości czasowej, co pozwala na efektywne modelowanie odbić i rozproszeń sygnału w środowiskach o złożonej architekturze [10]. Technika ta znajduje zastosowanie zwłaszcza w środowiskach miejskich, gdzie wielodrogowość sygnału stanowi kluczowe wyzwanie.

Wektorowe analizatory obwodów są nieodzownym narzędziem w badaniach propagacji fal milimetrowych, szczególnie w zakresie sub-THz. Umożliwiają one precyzyjny pomiar parametrów macierzy S, co pozwala na szczegółową analizę zmienności sygnału w dynamicznych warunkach propagacyjnych [7]. Praca opisuje wykorzystanie VNA do badania propagacji sygnałów w zakresie sub-THz (220–300 GHz), co pozwala na dokładne odwzorowanie warunków propagacyjnych w środowiskach o dużej gęstości przeszkód.

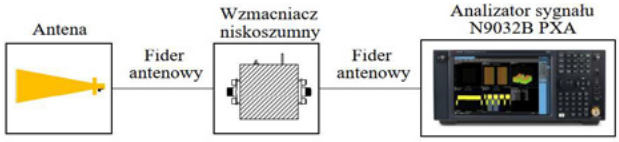
2.2. Charakterystyka stanowiska

W rozdziale przedstawiono część nadawczą i odbiorczą stanowiska do sondowania kanałów radiowych w zakresie fal milimetrowych. Cześć nadawcza stanowiska składa się z generatora sygnału mikrofalowego VXG model M9384B pracującego w zakresie częstotliwości od 1 MHz do 44 GHz z możliwością generacji sygnału o maksymalnej szerokości pasma wynoszącej 2 GHz, fidera antenowego, wzmacniacza, a także anteny.



Rys. 1. Nadawcza część stanowiska do sondowania kanałów radiowych

Na rysunku nr 2 została przedstawiona część odbiorcza stanowiska, w skład której wchodzi antena, wzmacniacz niskoszumny (LNA) oraz analizator sygnału N9032B PXA, który pozwala na rejestrację sygnału o szerokości pasma do 2 GHz w zakresie częstotliwości od 2 Hz do 44 GHz.



Rys. 2. Odbiorcza część stanowiska do sondowania kanałów radiowych

W celu umożliwienia realizacji różnych scenariuszy badawczych stanowisko zostało wyposażone w dwa typy wzmacniaczy mocy oraz dwa rodzaje wzmacniaczy niskoszumnych. Zdjęcia tych elementów pokazano na rys. 3. Podstawowe parametry tych elementów przedstawiono w tabeli nr 1. Parametry szczegółowe można znaleźć w dokumentacji technicznej [11-14].

W tabeli 2 zebrano podstawowe parametry anten wykorzystywane w badaniach.

Rys. 3. Różne typy wzmacniaczy

Tab. 1. Zestawienie wzmacniaczy oraz ich parametrów

Lp.	*Typ wzmacniacza*	*Parametry*
1	Wzmacniacz niskoszumny Model SBL-2434434025-2F2F-E3	Wzmocnienie 40 dB Zakres częstotliwości od 28 do 43,5 GHz
2	Wzmacniacz mocy Model SBP-2034434526-2F2F-S1	Wzmocnienie 40 dB Zakres częstotliwości od 20 do 44 GHz
3	Wzmacniacz niskoszumny Model SBL-1832735025-KFKF-S1	Wzmocnienie 50 dB Zakres częstotliwości od 18 do 26,5 GHz
4	Wzmacniacz mocy Model SBP-1832733030-KFKF-S1	Wzmocnienie 30 dB Zakres częstotliwości od 18 do 26,5 GHz

Tab. 2. Zestawienie różnych typów anten [15-18]

	Lp.	*Typ anteny*	*Parametry*
	1	Antena dookólna Model SAO-2734030810-KF-S1	Zakres częstotliwości od 26,5 do 40 GHz Zysk 7,5 dBi HPBW 360°
	2	Antena kierunkowa Model SAR-2507-28-S2	Zakres częstotliwości od 26,5 do 40 GHz Zysk 25 dBi HPBW 9°
	3	Antena kierunkowa Model SAR-2507-22-S2	Zakres częstotliwości od 33 do 50 GHz Zysk 25 dBi HPBW 9°
	4	Antena kierunkowa Model SAR-2507-42-S3	Zakres częstotliwości od 18 do 26,5 GHz Zysk 26 dBi HPBW 8°

3. POMIARY

Celem pomiarów z wykorzystaniem stanowiska sondującego jest wyznaczenie tłumienia środowiska $PL_{śr}$, którego zależność została opisana poniżej.

$$PL_{pom} = PL_{śr} - G_{SYS} \quad (1)$$

gdzie PL_{pom} oznacza tłumienie pomiarowe, $PL_{śr}$ określa tłumienie środowiska, a G_{SYS} opisuje zysk systemu.

Na zysk systemu składają się takie elementy jak tłumienie fidera, wzmocnienie wzmacniaczy oraz zysk antenowy obu anten.

$$G_{SYS}(f) = ATT_F(f) + G_{WZ}(f) + G_{AT}(f) \quad (2)$$

gdzie ATT_F oznacza tłumienie fidera, G_{WZ} opisuje wzmocnienie wzmacniaczy, a G_{AT} określa zysk antenowy obu anten.

Zatem problemem jest określenie wartości zysku systemu pomiarowego. Wartość zysku systemu możemy uzyskać poprzez pomiar poszczególnych elementów tworzących stanowisko pomiarowe.

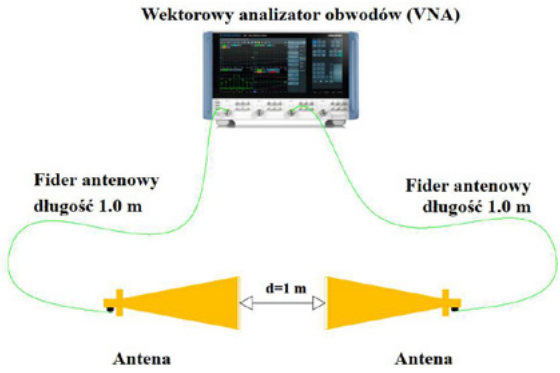
Wykorzystując wektorowy analizator obwodów (VNA) przeprowadzono pomiary parametrów macierzy S dla pięciu różnych scenariuszy pomiarowych oraz różnych częstotliwości dopasowanych do poszczególnych elementów układu, a także odległości pomiędzy antenami. W wyniku szczegółowych pomiarów z wykorzystaniem VNA uzyskano wartości parametru S_{21}, które opisują zysk poszczególnych elementów stanowiska sondującego.

Szczegółowe konfiguracje anten w poszczególnych scenariuszach pomiarowych, zakresy częstotliwości oraz obrane odległości pomiędzy antenami zostały opisane w tabeli nr 3.

Tab. 3. Zastosowane scenariusze pomiarowe

Nr	*Typ anteny*	*Typ anteny*	*Odległość d [m]*
1	Antena dookólna, model SA0-2734030810	Antena dookólna, model SA0273403810-KF-S1	0.5, 1, 1.5
	Zakres częstotliwości od 26,5 do 40 GHz		
2	Antena dookólna, model SA0-2734030810	Antena kierunkowa, model SAR-2507-28-S2	0.5, 1, 1.5
	Zakres częstotliwości od 26,5 do 40 GHz		
3	Antena kierunkowa, model SAR-2507-28-S2	Antena kierunkowa, model SAR-2507-28-S2	0.5, 1, 1.5
	Zakres częstotliwości od 26,5 do 40 GHz		
4	Antena kierunkowa, model SAR-2507-42-S3	Antena kierunkowa, model SAR-2507-42-S3	0.5, 1
	Zakres częstotliwości od 18 do 26,5 GHz		
5	Antena kierunkowa, model SAR-2507-22-S2	Antena kierunkowa, model SAR-2507-22-S2	0.5, 1, 1.5
	Zakres częstotliwości od 33 do 43 GHz		

Schemat połączeń pokazany na rysunku nr 4 ilustruje układ pomiarowy, w tym sposób podłączenia VNA do badanego układu, odległości pomiędzy antenami, a także długości kabli sygnałowych zastosowanych podczas pomiarów.



Rys. 4. Budowa stanowiska do pomiaru parametrów macierzy S

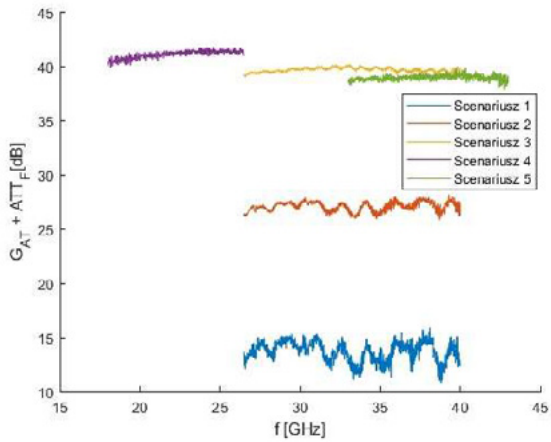
4. ANALIZA WYNIKÓW

Na podstawie szczegółowych pomiarów z wykorzystaniem VNA wykreślono poniższe wykresy. Na rysunku nr 5 został przedstawiony zysk anteny i tłumienie fidera dla pięciu scenariuszy badawczych bez uwzględnienia wzmacniaczy dla różnych odległości po odjęciu wartości tłumienia $FSPL$, które występuje na poszczególnych odległościach.

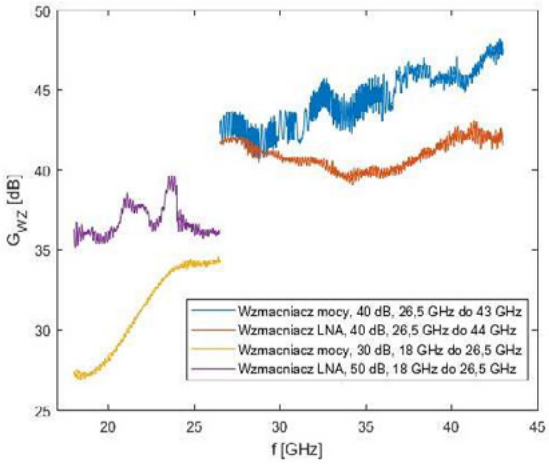
$$FSPL = 20log_{10}(d) + 20log_{10}(f) + 32{,}45 \quad (3)$$

gdzie d oznacza odległość pomiędzy antenami, a f określa częstotliwość.

W wyniku czego uzyskano zysk systemu antenowego z uwzględnieniem tłumienia fiderów antenowych. Na podstawie uzyskanych wyników można stwierdzić, że wraz ze wzrostem odległości charakterystyki zysku systemu nie ulegają zmianie.

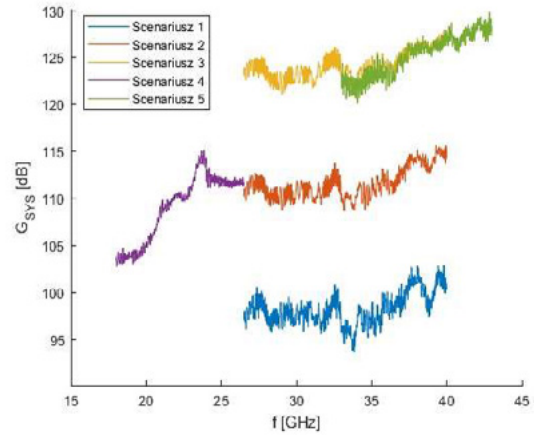


Rys. 5. Wykres przedstawiający G_{AT} zysk anteny i ATT_F tłumienie fidera antenowego dla pięciu scenariuszy badawczych bez wzmacniaczy po odjęciu wartości tłumienia, które występuje na poszczególnych odległościach



Rys. 6. Wzmocnienie wzmacniaczy G_{WZ} dla różnych zakresów częstotliwości

Uwzględniając zyski z pomiarów poszczególnych elementów stanowiska zamieszczonych w tabeli nr 1 i 2 wyznaczony został zysk systemu, którego wykres prezentowany jest na rysunku nr 7.



Rys. 7. Wykres G_{SYS}=f(f) dla pięciu scenariuszy badawczych

5. PODSUMOWANIE

Przeprowadzone badania i uzyskane na ich podstawie wyniki, mogą zostać wykorzystane w praktycznych realizacjach pomiarów, ponieważ dzięki nim na podstawie danych pomiarowych można wyznaczyć rzeczywiste wartości tłumienia mierzonego środowiska propagacji.

FINANSOWANIE

Praca powstała w ramach grantu badawczego ‘MubaMilWave’ nr 2021/43/I/ST7/03294 finansowanego przez Narodowe Centrum Nauki (NCN) oraz uczelnianego grantu badawczego nr UGB/22-059/2025/WAT, finansowanego przez Wojskową Akademię Techniczną (WAT).

LITERATURA

[1] Akada Takuma, Takahashi Shu, Fujimori Kazuhiro, Tanaka Toshiyasu. 2025. "A Proposal of Near-Field Measurement System Using Software Defined Radio for 5G Antenna Pattern Measurement." EUCAP 2025.

[2] Du Boyuan, Chen Jiaxin, Chen Peng. 2025. "Channel Measurement and Spatial Characteristics Analysis of mmWave Propagation in Maritime Evaporation Ducts." EUCAP 2025.

[3] Kourani Ali, Tuomela Juha, Kurvi Visa, Vähä-Savo Lauri, Haneda Katsuyuki. 2025. "Dynamic Wideband Channel Sounding in an Urban Macrocellular Scenario at 15 GHz." EUCAP 2025.

[4] Steinböck Gerhard, Käredal Johan, Olsson Bengt-Erik, Larsson Christina. 2025. "Empirical and Simulated Analysis of 28 GHz Propagation in Indoor Office Environments." EUCAP 2025.

[5] Prüller Richard, Langwieser Robert, Rupp Markus. 2025. "Indoor MIMO Channel Measurements of the Near-Field to Far-Field Transition in FR3." EUCAP 2025.

[6] Brennan Conor, Mbugua Allan Wainaina, Chen Yun, Hussain Sajjad. 2025. "Interpolation of Reflected and Diffracted Rays for Accelerated Ray Tracing Simulation." EUCAP 2025.

[7] Al-Dabbagh Mohanad Dawood, Humphreys David A., Kleine-Ostmann Thomas. 2025. "Investigating Time-Varying Signal Propagation at Sub-THz Frequencies Using a VNA." EUCAP 2025.

[8] Henderson James R., Hao Yang. 2025. "Machine Learning Accelerated Passive Electronically Scanned Phased Array Design." EUCAP 2025.

[9] Munoz Frédéric, Gougeon Gregory, Corre Yoann, D’Errico Raffaele. 2025. "Millimeter-Wave Backscattering Channel Modelling for Sensing Applications in an Industrial Scenario." EUCAP 2025.

[10] Oh Jun Hwa, Kwon Taek sun, Park Yuntae, Wi Sang-Hyuk. 2025."Pin-diode Based Low-loss Multi-state Electrical Phase Shifter for Sub-6 GHz Base Station." EUCAP 2025.

[11] https://www.eravant.com/24-25-to-43-5-ghz-40-db-gain-2-5-db-noise-figure-2-4-mm-f-connectors-low-noise-amplifier, Data dostępu: 01.06.2025.

[12] https://www.eravant.com/20-to-44-ghz-45-db-gain-26-dbm-p1db-2-4-mm-f-connectors-ka-band-power-amplifier, Data dostępu: 01.06.2025.

[13] https://www.eravant.com/18-to-26-5-ghz-50-db-gain-2-5-db-noise-figure-2-92-mm-k-f-connectors-k-band-low-noise-amplifier, Data dostępu: 01.06.2025.

[14] https://www.eravant.com/18-to-26-5-ghz-30-db-gain-30-dbm-p1db-2-92-mm-k-f-connectors-k-band-power-amplifier, Data dostępu: 01.06.2025.

[15] https://www.eravant.com/7-5-dbi-gain-26-5-to-40-ghz-2-92-mm-k-f-ka-band-omnidirectional-antenna, Data dostępu: 01.06.2025.

[16] https://www.eravant.com/25-dbi-gain-wr-28-ka-band-rectangular-horn-antenna, Data dostępu: 01.06.2025.

[17] https://www.eravant.com/25-dbi-gain-wr-22-q-band-rectangular-horn-antenna, Data dostępu: 01.06.2025.

[18] https://www.eravant.com/26-dbi-gain-wr-42-k-band-rectangular-horn-antenna, Data dostępu: 01.06.2025.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

POMIARY PARAMETRÓW KANAŁU RADIOWEGO W ZAKRESIE FAL MILIMETROWYCH DLA SIECI BAN WEWNĄTRZ POJAZDU

Measurements of Radio Channel at mmWave Band for Body-to-Vehicle Networks

Sławomir J. Ambroziak[1]; Agnieszka Czapiewska[1]; Krzysztof K. Cwalina[1]; Piotr Rajchowski[1]; Filipe D. Cardoso[2]; Manuel M. Ferreira[2]; Mariella Särestöniemi[3]; Anna Pietrenko-Dąbrowska[1]; Sławomir Kozieł[1,4]; Luis M. Correia[5]

[1] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, Polska
[2] Instituto Politécnico de Setúbal, Escola Superior de Tecnologia de Setúbal, INESC-INOV, Setúbal, Portugal
[3] Centre for Wireless Communications, Faculty of Information Technology and Electrical Engineering & Health Sciences and Technology, Faculty of Medicine, University of Oulu, Oulu, Finland
[4] Engineering Optimization and Modeling Center, Reykjavik University, Reykjavik, Iceland
[5] IST/INESC-INOV, University of Lisbon, 1049-001 Lisbon, Portugal

DOI: 10.15199/59.2025.4.67

Streszczenie: W referacie zaprezentowano pomiary parametrów kanału radiowego w zakresie fal milimetrowych, na częstotliwości 28 GHz i w paśmie 500 MHz, dla sieci BAN pracujących wewnątrz pojazdu. Opisano stanowisko pomiarowe oraz środowisko i scenariusze badawcze. Ponadto zaprezentowano analizę wstępnych wyników pomiarów profilu opóźnienia mocy, która wykazała, że największy wpływ na charakterystykę kanału radiowego ma wzajemna orientacja anten, a obecność pasażera nie ma na tę charakterystykę istotnego wpływu.

Abstract: This paper presents measurements of the radio channel at 28 GHz millimeter-wave band with a bandwidth of 500 MHz for body-to-vehicle networks. The measurement stand, environment and scenarios are described. The analysis of the preliminary results of the power delay profile is presented. It has been initially shown, that the orientation of the antennas has the greatest impact on the characteristics of the radio channel, while the presence of passengers has no significant effect.

Słowa kluczowe: kanał radiowy, sieci BAN, fale milimetrowe, modelowanie kanału radiowego.

Keywords: radio channel, Body Area Networks, milimeter waves, radio channel modellig.

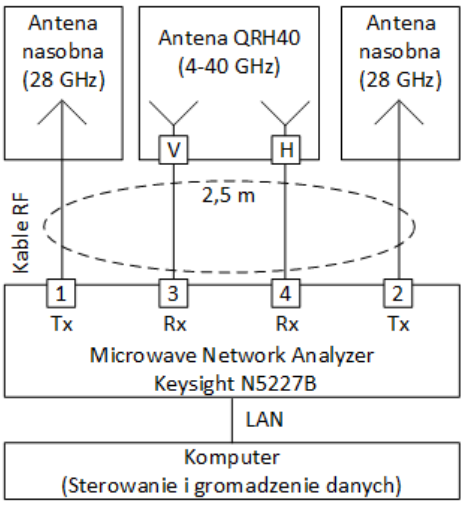
1. WSTĘP

Konieczność zapewnienia łączności pomiędzy urządzeniami wewnątrz pojazdu, a siecią BAN (*Body Area Network*) umieszczoną na ciele kierowcy lub pasażera i monitorująca jego różne parametry życiowe, wpisuje się w jeden ze scenariuszy użytkowych sieci 5G, a mianowicie V2X (*Vehicle-to-Everything*) [1]. Takie łącze radiowe typu *off-body* można zaklasyfikować jako łącze B2V (*Body-to-Vehicle*). Pasmo fal milimetrowych (*mmWave*) pokrywające zakres 24,25-52,6 GHz (FR2) jest rozważane również dla potrzeb 5G *New Radio* [2]. W celu zaprojektowania i optymalizacji niezawodnej sieci B2V niezbędne jest dogłębne poznanie charakterystyki kanału radiowego, tj. tłumienia systemowego, liczby składowych wielodrogowych, opóźnienia średniego oraz rozrzutu opóźnienia [3].

Niniejszy referat dotyczy sytuacji, w której użytkownik sieci BAN znajduje się wewnątrz pojazdu samotnie lub z pasażerem na bocznym siedzeniu. Te scenariusze zostały przebadane pod kątem charakterystyki kanału radiowego B2V w zakresie fal milimetrowych.

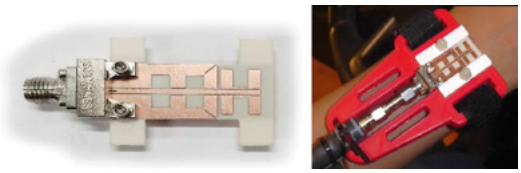
2. STANOWISKO POMIAROWE

Pomiary zrealizowano z wykorzystaniem mikrofalowego analizatora obwodów (MNA - *Microwave Network Analyser*) Keysight N5227B [4], który stanowi rdzeń stanowiska pomiarowego zaprezentowanego na Rys. 1. Na potrzeby pomiarów porty 1 i 2 analizatora działały jako nadajniki (Tx), natomiast porty 3 i 4 jako odbiorniki (Rx). Pozwoliło to na pomiar parametrów macierzy rozproszenia, tj. S31, S41, S32 i S42 w dziedzinie czasu, które mogą być rozważane jako profil opóźnienia mocy (PDP - *Power Delay Profile*) dla poszczególnych kanałów.



Rysunek 1: Schemat stanowiska pomiarowego

Jako antenę odbiorczą zastosowano dwupolaryzacyjną antenę QRH40 [5] (zysk 14 dBi i 30° wiązka główna dla 28 GHz niezależnie od polaryzacji). Natomiast w roli nasobnej anteny nadawczej zastosowano monopol planarny (Rys. 2) o zysku 11 dBi, szerokości wiązki głównej ±35° oraz charakteryzujący się współczynnikiem SAR (*Specific Absorption Rate*) poniżej limitu bezpieczeństwa IEEE wynoszącego 1.6 W/kg, jak wykazano w [6].



Rysunek 2: Nasobna antena nadawcza: prototyp (po lewej), zamontowana na nadgarstku (po prawej)

Połączenia anten z MNA zrealizowano przy pomocy kabli RF MWX061 [7] o długości 2,5 m. Po kalibracji stanowiska mierzone wartości składowych w profilu opóźnienia mocy odpowiadały tłumieniu sygnału między portem wejściowym anteny nadawczej, a portem wyjściowym anteny odbiorczej, czyli tłumieniu systemowemu [8].

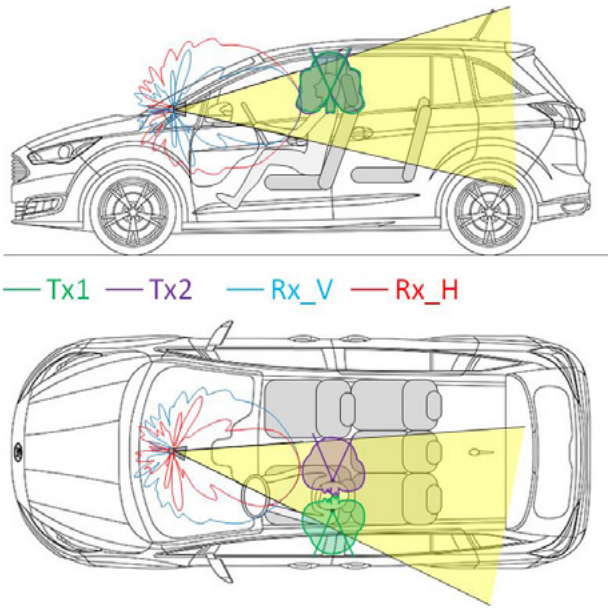
Badania przeprowadzono dla sygnału o częstotliwości 28 GHz i szerokości pasma 500 MHz (pasma n257 i n261 z zakresu FR2). Moc sygnału nadawanego przez porty Tx wynosiła 0 dBm. Przy 100 kHz szerokości pasma filtru IF (*Intermediate Frequency*) poziom szumów wynosił -88,9 dBm. Szerokość okna czasowego dla pomiaru PDP (Power Delay Profile) wynosiła 40 ns, przy 4001 punktach danych na pomiar. Dla każdego scenariusza pomierzono 200 profili opóźnienia mocy (100 dla scenariusza bez oddychania), przy czym średni czas pomiędzy pomiarami wynosił 276 ms. Należy podkreślić, że MNA realizuje pomiary w dziedzinie częstotliwości, zatem aby uzyskać PDP analizator wykonywał odwrotną transformatę Fouriera funkcji przenoszenia kanału radiowego z wykorzystaniem okna Kaisera-Bessela o współczynniku $\beta = 6$.

Za sterowanie pracą MNA odpowiedzialny był komputer z dedykowanym programem do automatyzacji pomiarów napisanym w G - standardowym języku LabVIEW [9]. Komputer także odpowiadał za zapisywanie wyników pomiarów na dysku - podczas badań zgromadzono ponad 5,6 GB danych.

3. ŚRODOWISKO I SCENARIUSZE

Pomiary zostały wykonane wewnątrz samochodu marki Ford Grand C-Max o wymiarach zewnętrznych $4,520 \times 1,828 \times 1,684$ m, który podczas badań miał zamknięte okna i był zaparkowany na podziemnym parkingu Politechniki Gdańskiej. Nie odnotowano istotnego wpływu innych osób i pojazdów na wyniki pomiarów. Analizator umieszczony był na tylnym siedzeniu, natomiast komputer - połączony z MNA za pomocą kabla Ethernet - znajdował się za pojazdem.

Podczas pomiarów dwie anteny nasobne były umiejscowione na użytkowniku-kierowcy w następujących konfiguracjach: na lewym i prawym nadgarstku (AB_L i AB_R), z lewej i prawej strony głowy (HE_L i HE_R) oraz na klatce piersiowej z polaryzacją pionową (V) i poziomą (H) (To_FV i To_FH), przy czym miejsca położenia anten opisano skrótami zaproponowanymi w [10]. Na miejsce instalacji dwupolaryzacyjnej anteny odbiorczej wybrano środek deski rozdzielczej oraz prawy górny róg przedniej szyby. Na Rys. 3 przedstawiono przykładową konfigurację anten oraz wzajemną orientację ich charakterystyk promieniowania.



Rysunek 3: Przykładowa wzajmena orientacja anten

Mając na względzie zachowanie użytkownika przebadane zostały trzy scenariusze: użytkownik nieruchomy oddychający normalnie, użytkownik nieruchomy wstrzymujący oddech oraz użytkownik poruszający się w sposób, jakby prowadził samochód (scenariusz quasi-dynamiczny). Każdy z powyższych scenariuszy został powtórzony dwukrotnie: bez pasażera i z pasażerem siedzącym po prawej stronie użytkownika-kierowcy. Podczas pomiarów, oprócz użytkownika siedzącego za kierownicą, zbadano także przypadek, w którym użytkownik stał na zewnątrz pojazdu w pobliżu drzwi kierowcy. Na Rys. 4 przedstawiono przykładowy scenariusz pomiarowy z użytkownikiem w fotelu kierowcy i pasażerem.

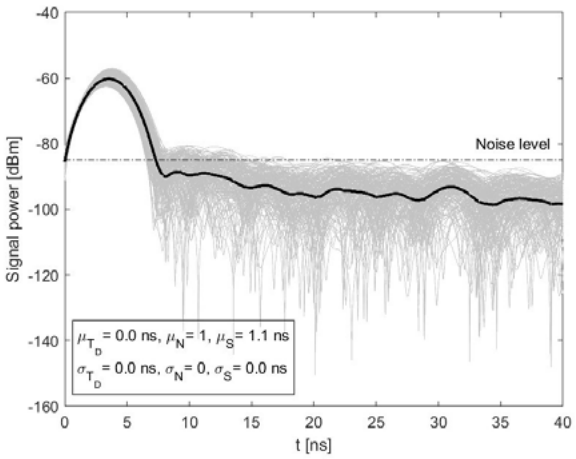
Rysunek 4: Przykładowy scenariusz pomiarowy

4. WSTĘPNE WYNIKI

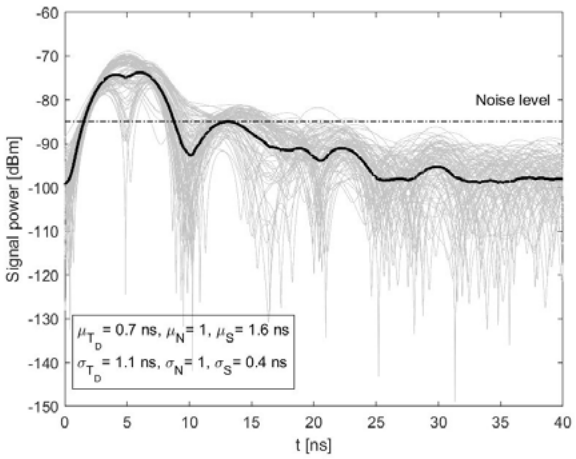
Wybrane wstępne wyniki pomiarów zostały przedstawione w formie wykresów, na których kolorem szarym zaprezentowano chwilowe profile opóźnienia mocy, a kolorem czarnym profile uśrednione (*Short-Term PDP* [11]).

Na potrzeby wstępnej analizy wybrano scenariusz z pionowo spolaryzowaną anteną odbiorczą zamontowaną na środku deski rozdzielczej oraz dwiema nasobnymi antenami nadawczymi zainstalowanymi na głowie użytkownika siedzącego na miejscu kierowcy i oddychającego normalnie (patrz Rys. 4). Na każdym z wykresów zaprezentowano również wartości średnie (μ) oraz odchylenia standardowe (σ) parametrów PDP, a mianowicie: opóźnienia średniego (T_D), liczby składowych wielodrogowych (N) oraz rozrzutu opóźnienia (S), obliczonych według [11].

Na Rys. 5 i Rys. 6 można zaobserwować wpływ umiejscowienia anteny nasobnej na PDP. Dla położenia HE_R (Tx2) widać wyraźnie jedynie składową bezpośrednią, podczas gdy dla położenia HE_L (Tx1) w PDP pojawia się również druga składowa odbita od elementów pojazdu, po lewej stronie użytkownika. Skutkuje to większymi wartościami T_D i S. Dodatkowo w drugim przypadku można zaobserwować niższe poziomy mocy wskutek większego tłumienia propagacyjnego. Należy podkreślić, że sytuacja, w której $\mu_{T_D} = 0$ i $\sigma_{T_D} = 0$, oznacza, że dany kanał jest bardzo stabilny i zawiera jedynie jedną składową wielodrogową.

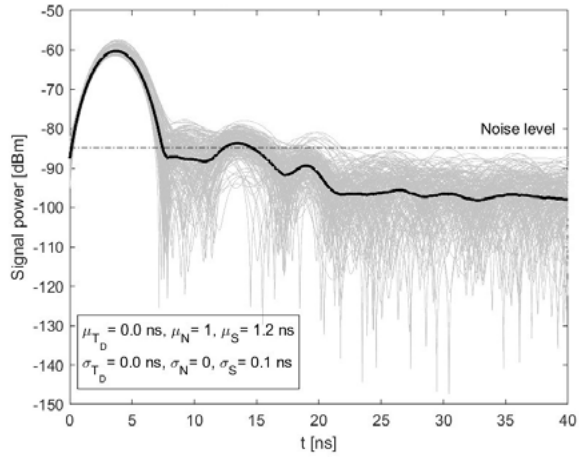


Rysunek 5: PDP dla HE_R - użytkownik nieruchomy, bez pasażera

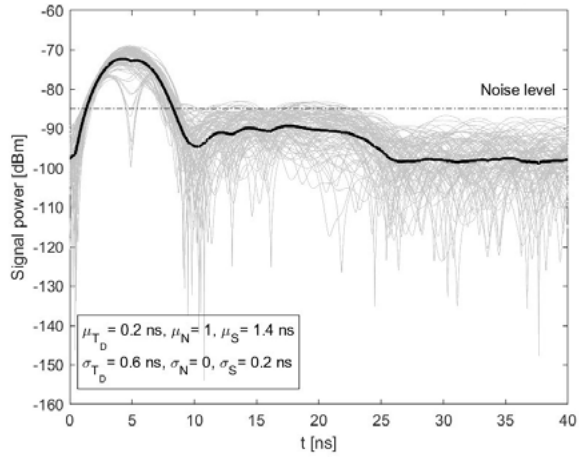


Rysunek 6: PDP dla HE_L - użytkownik nieruchomy, bez pasażera

Na Rys. 7 i Rys. 8 zaprezentowano PDP dla scenariusza z pasażerem, podczas gdy Rys. 9 i Rys. 10 pozwalają zaobserwować wpływ ruchu użytkownika w scenariuszu bez pasażera. Porównując Rys. 5 z Rys. 7 i Rys. 6 z Rys. 8 można zauważyć wpływ obecności pasażera: dla anteny HE_R widoczny jest wyższy poziom mocy późniejszych składowych wielodrogowych wskutek odbić od pasażera, z kolei dla anteny HE_L późne składowe wielodrogowe mają niższe moce wskutek ich tłumienia przez ciało pasażera. W pierwszym przypadku parametry czasowe PDP mają podobne wartości, ale w drugim są one istotnie niższe.

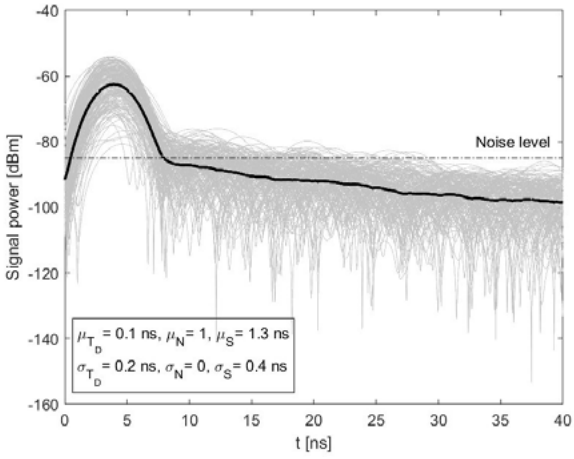


Rysunek 7: PDP dla HE_R - użytkownik nieruchomy, z pasażerem

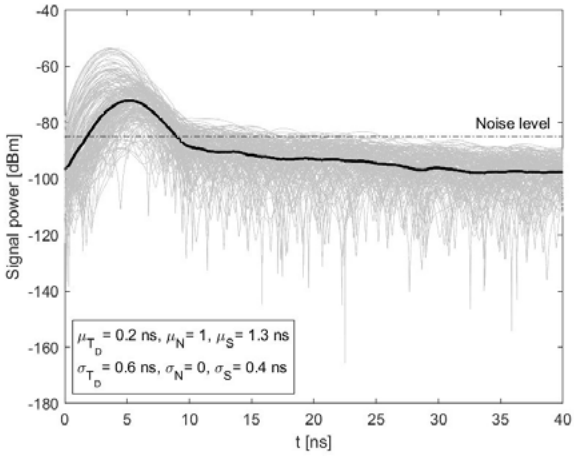


Rysunek 8: PDP dla HE_L - użytkownik nieruchomy, z pasażerem

Z kolei porównanie Rys. 5 z Rys. 9 i Rys. 6 z Rys. 10 pozwala zauważyć, że wpływ ruchu użytkownika jest również inny dla różnych miejsc położenia anteny nasobnej. Dla HE_R ruch użytkownika powoduje większe odchylenie standardowe T_D i S, natomiast dla HE_L otrzymano profil opóźnienia mocy o zupełnie innym kształcie, mniejszą liczbę składowych wielodrogowych oraz mniejsze wartości średnie parametrów czasowych PDP.



Rysunek 9: PDP dla HE_R - użytkownik ruchomy, bez pasażera



Rysunek 10: PDP dla HE_L - użytkownik ruchomy, bez pasażera

5. PODSUMOWANIE

W referacie zaprezentowano pomiary parametrów kanału radiowego dla łącza typu B2V w zakresie fal milimetrowych. Badania przeprowadzono dla sygnału o szerokości pasma 500 MHz na częstotliwości 28 GHz, jako jednej z możliwych częstotliwości przewidzianych dla sieci 5G. Opisano stanowisko pomiarowe, zawierające MNA oraz zastosowane anteny. Ponadto opisane zostało środowisko propagacyjne w postaci wnętrza samochodu osobowego, a także zrealizowane scenariusze pomiarowe.

Analiza wstępnych wyników pomiarów profilu opóźnienia mocy koncentrowała się na wpływie położenia anteny nasobnej na ciele użytkownika, na wpływie obecności pasażera oraz na wpływie ruchu użytkownika na wartości średnie i odchylenia standardowe parametrów czasowych PDP, tj. opóźnienia średniego i rozrzutu opóźnienia. Analiza ta wykazała, że wzajemna orientacja anten ma największy wpływ na charakterystykę kanału radiowego, natomiast obecność pasażera ma na tę charakterystykę wpływ nieistotny. Jednakże ten aspekt, podobnie jak wiele innych, wymaga głębszej analizy, co będzie przedmiotem dalszych prac.

6. PODZIĘKOWANIA

Zaprezentowane w referacie badania prowadzone były w ramach COST Action CA20120, *Intelligence-Enabling Radio Communications for Seamless Inclusive Interactions* (INTERACT), a także w ramach porozumienia o współpracy pomiędzy Wydziałem Elektroniki, Telekomunikacji i Informatyki Politechniki Gdańskiej, a Instituto Superior Técnico, University of Lisbon. Autorzy dziękują również Pani Hirze Manzoor za wkład w realizację badań pomiarowych.

LITERATURA

[1] A. V. Lopez, et al., "Opportunities and Challenges of mmWave NR," *IEEE Wireless Communications*, **26(2)**, April 2019, pp. 4–6.

[2] 3GPP TS 38.101-2, "NR; User Equipment (UE) Radio Transmission and Reception; Part 2: Range 2 Standalone," **v. 15.4.0**, Jan. 2019.

[3] S.J. Ambroziak, F.D. Cardoso, M.M. Ferreira, M. Särestöniemi and L.M. Correia, "People's Influence on Indoor Body Area Networks Channel Characteristics," *IEEE Access*, **12**, pp. 184285-184302, Dec. 2024.

[4] Keysight, "Keysight PNA Network Analyzer N5227B 900 Hz to 67 GHz," 2022, [online] Available: https://www.keysight.com/us/en/assets/9018-04327/technical-specifications/9018-04327.pdf.

[5] RF Spin, "Quad Ridged Horn Antenna – Model QRH40," [online] Available: https://www.rfspin.com/product/qrh40/.

[6] U. Ullah, S. Koziel and A. Pietrenko-Dabrowska, "Design and Characterization of a Planar Structure Wideband Millimeter-Wave Antenna With Wide Beamwidth for Wearable Off-Body Communication Applications," *IEEE Antennas and Wireless Propagation Letters*, **21**, 10, Oct. 2022, pp. 2070-2074, doi: 10.1109/LAWP.2022.3190567.

[7] Junkosha, "MWX0 SERIES," [online] Available: https://www.junkosha.com/en/products/EMF-00.

[8] ITU-R P.341-7, "The Concept of Transmission Loss for Radio Links," Aug. 2019, [online] Available: https://www.itu.int/rec/R-REC-P.341-7-201908-I/en.

[9] https://www.ni.com/pl-pl/shop/labview.html.

[10] M. Mackowiak, "Modelling MIMO Systems in Body Area Networks in Outdoors," *Ph. D. Dissertation*, IST-University of Lisbon, Lisbon, Portugal, 2013.

[11] ITU-R P.1407-8, "Multipath propagation and parameterization of its characteristics," Sep. 2021, [online] Available: https://www.itu.int/dms_pubrec/itu-r/rec/p/R-REC-P.1407-8-202109-I!!PDF-E.pdf.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

Projektowanie książki kodowej RIS z wykorzystaniem danych pomiarowych

Measurement-Based Codebook Design for RIS

Paweł Hatka[1]; Marcel Garczyk[1]; Paweł Płaczkiewicz[1]; Dawid Brząkała[1]; Mikołaj Wypych[1]; Piotr Skrzypczak[1]; Marcin Pakuła[1]; Maciej Niedźwiecki[1]; Paweł Sroka[2]; Michał Sybis[2]

[1] Politechnika Poznańska, imie.nazwisko@student.put.poznan.pl
[2] Politechnika Poznańska, Instytut Radiokomunikacji, imie.nazwisko@put.poznan.pl

DOI: 10.15199/59.2025.4.68

Streszczenie: Niniejsza praca przedstawia praktyczne podejście do projektowania książki kodów oraz doboru słów kodowych dla rekonfigurowalnej inteligentnej powierzchni (ang. Reconfigurable Intelligent Surfaces, RIS), w którym wykorzystuje się informacje o położeniu odbiornika w celu ograniczenia liczby słów kodowych uwzględnianych podczas optymalizacji wzorca odbicia RIS. Zaprezentowano procedurę tworzenia książki kodowej opartą na pomiarach, przeznaczoną dla scenariusza z mobilnym odbiornikiem. Uzyskane wyniki jednoznacznie wskazują na zasadność zaproponowanego podejścia.

Abstract: This work presents a practical approach to codebook design and codeword selection for reconfigurable intelligent surface (RIS), where position information is used to limit the number of codewords considered in the RIS reflection pattern optimization. A measurement-based codebook formulation procedure is shown for a scenario with a mobile receiver. The results clearly indicate that the proposed approach with codebook size reduction based on position information is valid.

Słowa kluczowe: RIS, pomiary, opracowanie książki kodowej

Keywords: RIS, measurements, codebook design

1. WSTĘP

Rekonfigurowalne inteligentne powierzchnie [6] cieszą się rosnącym zainteresowaniem w kontekście nowoczesnych systemów komunikacji bezprzewodowej [8], przede wszystkim ze względu na swój potencjał w zakresie kształtowania propagacji sygnału w sposób wcześniej niedostępny [1]. Urządzenia te, działające niemal pasywnie, mogą być integrowane z różnymi elementami infrastruktury, takimi jak elewacje budynków czy ściany wewnętrzne, umożliwiając m.in. kontrolowane odbicia fal elektromagnetycznych, wykraczające poza klasyczne reguły geometrii optycznej [2].

Jednym z głównych zastosowań technologii RIS jest poprawa warunków propagacji sygnału w sytuacjach, gdy bezpośrednia droga propagacyjna między nadajnikiem a odbiornikiem (ang. Line-of-Sight, LoS) jest niedostępna. Dzięki odpowiednio dobranym przesunięciom fazowym w elementach RIS możliwe jest utworzenie alternatywnego toru transmisji, co pozwala traktować tę powierzchnię jako sterowalny reflektor. Aby jednak w pełni wykorzystać jej możliwości, niezbędna jest precyzyjna znajomość stanu kanału (ang. Channel State Information, CSI), co w warunkach rzeczywistych stanowi poważne wyzwanie praktyczne.

W związku z trudnościami w uzyskaniu dokładnej informacji o stanie kanału, powszechnie stosuje się podejścia oparte na książkach kodowych, w których zamiast pełnych danych kanałowych wykorzystywany jest ograniczony zbiór wcześniej zdefiniowanych konfiguracji przesunięć fazowych [3]. Jednak zarówno projektowanie samej książki kodowej, jak i wybór optymalnej konfiguracji wiązki są zadaniami złożonymi, czasochłonnymi i wymagającymi znacznych zasobów obliczeniowych [3].

W niniejszej pracy rozważamy alternatywne podejście, oparte na informacji o położeniu urządzeń, inspirowane rozwiązaniem zaproponowanym w [5]. Zakłada ono, że dla znanych lokalizacji nadajnika, odbiornika i powierzchni RIS możliwe jest wygenerowanie ograniczonego podzbioru potencjalnych konfiguracji wiązki, co znacząco upraszcza proces wyboru.

Celem naszych badań jest opracowanie mapy środowiska radiowego (ang. Radio Environment Map, REM) [7], będącej połączeniem bazy danych i systemu zarządzania, która umożliwia wybór odpowiedniej książki kodów na podstawie informacji o lokalizacji [4]. Takie podejście ma na celu uproszczenie procesu pasywnego kształtowania wiązki oraz ograniczenie złożoności obliczeniowej systemu.

W ramach tej pracy przedstawiamy pierwsze kroki i pierwsze wyniki dotyczące opracowania i zastosowania proponowanego rozwiązania w środowisku miejskim. W eksperymencie wykorzystano rzeczywisty układ pomiarowy, w którym otwarto-źródłowa rekonfigurowalna inteligentna powierzchnia [10] została zainstalowana na terenie parkingu, umożliwiając komunikację między stacjonarnym nadajnikiem a poruszającym się odbiornikiem. Na podstawie zebranych danych analizowano moc sygnału odbieranego dla różnych konfiguracji przesunięć fazowych RIS (wzorów odbicia). Celem jest opracowanie książek kodów zależnych od położenia urządzeń dla danego scenariusza, które w kolejnym kroku mogą zostać zapisane w mapie REM.

2. SCENARIUSZ I KONFIGURACJA POMIAROWA

2.1 Procedura pomiarowa

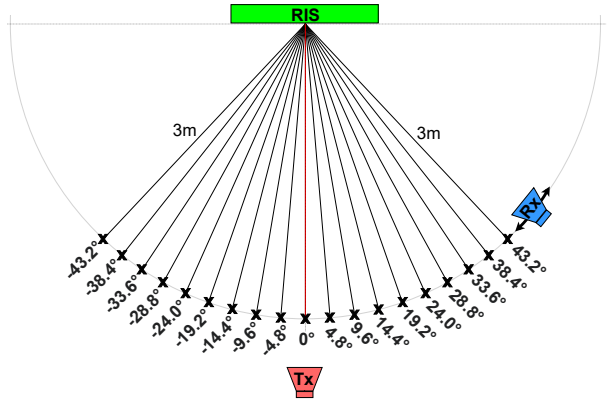
Pomiary przeprowadzono na parkingu, zgodnie z założeniami opracowanego scenariusza eksperymentalnego. W rozważanym scenariuszu pomiarowym nadajnik, wyposażony w antenę kierunkową AINFO LB-187-20-C-SF, pracujący z częstotliwością 5,5 GHz został umieszczony w odległości 5,27 m od powierzchni RIS, na wysokości 1,89 m. Antena nadawcza była precyzyjnie skierowana na środek połączonej powierzchni jednostek RIS, a cały system nadawczy obejmował również generator sygnałowy Rohde & Schwarz SMM100A, emitujący sygnał o mocy –10 dBm. Odbiornik, zintegrowany z analizatorem widma i zamontowany na wózku pomiarowym, przemieszczał się po obwodzie okręgu o promieniu 3 m wokół RIS, zachowując kątową rozdzielczość pomiarów równą 4,8°. Szczegółowe parametry generatora sygnałów zostały przedstawione w Tab. 1.

Tabela 1: Parametry generatora sygnałowego

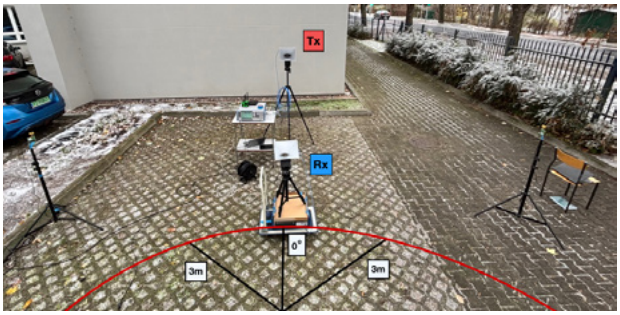
Parametr	Wartość
Moc nadawcza	–10 dBm
Częstotliwość środkowa	5,5 GHz
Tryb	CW

Pomiary mocy sygnału odbieranego zostały przeprowadzone w 19 punktach rozmieszczonych co 25 cm na obwodzie okręgu o promieniu 3 m wokół RIS, co odpowiada wspomnianej wcześniej rozdzielczości kątowej i zostało przedstawione na Rysunku 1. Odbiornik, zamontowany na wózku pomiarowym, był wyposażony w analizator widma Rohde & Schwarz FSV3000 oraz antenę odbiorczą AINFO LB-187-20-C-SF, umieszczoną na wysokości 0,92 m nad ziemią. W każdym z punktów pomiarowych rejestrowano moc odbieranego sygnału dla 27 wcześniej wybranych wzorców odbicia skonfigurowanych na jednostkach RIS. Celem pomiarów było wyznaczenie takiego wzorca, który zapewnia największą moc odbieranego sygnału w danym położeniu odbiornika. Punkt 0°, zaznaczony na Rysunkach 1 oraz 2, odpowiada sytuacji, w której odbiornik znajduje się w jednej linii z nadajnikiem i powierzchnią RIS. Szczegółowe parametry analizatora widma zostały przedstawione w Tab. 2.

Do precyzyjnego określania pozycji odbiornika podczas pomiarów zastosowano system lokalizacji szerokopasmowej (ang. Ultra WideBand, UWB) TREK1000 firmy Decawave. System ten wykorzystuje technologię pomiaru czasu przelotu (ang. Time of Flight, ToF) w paśmie UWB, co umożliwia wyznaczanie położenia z dokładnością rzędu kilkunastu centymetrów w warunkach rzeczywistych. W eksperymencie wykorzystano układ składający się z trzech zakotwiczonych w przestrzeni urządzeń referencyjnych oraz jednego urządzenia mobilnego umieszczonego na wózku pomiarowym. Dane lokalizacyjne były rejestrowane równolegle z pomiarami mocy sygnału odbieranego, co pozwoliło na precyzyjne przypisanie wyników pomiarów do konkretnych punktów trajektorii ruchu odbiornika.



Rysunek 1: Scenariusz pomiarowy



Rysunek 2: Widok scenariusz pomiarowego z perspektywy RISa

Tabela 2: Parametry analizatora widma

Parametr	Wartość
Zakres (Span)	0 Hz
RBW	500 Hz
Czas przebiegu	50 ms
Poziom odniesienia	–30 dBm
Częstotliwość środkowa	5,5 GHz
Poziom szumów	–110 dBm
Tryb	MaxHold

W ramach eksperymentu cztery jednostki RIS zostały połączone w jedną dużą powierzchnię, która następnie została zamontowana na słupie oświetleniowym znajdującym się na terenie parkingu. Powierzchnię umieszczono na wysokości 3 m nad poziomem gruntu i nachylono o 24,4° w płaszczyźnie elewacji. Połączenie jednostek miało na celu zwiększenie efektywnej powierzchni odbicia, a w konsekwencji poprawę mocy sygnału docierającego do odbiornika.

Cały proces pomiarowy został zautomatyzowany za pomocą oprogramowania napisanego w języku Python, co umożliwiło zdalne sterowanie analizatorem widma i generatorem sygnałowym, a także zmianę wzorców odbicia ustawianych na RIS. Pomiary przeprowadzono dla 27 arbitralnie wybranych wzorców odbicia.

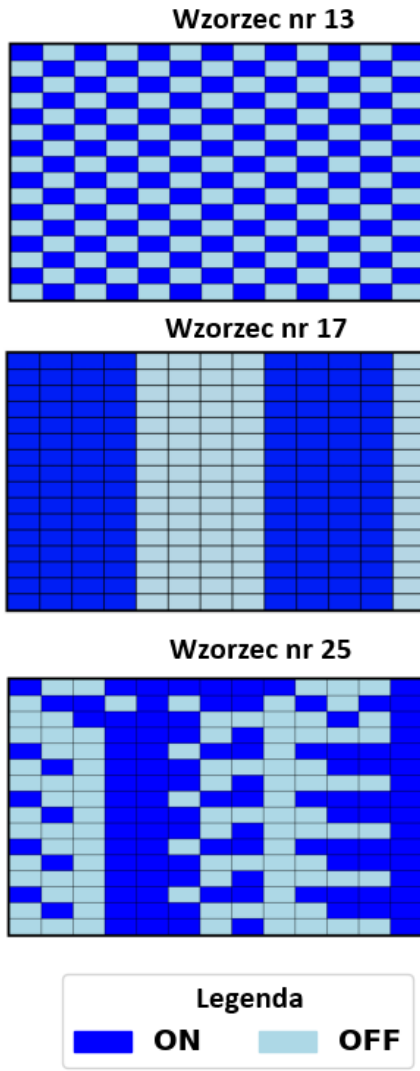
2.2 Sterowanie pojedynczą jednostką RIS

Sterowanie pojedynczą jednostką RIS o rozdzielczości 1-bitowej polega na przesłaniu sekwencji 256 bitów,

z których każdy odpowiadał jednemu elementowi antenowemu — określając, czy ma być on aktywny, czy nieaktywny. Elementy w stanie nieaktywnym nie modyfikowały fazy sygnału wynikającej z jego odbicia od powierzchni RIS-a, natomiast te w stanie aktywnym wprowadzały dodatkowe przesunięcie fazowe o ściśle określony kąt. W przypadku częstotliwości użytej w testach, wartość przesunięcia wynosiła około 180°, zgodnie z charakterystyką zastosowanej macierzy.

Można zatem powiedzieć, że sterowanie macierzą polega na wskazaniu, które elementy w danej grupie mają być aktywne, a które nie. Numeracja grup rozpoczyna się od indeksu 0, przy czym pierwsza grupa znajduje się w lewym górnym rogu macierzy. Konfiguracja przesyłana jest do RIS w postaci sekwencji liczb zapisanych w kodzie szesnastkowym, gdzie pozycja w sekwencji odpowiada numerowi grupy. Przykładowo:

- `0x0` (bitowo `0000`) — wszystkie elementy grupy są nieaktywne,
- `0xF` (bitowo `1111`) — wszystkie elementy grupy są aktywne,
- `0x3` (bitowo `0011`) — aktywne są dwa ostatnie elementy w grupie.



Rysunek 3: Przykładowe wzorce RIS wykorzystane w badaniach

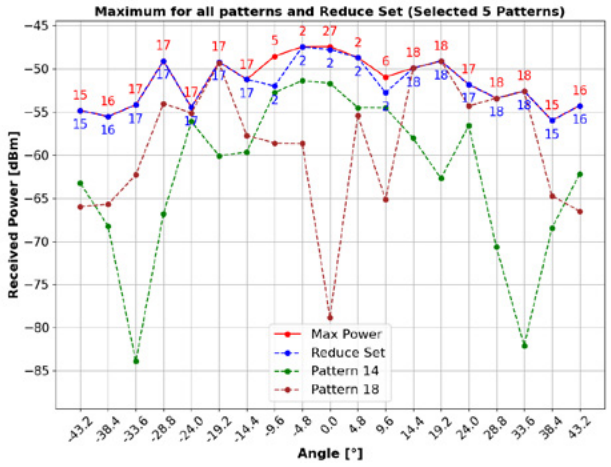
Przykładowa sekwencja `!0x800...000` oznacza aktywację wyłącznie pierwszego elementu w grupie 0 (elementem w lewym górnym rogu macierzy), a sekwencja `!0xF00...000` oznacza aktywację wszystkich elementów grupy 0 [9].

Przykładowe wzorce rozpatrywane w pracy zostały przedstawione na Rys. 3. Należy jednak zauważyć, że przedstawione wzorce są jedynie małym ułamkiem całego spektrum możliwych konfiguracji. W literaturze znaleźć można inne wzorce, które będą pozwalały na uzyskanie odmiennych charakterystyk Odbiciowych. Szczegółowe ich rozpoznanie i klasyfikacja stanowią jednak odrębny kierunek badań, wykraczający poza zakres niniejszej pracy.

3. WYNIKI

Uzyskane wyniki zaprezentowano na Rysunku 4, który przedstawia moc sygnału odbieranego w funkcji kąta położenia odbiornika względem powierzchni RIS. Analiza obejmuje dwa podejścia: Maximum Power (linia czerwona) oraz Reduced Set (linia niebieska). W przypadku metody Maximum Power, dla każdego kąta wybierany jest ten wzorzec odbicia, który – spośród całego zbioru testowanych konfiguracji – zapewnia największą moc sygnału odbieranego. Z analizy wynika, że niektóre wzorce (np. nr 17 i 18) wybierane są znacznie częściej, podczas gdy inne (np. nr 5 i 6) zostały użyte jedynie raz. Zjawisko to stanowi uzasadnienie dla zastosowania podejścia Reduced Set, w którym książka kodowa ograniczona jest do pięciu najczęściej wybieranych wzorców z metody Maximum Power.

Przy zastosowaniu Reduced Set, pogorszenie mocy odbieranego sygnału jest pomijalne, natomiast znacząco zmniejszony zostaje rozmiar przestrzeni przeszukiwań oraz częstotliwość przełączania wzorców, co przekłada się na bardziej efektywne rozwiązanie pod względem implementacyjnym.



Rysunek 4: Uzyskane wyniki mocy odebranej w funkcji kąta

Rysunek 4 zawiera także dwa dodatkowe przebiegi, odpowiadające rozkładom mocy dla wzorców nr 14 i 18. Profile te wyraźnie pokazują, że poszczególne konfiguracje RIS charakteryzują się różnymi maksymalnymi i minimalnymi poziomami mocy w zależności od kąta, co stwarza możliwość bardziej precyzyjnego kształtowania wiązki, w tym również ograniczania interferencji. Właściwość

ta może być szczególnie użyteczna w scenariuszach wielodostępowych, gdzie zakłócenia między użytkownikami mają istotny wpływ na jakość działania systemu. Pozwala przykładowo dobrać taki wzorzec, który zapewni maksymalizację mocy dla jednego z użytkowników jednocześnie minimalizując ją w kierunku innych.

4. WNIOSKI

W pracy zaprezentowano podejście do projektowania książki kodów dla rekonfigurowalnej inteligentnej powierzchni z wykorzystaniem informacji o położeniu, opracowane na podstawie rzeczywistych pomiarów przeprowadzonych w środowisku praktycznym. Eksperyment został zrealizowany na terenie parkingu, gdzie rozmieszczono nadajnik, odbiornik oraz powierzchnię RIS. Szczegółowo opisano zarówno przebieg eksperymentu, jak i zastosowaną konfigurację sprzętową i pomiarową.

Zaprezentowane wyniki jednoznacznie potwierdzają zasadność stosowania książek kodów o ograniczonym rozmiarze, dobieranych w oparciu o lokalizację urządzeń komunikacyjnych. Takie podejście pozwala istotnie zmniejszyć złożoność obliczeniową procesu kształtowania wiązki, przy jednoczesnym zachowaniu wysokiej jakości sygnału odbieranego.

W ramach dalszych prac planowane jest rozszerzenie badań o scenariusze z propagacją bez bezpośredniej widoczności (ang. Non-Line-of-Sight, NLOS), w których sygnał dociera do odbiornika jedynie poprzez odbicia od przeszkód takich jak budynki czy inne obiekty infrastrukturalne.

Dodatkowo zakłada się wykorzystanie RIS zamontowanego na platformie bezzałogowego statku powietrznego (ang. Unmanned Aerial Vehicle, UAV), co pozwoli na analizę działania systemu w warunkach dynamicznych i trójwymiarowej mobilności, obejmujących zarówno zmiany wysokości, jak i zmiany położenia w płaszczyźnie poziomej. Użycie UAV umożliwi symulację różnorodnych trajektorii ruchu oraz testowanie systemu w środowiskach trudno dostępnych lub niebezpiecznych dla tradycyjnej infrastruktury naziemnej. Planuje się również badanie wpływu parametrów lotu, takich jak prędkość czy zmienność kursu, na proces śledzenia sygnału i dostosowywania wzorców odbicia na RIS. Analiza tych rozszerzonych przypadków pozwoli lepiej zrozumieć efektywność generowania książek kodów opartych na informacji o położeniu w złożonych środowiskach propagacyjnych, torując drogę do bardziej odpornych i uniwersalnych rozwiązań.

Takie zaawansowane scenariusze umożliwią ocenę skuteczności projektowania książek kodowych w jeszcze bardziej złożonych i realistycznych warunkach propagacyjnych, uwzględniających mobilność nadawczo-odbiorczą, zmienne otoczenie i ograniczenia czasowe związane z przeliczaniem konfiguracji. Wyniki tych badań otworzą drogę do budowy bardziej elastycznych, skalowalnych i adaptacyjnych systemów RIS, które będą mogły znaleźć zastosowanie w nowoczesnych sieciach komunikacji bezprzewodowej, także w kontekście rozwoju technologii 6G i internetu rzeczy (ang. Internet of Things, IoT).

Praca została zrealizowana w ramach projektu nr 2021/43/*B*/*ST*7/01365 *ufundowanego przez Narodowe Centrum Nauki (NCN) w Polsce.*

LITERATURA

[1] Di, B., Zhang, H. 2022. "Intelligent Omni-Surfaces: Simultaneous Refraction and Reflection for Full-Dimensional Wireless Communications". IEEE Communications Surveys & Tutorials 24 (4): 1997-2028. doi: 10.1109/COMST.2022.3202813.

[2] Di Renzo, M, Debbah, M., Phan-Huy, D. et al. 2019. "Smart radio environments empowered by reconfigurable AI meta-surfaces: an idea whose time has come". J Wireless Com Network 2019, 129. https://doi.org/10.1186/s13638-019-1438-9

[3] An, J., et al. 2024. "Codebook-Based Solutions for Reconfigurable Intelligent Surfaces and Their Open Challenges". IEEE Wireless Communications, 31(2), 2024. https://doi.org/10.1109/MWC.010.2200312

[4] Yu, Z., et al. 2024. "Environment-Aware Codebook Design for RIS-Assisted MU-MISO Communications: Implementation and Performance Analysis". IEEE Transactions on Communications, 72(12), 2024. https://doi.org/10.1109/TCOMM.2024.3415619

[5] Nor, A. M., Fratu, O., Halunga, S. 2023. "Positioning Information-Based Codebook for Reconfigurable Intelligent Surface Passive Beamforming". IEEE Open Journal of the Communications Society, 4, 2023. https://doi.org/10.1109/OJCOMS.2023.3334474

[6] ETSI. 2023. "Reconfigurable Intelligent Surfaces (RIS); Use Cases, Deployment Scenarios and Requirements". ETSI GR RIS 001 V1.1.1 (2023-04). https://www.etsi.org/committee/1966-ris, accessed online: January 3, 2025.

[7] Wei, Z., Zhang, Q., Feng, Z., Li, W., Gulliver, T. A. 2013. "On the construction of Radio Environment Maps for Cognitive Radio Networks". 2013 IEEE Wireless Communications and Networking Conference (WCNC). https://doi.org/10.1109/WCNC.2013.6555304,

[8] E. Shi et al., "RIS-Aided Cell-Free Massive MIMO Systems for 6G: Fundamentals, System Design, and Applications," in Proceedings of the IEEE, vol. 112, no. 4, pp. 331-364, April 2024, doi: 10.1109/JPROC.2024.3404491.

[9] OpenSourceRIS https://github.com/mheinri/OpenSourceRIS

[10] Heinrichs, M., Sezgin, A., Kronberger, R. 2023. "Open Source Reconfigurable Intelligent Surface for the Frequency Range of 5 GHz WiFi". 2023 IEEE International Symposium On Antennas And Propagation (ISAP), Kuala Lumpur, Malaysia, pp. 1-2. https://doi.org/10.1109/ISAP57493.2023.10389095.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

ZESTAW BADAWCZY DO POMIARU PARAMETRÓW CHARAKTERYSTYCZNYCH ŁĄCZA JONOSFERYCZNEGO

RESEARCH SET FOR MEASURING CHARACTERISTIC PARAMETERS OF THE IONOSPHERIC LINK

Bogdan Uljasz
Wojskowa Akademia Techniczna im. Jarosława Dąbrowskiego, Wydział Elektroniki,
Instytut Systemów Łączności, ul. Gen. Sylwestra Kaliskiego 2, 00-908 Warszawa,
bogdan.uljasz@wat.edu.pl

DOI: 10.15199/59.2025.4.69

Streszczenie: **Badania testowymi sygnałami transmisyjnymi zjawisk występujących podczas transmisji sygnału radiowego przez rzeczywisty kanał jonosferyczny ma istotne znaczenie dla modelowania i analizy możliwości transmisji fal radiowych w jonosferze. W celu wykonania możliwie najbardziej wiarygodnych pomiarów należy przygotować zestaw badawczy wprowadzający pomijalne w analizie otrzymanych wyników badań błędy. W artykule opisano podstawowe zjawiska i parametry charakterystyczne dla łącza jonosferycznego. Opisano budowę stanowiska nadawczego oraz odbiorczego. Opisano również niezbędne czynności dla minimalizacji błędów wprowadzanych przez układ badawczy. Na koniec przedstawiono testy weryfikujące możliwości pomiarowe stanowiska w zakresie parametrów częstotliwościowych.**

Abstract: **Studying the phenomena occurring during radio signal transmission through a real ionospheric channel using test transmission signals is of great importance for modeling and analyzing the possibilities of radio wave transmission in the ionosphere. To achieve the most accurate measurements, it is essential to design a research setup that minimizes errors in the analysis of the results. The article discusses the main phenomena and parameters characteristic of the ionospheric link. It describes the construction of the transmitting and receiving stations and outlines the necessary measures to reduce errors in the research system. Finally, tests verifying the measurement capabilities of the station in terms of frequency parameters were presented.**

Słowa kluczowe: **propagacja fal radiowych w jonosferze, testy łącza jonosferycznego, pomiary.**

Keywords: **radio wave propagation in the ionosphere, ionospheric link tests, measurements.**

1. WSTĘP

Właściwości transmisyjne ośrodka jakim jest jonosfera zależą od wielu parametrów łącza radiowego, które jest zmienne w czasie. Dla opisania kanału jonosferycznego niezbędne jest przebadanie i na podstawie wyników badań opracowanie wartości opisującej matematycznie transmitancję kanału. Uzyskane rezultaty wraz z danymi sondowania jonosferycznego [11] posłużą do opracowania modelu propagacji jonosferycznej. Celem artykułu jest przedstawienie zestawu badawczego umożliwiającego pomiar parametrów charakterystycznych sygnału radiowego, które pozwolą na opis transmitancji kanału.

Parametry transmisyjne ośrodka propagacji jakim jest jonosfera cechuje ich zmienność o losowym charakterze, która występuje zarówno w krótkich jak i długich przedziałach czasu. Fakt ten wynika z losowych zmian profilu gęstości elektronów w jonosferze. Zatem podstawą dla modelowania rzeczywistych warunków propagacji fal radiowych w jonosferze jest odwzorowanie statystycznych zmian charakterystycznych parametrów na podstawie transmisji poprzez łącze jonosferyczne. Zmiany parametrów sygnału na wyjściu kanału jonosferycznego, pozwolą na określenie parametrów statystycznych niezbędnych do modelowania kanału.

2. ZJAWISKA I PARAMETRY CHARAKTERYSTYCZNE DLA ŁACZA JONOSFERYCZNEGO

Fale radiowe w radiokomunikacji lądowej rozchodzą się w atmosferze, obszarze otaczającym Ziemię. Właściwości atmosfery zależą od wysokości jej warstw. Obszar atmosfery powyżej 60 km nad poziomem morza nosi nazwę jonosfery. Składa się on ze zjonizowanych gazów, powstałych na skutek promieniowania słonecznego (oraz w pewnym stopniu promieni kosmicznych, oddziaływania strumieni meteorów bombardujących atmosferę, itp.). Zjonizowany gaz wpływa w określony sposób na właściwości elektryczne i tym samym oddziałuje na propagację fali elektromagnetycznej w jonosferze. Jonosferę można podzielić na pewne obszary zwane warstwami przechodzącymi płynnie jedna w drugą. Wyróżnia się je wówczas jeśli ze wzrostem wysokości gęstość jonizacji najpierw rośnie, a później maleje.

Wyróżnione główne warstwy oznaczono literami D, E i F. Występuje niekiedy dodatkowa warstwa E_s, natomiast warstwa F w ciągu dnia rozdwaja się na dwie warstwy F1 i F2. Warstwy jonizacyjne mają zdolność załamywania (refrakcji) i odbijania fal radiowych. Właściwości te są zależne od warunków fizycznych panujących w danej warstwie oraz od długości fali.

Fizyczną przyczyną załamywania fal w jonosferze jest zmiana gęstości jonizacji z wysokością, określana gradientem, czyli przyrostem gęstości jonizacji lub gęstości elektronowej przypadającej na jednostkę długości. W jonosferze fale radiowe są również tłumione. Tłumienie to jest zależne od gęstości i długości fali, nosi nazwę absorpcji fali, która określana jest współczynnikiem tłumienia (1).

$$A = j \cdot q \cdot g \cdot S_o \cdot \mathrm{Sec}(\varphi)\,[\mathrm{V/V}] \qquad (1)$$

gdzie :
a - współczynnik tłumienia;
j - współczynnik zmian sezonowych (wartość 1 do 1.3);
q - współczynnik zmian absorpcji w funkcji zmian aktywności słonecznej;
g - średnia wartość współczynnika dobowych zmian absorpcji na danej trasie;
S_o - funkcja częstotliwości i odległości;
φ - kąt wejścia w warstwę jonosferyczną fali.

Warstwy jonosferyczne nie są w rzeczywistości jednorodne, wykazują niejednorodności, które nazywa się obłokami. Obłoki te są nieustannie w stadium powstawania i zanikania. W jonosferze mają miejsca ruchy związane z przypływami i odpływami wywołanymi oddziaływaniem Słońca i Księżyca, przy czym w grę wchodzą nie tylko siły grawitacyjne, ale również wpływy termiczne. Warstwy jonosferyczne odbijają tylko pewien zakres długości fal. W jonosferze nie ma stanu ustalonego, a jej właściwości lokalne mogą być opisane tylko metodami statystycznymi.

W kanale jonosferycznym występują takie zjawiska jak zaniki, dyspersja sygnału w dziedzinie czasu i częstotliwości, efekt Dopplera. Zaniki związane są z wielodrogowoscią spowodowaną występowaniem pojedynczych i wielokrotnych odbić od warstw jonosfery, jak i możliwości odbić pośrednich od powierzchni Ziemi. Dyspersja w dziedzinie czasu i częstotliwości oraz efekt Dopplera związane są z ciągłą nieregularną zmianą jonizacji a tym samym zmianą długości każdej ścieżki propagacyjnej w funkcji czasu. Zmiany te mogą być szybkie i powolne, krótkotrwałe i długoterminowe, wąskopasmowe oraz szerokopasmowe. Wartości zmian dla każdej ze ścieżek nie muszą być ze sobą skorelowane a wręcz są nieskorelowane.

W literaturze [3-6, 10, 12-15] można znaleźć przybliżone wartości charakterystyczne parametrów dla poszczególnych ścieżek propagacyjnych. Wyniki analizy podanej wyżej literatury umieszczono w tabeli 1.

Tab. 1. Parametry kanału zgodne z CCIR dla kanałów o różnej jakości (88 km)

Parametr	Jednostka	dobry	średni	zły
Opóźnienie różnicowe	[ms]	0,5	1	2
Rozproszenie w dziedzinie czasu	[ms]	0,03		2
Rozproszenie Dopplera	[Hz]	0,1	0,5	1

Szczególnym rodzajem zaników jest zjawisko zaniku powszechnego, inaczej zwane zjawiskiem Mogel-Dellingera. Polega na nagłym wzroście jonizacji warstwy D, co powoduje nagły wzrost absorpcji i zanik odbioru fal krótkich, na całej półkuli ziemskiej oświetlonej przez Słońce. Zanik powszechny jest normalnie krótkotrwały, czas jego trwania zawiera się od kilku minut do 2...3 godzin. Przyczyną zaników są wybuchy na Słońcu. Można je uznać za deterministyczne w pewnym przedziale czasu.

Zaniki są główną cechą propagacji jonosferycznej, są one ciągłe o charakterze przypadkowej zmiany poziomu odbieranego sygnału. Głębokość ich może osiągać wartość dziesiątków decybeli, a częstotliwość zaników waha się od części herca do dziesiątków herców. Zaniki mogą mieć charakter synchroniczny lub selektywny. Przy zaniku synchronicznym natężenie pola wszystkich składowych fali zmodulowanej ulega jednakowym wahaniom (jest to inaczej zwany zanik szerokopasmowy). Przy zanikach selektywnych przeciwnie, zmiany natężenia pola poszczególnych składowych widma są różne. Zaniki synchroniczne związane są z tłumieniem warstw jonosferycznych. Ich zmiana zależna jest od fluktuacji tłumienia poszczególnych warstw (szczególnie warstwy F2, która jest bardzo niestabilna).

Wyniki pomiarów powinny pozwolić na ocenę statystyczną wartości mocy odbieranego sygnału i zaników, dyspersji czasu i dyspersji częstotliwości, przesunięcia Dopplera i czasu propagacji, opóźnień między ścieżkowych. Przeprowadzona analiza na zebranych danych będzie związana również ze zmianami tych parametrów w funkcji czasu. Istotne będą podczas badań struktury czasowo – częstotliwościowe sygnału transmitowanego.

3. STANOWISKO NADAWCZE

Stanowisko nadawcze składa się z:

- Transceiver firmy ICOM typu IC-7610 (maksymalna moc znamionowa 100 W) (rys. 1);
- Wzorca wysokostabilnej częstotliwości stabilizowanego sygnałem GPS typu RigExpert GPSDO (rys. 3) wraz z aktywną anteną GPS (opcjonalnie z generatora rubidowego o częstotliwości wzorcowej 10 MHz i stabilności nie gorszej niż 1 ppb, przy wykorzystaniu jednoczesnym dwóch stanowisk odbiorczych pomiarowych);
- Zasilacza GZV-4000 (rys. 2);
- Anteny WD-330.

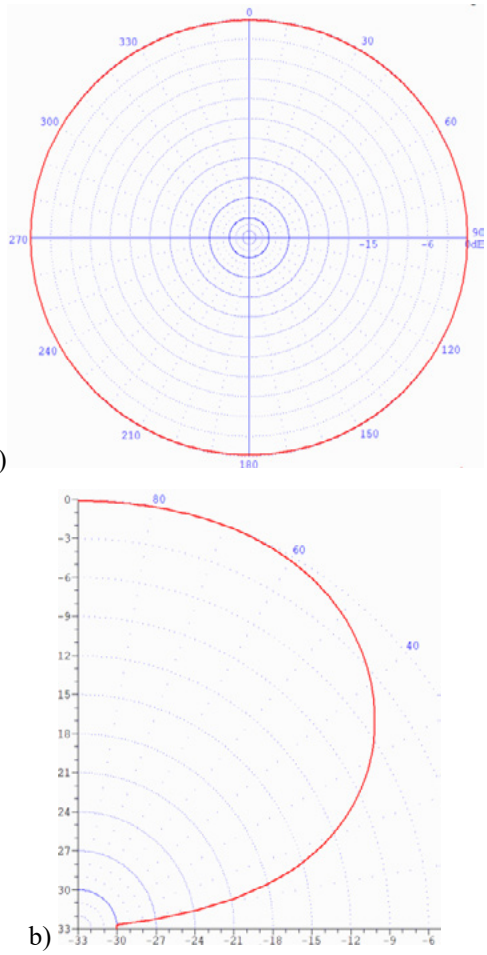
Rys. 1. Widok Transceiwer'a IC-7610 pełniącego rolę nadajnika sygnału pomiarowego [2]

Rys. 2. Widok zasilacza Transceiwer'a IC-7610

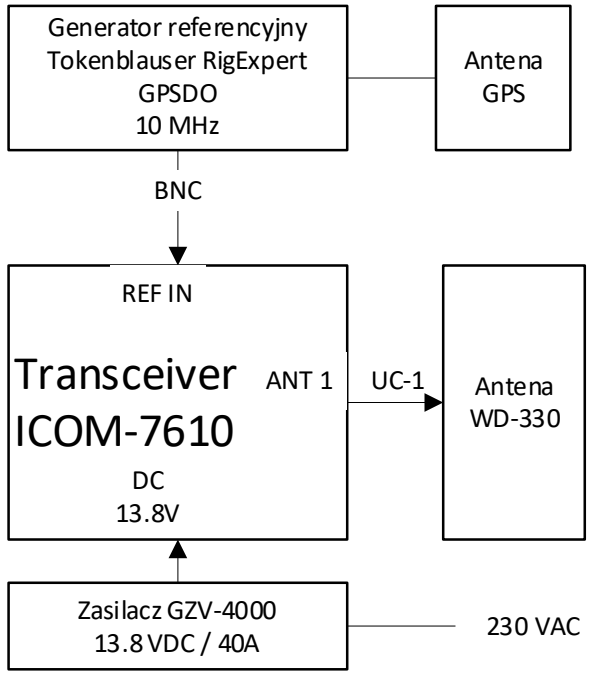
Do zasilania nadajnika wykorzystywany jest zasilacz GZV-4000 o napięciu ustawionym na wartość 13,8 VDC i maksymalnym natężeniu prądu zasilającego 40 A. Generator referencyjny GPSDO Tokenblauser pozwala na uzyskanie stabilności częstotliwości wzorcowej 10 MHz lepszej niż 1ppb. Wzorzec częstotliwości ma dodatkowe cztery wyjścia zapewniające uzyskanie innych wzorcowych częstotliwości o stabilności jak dla częstotliwości 10 MHz, co będzie przydatne na stanowisku odbiorczym.

Rys. 3. Widok generatora referencyjnego Rigexpert GPSDO Tokenblauser [7]

Antena nadawcza, Dipol szerokopasmowy WD-330 został rozwieszony nad dachem budynku na wysokości ok. 4 m (od 1/15 do 1/5 długości fali planowanych podczas późniejszych badań). Wysokość zawieszenia anteny pozwala na zmianę charakterystyki anteny do pracy w układzie typu NVIS (wysokość zgodnie z [8-9] powinna wynosić od 1/4 do 1/20 długości fali).



Rys. 4. Przykładowa charakterystyka anteny typu dipol rozwieszonej na wysokości 4 dla częstotliwości 6 MHz: a – pozioma, b – pionowa [16]



Rys. 5. Schemat połączeń funkcjonalnych elementów stanowiska nadawczego

Antena WD-330 połączona jest z nadajnikiem IC-7610 kablem koncentrycznym o długości 30 m obustronnie ze złączami UC-1. Antena aktywna GPS wzorca częstotliwości Rigexpert GPSDO Tokenblauser umieszczona została na dachu budynku.

4. STANOWISKO ODBIORCZE

Stanowisko odbiorcze składa się z:

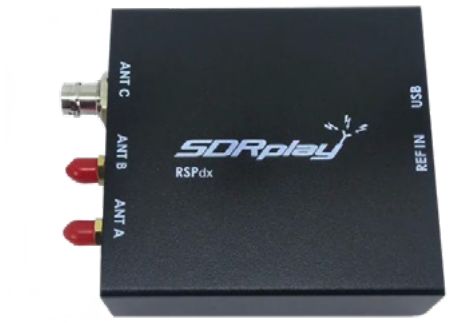
- Odbiornika SDR RSPdx (rys. 6),
- Wzorca wysokostabilnej częstotliwości stabilizowanego sygnałem GPS typu RigExpert GPSDO (rys. 3),
- Komputera sterującego odbiornikiem z zainstalowaną aplikacją SDR Play i Time Lab,
- Anteny BB6W.

Na stanowisku wykorzystany jest odbiornik szerokopasmowy SDR - RSPdx brytyjskiego producenta SDRplay. Zakres odbieranych częstotliwości przez odbiornik mieści się w granicach od 0,001MHz do 2000MHz. Odbiornik posiada maksymalną szerokość analizy widmowej w czasie rzeczywistym równą 10MHz. Interfejs użytkownika dla odbiornika oraz sterowanie zapewnia aplikacja SDR Uno.

Aplikacja SDR Uno pozwala na rejestrację sygnału IF odbiornika o szerokości pasma do 2 MHz. Przesyłanie sygnału do komputera sterującego odbywa się poprzez interfejs USB 2.0. Sygnał jest rejestrowany na dysku PC w postaci pliku w formacie Wave stereo.

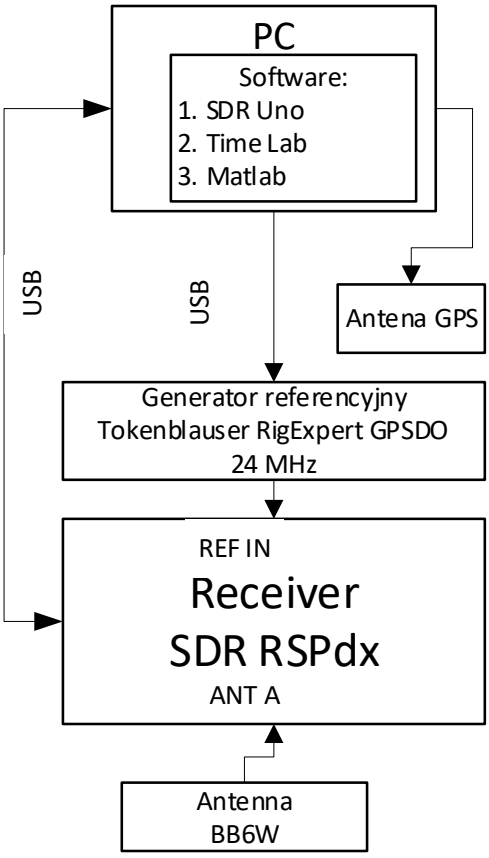
Dokładność wewnętrznego wzorca częstotliwości odbiornika SDR - RSPdx jest lepsza niż 0,5 ppm. Odbiornik umożliwia wykorzystanie zewnętrznego generatora częstotliwości wzorcowej o częstotliwości 24 MHz.

Z zewnętrznym wzorcem stabilność częstotliwości osiągana przez odbiornik jest lepsza niż 1 ppb.



Rys. 6. Widok odbiornika SDR RSPdx od góry wraz z opisami interfejsów [1]

Na stanowisku odbiorczym jako generator referencyjny wykorzystywany jest drugi wzorzec częstotliwości GPSDO Tokenblauser synchronizowany sygnałem GNSS. Urządzenie posiada 5 wyjść sygnałów wzorcowej częstotliwości, na czterech z nich można ustawić wymaganą przez odbiornik częstotliwość równą 24 MHz. Ustawiono ją na wyjściu Out A. Moduł Rigexpert GPSDO Tokenblauser pozwala na zwiększenie stabilności częstotliwości odbiornika SDR – RSPdx do wartości lepszych niż 1ppb (0,001ppm).



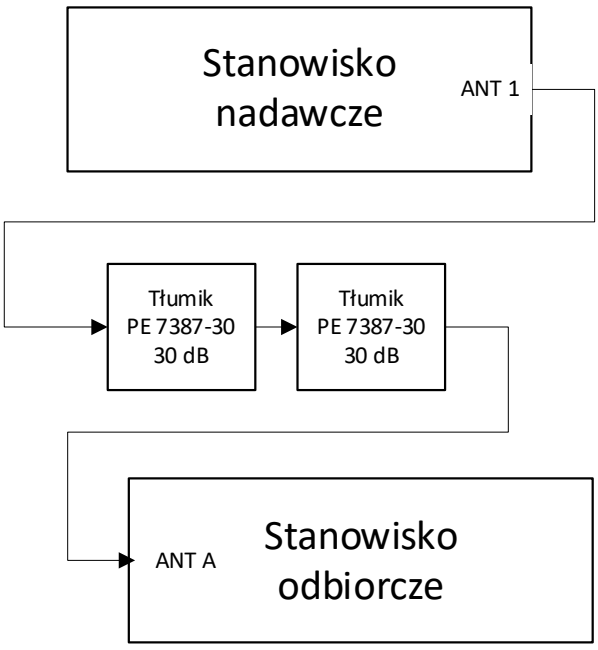
Rys. 7. Schemat połączeń funkcjonalnych elementów stanowiska odbiorczego

Szerokopasmową antenę stanowiska odbiorczego Diamond BB6W należy rozwiesić na wysokości od 1 do 4 m równolegle nad Ziemią, co odpowiada wartością od 1/4 do 1/20 długości fali. Przy takiej konfiguracji antena pracuje w trybie NVIS [8-9]. Antena BB6W połączona jest z odbiornikiem SDR RSPdx kablem koncentrycznym o długości 30 m obustronnie ze złączami UC-1. Po stronie odbiornika zastosowano przejściówkę zapewniającą połączenie z wejściem ANT A odbiornika. Antena aktywna GPS wzorca częstotliwości Rigexpert GPSDO Tokenblauser umieszczona powinna być tak, aby zapewniała odbiór z co najmniej 4 satelitów. Wartość SNR na wyświetlaczu GSPDO powinna być większa od 30 dB.

5. WERYFIKACJA STANOWISKA BADAWCZEGO

Stanowisko badawcze zostało skompletowane i uruchomione w celu weryfikacji poprawności działania oraz sprawdzenie błędów wprowadzanych przez zestaw pomiarowy, które zostaną później uwzględnione przy opracowywaniu wyników z pomiarów badawczych łącza jonosferycznego.

Na stanowiskach nadawczym i odbiorczym przed rozpoczęciem badań należy wykonać weryfikację parametrów sygnału z wzorca częstotliwości referencyjnej dla Transceiver'a oraz dla odbiornika SDR-RSPdx. Testy pozwolą na ocenę błędu częstotliwości wprowadzaną przez układy pomiarowe. Aplikacja Time Lab służy do takiej weryfikacji stabilności źródła częstotliwości jakim jest GPSDO. Może być to ważne podczas podejrzenia możliwości zakłóceń sygnału GPS.



Rys. 8. Połączenie stanowisk nadawczego z odbiorczym dla weryfikacji możliwości użytkowych

Na rysunku 9 przedstawiono wykonane za pomocą aplikacji Time Lab pomiary wartości i dryftu częstotliwości wzorca częstotliwości w funkcji czasu względem wartości początkowej pomiaru. Dla badań najbardziej istotna jest wartość z największym błędem, która daje widzę o maksymalnym błędzie wprowadzanym przez posiadany wzorzec.

Origin	Drift (per sec)	Drift (per hr)
−6.38E−11	+2.44E−14	+8.77E−11
−4.52E−12	+2.13E−15	+7.68E−12
+2.66E−8	−2.52E−11	−9.06E−8

Avg Time (s)	Freq (Hz) at 3599s	Error
1	10 000 000 . 001 440 000	+1.44E−10
3	10 000 000 . 001 456 667	+1.46E−10
10	10 000 000 . 000 996 999	+9.97E−11
30	10 000 000 . 000 830 000	+8.30E−11
100	9 999 999 . 998 961 801	−1.04E−10
300	9 999 999 . 999 937 633	−6.24E−12
1 000	9 999 999 . 999 905 361	−9.46E−12
3 000	9 999 999 . 999 908 080	−9.19E−12

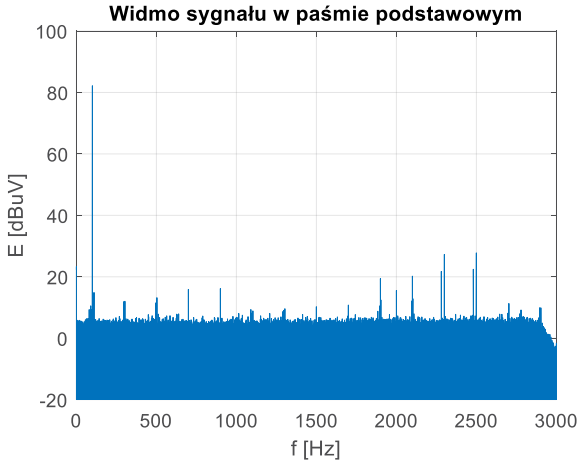
Rys. 9. Zmiany wartości i dryft częstotliwości wzorcowej 10 MHz na wyjściu GSPDO (kolor niebieski oznacza jedną sesję pomiarową, różowy i zielony z wcześniejszych sesji pomiarowych))

W czasie badań kanału jonosferycznego ze względu na znaczną odległość między stanowiskiem odbiorczym a stanowiskiem nadawczym wykorzystywane będą dwa niezależne wzorce GPSDO stabilizowane sygnałem z GPS. Należy wówczas uwzględniać maksymalny błąd dla obu urządzeń.

Podczas pomiarów mocy sygnału odbieranego i jego zmian należy w stanowisku odbiorczym wyłączyć wszystkie układy do automatycznej regulacji wzmocnienia. W niniejszym artykule nie przedstawiono wyników testów weryfikujących pomiar zmian mocy na stanowisku odbiorczym.

W czasie testów przeprowadzanych na stanowisku testowym przedstawionym na rys. 8 sygnał odbierany przez SDR RSPdx zapisywany jest w postaci próbek IQ w pliku Wave. Obróbka danych IQ realizowana jest następnie w środowisku Matlab.

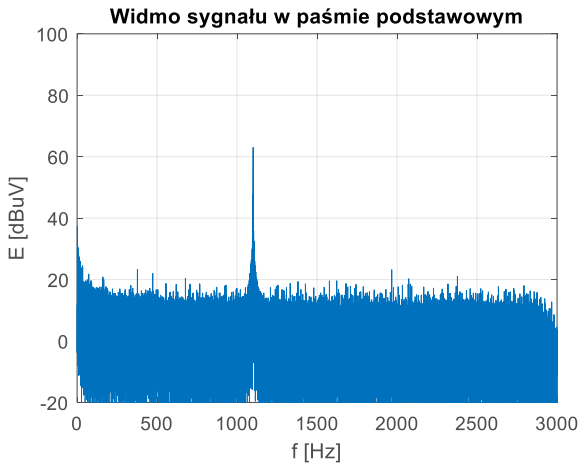
Na rysunku 10 przedstawiono wynik analizy poprawności działania zestawu badawczego zestawionego zgodnie z rysunkiem 8.



Rys. 10. Przykładowy wynik analizy graficznej przedstawiający widmo sygnału w paśmie podstawowym (Wyniki dla sygnału IQ zapisanego w pliku: SDRuno_20241203_174034Z_23090kHz.wav)

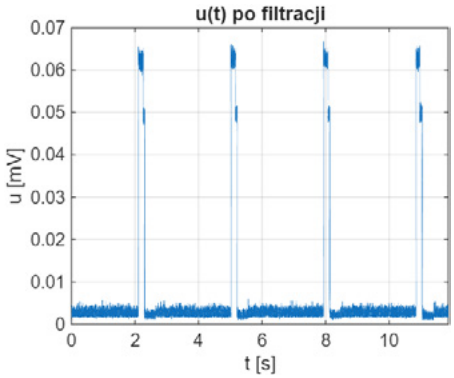
Częstotliwość w odbiorniku była zmieniona względem nadajnika o 100 Hz. Błąd częstotliwości dla przykładowego pomiaru podczas testów zestawu badawczego wyniósł 2,37e-10 co w wartościach bezwzględnych daje 5,5e-3 Hz dla sygnału o częstotliwości nośnej 23090 kHz. Wartość SWR na mierniku w nadajniku wynosił około 1,1:1. Na rys. 10 można zauważyć ponad szumem sygnały harmoniczne o poziomie poniżej 50 dB.

W ramach weryfikacji stanowiska przeprowadzono testy z wykorzystaniem anten na dystansie ok. 1 km w terenie silnie zurbanizowanym. Wyniki graficzne analizy wyników pomiarów przedstawiono na rysunkach 11 i 12.



Rys. 11. Przykładowy wynik analizy graficznej przedstawiający widmo sygnału w paśmie podstawowym (dla częstotliwości fali nośnej 5799 kHz z wykorzystaniem kompletnego stanowiska badawczego; plik SDRuno_20250506_103837Z_5788kHz; bez sygnału wzorcowego z GPSDO)

Częstotliwość w odbiorniku była zmieniona względem nadajnika o 1100 Hz. Błąd częstotliwości dla przykładowego pomiaru podczas testów zestawu badawczego wyniósł 6,77e-7 co w wartościach bezwzględnych daje 4,71 Hz dla sygnału o częstotliwości nośnej 5788 kHz. Wartość SWR na mierniku w nadajniku wynosił około 1,05:1.



Rys. 12. Przykładowy wynik analizy graficznej dla testowego sygnału ASK po demodulacji (dla częstotliwości fali nośnej 5799 kHz z wykorzystaniem kompletnego stanowiska badawczego; plik SDRuno_20250506_103837Z_5788kHz)

Podczas badań na odległościach, do ok. 400 - 600 km, należy pamiętać o dostosowaniu posiadanych dipoli do pracy jako anteny NVIS. W ramach przygotowania do badań przeprowadzono dodatkowo testy na odległości 14 km weryfikujące poprawność zawieszenia anten do pracy NVIS. Wyniki tych testów nie zostały zamieszczone w artykule.

6. PODSUMOWANIE

W niniejszym artykule skupiono się na przedstawieniu weryfikacji modelu pod względem pomiarów częstotliwości. Uzyskiwana dokładność pomiaru częstotliwości powinna być lepsza niż 1*10-9, świadczą o tym wynik pomiarów podczas testowania stanowiska pomiarowego. Przy fali nośnej 6 MHz daje to dokładność lepszą niż 0,006 Hz. Pozwala to na ocenę częstotliwości przesunięcia Dopplera oraz dyspersji częstotliwości.

Opisany system badawczy umożliwia pomiar parametrów odbieranego sygnału, takich jak: amplituda chwilowa, częstotliwość chwilowa, dyspersja w funkcji czasu i częstotliwości. Do przeprowadzenia badań dla szerszego zakresu parametrów potrzebne są jeszcze odpowiednio spreparowane sygnały testowe o odpowiedniej strukturze.

PODZIĘKOWANIA

Praca ta została sfinansowana przez Wojskową Akademię Techniczną z funduszów projektu Europejskiej Agencji Obrony EDA Contract No B.PRJ.RT.832 pt. „LOng RAnge COmmunications”.

LITERATURA

[1] SDRplay Wielka Brytania, 2024, User Manual v1.42 built 1710, SDR UNO

[2] Icom Inc. Japonia, 2021, Basic manual IC-7610

[3] ITU, 1997, HF field-strength measurement, Rec. ITU-R P.845-3

[4] ITU, 1994, Propagation factors affecting frequency sharing in HF terrestrial systems, Rec. ITU-R P.1060

[5] ITU, 2023, Guide to the application of the propagation methods of Radiocommunication Study Group 3, Rec. ITU-R P.1144-12

[6] ITU, 2022, Ground-wave propagation prediction method for frequencies between 10 kHz and 30 MHz, Rec. ITU-R P.368-10

[7] Rig Expert Ukraina Ltd. , Ukraina, 2025, User’s manual, Tokenblauster GPSDO

[8] SP5MXF, Antena NVIS, http://sp5mxf.com/antena-nvis/

[9] Thompson L. (W5IFQ), NVIS Antenna Theory & Practice, 07.09.2024, (http://eny.arrl.org/ARES/FILES/SET/NVIS-Antenna-Theory-and-Practice-AARC.pdf)

[10] Uljasz B., 1992, Symulator kanału krótkofalowego, Praca magisterska, WAT

[11] Walczyna T., Wojtuń J., Ziółkowski C., Uljasz B., 2024, Analityczny model profilu gęstości elektronów jako wynik aproksymacji danych pomiarowych z jonogramów, KRiT 2024, Poznań

[12] Ruchirapha T., 2003, An Evaluation of HF Modem Performance, The University of Adelaide

[13] Vogler L.E., Hoffmeyer J.A., 1988, A New Approach to HF Channel Modeling and Simulation, Part III: Deterministic Model, U.S. Department of Commerce

[14] Vogler L.E., Hoffmeyer J.A., 1992, A New Approach to HF Channel Modeling and Simulation, Part III: Transfer Function, U.S. Department of Commerce

[15] Vogler L.E., Hoffmeyer J.A., 1990, A New Approach to HF Channel Modeling and Simulation, IEEE

[16] Hand G., 2005, “HFANT” – przykładowa antena typ 23 ITSA-1 Horizontal Dipole

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

ANALIZA MOŻLIWOŚCI WDROŻENIA USŁUG SATELITARNYCH W SIECI KOMÓRKOWEJ 5G I 6G

ANALYSIS OF THE POSSIBILITIES OF IMPLEMENTING SATELLITE SERVICES IN 5G AND 6G MOBILE NETWORK

Ryszard J. Zieliński

[1] Politechnika Wrocławska, Ryszard.zielinski@pwr.edu.pl

DOI: 10.15199/59.2025.4.70

Streszczenie: **W artykule przedstawiono analizy związane z łączem satelitarnym zapewniającym podobne warunki odbioru sygnału nadawanego przez satelitę do tych nadawanych przez stację bazową gNB sieci komórkowej 5G i 6G. Analizy są ściśle związane z wprowadzeniem do opisu standardu 5G (w wydaniu 17) możliwości stosowania m.in. łączy satelitarnych. Przedstawiono zakresy częstotliwości przeznaczone do wykorzystania w tego typu zastosowaniach, uproszczony bilans energetyczny łącza do terminali naziemnych oraz warunki jakie należy spełnić, by zapewnić poprawną realizację usług.**

Abstract: **The article presents analyses related to a satellite link providing similar conditions for receiving a signal transmitted by a satellite to those transmitted by the gNB base station of the 5G and 6G cellular network. The analyses are closely related to the introduction of the possibility of using, among others, satellite links to the description of the 5G standard (in the 17th release). The frequency ranges intended for use in such applications, a simplified energy analysis of the link to terrestrial terminals and the conditions that must be met to ensure correct implementation of services are presented.**

Słowa kluczowe: **łącza satelitarne, nienaziemne sieci, sieci 5G i 6G.**

Keywords: **Satellite Links, Non Terrestrial Networks (NTN), 5G and 6G Networks.**

1. WSTĘP

Jeszcze do niedawna realizacja masowego świadczenia szerokopasmowego dostępu do internetu i do szeregu usług wymagających dużych szybkości transmisji ze względów technicznych możliwa była tylko w sieci naziemnej. Monopol ten wraz z rozwojem technologii teleinformatycznych został złamany poprzez rozwój systemów satelitarnych stosujących satelity na niskich orbitach. W roku 2015 firma OneWeb podejmuje się budowy systemu satelitarnej łączności szerokopasmowej i budowy konstelacji biegunowej składającej się z 648 satelitów rozmieszczonych na 18 orbitach na wysokości 1200 km. W systemie zaplanowano uzyskanie szybkości 50 Mb/s przy dostępie do internetu i opóźnienia na poziomie 30 ms. Obecnie Eutelsat OneWeb deklaruje szybkości 32 Mb/s do satelity i 155 Mb/s w kierunku do terminali z opóźnieniem 5 do 10 ms. Nastawiona jest głównie na obsługę instytucjonalną przy użyciu małych anten. Na znacznie większą skalę działa Starlink oferując użytkownikom usługi szerokopasmowe o szybkości na poziomie 100 Mb/s przy użyciu bardzo małych terminali wyposażonych w anteny z elektronicznie sterowaną wiązką o dużym zysku. W obu wymienionych przypadkach do świadczenia usług szybkiej transmisji danych trzeba się wyposażyć w specjalizowany terminal. Kolejnym wielkim krokiem w rozwoju będzie świadczenie usług transmisji danych przy użyciu standardowych terminali sieci 5G a w przyszłości 6G. W tym jednak przypadku podstawowym ograniczeniem są parametry nadawczo odbiorcze terminala. Głównie ograniczenia dotyczą mocy nadajnika oraz zysku anteny, w którą typowe terminale są wyposażone. Z tego względu zapewnienie odpowiedniego bilansu energetycznego łącza leży po stronie satelity. Tego typu satelity są obecnie testowane na orbicie. Celem artykułu jest określenie parametrów łącza satelitarnego, które zapewni porównywalne warunki działania terminali w stosunku do parametrów łącza naziemnego ze stacją bazową gNB.

2. CZĘSTOTLIWOŚCI PRZYPISANE SIECIOM NTN

W dokumentach 3GPP przewidziano kilka pasm do komunikacji terminali poprzez satelity. Ze względu na straty propagacyjne pasmem najlepiej nadającym się do łączności pomiędzy satelitą i terminalem użytkownika UE (User Equipment) jest zakres częstotliwości w pobliżu 2 GHz. Wyróżniono w nim dwa pasma n255 i n256. Parametry tych pasm zestawiono w tabeli 1.

Tab. 1. Parametry częstotliwościowe łącza satelita – UE w paśmie częstotliwości poniżej 6 GHz (FR 1)

Nr pasma	Zakres częstotliwości [MHz]		Typ dupleksu
	łącze do satelity	łącze do UE	
n255	1626,5 – 1660,5	1525 - 1559	FDD
n256	1980 - 2010	1170 - 2200	

W paśmie S (n255) zarezerwowano 34 MHz do łączności w obu kierunkach, natomiast w paśmie L (n256) zarezerwowano 30 MHz. W pasmach tych maksymalna szerokość częstotliwościowa kanału transmisyjnego jest obecnie ograniczona do 20 MHz. W tych dwóch pasmach można stosować odległości podnośnych SCS (Subcarrier Spacing) równe 15 kHz, 30 kHz i 60 kHz dla kanałów

transmisyjnych o szerokości 10 MHz, 15 MHz i 20 MHz. Dla kanału o szerokości 5 MHz można jedynie stosować odległości SCS równe 15 kHz. Łącznie do dyspozycji systemu NTN jest 64 MHz, co wydaje się zbyt małym pasmem do świadczenia usług szerokopasmowych eMBB (Enhanced Mobile Broadband). Natomiast może zostać użyte do przesyłania krótkich wiadomości, realizując jedną z usług mMTC (Massive Machine-Type Communications) wspierającą tzw. internet rzeczy IoT. Ze względu na występujące opóźnienia w łączu satelitarnym tego typu rozwiązania nie można użyć do świadczenia usług wymagających bardzo małych opóźnień w grupie usług typu URLLC (Ultra Reliable and Low Latency Communications).

Do łączności pomiędzy terminalami naziemnymi a satelitą rozważane są w ramach pasm FR 2 powyżej 6 GHz dwa podpasma. Niższe przedstawiono w tabeli 2 a wyższe w tabeli 3.

Tab. 2. Parametry częstotliwościowe łącza satelita – UE w paśmie częstotliwości powyżej 6 GHz (FR 2.0)

Nr pasma	Zakres częstotliwości [GHz]		Typ dupleksu
	łącze do satelity	łącze do UE	
N510	17,70 – 20,20	27,50 – 28,35	FDD
N511	17,70 – 20,20	28,35 – 30,00	
N512	17,70 – 20,20	27,50 – 30,00	

Pasma n510 i n511 dedykowane są USA natomiast n512 Europie.

Tab. 3. Parametry częstotliwościowe łącza satelita – UE w paśmie częstotliwości powyżej 6 GHz (FR 2.1)

Nr pasma	Zakres częstotliwości [GHz]		Typ dupleksu
	łącze do satelity	łącze do UE	
n257	26,50 – 29,50	26,50 – 29,50	TDD
n258	24,25 – 27,50	24,25 – 27,50	
n259	39,50 – 43,50	39,50 – 43,50	
n260	37,00 – 40,00	37,00 – 40,00	
n261	27,50 – 28,35	27,50 – 28,35	
n262	47,20 – 48,20	47,20 – 48,20	

Biorąc pod uwagę duży wpływ na bilans energetyczny łącza satelitarnego częstotliwości i czynników atmosferycznych takich jak opady istotne dla pasma wyższego, pozostałe rozważania, ze względu na ograniczoną liczbę stron referatu, zostaną ograniczone do łączy satelitarnych stosujących częstotliwości z niższego pasma.

3. BILANS ENERGETYCZNY ŁĄCZA SATELITARNEGO

Jednym z krytycznych warunków świadczenia usług satelitarnych jest bilans energetyczny łącza satelita – terminal naziemny. Wynika on z konieczności zapewnienia odpowiedniego stosunku mocy sygnału użytecznego do mocy szumu i zakłóceń. CNIR (Carrier to Noise and Interference Ratio) na wejściu odbiornika. W analizie łączy satelitarnych częściej posługuje się stosunkiem mocy sygnału użytecznego C do widmowej gęstości mocy szumów na wejściu odbiornika N_0. Upraszczając analizę jedynie do uwzględnienia poziomu szumu własnego odbiornika wzór na stosunek mocy sygnału użytecznego do widmowej gęstości mocy szumu CN_0R (Carrier to Noise density Ratio) można wyznaczyć ze wzoru (1).

$$\frac{C}{N_0} = \frac{EIRP_{nad}}{\text{straty propagacji}} \cdot \frac{\text{zysk przy odbiorze}}{\text{temperatura szumowa}} \cdot \frac{1}{k} \quad (1)$$

w którym; $EIRP_{nad}$ to zastępcza moc promieniowana izotropowo przez nadajnik a k to stała Boltzmanna $1{,}38 \cdot 10^{-23}$ J/K (-228,6 dB). Przy analizie łącza do terminala doręcznego zarówno zysk anteny odbiorczej jak również temperatura szumowa systemu odbiorczego są ustalone. Zgodnie z [1] można przyjąć, że terminale UE (User Equipment) mają współczynnik szumów F=9 dB. Na tej podstawie możemy wyznaczyć temperaturę szumową odbiornika na wejściu $T_R=(F-1)\cdot T_0$ (T_0=290 K) równą 2014 K. Uwzględniając temperaturę otoczenia T_0=290 K oraz zakładając, że straty pomiędzy wyjściem anteny odbiorczej i wejściem odbiornika są pomijalne otrzymano zastępczą temperaturę szumową systemu odbiorczego UE równą 2304 K. Można również przyjąć, że anteny instalowane w terminalach mobilnych ze względu na ograniczenie miejsca i ich prostą budowę wykazują zysk G= 0 dBi. Zatem, w zgodzie z zależnością (1) można przyjąć że za odpowiedni stosunek mocy sygnału użytecznego do widmowej gęstości mocy szumów odpowiadać będą dwa parametry: $EIRP_{nad}$ oraz starty propagacyjne.

4. STRATY PROPAGACYJNE ŁĄCZA SATELITA - TERMINAL NAZIEMNY

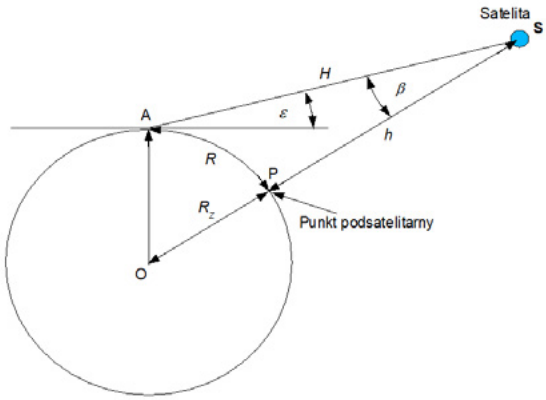
Wielkość strat propagacyjnych w łączu satelita terminal naziemny UE zależy od długości tego łącza oraz modelu propagacyjnego przyjętego do opisu środowiska propagacyjnego. Jednym z prostszych modeli jest model wolnej przestrzeni. W zależności od warunków odbioru sygnału przez UE model ten uzupełnia się o elementy związane ze stratami powodowanymi przez gazy atmosferyczne, opady, scyntylacje, efekty związane z cieniowaniem sygnału i niedopasowaniem polaryzacyjnym. Efekty te wprowadzają dodatkowe straty do bilansu. Zgodnie z [1] ich wielkości przedstawiono w tabeli 4.

Tab. 4. Dodatkowe straty uwzględniane w bilansie łącza satelita – UE [1]

Straty dodatkowe	L_{dod} [dB]
Scyntylacje	2,2
Gazy atmosferyczne	0,1
Straty polaryzacyjne	3
Cieniowanie	3
Sumarycznie	8,3

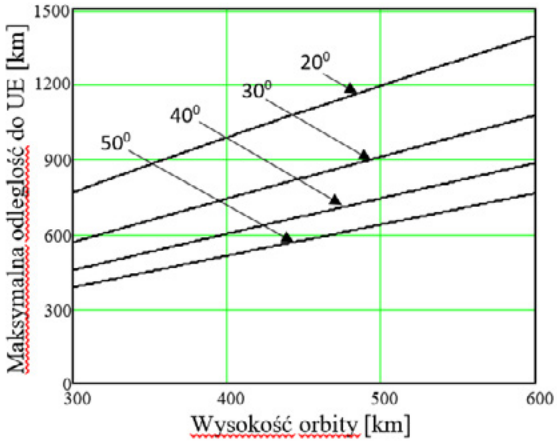
Biorąc pod uwagę pasma w okolicach 2 GHz można przyjąć niewielkie straty związane z występowaniem opadów – analiza dla tzw. czystego nieba (Clear Sky Conditions). Natomiast odległość H do wybranych punktów obsługiwanego przez satelitę obszaru ograniczoną minimalnym kątem elewacji ε osi wiązki anteny odbiorczej do satelity można wyznaczyć ze wzoru (2) (oznaczenia jak na rys. 1)

$$H(h,\varepsilon) := Rz \cdot \frac{\sin\left(\frac{\pi}{2} - \varepsilon - \operatorname{asin}\left(\frac{Rz}{Rz+h}\cos(\varepsilon)\right)\right)}{\left(\frac{Rz}{Rz+h}\cos(\varepsilon)\right)} \quad (2)$$



Rys. 1. Geometria do wyznaczenia długości łącza satelitarnego [2]

w którym R_Z = 6378 km (promień Ziemi) a h jest wysokością orbity kołowej w [km]. Minimalny kąt elewacji anteny pod którym satelita obsługuje UE wyznacza maksymalną odległość H pomiędzy satelitą i UE. Na rys. 2 przedstawiono te odległości w funkcji wysokości orbity h i kąta elewacji anteny ε równego: 20^0, 30^0, 40^0 ı 50^0. Kąt ten ogranicza również rozmiar obszaru obsługiwanego przez satelitę i czas, w którym połączenie może być aktywne. Przykładowo, dla satelitów konstelacji Starlink umieszczonych na orbicie o wysokości 550 km wynosi on 40^0.



Rys. 2. Maksymalna długość łącza satelita – UE w funkcji wysokości orbity dla wybranych kątów elewacji anteny ε

Odległości satelity H do granicy obsługiwanego obszaru dla zadanej wysokości orbity h i kąta elewacji anteny ε zestawiono w tabeli 5.

Tab 5. Odległości satelita – UE [km] dla granicy obszaru obsługi dla wybranych wysokości orbity i kątów elewacji wiązki anteny

Wysokość orbity [km] Kąt elewacji	300	400	500
20^0	764	984	1193
30^0	564	739	910
40^0	453	598	741
50^0	386	512	637

Na podstawie wcześniej wyznaczonych maksymalnych długości łączy możemy przy użyciu modelu propagacji w wolnej przestrzeni wyznaczyć straty L_{fs} (wzór 3),

$$L_{fs}(d, f) = 32{,}45 + 20 \cdot \log_{10}(d) + 20 \cdot \log_{10}(f) \qquad (3)$$

w którym d jest w [km] a f w [MHz].

Przyjmując maksymalną częstotliwość z pasma n256 równą 2200 MHz do łączności z UE możemy wyznaczyć wielkość strat (tabela 6).

Tab. 6. Maksymalne straty propagacyjne L_{fs} [dB] wolnej przestrzeni przy częstotliwości 2200 MHz

Wysokość orbity [km] Kąt elewacji	300	400	500
20^0	157,0	159,2	160,8
30^0	154,3	156,7	158,5
40^0	152,4	154,8	156,7
50^0	151,0	153,5	155,4

Przekształcając zależność (1) do postaci logarytmicznej i wyznaczając $EIRP_{nad}$ uzyskamy wzór (4) opisujący minimalny poziom zastępczej mocy promieniowanej izotropowo przez antenę nadawczą na pokładzie satelity zapewniający uzyskanie odpowiedniej wielkości stosunku (C/N_0) wymaganego do świadczenia wybranych usług.

$$EIRP_{nad} = C/N_0 + L - G_{odb} + T - 228{,}6 \qquad (4)$$

w którym $L = L_{fs} + L_{dod}$ i wszystkie wielkości wyrażone są w decybelach.

5. WARUNKI POPRAWNEGO ODBIORU SYGNAŁU PRZEZ TERMINAL MOBILNY

Odbiorniki stosowane w systemach 5G adaptują się pod względem świadczonych usług do zewnętrznych warunków panujących wokół anteny i sygnałów pojawiających się na ich wejściu. W zależności od poziomu sygnału użytecznego w stosunku do szumu własnego odbiornika zależnego od szerokości pasma i poziomów sygnałów zakłócających dobierana jest optymalna modulacja i sprawność kodowania. Nie wchodząc w szczegóły bardzo złożonych algorytmów optymalizujących parametry odbiornika można do oszacowania wymaganej zastępczej mocy promieniowanej przez anteną na pokładzie satelity przyjąć następujące założenia związane z warunkami odbioru, które zestawiono w tabeli 7, w której również wyznaczono stosunek C/N_0 dla kanału o szerokości 5 MHz, stosując zależność (5).

$$C/N_0 = C/N + 10 \cdot \log_{10} B \qquad (5)$$

Tab. 7. Jakość sygnału odbieranego przez UE w kanale o szerokości 5 MHz [3]

Jakość odbioru	C/N [dB]	C/N_0
znakomita	> 20	> 87
dobra	13 - 20	80 - 87
w środku komórki	0 - 13	67 - 80
na brzegu komórki	< 0	< 67

Stosując zależność (4) można wyznaczyć wymaganą $EIRP_{nad}$ dla wybranych warunków odbioru i strat propagacyjnych wynikających z długości łącza radiowego będącego funkcją minimalnego kąta elewacji anteny ε i wysokości orbity satelity h. Do obliczeń przyjęto dwie wartości C/N_0: 62 dB równoważne -5 dB dla C/N – minimalna wartość umożliwiająca dekodowanie sygnału i 80 dB – minimalna wartość zapewniająca dobre warunki odbioru.

Tab. 8. $EIRP_{nad}$ anteny na satelicie dla C/N_0 równego 62,0/80,0 dB[Hz]

Wysokość orbity [km] Kąt elewacji	300	400	500
20^0	32,3/50,3	34,5/52,5	36,1/54,1
30^0	29,6/47,6	32,0/50,0	33,8/51,8
40^0	27,7/45,7	30,1/48,1	32,0/50,0
50^0	26,3/44,3	28,8/46,8	30.7/48,7

Biorąc pod uwagę najbardziej krytyczne warunki związane ze stratami propagacyjnymi oraz możliwością obsługi obszaru przez antenę wielowiązkową do analizy wybrano satelity na wysokości 300 km z minimalnym kątem elewacji anteny naziemnej równym 50^0. Dla tego wariantu wymagane EIRP= 44,3 dBW. Zakładając, że moc nadajnika na pokładzie satelity w kanale o szerokości 5 MHz wynosi 50 W (17 dBW) możemy wyznaczyć minimalny zysk energetyczny anteny G=27,3 dBi.

6. STRUKTURA SYSTEMU NADAWCZEGO NA POKŁADZIE SATELITY

Warunki dobrego odbioru sygnału satelitarnego wymuszają zastosowanie na pokładzie satelity anteny o zysku 27,3 dBi. Dla anteny o takim zysku można wyznaczyć 3 dB szerokość wiązki. Do obliczenia tej szerokości wiązki Θ można zastosować uproszczoną zależność (6), zakładając optymalny rozkład prądu w aperturze anteny. Dla zysku równego 537 W/W otrzymujemy szerokość wiązki antenowej równą $7,3^0$. Dla tej szerokości wiązki można wyznaczyć promień obszaru pokrycia R.

$$\Theta(G) := \sqrt{\frac{29000}{G}} \quad [\text{deg}] \qquad (6)$$

W tym celu najpierw wyznaczony zostanie kąt elewacji anteny naziemnej ε ze skraju obszaru pokrycia (7).

$$\varepsilon(\beta, h) := \operatorname{acos}\left[\sin(\beta)\cdot\frac{(Rz+h)}{Rz}\right] [\text{deg}] \qquad (7)$$

Dla kąta nadiru $\beta=\Theta/2=3,65^0$ i wysokości orbity h=300 km kąt elewacji anteny ze skraju obszaru pokrycia wynosi $86,18^0$. Kąt środkowy R wyznaczający promień obszaru pokrycia wyrażony zależnością (8) jest równy $0,172^0$ i odpowiada promieniowi obszaru pokrycia równym 19,15 km.

$$R(\varepsilon, h) := \operatorname{acos}\left[\frac{Rz}{(Rz+h)}\cdot\cos\left(\varepsilon\cdot\frac{\pi}{180}\right)\right]\cdot\frac{180}{\pi} - \varepsilon \qquad (8)$$

Cały obszar wyznaczony przez kąt elewacji anteny równy 50^0, czyli dla kąta nadiru $\beta=47,87^0$ ma promień równy 236,8 km. W obszarze tym może zmieścić się ok. 153 komórki tworzone przez tak wąskie wiązki promieniowania anteny satelitowej. Korzystając z przekształconej zależności [2] na liczbę komórek tworzących tzw. pierścienie wokół komórki centralnej (9):

$$n_k=1+3n(n+1) \quad (n=0, 1, 2, \ldots) \qquad (9)$$

w którym n określa liczbę pierścieni wchodzących w skład obszaru pokrycia, dla 7 pierścieni (n=7) otrzymujemy 169 komórek (wiązek antenowych). Tworzenie tak dużej liczby wiązek antenowych jest trudne ale możliwe przy użyciu matryc antenowych sterowanych fazowo. Rozwiązanie takie wymaga również zastosowania bardzo szybkiego i wydajnego systemu przenoszenia połączenia z komórki do komórki ze względu na bardzo krótki czas przebywania terminala UE w szybko przemieszczającej się wiązce w przypadku przemieszczającej się struktury komórek na powierzchni Ziemi. Innym rozwiązaniem jest utrzymanie stacjonarnej struktury komórek przez przemieszczającego się satelitę. W tym przypadku znacznie komplikuje się budowa anteny, która musi odpowiednio sterować wiązkami. Należy zwrócić uwagę, że zasilanie każdej wiązki anteny mocą 50 W powoduje, że całkowita moc promieniowana przez antenę składającą się ze 153 komórek wynosi 7,65 kW. Moc całkowitą można zmniejszyć, stosując rozwiązanie znane z systemu Teledesic. W systemie tym jedna wiązka antenowa obsługiwała dziewięć komórek w sekwencyjny sposób. Wtedy całkowita moc promieniowana przez antenę wynosi 850 W, co jest wielkością realizowalną. Próby budowy tego typu systemów satelitarnych weszły w stadium eksperymentalne. Przykładem mogą być satelity BlueWalker firmy AST SpaceMobile wyposażone w anteny o powierzchni 64 m^2 (blok 1) i 128 m^2 (blok 2) [4].

LITERATURA

[1] 3GPP TR 36.763 V17.0.0 (2021-06), Study on Narrow-Band Internet of Things (NB-IoT) / enhanced Machine Type Communication (eMTC) support for Non-Terrestrial Networks (NTN), (Release 17).

[2] Zieliński Ryszard J. 2009. „Satelitarne systemy teleinformatyczne", WNT, Warszawa.

[3] Boiarynow S.,: "Jakość sygnału LTE/5G – parametry jakości sygnału LTE I 5G" https://support.zyxel.eu/hc/pl/articles/360005188999-Jako%C5%9B%C4%87-sygna%C5%82u-LTE-5G-parametry-jako%C5%9Bci-sygna%C5%82u-LTE-i-5G.

[4] https://ast-science.com/spacemobile-network/bluewalker-3/.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

ANALIZA TRANSMISJI RADIOWEJ KOMERCYJNYCH BSP

ANALYSIS OF THE RADIO TRANSMISSION OF COMMERCIAL UAVS

Jarosław Magiera

Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk
jaroslaw.magiera@pg.edu.pl

DOI: 10.15199/59.2025.4.71

Streszczenie: **Artykuł dotyczy analizy czasu trwania i szerokości pasma transmisji komercyjnych bezzałogowych statków powietrznych (BSP), w kontekście przydatności takiej informacji dla potrzeb identyfikacji producenta lub modelu statku. Została przedstawiona metodyka przetwarzania odbieranego sygnału, umożliwiająca estymację parametrów transmisji z pożądaną dokładnością. Następnie, zaprezentowano wyniki analizy rozkładu parametrów transmisji dla ośmiu modeli BSP, których sygnały zarejestrowano w komorze bezodbiciowej.**

Abstract: **The paper focuses on the analysis of the duration and transmission bandwidth of commercial unmanned aerial vehicles (UAVs), in the context of the usefulness of such information for the identification of the manufacturer or model of the vessel. A methodology for processing the received signal, enabling the estimation of transmission parameters with the desired accuracy, is presented. Next, the results of the analysis of the distribution of transmission parameters for eight models of BSPs whose signals were recorded in an anechoic chamber are presented.**

Słowa kluczowe: **Analiza czasowo-częstotliwościowa, Bezzałogowe statki powietrzne, Klasyfikacja nadajnika, Sygnatura radiowa.**

Keywords: **Transmitter classification, RF signature, Time-frequency analysis, Unmanned aerial vehicles.**

1. WSTĘP

Bezzałogowe statki powietrzne (BSP) w ostatnich latach stały się obiektem szczególnego zainteresowania w kontekście ich użycia jako środków do walki zbrojnej. Trwające konflikty ukazują, że zastosowanie znajdują nie tylko kosztowne produkty przeznaczone na rynek wojskowy, ale także tanie drony komercyjne, mogące przenosić ładunki wybuchowe lub być używane jako narzędzia zwiadowcze [1]. Z tego powodu następuje intensywny rozwój systemów przeciwdronowych, których zadaniem jest wczesna detekcja i często neutralizacja (mechaniczna lub elektroniczna) niepożądanych BSP [2].

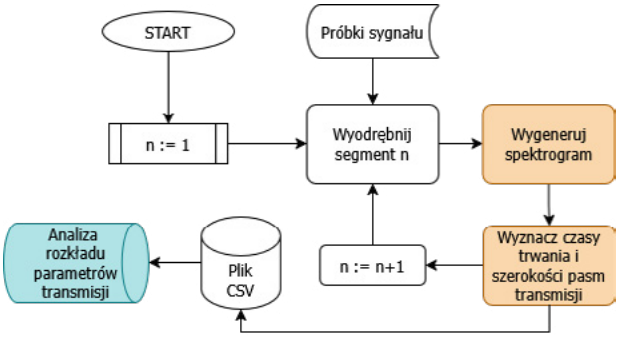
Systemy detekcji BSP obejmują m.in. rozwiązania wizyjne, akustyczne, radarowe oraz radiowe [3]. Detekcja radiowa bazuje na analizie transmisji sygnałów radiowych emitowanych przez BSP, które mogą przenosić np. strumień wideo, dane telemetryczne oraz dane identyfikacyjne. Analiza ta może obejmować samą detekcję transmisji, jak również klasyfikację nadajnika na podstawie sygnatury radiowej[4].

Duża część dostępnych na rynku cywilnym dronów realizuje transmisje cyfrowe o charakterze impulsowym. Informacja o czasie trwania i szerokości pasma tych transmisji może być potencjalnie wykorzystana do identyfikacji modelu BSP.

W niniejszym referacie przedstawiono metodykę i wyniki analizy sygnałów komercyjnych BSP, których indywidulane transmisje zarejestrowano w komorze bezodbiciowej. Celem tego badania jest określenie, czy poszczególne modele dronów wykazują charakterystyczne cechy transmisji, umożliwiające ich wstępną klasyfikację.

2. ANALIZA SYGNAŁU BSP

Analiza czasowo-częstotliwościowa sygnału może być wykonana m.in. na podstawie jego spektrogramu wygenerowanego przy użyciu krótkoczasowej transformacji Fouriera (STFT).

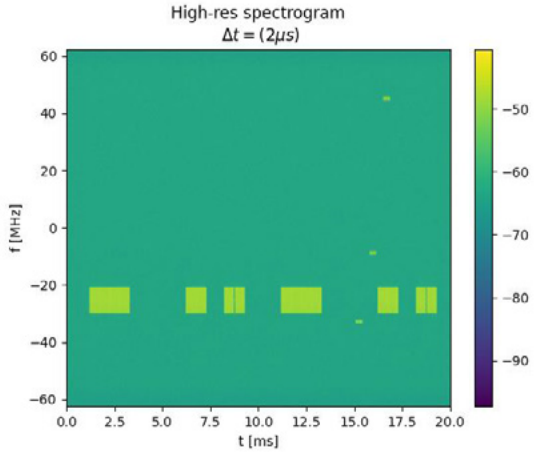


Rys. 1. Procedura przetwarzania odebranego sygnału.

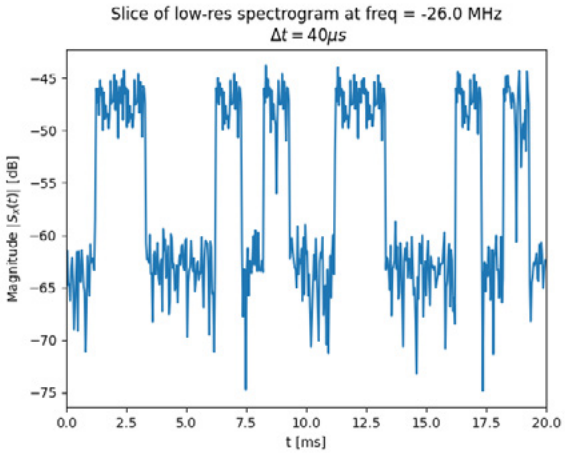
Procedura przetwarzania została przedstawiona na rys. 1. Próbki odebranego sygnału są dzielone na segmenty, dla których obliczane są spektrogramy. W każdym spektrogramie przeprowadzana jest detekcja transmisji oraz estymacja ich czasu trwania i szerokości pasma. Wyniki estymacji są zapisywane do pliku celem późniejszej analizy statystycznej.

Dla poprawności detekcji istotna jest rozdzielczość spektrogramu oraz sposób jego generowania. Na podstawie wstępnych obserwacji założono rozdzielczość częstotliwości 100 kHz oraz rozdzielczość czasową 40 µs. Aby uzyskać lepszej jakości spektrogram, najpierw jest on generowany z rozdzielczością 2 µs, a następnie jest on kumulowany w dziedzinie czasu po 20 próbek. Na rys. 2 pokazano spektrogram przykładowego segmentu sygnału

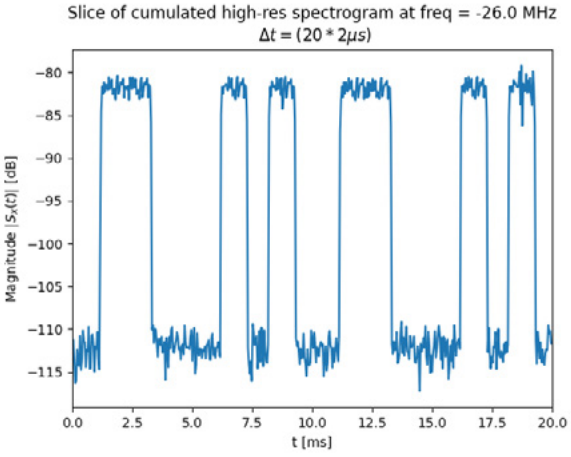
o długości 20 ms. Można na nim zauważyć 6 transmisji szerokopasmowych oraz 3 krótkie i wąskopasmowe.



Rys. 2. Spektrogram przykładowego fragmentu transmisji BSP.



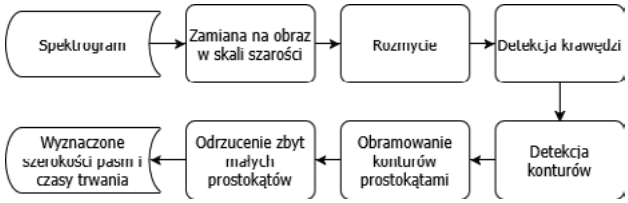
Rys. 3. Przekrój spektrogramu o niskiej rozdzielczości czasowej.



Rys. 4. Przekrój skumulowanego spektrogramu o wysokiej rozdzielczości czasowej.

Rys. 3 prezentuje przekrój spektrogramu o niskiej rozdzielczości (40 μs) na częstotliwości -26 MHz. Z kolei rys. 4 pokazuje analogiczny przerój dla skumulowanego po 20 próbek spektrogramu o wyższej rozdzielczości (2 μs). Zauważalna jest istotna różnica na korzyść spektrogramu skumulowanego. W jego przypadku przebieg jest mniej zaszumiony, a chwile początków i końców transmisji są bardziej jednoznaczne. Wynika to z tego, że spektrogram skumulowany jest obliczany na podstawie 20-krotnie większej liczby próbek niż spektrogram niskorozdzielczy.

Kolejnym etapem analizy jest estymacja parametrów transmisji. W tym celu wykorzystano otwartą bibliotekę OpenCV służącą do analizy i przetwarzania obrazu [5]. Procedura została przedstawiona na rys. 5.



Rys. 5. Przetwarzanie z użyciem biblioteki OpenCV.

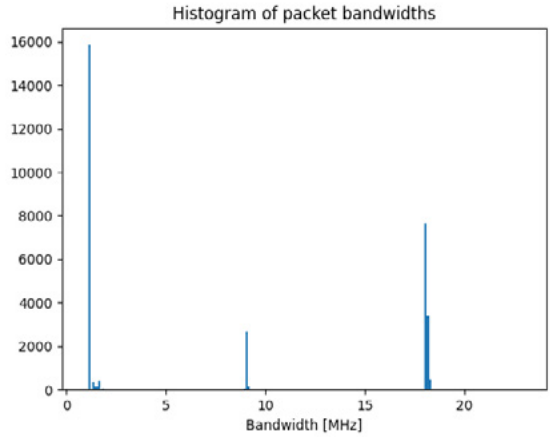
W pierwszej kolejności moduł spektrogramu (w skali dB) jest zamieniany na obraz w 8-bitowej skali szarości. Kolejnym krokiem jest rozmycie obrazu filtrem gaussowskim, co ma na celu zredukowanie ziarnistości. Po tym jest stosowany algorytm Canny'ego, który dokonuje detekcji krawędzi pionowych i poziomych. Wśród wykrytych krawędzi wybierane są tylko te, które tworzą kontury (zamknięte obszary), a następnie dla każdego konturu jest wyznaczany obejmujący go możliwie najmniejszy prostokąt. Na podstawie wymiarów tego prostokąta estymowana jest szerokość pasma i czas trwania transmisji. Przyjęto że minimalna szerokość pasma to 500 kHz, a minimalny czas trwania to 200 μs, w związku z czym prostokąty o boku krótszym niż 5 pikseli są odrzucane.

3. ANALIZA STATYSTYCZNA

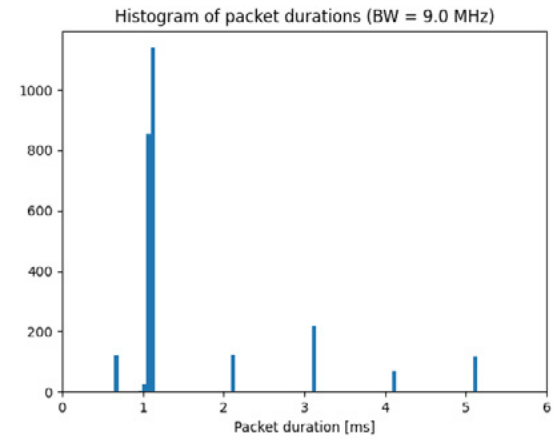
Analizowane dane stanowią zbiór plików binarnych z próbkami sygnałów zarejestrowanych w komorze bezodbiciowej dla ośmiu modeli BSP. Rejestracji dokonywano równocześnie w pasmach 2,4 GHz i 5,8 GHz z próbkowaniem 125 MHz. Uwzględniono przy tym różne scenariusze, np. start, lot, spoczynek. Dla każdego scenariusza przeprowadzono 32-sekundową rejestrację.

Zarejestrowane pliki poddano procedurze przetwarzania opisanej w p. 2. Wyniki poddano analizie statystycznej. W pierwszej kolejności generowano histogram szerokości pasm. Przykład takiego histogramu, wyznaczonego dla modelu DJI Mini 2, przedstawiono na rys. 6. Można na nim zauważyć skupienie wyników wokół wartości 1,2 MHz, 9 MHz i 18 MHz.

Dla każdego ze skupisk związanych z daną szerokością pasma wygenerowano analogiczne histogramy czasu trwania. Przykładowy histogram dla tego samego modelu drona, dla pasma 9 MHz przedstawiono na rys. 7. W tym wypadku również widać skupiska wokół określonych wartości. Duże skupienie wyników wskazuje że estymacja parametrów jest realizowana z dokładnością zbliżoną do założonej rozdzielczości spektrogramu.



Rys. 6. Histogram szerokości pasm DJI Mini 2.



Rys. 7. Histogram czasów trwania transmisji o szerokości pasma 9 MHz dla DJI Mini 2.

W tab. 1 przedstawiono listę modeli BSP, dla których przeprowadzono rejestrację, wraz z liczbą transmisji, które wykryto we wszystkich przetworzonych plikach. Różnice w liczbie transmisji wynikają zarówno z odmienności standardów transmisji dla poszczególnych dronów, jak również z faktu, że dla niektórych modeli nie udało się przeprowadzić niektórych scenariuszy np. z uwagi na niedostępność sygnału GPS.

Tab. 1. Przebadane modele BSP oraz liczby wykrytych transmisji.

Lp.	*Model BSP*	*L. transmisji*
1	DJI Mavic 3	87748
2	DJI Matrice 300	15535
3	DJI Air 2S	51143
4	DJI Mini 2	32057
5	DJI Matrice 30	25370
6	Autel Evo Nano	66971
7	Autel Evo Lite	49606
8	Autel Evo II Pro	129190

W tab. 2 i 3 zaprezentowano powiązanie pomierzonych parametrów transmisji z modelami BSP, dla których takie transmisje stwierdzono. Numery w tabeli odpowiadają numerom w pierwszej kolumnie tab. 1.

Tab. 2. Parametry transmisji modeli DJI.

	Szerokość pasma [MHz]				
Czas trwania [ms]	**1,2**	**2,2**	**9**	**18**	**36**
0,5	4	1,2,3,5			
0,6			3		
0,7			1,2,4,5		
1,1			1,3,4	2,3,4	1
2,1			1,2,3,4,5	2,3,4	1,5
3,1			1,2,3,4,5	2,3,4	1,5
4,1			1,2,3,4,5	2,3,4	1,5
5,1			1,2,3,4,5	2,3,4	1,5

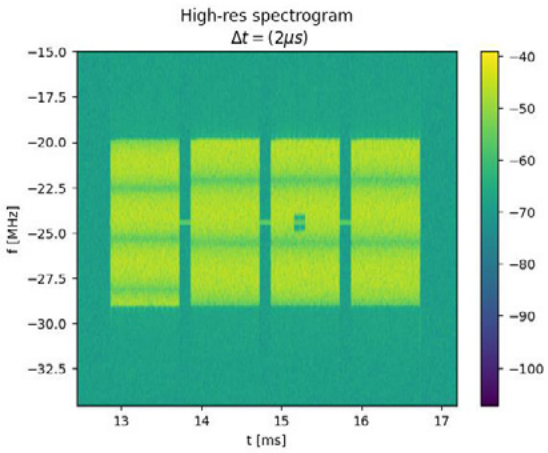
Tab. 3. Parametry transmisji modeli Autel.

	Szerokość pasma [MHz]				
Czas trwania [ms]	***1,2***	***1,8***	***9***	***13,6***	***18***
0,2		6			
0,3	6		6		
0,6				6	
0,8	7,8	6			
0,9			6,7		8
1,3	6				

Drony DJI stosują autorski standard transmisji Ocusync, przy czym każdy z badanych modeli implementował inną wersję tego standardu.. W wypadku tych dronów można zauważyć, że te same czasy trwania i szerokości pasm są stosowane w wielu modelach. W szczególności wspólne dla wszystkich są transmisje w paśmie 9 MHz o czasach trwania od 2,1 ms do 5,1 ms z krokiem co 1 ms. Największy odsetek stanowią transmisje o czasie trwania 0,5 ms (od 35% do 50% wszystkich transmisji, w zależności od BSP), które przenoszą dane sterujące lub telemetryczne.

Dla dronów Autel, podobieństwo stwierdzono w przypadku modeli Lite i Evo II Pro, które stosują te same dwie długości transmisji, przy czym dla transmisji o długości 0,9 ms pierwszy model stosuje pasmo 9 MHz a drugi 18 MHz. Z kolei model Evo Nano stosuje transmisje o siedmiu różnych zestawach parametrów, przy czym tylko jeden z nich jest zgodny z transmisją innego modelu (Evo Lite, 0.9 ms, 9 MHz).

Należy wspomnieć, że parametry niektórych transmisji dronów Autel zostały estymowane błędnie z uwagi na ich seryjny charakter. Modele Autel często nadają w seriach po 3-5 pakietów, pomiędzy którymi transmisja nie jest zupełnie przerywana, a jedynie następuje przerwa w modulacji fali nośnej. Taki przypadek przedstawiono na rys. 8. Niemniej, takie błędne estymaty zostały zidentyfikowane i nieuwzględnione w wynikach.



Rys. 8. Spektrogram transmisji seryjnej Autel Evo Lite.

4. PODSUMOWANIE

Uzyskane wyniki pokazują, że możliwe jest rozróżnienie BSP różnych producentów. Wśród analizowanych modeli, pomiędzy DJI a Autel nie stwierdzono żadnego typu transmisji która miałaby zarazem taki sam czas trwania i szerokość pasma.

Rozróżnienie modeli danego producenta jest mniej jednoznaczne, zwłaszcza w przypadku DJI gdzie drony emitują transmisje o podobnych parametrach. Niemniej, dla każdego z analizowanych modeli zidentyfikowano cechę odróżniającą go od pozostałych:

- Mini 2 jako jedyny nadaje z szerokością pasma 1.2 MHz,
- Air 2S jako jedyny nadaje [9 MHz, 0.6 ms],
- Mavic 3 jako jedyny nadaje zarazem 36 MHz oraz [9 MHz, 1.1 ms],
- Matrice 30 jako jedyny nie nadaje z czasem 1.1 ms,
- Matrice 300 jako jedyny nadaje zarazem [9 MHz, 0.7 ms] oraz 2.2 MHz i 18 MHz.

Natomiast dla modeli Autel stwierdzono co następuje:

- Evo Nano jako jedyny nadaje z szerokością pasma 13.6 MHz,
- Evo Lite jako jedyny nadaje zarazem 9 MHz oraz [1.2 MHz, 0.8 ms],
- Evo II Pro jako jedyny nadaje zarazem 18 MHz oraz [1.2 MHz, 0.8 ms].

Jednocześnie należy stwierdzić, że zbiór przebadanych modeli BSP był względnie niewielki i nie można wykluczyć, że wyżej wymienione cechy mogą być właściwe również dla innego, nieuwzględnionego tu modelu. Dlatego też przewidziana jest kontynuacja badań w zakresie analizy innych parametrów, takich jak np. odstęp pomiędzy transmisjami lub sekwencja przeskoków częstotliwości obserwowanych przy transmisji pakietów wąskopasmowych.

Inną cechą, która może być wykorzystana w klasyfikatorze BSP jest obecność stałych, charakterystycznych fragmentów transmisji, jak chociażby „wycięcie” widoczne w trzecim pakiecie na spektrogramie Autel Evo Lite z rys. 8. W przypadku dronów DJI także stwierdzono charakterystyczne elementy, będące symbolami synchronizacyjnymi, modulowanymi sekwencjami Zadoff-Chu.

Dalsze prace powinny obejmować również rejestrację emisji w środowisku rzeczywistym, co pozwoliłoby określić dokładności estymacji parametrów transmisji w obecności zaników i zakłóceń od innych emisji obecnych w pasmach nielicencjonowanych.

LITERATURA

[1] Molloy, Oleksandra. "Drones in Modern Warfare: Lessons Learnt from the War in Ukraine." Australian Army Research Centre (2024).
[2] Wang, Jian, Yongxin Liu, Houbing Song. "Counter-unmanned aircraft system (s)(C-UAS): State of the art, challenges, and future trends." *IEEE Aerospace and Electronic Systems Magazine* 36.3 (2021): 4-29.
[3] Guvenc, Ismail, et al. "Detection, tracking, and interdiction for amateur drones." *IEEE communications magazine* 56.4 (2018): 75-81.
[4] Nguyen, Phuc, et al. "Matthan: Drone presence detection by identifying physical signatures in the drone's RF communication." *Proceedings of the 15th annual international conference on mobile systems, applications, and services*. 2017.
[5] https://opencv.org/

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

RADIOKOMUNIKACJA MORSKA W ŚWIETLE PRZEPISÓW IMO

IMO CLASSIFICATION OF SEA RADIOKOMMUNIKATION SYTEMS

Ryszard J Katulski

Politechnika Gdańska, Gdańsk, rjkat@eti.pg.edu.pl

DOI: 10.15199/59.2025.4.72

Streszczenie**: W pracy przedstawiono obowiązkowe wyposażenie statków morskich w urządzenie radiowe i radionawigacyjne, niezbędne dla potrzeb efektywnego zarządzania rejsem i bezpiecznego nawigowania nierzadko w trudnych warunkach pogodowych. Scharakteryzowano cztery kategorie obszarów pływania i w powiązaniu z tym opisano standardy tych urządzeń**

Abstract**: In this paper the obligatory ship radio and navigation equipments have been presented due to four categories of sea regions.**

Słowa kluczowe**: ship radio and radio navigation obligatory equipment, IMO regulations, GMDSS.**

1. INTRODUCTION

Przedmiotem referatu są systemy radiokomunikacyjne stosowane w praktyce morskiej z punktu widzenia obowiązkowego wyposażenia jednostek pływających, zwłaszcza statków handlowych, w urządzenia do realizacji łączności radiowej i radionawigacyjnej, co warunkuje ich dopuszczenie do pływania na morzu przez przepisy Międzynarodowej Organizacji Morskiej IMO (International Maritime Organization). Przepisy te odnoszą się do statków o wyporności poczynając od 300 DWT, tzw. tony rejestrowe. Nie stosowanie się do tych przepisów obwarowane jest Kodeksem Karnym, czego wyrazem jest nie wydanie zezwolenia na wpłynięcie do portu.

W drugiej części tego opracowania opisano Globalny Morski System Łączności Alarmowej i Bezpieczeństwa znany pod skrótową nazwą GMDSS (Global Maritime Distress and Safety System), którego ważnym elementem jest łączność satelitarna.

2. KLASYFIKACJA REGIONÓW PŁYWANIA

2.1. Region A1

Region A1 to obszar przybrzeżny położony pomiędzy linią brzegową lądu do odległości 30 mil morskich, tj. około 50 km od tej linii. Jest to obszar, który jest (powinien być) w pełni kontrolowany przez służby graniczne danego państwa. Nieuprawnione wpłynięcie do tego obszaru jest traktowane jako naruszenie granicy państwowej z wszystkimi konsekwencjami takiego stanu rzeczy. Wszystkie obiekty wpływające do tego obszaru podlegają przepisom prawnym tego państwa. Kapitan każdego statku jest zobowiązany do wykonywania poleceń służb granicznych, w tym umożliwiania wejścia na pokład funkcjonariuszom tych służb.

W przypadku nie dostosowania się do tych wymagań, sytuację taką traktuje się jako stan wyjątkowy i powiadamia się w trybie alarmowym siły zbrojne z Marynarką Wojenną włącznie.

2.2. Region A2

Z kolei Region A2 to obszar położony w odległości powyżej 30 mM od brzegu, rozciągający się do 150 mM w głąb morza, tj. do około 300 km od linii brzegowej. Zazwyczaj jest to obszar o charakterze międzynarodowym podlegający wyłącznie przepisom IMO.

2.3. Region A3

Natomiast Region A3 to obszar położony w odległości powyżej 150 mM od brzegu, który stanowią obszary otwartych mórz i oceanów. Jest to obszar jednoznacznie o charakterze międzynarodowym, podlegający także przepisom IMO.

2.4. Region A4

W porównaniu do powyżej opisanych Regionów pływania, poprzez które przebiegają trasy żeglugi handlowej, Region A4 stanowią obszary podbiegunowe. Jak sama ich nazwa wskazuje, są one położone wokół biegunów: północnego i południowego, gdzie ze względu na warunki nie przebiegają trasy żeglugi handlowej.

3. WYMAGANIA IMO

3.1. Region A1

Statki pływające tylko w Regionie A1 powinny być wyposażone w następujący zestaw urządzeń radiowych i radionawigacyjnych, tzn. w:

- radiostację VHF do obsługi kanałów radiowych w paśmie od 156 MHz do 162,05 MHz, która powinna obsługiwać wydzielone kanały do pracy w niebezpieczeństwie i dla potrzeb bezpieczeństwa, tzn. do nadawania i odbioru sygnałów alarmowych.

3.2. Regiony A1 i A2

Statki pływające z oczywistych względów w obu Regionach, tzn. w Regionie A1 i w Regionie A2, powinny być wyposażone w:

- radiostację VHF do obsługi kanałów radiowych w paśmie od 156 MHz do 162,05 MHz, która powinna obsługiwać wydzielone kanały do pracy w niebezpieczeństwie i dla potrzeb bezpieczeństwa, oraz w:
- radiostację pośrednio-falową do obsługi kanałów radiowych w paśmie od 1605 kHz do 4000 kHz, która również powinna obsługiwać wydzielone kanały do pracy w niebezpieczeństwie i dla potrzeb bezpieczeństwa.

3.3. Regiony A1, A2 i A3

Statki dalekomorskie pływające w Regionach A1, A2 i A3, powinny być wyposażone w:
- urządzenia radiokomunikacyjne właściwe dla Regionów A1 i A2, oraz w
- radiostację do łączności satelitarnej w systemie INMARSAT/COSPAR-SARSAT [1, 2], oraz także
- radiostację krótkofalową, w przypadku żeglugi w obszarach podbiegunowych.

4. SYSTEMY ŁĄCZNOŚCI DLA POTRZEB RATOWNICTWA MORSKIEGO

Oprócz powyżej opisanego wyposażenia, w celu poprawienia bezpieczeństwa żeglugi są eksploatowane dwa podsystemy radiokomunikacyjne przeznaczone do rozgłaszania ostrzeżeń nawigacyjnych i meteorologicznych oraz prognoz pogodowych, tj:
- NAVTEX (Navigational Telex*)* oraz
- SAFETY NET.

Pierwszy z wymienionych systemów pracuje w paśmie fal średnich, na częstotliwości 518 kHz. Odbiór komunikatów odbywa się przy użyciu specjalnych odbiorników radiotelegraficznych, których wyposażenie jest obligatoryjne dla wszystkich statków objętych Konwencją SOLAS (Safety of Life At Sea). Są to tzw. jednostki pływające *konwencyjne*, niezależnie od obszaru pływania. Nasłuch przy użyciu tych odbiorników odbywa się całodobowo, zaś odbiór komunikatów jest automatyczny.

Natomiast drugi z tych podsystemów wchodzi w skład satelitarnego systemu *wywołania grupowego* o nazwie EGC (Enhanced Group Calling), realizowany przez system INMARSAT do wybranej grupy statków, na potrzeby autoryzowanych instytucji.

5. RADIOWE ŚRODKI RATUNKOWE

Wszystkie statki konwencyjne muszą być wyposażone w tzw. *radiowe środki ratunkowe*, co jest obligatoryjne niezależnie od obszaru pływania. Do środków tych zalicza się *radiopławę awaryjną* oraz *transponder radarowy*.

Pierwszy z nich, tzn. *radiopława awaryjna,* - tzw. EPIRB (Emergency Positioning Indicating Radio Beacon) to specjalny nadajnik radiowy przeznaczony do powiadamiania o wystąpieniu zagrożenia. Nadajnik ten może pracować w następujących zakresach częstotliwości:
- w paśmie 156 MHz, z zastosowaniem DSC (Digital Selective Calling), do stosowania w obszarze A1,
- w paśmie 406 MHz, we współpracy z niskoorbitowym systemem satelitarnym COSPAS-SARSAT, dzięki temu może być stosowany we wszystkich obszarach pływania,
- w paśmie 1,6 GHz, we współpracy z systemem geostacjonarnym INMARSAT, do stosowania także globalnie, bez obszarów podbiegunowych.

Przy czym, najbardziej uniwersalne pod względem użyteczności są radiopławy współpracujące z systemem COSPAS/SARSAT, co przesądza o ich powszechnym stosowaniu.

Zadaniem każdej radiopławy jest wysyłanie sygnału zawierającego następujące informacje:
- pozycję katastrofy morskiej i dane identyfikacyjne zagrożonego statku, oraz
- inne dodatkowe informacje, które mogą ułatwić prowadzenie akcji ratunkowej.

Uruchomienie radiopławy może być ręczne lub automatyczne po jej spłynięciu na wodę. Po takim uaktywnieniu nadajnik radiopławy wysyła w kierunku satelity sygnał alarmowy poprzez antenę o hemisferycznej charakterystyce promieniowania. Zadaniem tego nadajnika jest pomoc w wykrywaniu i lokalizowaniu rozbitków na morzu. Ponadto, po uruchomieniu nadajnika uaktywnia się również transponder radarowy zainstalowany na radiopławie.

Radiopława stanowi układ odbiorczo-nadawczy, który po odebraniu sygnału radarowego wysłanego z innego statku lub samolotu uczestniczącego w akcji ratunkowej wysyła krótkotrwały sygnał odzewowy z szybkim przesuwaniem częstotliwości w paśmie od 9,2 GHz do 9,5 GHz. Taki sygnał odebrany przez urządzenie radarowe jest zobrazowany na jego ekranie w postaci równo oddalonych kropek, tworząc tzw. kreską namiarową wskazującą położenie radiopławy i tym samym miejsce katastrofy.

6. ŁĄCZNOŚĆ ALARMOWA (GMDSS)

Przed wprowadzeniem systemu GMDSS w praktyce morskiej obowiązywał system alarmowania, który polegał na tym, że sygnał alarmowy ze statku był nadawany dookólnie (do wszystkich statków) na częstotliwości 500 kHz. Wymagało to obligatoryjnego zainstalowania na statku automatycznego odbiornika tego sygnału, tzw. *autoalarmu*, przeznaczonego do całodobowego nasłuchu na powyższej częstotliwości. Dopiero po odebraniu sygnału alarmowego przez autoalarm, informację o tym przekazywano do ratowniczych służb lądowych. W takiej sytuacji jako pierwsze akcję ratunkową podejmowały statki znajdujące się w pobliżu miejsca katastrofy – jeśli się znajdowały?!

Podstawowym zadaniem systemu GMDSS jest alarmowanie o zagrożeniu przy użyciu sygnału wysyłanego z morza do regionalnego centrum ratownictwa morskiego. W Polsce centrum takie mieści się w Gdyni, zaś jego nazwa to: *Morskie Ratownicze Centrum Koordynacyjne* MRCK, które jest wspomagane przez Pomocnicze Centrum Koordynacyjne PCK znajdujące się w Świnoujściu.

Podlegający polskiej administracji system łączności alarmowej opiera się na sieci radiowej WAN (Wireless

Aeria Network) TPSA, do której jest podłączone MRCK oraz dwie stacje radiokomunikacyjne, tzn.:
- nadbrzeżna stacja Witowo-Radio, położona w pobliżu Gdyni, oraz
- tzw. ekspozytury polskiej służby SAR (Search and Rescure), rozlokowane wzdłuż wybrzeża, od Świnoujścia do Krynicy Morskiej, które odbierają sygnały alarmowe nadawane z morza przy użyciu naziemnych środków łączności, w tym z:
- naziemnych (morskich) radiopław awaryjnych, oraz z
- satelitarnych radiopław awaryjnych, w tym przypadku za pośrednictwem naziemnej stacji satelitarnej.

7. ŁĄCZNOŚĆ OPERACYJNA

Łączność operacyjna służy do obsługi sieci rozlokowanych wzdłuż wybrzeża ekspozytur polskie służby SAR, będącymi zespołami nadawczo-odbiorczymi. Zespoły te sterowane są z wydzielonego komputera znajdującego się na stanowisku operacyjnym w centrum koordynacyjnym MRCK.

8. STAN AKTUALNY

Zatem jeśli zagrożony statek znajduje się na otwartym akwenie morskim lub oceanicznym to alarmuje i prowadzi łączność przy użyciu środków satelitarnych – jeśli nimi dysponuje. Wysyła także dookólnie sygnał alarmowy do wszystkich statków, przy użyciu naziemnych środków łączności, przy użyciu wywołania w niebezpieczeństwie.

Skuteczność tego zależy od kilku czynników, tzn. od:
- stanu lądowej infrastruktury telekomunikacyjnej,
- jakości usług serwisowych w portach, do których zawija statek,
- wyposażenia statku w sprawne urządzenia telekomunikacyjne, oraz od
- kwalifikacji operatora, zwłaszcza statkowego.

W świetle wymagań GMDSS-u armatorzy statków mają obowiązek zapewnienia serwisu urządzeń statkowych w portach, do których zawija statek. Jednakże często w praktyce z uwagi na koszty ma to charakter fikcyjny. W rezultacie sprzęt telekomunikacyjny na statku ulega szybkiemu starzeniu.

Poza tym, średnio rzecz biorąc, poziom kwalifikacji statkowych operatorów urządzeń GMDSS w praktyce jest relatywnie niski. Zwłaszcza jeśli się weźmie pod uwagę zlikwidowanie stanowiska radiooficera, którego obowiązki przejął główny elektryk.

W praktyce sprowadza się to do instalowania na statkach najprostszych (najtańszych wersji) terminali satelitarnych i korzystania z urządzeń telekomunikacyjnych niesatelitarnych. Powoduje to obniżenie poziomu tzw. *bezpieczeństwa radiowego*.

9. SYSTEMY ŁĄCZNOŚĆI DLA POTRZEB OCHRONY ŻEGLUGI

W praktyce morskiej znajdują zastosowanie następujące systemy przeznaczone dla potrzeb ochrony żeglugi:
- automatyczny system identyfikacyjny, w skrócie AIS,
- statkowy system alarmowania, w skrócie SSAS, oraz
- system identyfikacji i śledzenia, w skrócie LRIT.

9.1. AIS

AIS to system cyfrowy działający w strefie A1, oparty na radiokomunikacji VHF bliskiego zasięgu, tzn. pracujący w paśmie od 156 MHz do 162 MHz, a dokładnie dla potrzeb tego systemu przydzielono dwa następujące simpleksowe kanały częstotliwościowe:
- AS1, przy częstotliwości 161,975 MHz, oraz
- AS2, przy częstotliwości 162,025 MHz,

przy czym możliwe jest użytkowanie obu kanałów jednocześnie w trybie dupleksowym.

System ten służy do automatycznego przekazywania zestawu podstawowych informacji związanych z żeglugą statku. Informacje te są przesyłane ze statku do lądowych stacji radiokomunikacyjnych, nazywanych bazowymi. Z kolei stamtąd są przekazywane do odpowiednich służb zajmujących się obsługą i bezpieczeństwem statku, jak np. Służba SAR lub Straż Graniczna.

Przekazywany w ten sposób zestaw informacji zawiera:
- identyfikator statku,
- parametry nawigacyjne statku, tzn. aktualną pozycję, kurs i prędkość statku,
- rodzaj przewożonego ładunku.

Dane te są przekazywane z szybkością 9,6 kbit/s i zapisywane w odpowiednich bazach danych, z których w zależności od potrzeb są pobierane i przetwarzane do różnych celów, np. do aktualnego zobrazowania rejsu statku na mapie elektronicznej.

W praktyce można wyróżnić trzy podstawowe tryby pracy systemu AIS zależne od akwenu, w którym znajduje się statek, tzn.
- autonomiczny i ciągły na wodach otwartych,
- przypisany do obszaru, w którym statek jest monitorowany, oraz
- odzewowy na żądanie kompetentnych władz.

Należy podkreślić, że geneza wprowadzenia systemu AIS wynikała z potrzeby poprawy obsługi nawigacyjnej statku, z punktu widzenia ochrony przed akcjami terrorystycznymi, czemu został przyporządkowany późniejszy rozwój tego systemu.

System AIS na obszarze przybrzeżnym Bałtyku został w pełni pokryty zintegrowaną siecią stacji bazowych poszczególnych krajów, przy czym podstawowym warunkiem tego była kompatybilność standardów poszczególnych sieci narodowych oraz jednolite formaty kodowania informacji.

Tworzenie polskiej sieci rozpoczęto w 2004 roku, poczynając od pierwszej stacji bazowej na Helu. Aktualnie na polskim wybrzeżu pracuje kilka takich stacji bazowych.

9.2. SSAS

Natomiast system SSAS jest przeznaczony do poufnego informowania armatora statku i służb oraz odpowiednich instytucji o wystąpieniu na statku zagrożenia terrorystycznego.

Ważnym elementem tego systemu są dwa ukryte niezależne przyciski alarmowe przeznaczone do jego uruchomienia. Jeden z tych przycisków powinien znajdować się w pomieszczeniu, gdzie prowadzona jest nawigacja statku. Natomiast drugi przycisk w innym poufnym miejscu. Poza tym instalacja wszystkich elementów powinna być wykonana w sposób dyskretny, tzn. nie powinna być widoczna dla otoczenia.

System ten powinien mieć również funkcję testowania uruchamianą przy użyciu przeznaczonego do tego celu trzeciego przycisku, także poufnie zainstalowanego.

Zatem zadziałanie opisywanego systemu rozpoczyna się z chwilą uruchomienia jednego z tych przycisków. W następstwie tego poprzez statkowy terminal satelitarny jest wysyłany sygnał alarmowy, który następnie poprzez satelitę i stację naziemną, za pośrednictwem lądowego usługodawcy łączności alarmowej zostanie przekazany do instytucji zobowiązanej do ochrony statków.

Dla potrzeb systemu SSAS można wykorzystać dowolny podsystem satelitarnej radiokomunikacji morskiej, np. system Inmarsat lub Irydium.

W tym miejscu należy podkreślić, że zgodnie z wymaganiami kodeksu ISPS (International Ship and Port Facility Security Code), na statku powinien funkcjonować specjalny oficer ochrony, tzw. SSO (Ship Security Oficer), którego zadaniem jest zorganizowanie i nadzorowanie statkowego systemu ochrony. Rolę tego oficera powinno się powierzać poufnie wybranemu członkowi załogi statku.

9.3. LRIT

Inicjatywa budowy systemu LRIT została podjęta w Stanach Zjednoczonych. Działanie tego systemu miało obniżyć ryzyko zagrożenia żeglugi przez czynniki terrorystyczne. Z założenia system ten ma umożliwiać śledzenie każdego statku zwłaszcza podczas rejsu. Przez co rozumie się:

- wykrycie statku i jego identyfikację,
- ciągły nadzór nad tym statkiem,
- klasyfikację wykrytych statków z punktu widzenia zagrożenia.

Oznacza to, że każdy statek konwencyjny jest (powinien być) monitorowany przez system LRIT, który to system jest (powinien być) kompatybilny ze standardami IMO i pracuje pod nadzorem IMSO (International Maritime Satellite Organization), w oparciu o rozproszone (trzy) bazy danych.

Oprócz zbierania danych system LRIT realizuje dystrybucję tych danych do określonych odbiorców, przede wszystkim do:

- służb administracyjnych państwa flagi statku oraz państw przybrzeżnych,
- administracji portowych,
- służb SAR-owskich oraz służb monitoringu środowiska w poszczególnych krajach.

Zatem, system LRIT stanowi rozbudowaną strukturę monitorująco-kontrolną. .

10. RADIOKOMUNIKCJA MORSKA W UJĘCIU PRAKTYCZNYM

10.1. Podsystem VHF

Jest to najbardziej eksploatowany podsystem radiokomunikacji morskiej, który wykazuje dużą przydatność do opisanych zadań systemu GMDSS, zwłaszcza dla zapewnienia łączności w niebezpieczeństwie i dla potrzeb bezpieczeństwa w strefie A1.

Poza tym jest to system do komunikowania się pomiędzy jednostkami pływającymi na terenie portów i w Regionie A1. Z tego tytułu jest to system bardzo obciążony, co wpływa niekorzystnie na bezpieczeństwo żeglugi w takich warunkach. Sprawą zasadniczą jest bardziej efektywne wykorzystanie zasobów częstotliwościowych, polegające na zamianie kanałów dupleksowych na simpleksowe z 25 kHz na 12,5 kHz, co odnosi się do sygnałów mowy i telegrafii. Na tym tle atrakcyjne jest stosowanie systemu cyfrowego z kanałami wąskopasmowymi o szerokości 6,25 kHz lub standardu TETRA (Trans-European Truncked Radio).

System TETRA zapewnia szeroki zestaw profesjonalnych usług i jest przydatny dla realizacji łączności najogólniej przybrzeżnej, co jest także ważne z punktu widzenia wsparcia ochrony naszej strefy przybrzeżnej, co ma znaczenie nie tylko gospodarcze lecz i także militarne.

10.2. Radioteleks

Radioteleks morski, tzn. automatyczna telegrafia wąskopasmowa miała w systemie GMDSS zastąpić telegrafię Morse'a. W pierwszej połowie lat 90-tych, kiedy to miało miejsce apogeum stosowania telegrafii Morse'a, stacje nadbrzeżne obsługiwały tę łączność w trybie automatycznym, przy pomocy komputerów klasy PC, co umożliwiało realizację wielu dodatkowych usług, np. konwersję wiadomości tekstowych w trybie najpierw telefaksowym lub później w trybie e-mailowym. Jednakże z upływem czasu, zaczęto stosować tanie terminale satelitarne klasy Inmarsat C, przeznaczone do transmisji danych tekstowych.

Należy jednak mieć na uwadze, że w systemie GMDSS stosuje się radioteleks do odbioru ostrzeżeń w systemie NAVTEX.

W takim stanie rzeczy zaczęła funkcjonować sieć kilkudziesięciu stacji nadbrzeżnych rozmieszczonych na całym świecie, która obsługuje ponad 10 tysięcy statków, oferując różne usługi, jak śledzenie statków i pocztę elektroniczną.

Dodatkowym czynnikiem, który należy wziąć pod uwagę, to zainteresowania tym sposobem transmisji ze strony sił morskich NATO.

10.3. Radiotelefonia MF/HF

Jest to podstawowy środek łączności w niebezpieczeństwie i dla potrzeb bezpieczeństwa stosowany w ośrodkach ratownictwa morskiego. Poza tym łączność tą stosuje się także do komunikacji eksploatacyjnej pomiędzy statkami, zwłaszcza w paśmie MF, w Regionach A1 i A2.

Poza tym należy mieć na uwadze, że łączność w paśmie MF, w porównaniu do łączności w paśmie VHF, jest bardziej uniwersalna zasięgowo.

10.4. Cyfrowe selektywne wywołanie

Jest to także ważny element współczesnej radiokomunikacji morskiej, umożliwiający automatyzację w nawiązywaniu połączeń. Jednakże wymaga to uproszczenia protokołu komunikacyjnego oraz procedury prowadzenia takiej łączności.

10.5. Łączność satelitarna

W warunkach morskich (zwłaszcza na obszarach oceanicznych) nie mamy bezpośredniego dostępu do lądowej infrastruktury telekomunikacyjnej. Zatem, w takim stanie rzeczy istotne znaczenie ma dostęp do satelity telekomunikacyjnego pośredniczącego w łączności z lądową infrastrukturą telekomunikacyjną. Jest to ważny rodzaj współczesnej radiokomunikacji morskiej, umożliwiający globalny dostęp, za wyjątkiem obszarów podbiegunowych, do systemu GMDSS.

Kluczową rolę w tym spełnia system INMARSAT (International Maritime Satellite System), który stał się światowym liderem w dostarczaniu usług bezprzewodowej łączności telefonicznej i transmisji danych. Aktualnie w ramach tego systemu funkcjonują następujące podsystemy składowe:

- INMARSAT-B, system cyfrowy, dostarcza usługi telefonii i łączności telefaksowej oraz transmisję danych z prędkością do 64 kbit/s oraz zapewnia łączność alarmową w systemie GMDSS,

- INMARSAT-M, który podobnie jak powyżej, zapewnia łączność użytkowników publicznej sieci telefonicznej z użytkownikami stacji ruchomych na lądzie i na morzu, w tym ostatnim przypadku także w systemie GMDSS,

- INMARSAT-Mini-M, w którym stosowane są zminiaturyzowane terminale użytkowników,

- INMARSAT-E, w którym funkcjonują satelitarne radiopławy awaryjne EPIRB, oraz

- INMARSAT-C, który zapewnia następujące rodzaje usług:
  - teleks,
  - faksymiliografię,
  - transfer plików o rozmiarze do 32 kB,
  - wysyłanie i odbiór wiadomości e-mailowych,
  - przesyłanie informacji w relacji statek-statek.

10.6. System telefonii komórkowej na morzu

Jest to także ważny element współczesnej radiokomunikacji morskie w Regionie A1, obsługujący lądowy i morski obszar przybrzeżny.

10.7. Podsumowanie

Systemy radiokomunikacji morskiej umożliwiają dostęp do informacji zarówno służbowej jak i poza służbowej w złożonych, często skrajnie niebezpiecznych, warunkach pracy na morzu. Znaczenie tego faktu jest fundamentalne i nie da się tego przecenić, zarówno z punktu widzenia gospodarczego jak i również bezpieczeństwa i obronności.

LITERATURA

[1] Czajkowski J. 1994, „Inmarsat-C oraz -M a System GMDSS", PWP Skryba, Gdańsk, 30-64, 88

[2] Felski A., Urbański J. 1997, „Satelitarne systemy nawigacji i bezpieczeństwa żeglugi". AMW, Gdynia, 176-180

[3] Katulski R. J., 1998, „Analiza i badania anten do satelitarnej telekomunikacji morskiej". WAT/PG, 2558/98, Warszawa - Gdańsk, 1-177

[4] Katulski R. J., 2009, „Propagacja fal radiowych w telekomunikacji bezprzewodowej". WKiŁ, Warszawa, 1-230

[5] Matyszkiel R., 2020, „Bezpieczeństwo morskiej łączności bezprzewodowej". WAT, Warszawa

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

INFOLIGHT: WIELOSENSOROWA INTELIGENTNA PLATFORMA STEROWANIA OŚWIETLENIEM

INFOLIGHT: A MULTI-SENSOR INTELLIGENT LIGHTING CONTROL PLATFORM

Tomasz Śmiałkowski[1]; Wojciech Słowiński[2]; Piotr Szczuko[3]

[1] TSTRONIC, Gdańsk, tomasz.smialkowski@tstronic.eu
[2] REVA-SILED, Gdańsk, wojciech.slowinski@revasiled.pl
[3] Wydział Elektroniki, Telekomunikacji i Informatyki, Politechnika Gdańska, piotr.szczuko@pg.edu.pl

DOI: 10.15199/59.2025.4.73

Streszczenie: **Referat przedstawia wielosensorowe rozwiązanie do inteligentnego sterowania oświetleniem, wykorzystujące przetwarzanie brzegowe i analizę danych z czujników. System ma na celu efektywne zarządzanie infrastrukturą oświetleniową, oszczędność energii i poprawę jakości życia. W urządzeniu brzegowym realizowane jest przetwarzanie danych i sterowanie, a poprzez warstwę pośredniczącą system łączy się z zasobami chmurowymi. Zaproponowano i przetestowano modułową konstrukcję, która umożliwia rozbudowę funkcji i dobór czujników do każdej lampy.**

Abstract: **This paper presents a multi-sensor solution for intelligent lighting control, utilizing edge processing and sensor data analysis. The system aims to effectively manage infrastructure, reduce energy consumtion, and improve quality of life. Data processing and control are performed at the edge device level, and the system connects to cloud resources via an intermediate layer. A modular design has been proposed and tested, allowing for function expansion and sensor selection for each lamp.**

Słowa kluczowe: **Inteligentne miasto, Internet Rzeczy, Urządzenia brzegowe, Zanieczyszczenie światłem**

Keywords: **Edge Computing, Internet of Things, Light pollution, Smart City**

1. WPROWADZENIE

Miasta i aglomeracje stoją obecnie przed wyzwaniami związanymi z efektywnym zarządzaniem infrastrukturą, oszczędnością energii oraz poprawą jakości życia mieszkańców. Jednym z kluczowych obszarów rozwoju koncepcji Smart City jest inteligentne oświetlenie uliczne, które dzięki integracji z systemami informatycznymi i wykorzystaniu różnorodnych sensorów może pełnić funkcje znacznie wykraczające poza dostarczanie właściwego poziomu światła o wskazanej porze.

Poprawa jakości życia mieszkańców miast wiąże się z zapewnieniem bezpieczeństwa, komfortu oraz zdrowego środowiska. Oświetlenie uliczne, odpowiednio dostosowane do warunków otoczenia, może znacząco wpłynąć na poczucie bezpieczeństwa w przestrzeni publicznej, ograniczenie przestępczości, a także na zmniejszenie zanieczyszczenia światłem [10] i zużycia energii. Cele te osiągnąć można poprzez zastosowanie systemów, które reagują na zmieniające się warunki środowiskowe, oświetleniowe, widoczność, charakter i natężenie ruchu pieszych i pojazdów, rodzaj otoczenia (zabudowania, ciągi piesze lub ulice) i różne dynamiczne zdarzenia.

Odpowiedni poziom automatyzacji i inteligencji w INFOLIGHT osiągany jest poprzez akwizycję i analizę danych z różnych sensorów: czujników światła, hałasu, ruchu, jakości powietrza, obrazu i detekcji mikrofalowej.

Dzięki lokalnemu przetwarzaniu danych w urządzeniu brzegowym, możliwe jest automatyzowanie podstawowych funkcji oraz analizowanie strumieni danych w bezpieczny sposób, bez konieczności przesyłania ich do chmury. Daje to możliwości do dynamicznego dostosowania natężenia światła do poziomu jasności otoczenia i obecności pieszych lub pojazdów; zmiany barwy światła w zależności od pory dnia lub warunków atmosferycznych i charakterystyki otoczenia; oszczędność energii i zmniejszanie zanieczyszczenia światłem w godzinach nocnych przy braku ruchu; wykrywanie zdarzeń i przekazywanie do jednostek zarządzających.

W niniejszym referacie przedstawiono architekturę i funkcjonalność wielosensorowej platformy INFOLIGHT, zaprojektowanej jako uniwersalne rozwiązanie do sterowania oświetleniem z wykorzystaniem Internetu Rzeczy, przetwarzania brzegowego oraz sztucznej inteligencji.

2. ARCHITEKTURA SYSTEMU

System INFOLIGHT został zaprojektowany jako skalowalna i modularna platforma sterowania infrastrukturą oświetleniową w środowisku Smart City. Jego architektura obejmuje kluczowe elementy, opisane w kolejnych podrozdziałach:

- Urządzenia brzegowe nazwane InfoLight Edge Device ILED, zintegrowane ze słupami oświetleniowymi, wyposażone w sensory, moduły komunikacyjne i akceleratory obliczeń dla analizy danych.

- Serwer brzegowy InfoLight Fog Server ILFS, tj. lokalny węzeł przetwarzania danych, komunikacji i zarządzania.

- Warstwę chmurową, zawierającą systemy analityczne, zarządzania i integracji z miejskimi usługami cyfrowymi.

2.1. Urządzenie brzegowe ILED

Sterownik ILED pełni funkcję inteligentnego węzła Internetu Rzeczy, który: 1) steruje parametrami świecenia (natężenie, barwa, kształt strumienia), 2) zbiera dane z czujników środowiskowych, akustycznych, optycznych i ruchu, 3) przetwarza dane lokalnie z wykorzystaniem akceleratorów AI (np. Coral Edge TPU [6], 4) komunikuje

się z serwerem ILFS i chmurą przez LTE, 5G, a z urządzeniami IoT poprzez Zigbee, WiFi Mesh lub BLE.

Modularna konstrukcja ILED oparta jest na standardzie Zhaga Book 18 (SR – Sensor Ready) [16], co umożliwia łatwą wymianę i rozbudowę funkcjonalności. Opracowane moduły sensorów mogą być montowane w różnych konfiguracjach na jednej oprawie lampy wspierającej ten standard.

2.2. Serwer brzegowy ILFS

ILFS to przemysłowy komputer zlokalizowany w szafie ulicznej lub na słupie, którego zadaniem jest: 1) agregowanie danych z wielu urządzeń ILED, 2) lokalne zarządzanie i analiza, 3) funkcja bramy do warstwy chmurowej, 4) obsługa komunikacji w wybranych standardach, np. V2X z pojazdami autonomicznymi [1].

ILFS zapewnia redundancję i niskie opóźnienia dzięki architekturze Fog Computing. Wyposażony jest w moduły Jetson TX2 (dla trenowania modeli uczenia maszynowego i wnioskowania) [11] oraz Coral (dla wnioskowania), a także w systemy komunikacyjne LTE/5G, WiFi, Ethernet i V2X.

2.3. Warstwa komunikacyjna

System INFOLIGHT wspiera najważniejsze standardy komunikacyjne w tym: 1) zewnętrzne: LTE, 5G, WiFi 5 GHz, Ethernet; 2) lokalne: Zigbee, Bluetooth Low Energy, WiFi Mesh oraz 3) sterujące: DALI, DALI-2, D4i (dla zasilaczy LED) [7]. W warstwie protokołów zastosowano autorski InfoLight Fog Protocol ILFP, oparty na MQTT i JSON. Obsługuje on wybrane wzorce komunikacyjne (rozdział 4.2).

Protokół wspiera integrację z otwartymi modelami danych, aby zapewnić interoperacyjność z innymi systemami miejskimi, np. zarządzania ruchem, bezpieczeństwa, energetyki. Wykorzystywane są standardy:

- FIWARE Smart Data Models [8] – dla integracji z platformami miejskimi.

- OMA LwM2M (Lightweight M2M) [12] – dla zarządzania urządzeniami IoT.

- uCIFI Device Data Model [13] – zgodność z Zhaga D4i i systemami zarządzania oświetleniem.

3. SENSORY I FUNKCJE SYSTEMU

Tworząc system INFOLIGHT, przewidziano wielomodalność, tj. różne lampy w jednej instalacji mogą mieć różne zestawy sensorów i funkcji (rys. 1). Dzięki temu możliwe jest optymalne rozmieszczenie urządzeń o właściwych konfiguracjach w przestrzeni miejskiej, redukcja kosztów i zwiększenie efektywności systemu. Ponadto dzięki takiemu podejściu możliwe jest zbieranie danych z otoczenia, ich kontekstowa interpretacja i dostosowanie parametrów pracy oświetlenia do aktualnych warunków środowiskowych i sytuacyjnych.

3.1. Typy wykorzystywanych sensorów

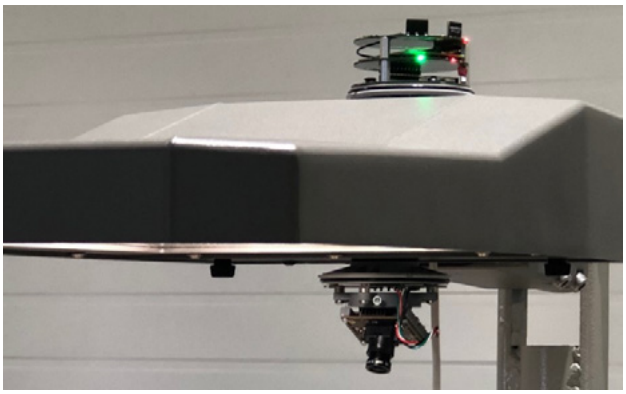
System INFOLIGHT obsługuje szeroką gamę sensorów wykonanych w formie modułów, będących unikatowym rezultatem autorskich prac badawczo-rozwojowych:

ILED_MAIN to moduł główny, zawierający, poza układami zasilania i kontrolerem, czujnik temperatury, czujnik światła otoczenia (luxometr), 3-osiowy akcelerometr (detekcja wibracji, przechyłów, kolizji) (rys. 2-3).

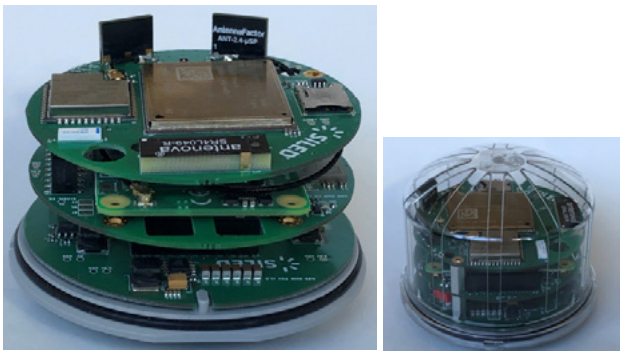
ILED_CAM jest modułem kamery wizyjnej RGB z opcją rejestracji w podczerwieni, co daje możliwość detekcji ruchu, klasyfikacji obiektów, analizy zdarzeń (np. kolizje, obecność pieszych) w różnych warunkach oświetleniowych i środowiskowych.

ILED_SOUND to moduł sondy akustycznej, korzystający z mikrofonów MEMS z interfejsem I2S, dający możliwość analizy poziomu hałasu, detekcji sygnałów alarmowych (syreny, krzyki), transmisji strumienia audio lub lokalnej ekstrakcji anonimowych cech sygnału i ich przekazywanie do jednostki zarządzającej [9].

ILED_WEATHER jest modułem środowiskowym, posiadającym czujniki temperatury, wilgotności, ciśnienia, poziomu światła widzialnego i podczerwonego.



Rys. 1. Mocowanie modułów na oprawie SILED CELEBRA. Górny – ILED_MAIN, dolny – ILED_CAM



a) *b)*

Rys. 2. a) ILED_MAIN (bez osłony), zawierający moduły (od góry): modem, kontroler, zasilanie; b) ILED_MAIN w osłonie.

Przedstawioną pulę rozszerzać można o czujniki jakości powietrza (PM2.5, PM10, NOx, SOx).

Każdy z tych modułów może być instalowany niezależnie, co pozwala na tworzenie zróżnicowanych konfiguracji w zależności od potrzeb konkretnej lokalizacji.

3.2. Analiza danych

Dane z sensorów są przetwarzane lokalnie w urządzeniu ILED. Wykorzystano modele uczenia maszynowego, m.in. YOLO do analizy obrazu z kamery w celu

wykrywania obecności pieszych i pojazdów [14], klasyfikacji typów zdarzeń (np. wypadek, wtargnięcie na jezdnię), sieci splotowe do klasyfikacji zdarzeń akustycznych [9]. Wykonano algorytmy oceny hałasu pojazdów [4] oraz precyzyjnej predykcji ruchu pojazdów na podstawie analizy danych z czujników INFOLIGHT [5]. Ponadto stosowana jest jednoczesna detekcja ruchu i hałasu, która wskazuje na możliwe zdarzenie drogowe.

Dzięki lokalnemu przetwarzaniu możliwa jest automatyzacja działania bez konieczności przesyłania dużych ilości danych do chmury. ILED wysyła wyłącznie informacje o wykrytych zdarzeniach a strumień wideo, dostępny jest wyłącznie na żądanie operatora.

3.3. Adaptacyjne sterowanie oświetleniem

Na podstawie analizy danych sensorycznych system automatycznie dostosowuje parametry pracy elementów wykonawczych lampy.

Natężenie światła regulowane jest zgodnie z obowiązującymi przepisami, dotyczącymi dostosowania do pory dnia oraz klasy drogi. Choć obecnie kategoria drogi to parametr narzucany z góry w sposób statyczny, to w działaniu opisywanego prototypowego rozwiązania informacje z czujników i z modelu oceny natężenia ruchu użyte są do dynamicznego określania klasy drogi. Uzyskiwana jest innowacyjna funkcja dostosowania poziomu i barwy światła do warunków na drodze.

Zanieczyszczenie światłem, które wpływa na komfort i samopoczucie, jest niwelowane poprzez zastosowanie dynamicznego sterowania wiązkami światła o dopasowanych temperaturach barwowych. Z jednej oprawy emitowane jest w kierunku zabudowań mieszkalnych światło ciepłe, a na drogę i trakty piesze – światło zimne, poprawiające widoczność.

Dostępny system sterowania obsługuje reguły definiowane w module zarządzającym i opisane działanie może zostać dopasowane do potrzeb.

4. ZARZĄDZANIE DANYMI I KOMUNIKACJA

W architekturze INFOLIGHT wykorzystana jest rozproszona analiza danych: przetwarzanie i podejmowanie decyzji odbywa się i w chmurze, i lokalnie, na poziomie urządzeń brzegowych (ILED) oraz serwerów pośredniczących (ILFS). Takie podejście zapewnia niskie opóźnienia działania kluczowych funkcji i odporność na awarie sieciowe. Przewidziano tryby awaryjne, które w przypadku braku komunikacji z pozostałymi elementami systemu, aktywują w lampie tryb oparty w pełni na deterministycznych regułach definiowanych w przepisach prawa.

4.1. Architektura przetwarzania danych

Zaimplementowano model Fog-to-Cloud (F2C), w którym przetwarzanie danych odbywa się na trzech poziomach: 1) Edge (ILED) zintegrowany z lampami, wyposażony w sensory i akceleratory obliczeń; 2) Fog (ILFS) to serwery brzegowe agregujące dane z wielu jednostek ILED, realizujące lokalne wnioskowanie i zarządzanie; 3) chmurowa warstwa centralna, odpowiedzialna za globalne zarządzanie, długoterminową analizę, umożliwiająca uczenie modeli, integrację z systemami miejskimi. Taka architektura umożliwia dynamiczne dzielenie obliczeń i elastyczne zarządzanie zasobami.

4.2. Protokół komunikacyjny ILFP

Wszystkie warstwy komunikacyjne są zintegrowane za pomocą autorskiego protokołu ILFP, opartego na MQTT 3.1.1 [3] i JSON Schema [2]. Protokół wspiera następujące typy wiadomości:

- TELEMETRY/EVENT – dane telemetryczne (np. odczyty sensorów),

- REQUEST – żądania podania wymaganych danych,

- COMMAND - żądanie wykonania określonych działań; polecenie jest potwierdzane (np. zmiana jasności),

- NOTIFICATION – powiadomienia systemowe o zmianie statusu.

ILFP wykorzystuje strukturę tematów MQTT zorganizowaną hierarchicznie według lokalizacji, typu urządzenia i identyfikatora.

5. WDROŻENIE LAMPY TESTOWEJ

W celu weryfikacji założeń projektowych oraz przetestowania funkcjonalności platformy INFOLIGHT w warunkach rzeczywistych, opracowano i wykonano prototyp inteligentnej lampy ulicznej. Lampa ta stanowi kompletny węzeł systemu – wyposażony w sensory, moduły komunikacyjne, układy wykonawcze (panele diodowe i sterowaną optykę) oraz interfejsy zgodne ze standardami przemysłowymi.

Lampa testowa została zbudowana w oparciu o oprawę przystosowaną do montażu modułów w standardzie Zhaga Book 18. Oprawa została wyposażona w jedno złącze Zhaga SR na górnej powierzchni (dla modułu komunikacyjnego) oraz dwa na dolnej powierzchni (dla kamery i sondy akustycznej). W prototypie zastosowano następujące moduły:

- ILED_MAIN z komputerem Raspberry Pi CM4, akceleratorem Coral Edge TPU, czujnikami środowiskowymi i interfejsem komunikacji LTE.

- ILED_LINK – moduł sterujący zasilaczami LED w standardzie DALI/D4i, z konwerterem USB/DALI i układem zasilania.

- ILED_CAM – moduł kamery Arducam B0200 (2 MPx, IR), z kompresją H.264.

- ILED_SOUND – moduł sondy akustycznej z mikrofonami MEMS i interfejsem I2S/USB.

- ILED_WEATHER – moduł czujników środowiskowych (temperatura, wilgotność, UV, VOC, IMU).

Moduły zostały połączone wewnętrzną magistralą ILED BUS, która zastępuje standard DALI w komunikacji między modułami, zachowując jednak zgodność mechaniczną i elektryczną ze złączem Zhaga.

Lampa została wyposażona w specjalnie zaprojektowany panel LED INFOLIGHT 16x4, składający się z 16 punktów świetlnych, z których każdy zawiera zestaw diod

CSP (Chip-Scale Package) o różnych temperaturach barwowych. Panel umożliwia dynamiczne sterowanie jasnością i barwą światła (np. 2700K, 4000K, 5700K), zmianę kształtu strumienia świetlnego (np. rozkład wzdłużny lub poprzeczny), testowanie różnych wariantów optycznych z wykorzystaniem soczewek STRADA. Zasilanie panelu realizowane jest przez zasilacze DALI DT8, sterowane przez moduł ILED_LINK.

Lampa testowa została podłączona do serwera ILFS przez sieć WiFi 5 GHz (802.11n) oraz alternatywnie przez modem LTE/5G. Dane z sensorów były przetwarzane lokalnie (klasyfikacja obiektów w kamerze, analiza hałasu), a następnie przesyłane do ILFS w formacie JSON przez protokół ILFP. Testowano również funkcje systemowe: zdalnej aktualizacji oprogramowania; dynamicznej konfiguracji parametrów pracy; lokalnego wnioskowania i reakcji na zdarzenia (np. zwiększenie jasności po wykryciu ruchu). Testy wykazały wysoką responsywność systemu (czas reakcji < 1 s), skuteczne działanie lokalnych algorytmów AI (detekcja pieszych, pojazdów i klasyfikacja zdarzeń akustycznych) [9][14], stabilność komunikacji w warunkach miejskich, elastyczność konfiguracji i łatwość rozbudowy funkcji.

Platformę sterownika ILED wykorzystano również do wdrożenia algorytmu detekcji awarii w systemie oświetleniowym poprzez analizę zapisów z liczników energii elektrycznej [15].

6. PODSUMOWANIE

Lampa testowa potwierdziła zasadność przyjętej architektury modularnej oraz efektywność zastosowania przetwarzania brzegowego w systemach oświetleniowych nowej generacji.

Zastosowanie lokalnego przetwarzania pozwoliło na zwiększenie niezawodności oraz na minimalizację opóźnień, eliminując konieczność stałej komunikacji z warstwą chmurową. Dopasowanie możliwości obliczeniowych urządzenia brzegowego do specyfiki pozyskiwanych danych, umożliwiło analizę strumieni audio i wideo bez konieczności ich przesyłania, co wpływa na bezpieczeństwo i zachowanie prywatności osób postronnych.

Dzięki modułowej architekturze uzyskano łatwość dostosowania oferowanych funkcji do specyfiki lokalnych warunków.

PODZIĘKOWANIA

This research was funded by the Polish National Centre for Research and Development (NCBR) through the European Regional Development Fund entitled INFOLIGHT Cloud-Based Lighting System for Smart Cities under Grant POIR.04.01.04-00-075/19.

LITERATURA

[1] A. Alalewi, I. Dayoub, S. Cherkaoui. 2021. "On 5G-V2X Use Cases and Enabling Technologies: A Comprehensive Survey". IEEE Access. 9: 107710–107737. DOI: 10.1109/ACCESS.2021.3100472

[2] A. Wright, H. Andrews, B. Hutton, G. Dennis. 2022. „JSON Schema: A Media Type for Describing JSON Documents". https://json-schema.org/draft/2020-12/json-schema-core (Accessed June 1st 2025)

[3] Andrew Banks, Rahul Gupta. 2015. „MQTT 3.1.1 Specification" http://docs.oasis-open.org/mqtt/mqtt/v3.1.1/mqtt-v3.1.1.html (Accessed June 1st 2025)

[4] Andrzej Czyżewski, Tomasz Śmiałkowski. 2021. „Selective monitoring of noise emitted by vehicles involved in road traffic". Journal of the Acoustical Society of America (149): 141. DOI: 10.1121/10.0005345

[5] Andrzej Sroczyński, Andrzej Czyżewski. 2023. „Road traffic can be predicted by machine learning equally effectively as by complex microscopic model". Scientific Reports (13): 14523. DOI: 10.1038/s41598-023-41902-y

[6] Coral Google TPU product webpage. https://coral.ai/products/ (Accessed June 1st 2025).

[7] DALI Alliance. „D4i – the DALI standard for intelligent, IoT-ready luminaires". https://www.dali-alliance.org/d4i/ (Accessed June 1st 2025).

[8] „FIWARE - A curated framework of Open Source Platform components to accelerate the development of Smart Solutions" https://www.fiware.org/about-us/ (Accessed June 1st 2025)

[9] J. Kotus, G. Szwoch. 2021. „Estimation of Average Speed of Road Vehicles by Sound Intensity Analysis". MDPI Sensors. DOI: 10.3390/s21165337

[10] K. M. Zielinska-Dabkowska, E. S. Schernhammer, J. P. Hanifin, G.C. Brainard. 2023. „Reducing nighttime light exposure in the urban environment to benefit human health and socjety". Science 380(6650): 1130-1135. DOI: 10.1126/science.adg5277

[11] NVidia. „Jetson TX2 module". https://developer.nvidia.com/embedded/jetson-tx2 (June 1st 2025)

[12] Open Mobile Alliance. „LwM2M Documentation". https://www.openmobilealliance.org/specifications/lwm2m (Accessed June 1st 2025)

[13] Open Mobile Alliance. „uCIFI and OMA: Driving Open IoT Standards". https://www.openmobilealliance.org/solutions/smart_cities (Accessed June 1st 2025)

[14] Sebastian Cygert, Andrzej Czyżewski. 2022. „Robust Object Detection with Multi-input Multi-output Faster R-CNN". Proc. Intl. Conf. Image Analysis and Processing. Springer. DOI: 10.1007/978-3-031-06427-2_48

[15] T. Śmiałkowski, A. Czyżewski. 2023. „Autoencoder application for anomaly detection in power consumption of lighting systems". IEEE Access, 124150-124162. DOI: 10.1109/ACCESS.2023.3330135

[16] Zhaga Consortium. „Book 18: Smart interface between outdoor luminaires and sensing / communication modules". https://www.zhagastandard.org/books/overview/smart-interface-between-outdoor-luminaires-and-sensing-communication-modules-18.html (Accessed June 1st 2025).

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

DOBÓR POSTACI SYGNAŁU SYNCHRONIZACYJNEGO NPSS W NB-IOT PRZY UŻYCIU ALGORYTMU GENETYCZNEGO

CHANGING THE FORM OF NPSS SYNCHRONIZATION SIGNAL IN NB-IOT USING A GENETIC ALGORITHM

Olga Błaszkiewicz; Piotr Rajchowski; Jarosław Sadowski

Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk
{olga.blaszkiewicz, piotr.rajchowski, jaroslaw.sadowski1}@pg.edu.pl

DOI: 10.15199/59.2025.4.74

Streszczenie**: W artykule zaprezentowano użycie algorytmu genetycznego na potrzeby modyfikacji postaci sygnału synchronizacyjnego NPSS (Narrowband Primary Synchronization Signal) w NB-IoT (Narrowband Internet of Things) w celu wygenerowania sygnału synchronizacyjnego o lepszych właściwościach korelacyjnych w porównaniu do jego ustandaryzowanej postaci.**

Abstract**: The article presents new form of NPSS (Narrowband Primary Synchronization Signal) in NB-IoT (Narrowband Internet of Things) which was generated based on genetic algorithm approach. Presented form of NPSS has better correlation properties in comparison to its standard version.**

Słowa kluczowe**: algorytm genetyczny, NB-IoT, NPSS, synchronizacja**

Keywords**: genetic algorithm, NB-IoT, NPSS, synchronization**

1. WSTĘP

W celu nawiązania połączenia pomiędzy terminalem użytkownika a stacją bazową niezbędne jest przeprowadzenie procedury synchronizacji częstotliwościowo-czasowej [2]. W przypadku interfejsu radiowego NB-IoT (Narrowband Internet of Things) na potrzeby synchronizacji używane są dwa typy sygnałów: NPSS (Narrowband Primary Synchronization Signal) oraz NSSS (Narrowband Secondary Synchronization Signal) [1]. Szczegółowe informacje dotyczące postaci sygnałów synchronizacyjnych, ze szczególnym uwzględnieniem sygnału synchronizacyjnego NPSS, zostały zawarte w [3].

W przytoczonym artykule wykazano, że postać sekwencji synchronizacyjnej ze standardu, dla przyjętej miary jakościowej, nie jest optymalna, ale wykonane badania nie objęły wszystkich możliwości zmian parametrów charakteryzujących sekwencję synchronizacyjną, na podstawie której generowany jest sygnał synchronizacyjny.

Niniejszy artykuł stanowi zatem kontynuację przedstawianych w tekście [3] treści i zawiera rozszerzenie dotyczące poszukiwania optymalnej postaci sygnału synchronizacyjnego NPSS w interfejsie radiowym NB-IoT dla zdefiniowanej miary jakościowej przy użyciu algorytmu genetycznego.

W związku z powyższym, artykuł został podzielony na trzy główne rozdziały: rozdział drugi prezentujący opis algorytmów genetycznych, rozdział trzeci, w którym przedstawiono kolejne kroki związane z implementacją algorytmu genetycznego na potrzeby modyfikacji ustandaryzowanej postaci sygnału synchronizacyjnego NPSS w interfejsie radiowym NB-IoT w celu uzyskania sygnału o lepszych właściwościach korelacyjnych, a także rozdział czwarty, w którym zawarto podsumowanie przeprowadzonych badań i zawarto opis przyszłych prac.

2. ALGORYTMY GENETYCZNE I ICH PODSTAWOWE ELEMENTY

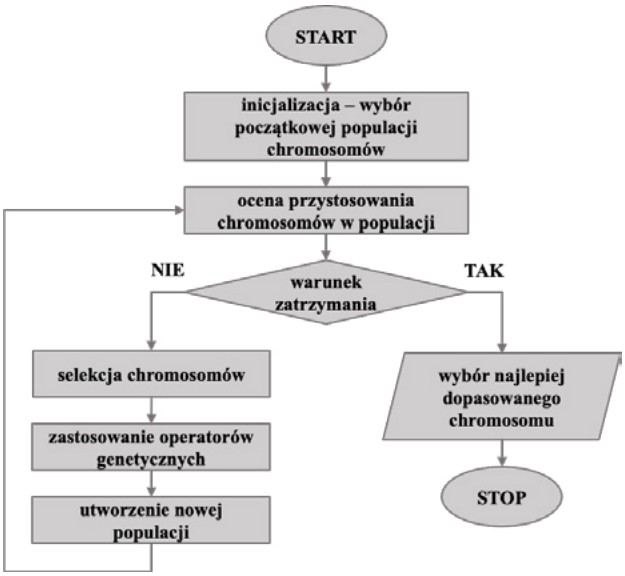
W związku z faktem, że zagadnienie poszukiwania sekwencji można potraktować jako problem wymagający rozwiązania optymalizacyjnego przyjęto, że w celu jego rozwiązania można posłużyć się algorytmami genetycznymi [5, 7].

Algorytmy genetyczne to w najprostszym ujęciu metody poszukiwania rozwiązań wzorowane na kształt ewolucji biologicznej. Zatem celem ich stosowania jest znalezienie najlepiej przystosowanych osobników w danej populacji, a używana nomenklatura jest odzwierciedleniem określeń z obszaru genetyki. Stąd w algorytmach genetycznych używa się określeń takich jak m.in. populacja, osobnik, chromosom czy gen.

Populacja stanowi zbiór osobników o określonej liczebności, z których każdy składa się z chromosomów, które z kolei stanowią uporządkowany zbiór genów. Poza podstawowymi, wymienionymi powyżej, pojęciami, w przypadku algorytmów genetycznych szczególną uwagę należy zwrócić na funkcję przystosowania, której wynik stanowi o warunku zatrzymania procesu poszukiwania optymalnego rozwiązania. Funkcja ta pozwala na ocenę przystosowania danego osobnika w analizowanej populacji w aktualnie analizowanej iteracji [7, 8]. Ogólny schemat blokowy działania algorytmu genetycznego został przedstawiony na rys. 1.

Działanie algorytmu genetycznego rozpoczyna się od kroku inicjalizacji, przez który należy rozumieć zdefiniowanie początkowej populacji chromosomów. Krok ten realizowany jest na drodze losowania zadanej liczby osobników czy też ich ręcznego wyboru [5, 7, 8]. Następnie dokonywana jest ocena przystosowania, określonej

w pierwszym kroku grupy chromosomów. W tym celu wyliczana jest wartość zdefiniowanej funkcji przystosowania dla każdego chromosomu z populacji.



Rys. 1. Schemat blokowy algorytmu genetycznego [8]

W zależności od przyjętej wartości funkcji przystosowania (np. najlepiej dopasowany chromosom, to chromosom, dla którego wyliczona wartość funkcji przystosowania osiąga maksimum) algorytm może zakończyć swoje działanie lub też kontynuować je z uwzględnieniem konieczności dokonania modyfikacji w analizowanej populacji.

Mając na uwadze powyższe, na podstawie uzyskanych wartości funkcji przystosowania dokonuje się selekcji chromosomów, następnie stosuje się na nich operatory genetyczne i na tej podstawie tworzy nową populację osobników, którą ponownie poddaje się ocenie przystosowania.

Przez selekcję chromosomów rozumie się wybór takich chromosomów, które charakteryzują się najlepszym przystosowaniem w analizowanej iteracji. Wybór ten podyktowany jest chęcią, zgodnie z zasadami selekcji naturalnej, zmaksymalizowania wyodrębnienia najlepszego, w ujęciu globalnym, osobnika. Kolejnym krokiem jest zastosowanie operatorów genetycznych, czyli operatora krzyżowania oraz operatora mutacji.

O ile mutacja to przypadkowa zmiana genu w obrębie chromosomu, o tyle krzyżowanie jest operatorem, dzięki któremu dochodzi do znacznych zmian postaci chromosomów. Pierwszym krokiem operacji krzyżowania jest wybór osobników, które będą ze sobą krzyżowane, następnie ustalany jest punkt krzyżowania, czyli miejsce w chromosomie, w którym następuje zamiana genów. W najprostszym ujęciu dochodzi do krzyżowania jednopunktowego, co oznacza, że chromosom potomka stanowi kombinację genów na pozycjach od pierwszej do punktu krzyżowania od jednego z rodziców i od punktu krzyżowania do ostatniego genu od drugiego z rodziców i na odwrót dla drugiego z potomków. Dzięki operacji krzyżowania uzyskuje się zatem parę potomków łączących cechy swoich poprzedników. Utworzona nowa populacja ponownie jest sprawdzana pod kątem przystosowania [7, 8].

Proces ten powtarzany jest, w zależności od przyjętego kryterium, skończoną liczbę razy albo tak długo, aż nakład obliczeniowy będzie niewspółmiernie większy względem uzyskiwanego zysku (np. nieznaczna zmiana wartości maksymalnej funkcji przystosowania).

3. PROPOZYCJA MODYFIKACJI POSTACI SYGNAŁU SYNCHRONIZACYJNEGO NPSS

Zgodnie z przytoczonym w rozdziale drugim twierdzeniem, poszukiwanie innej postaci sekwencji synchronizacyjnej można rozpatrywać w kategorii problemu optymalizacyjnego. W przypadku analizowania postaci sekwencji synchronizacyjnej (na podstawie której generowany jest sygnał synchronizacyjny) optymalizacji podlega zatem wartość funkcji przystosowania, która jest bezpośrednio powiązana z przeznaczeniem tej sekwencji.

Na potrzeby wygenerowania sygnału synchronizacyjnego typu NPSS korzysta się z zależności opisanej w [1]. Na jej podstawie generowanych jest 11 symboli Zadoff-Chu, w dziedzinie częstotliwości, które powielane są w dziedzinie czasu (zależność (1)).

$$d_{l_{NB}}(n_{NB}) = S(l_{NB}) \cdot e^{-j\frac{\pi u_{NB} n_{NB}(n_{NB}+1)}{11}}, \qquad (1)$$

gdzie:

$n_{NB} = 0, 1, \ldots, 10,$

$u_{NB} = 5,$

$l_{NB} = 0, 1, \ldots, 10,$

$S(l_{NB}) = [1, 1, 1, 1, -1, -1, 1, 1, 1, -1, 1].$

W analizowanym przypadku rozważany jest zatem sygnał synchronizacyjny, którego położenie w sygnale odebranym jest dookreślane na podstawie uzyskiwanej wartości funkcji korelacji. W zależności od przyjętej metody synchronizacji odbiornika może to być korelacja skrośna sygnału odebranego z lokalnie wygenerowanym w odbiorniku sygnałem wzorcowym czy też autokorelacja odebranego sygnału. Niezależnie jednak od zastosowanego typu korelacji, moment wystąpienia sygnału synchronizacyjnego może być określony na podstawie detekcji wartości maksymalnej funkcji korelacji czy też detekcji prążka znajdującego się powyżej przyjętego progu decyzyjnego. Stąd przyjęta, na potrzeby implementacji algorytmu genetycznego, funkcja przystosowania związana jest z minimalizacją wartości stosunku procentowego (SP) prążka o maksymalnej wartości lokalnej funkcji korelacji sygnału do prążka o maksymalnej wartości globalnej tej funkcji. Maksimum lokalne w przytoczonej mierze należy rozumieć jako maksymalną wartość funkcji korelacji, w której wyzerowano wartość prążka o maksymalnej wartości globalnej.

Implementacja algorytmu genetycznego na potrzeby poszukiwania najlepszego rozwiązania dla zdefiniowanej miary jakościowej została przeprowadzana zgodnie z krokami wyszczególnionymi na schemacie zaprezentowanym na rys. 1.

Proces poszukiwania najlepszego rozwiązania rozpoczęto od inicjalizacji początkowej populacji, której chromosomy zastępowały kolejno dotychczasowy składnik funkcji wykładniczej zaprezentowanej w zależności (1). Na podstawie wcześniejszych analiz [3] przyjęto także założenie, że binarna sekwencja zostanie zastąpiona 11-elementowym ciągiem Barkera.

Zmodyfikowana postać zależności (1) z uwzględnieniem wprowadzonych zmian została ujęta w zależności (2).

$$dAG_{l_{NB}}(n_{NB}) = S_{Barker}(l_{NB}) \cdot e^{-jg_{NB}}, \qquad (2)$$

gdzie:

g_{NB} = gen,

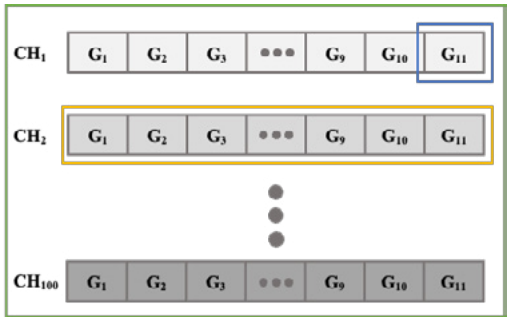
$n_{NB} = 0, 1, \ldots, 10,$

$l_{NB} = 0, 1, \ldots, 10,$

$S_{Barker}(l_{NB}) = [1, 1, 1, -1, -1, -1, 1, -1, -1, 1, -1].$

Na początku wygenerowano zatem losowo 100 chromosomów zawierających po 11 genów, a każdemu z nich losowo przyporządkowano wartość z przedziału od 0 do 2π.

Graficznie ujęto to na rys. 2, gdzie w niebieskiej ramce zaprezentowano jeden z genów, w żółtej ramce jeden chromosom składający się z 11 genów oraz w ramce zielonej całą wygenerowaną populację. W analizowanym przypadku osobnik zawiera jeden chromosom.



Rys. 2. Wybrane elementy algorytmu genetycznego

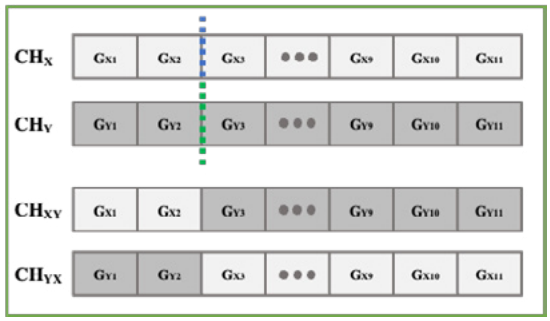
W kolejnym kroku dla każdego z chromosomów wygenerowano sygnał NPSS oraz na jego podstawie określono wartość odpowiadającej mu miary SP. Następnie osobniki zostały posortowane rosnąco zgodnie z wartościami przyjętej miary jakościowej (im mniejsza wartość, tym lepszy osobnik) i podzielone na cztery grupy (grupa najlepszych osobników, dwie grupy gorszych osobników i grupa najgorszych osobników w danej iteracji). Przy czym, do każdej kolejnej iteracji trafiała: niezmieniona grupa najlepszych osobników, grupa najlepszych osobników poddana mutacji, grupa osobników poddana krzyżowaniu jednopunktowemu i zupełnie nowa grupa losowo wybieranych osobników, która zastępowała odrzuconą grupę najgorszych osobników.

Grupa najlepszych osobników została poddana operacji mutacji, przez którą w analizowanym przykładzie rozumie się zmianę wartości aktualnego genu o wartość z zakresu od -π/10 do π/10.

Osobniki z grupy pierwszej były krzyżowane jednopunktowo z osobnikami z grupy drugiej lub trzeciej. W przypadku tego operatora genetycznego losowane były następujące elementy:

- miejsce cięcia jednopunktowego,
- numer osobnika ze zbioru najlepszych osobników,
- numer grupy, spośród której ma być drugi osobnik (druga lub trzecia grupa),
- numer osobnika z wylosowanej grupy,
- kolejność genów.

W celu zobrazowania przeprowadzonej operacji krzyżowania jednopunktowego jego schematyczne ujęcie przedstawiono na rys. 3.



Rys. 3. Schematyczne ujęcie zastosowania operacji krzyżowania jednopunktowego

Dodatkowo, osobniki po operacji krzyżowania poddawane były również operacji mutacji zgodnie z opisanym powyżej sposobem jej wprowadzenia.

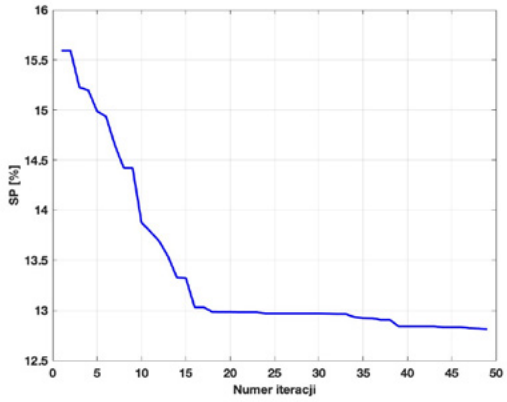
Dotychczas przeprowadzane badania, których numeryczna prezentacja została zaprezentowana w artykule [3] wykazały, że wartość przyjętej miary jakościowej SP zmienia się w zależności od modyfikowanej wartości parametru charakteryzującego sekwencję synchronizacyjną. Stąd też oddzielna modyfikacja parametrów dawała różny zakres zmian tego parametru.

Zakres zmian uzyskiwanej wartości SP w zależności od zmienionej wartości parametru charakteryzującego sekwencję przedstawia się następująco:

- zmiana wartości root index wpłynęła na uzyskiwanie wartości SP w zakresie od 18,56% do 32,07%,
- zmiana wartości dodajnika wpłynęła na uzyskiwanie wartości SP w zakresie od 28,38% do 42,91%.

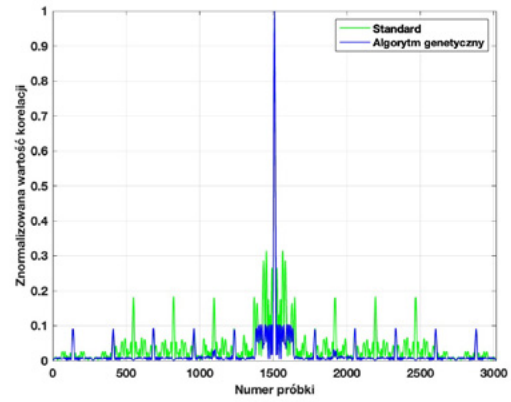
Zaobserwowano również zmiany wartości miary SP przy zmianie sekwencji binarnej, ale z związku z faktem, że 11-elementowy ciąg Barkera charakteryzuje się najlepszymi właściwościami korelacyjnymi [4, 6] spośród innych sekwencji tej długości, a w aktualnej postaci algorytmu genetycznego postać tej sekwencji nie ulega zmianie, to określonej wartości miary SP nie przyjęto za wartość referencyjną, do której porównywano wyniki uzyskane na podstawie algorytmu genetycznego.

Na potrzeby niniejszych badań przeprowadzono 1000 symulacji, z których każda była ograniczona warunkiem zatrzymania równym 50 iteracjom. Przykładowy zakres zmian parametru SP dla jednej z symulacji zaprezentowano na rys. 4.



Rys. 4. Przykładowy wynik działania skryptu

Przy użyciu zaimplementowanego algorytmu genetycznego uzyskano wartość parametru SP równą 10,37%. Uzyskaną funkcję korelacji dla sekwencji ze standardu [1] oraz uzyskaną na podstawie przeprowadzonych badań zobrazowano na rys. 5.



Rys. 5. Funkcja korelacji dla dwóch sekwencji synchronizacyjnych

Zbiór poszczególnych symboli, na podstawie których został wygenerowany sygnał synchronizacyjnych zaprezentowano w tab. 1.

Tab. 1. Wartości kolejnych symboli sekwencji

Numer porządkowy	*Wartość*
symbol 1	0,3436 – 0,9391i
symbol 2	0,3977 – 0,9175i
symbol 3	0,7992 + 0,6010i
symbol 4	0,8755 – 0,4831i
symbol 5	0,4565 – 0,8897i
symbol 6	0,2343 – 0,9722i
symbol 7	-0,3818 – 0,9242i
symbol 8	-0,5411 – 0,8409i
symbol 9	-0,9814 – 0,1920i
symbol 10	-0,8405 + 0,5419i
symbol 11	0,1337 – 0,9910i

4. PODSUMOWANIE I PRZYSZŁE PRACE

W niniejszym artykule poruszono zagadnienia związane ze zmianą postaci sygnału typu PSS używanego w standardzie NB-IoT na potrzeby przeprowadzenia procesu synchronizacji. W tym celu posłużono się algorytmem genetycznym, na podstawie którego generowano sygnały synchronizacyjne poddawane analizie porównawczej z ustandaryzowaną postacią sygnału NPSS. Za miarę jakości uzyskiwanej sekwencji przyjęto wartość współczynnika SP (im mniejsza wartość, tym lepiej).

Analiza uzyskanego wyniku pozwala stwierdzić, że użycie algorytmu genetycznego na potrzeby rozwiązania problemu optymalizacyjnego dla zadanej miary jest podejściem pozwalającym na osiągnięcie pożądanych wyników.

Wygenerowana przy użyciu tego algorytmu sekwencja, o potencjalnie lepszych właściwościach korelacyjnych, wymaga jednak programowej weryfikacji oznaczającej konieczność przeprowadzenia porównania wartości błędów czasu detekcji sygnału synchronizacyjnego NPSS w interfejsie NB-IoT z uwzględnieniem wpływu kanału radiowego. Poza wartością błędu czasu detekcji tego sygnału ważny jest również czas niezbędny do przeprowadzenia całego procesu synchronizacji dla metod znanych z literatury przedmiotu. Zatem element ten również wymaga symulacyjnego sprawdzenia a tym samym stanowi zapowiedź przyszłych prac.

LITERATURA

[1] 3GPP TS 36.211 version 14.14.0 Release 14. 2020
[2] Błaszkiewicz Olga, Jarosław Magiera, Alicja Olejniczak, Piotr Rajchowski, Jarosław Sadowski, Jacek Stefański, Krzysztof Cwalina. 2024. Framework interfejsu radiowego NB-IoT. Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne
[3] Błaszkiewicz Olga, Jarosław Sadowski, Piotr Rajchowski. 2024. Analiza postaci sygnału synchronizacyjnego NPSS w NB-IoT. Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne
[4] Holubnychyi Alexei. 2013. Generalized Binary Barker Sequences and Their Application to Radar Technology. 2013 Signal Processing Symposium (SPS)
[5] Kramer Oliver. 2017. Genetic Algorithm Essentials. Springer International Publishing AG
[6] Matsuyuki Shota, Akio Tsuneda. 2018. A Study on Aperiodic Auto-Correlation Properties of Concatenated Codes by Barker Sequences and NFSR Sequences. 2018 International Conference on Information and Communication Technology Convergence (ICTC)
[7] Michalewicz Zbigniew. 1992. Genetic Algorithms + Data Structures = Evolution Programs. Springer Berlin, Heidelberg
[8] Rutkowska Danuta, Maciej Piliński, Leszek Rutkowski. 1997. Sieci neuronowe, algorytmy genetyczne i systemy rozmyte. Wydawnictwo Naukowe PWN Sp. z o.o. Warszawa

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

KONCEPCJA STANOWISKA POMIAROWEGO DO BADANIA MOŻLIWOŚCI LOKALIZACJI EMITERÓW EM W ŚRODOWISKU ZURBANIZOWANYM
THE CONCEPT OF MEASUREMENT SETUP FOR TESTING THE ABILITY TO LOCATE EM EMITTERS IN AN URBAN ENVIRONMENT

Jaromir Gajewski[1]; Jerzy Łopatka[2];

[1] Wojskowa Akademia Techniczna, Warszawa, jaromir.gajewski@wat.edu.pl
[2] Wojskowa Akademia Techniczna, Warszawa, jerzy.lopatka@wat.edu.pl

DOI: 10.15199/59.2025.4.75

Streszczenie: **Niniejszy artykuł opisuje koncepcję stanowiska pomiarowego do badania zdolności lokalizacji urządzeń promieniujących falę EM w środowisku zurbanizowanym. Porusza on problematykę określania położenia w środowisku, w którym w wyniku przeszkód terenowych kierunek nadejścia fali EM nie jest tożsamy z faktycznym azymutem pracującego nadajnika. Opisane jest w nim rozwiązanie niniejszego problemu poprzez zastosowanie technik wstecznego śledzenia promieni oraz uczenia maszynowego.**

Abstract: **The following article presents the concept of developing a measurement setup for evaluating the ability to locate EM emitters in an urbanized environment. It raises the issue of localization in an environment where, due to terrain obstacles, the direction of arrival of the EM wave is different from actual direction to the working emitter. It describes the usage of reverse raytracing and machine learning to solve this issue.**

Słowa kluczowe: **DOA, lokalizacja, MUSIC, uczenie maszynowe, wsteczne śledzenie promieni**

Keywords: **DOA, localization, machine learning, MUSIC, reverse raytracing.**

1. WSTĘP

1.1. Teoria Namierzania

Lokalizacja źródeł emisji radiowej stanowi fundamentalny element w wielu obszarach wiedzy, takich jak telekomunikacja, systemy nawigacyjne czy szeroko rozumiane bezpieczeństwo publiczne [1]. W środowiskach silnie zurbanizowanych, gdzie występuje wysoka gęstość zabudowy oraz liczne przeszkody terenowe, propagacja fal radiowych ulega znacznym zakłóceniom, co istotnie utrudnia precyzyjną lokalizację urządzeń nadawczych. Konwencjonalne metody pozycjonowania, oparte na pomiarach odległości bądź kąta nadejścia sygnału, często wykazują ograniczoną dokładność w wyniku występowania zjawisk wielodrogowości i odbić. Problem ten nabiera szczególnego znaczenia w kontekście działań militarnych [2], gdzie rozpoznanie radioelektroniczne stanowi kluczowy komponent wczesnego etapu działań bojowych, dostarczając informacji o dyslokacji jednostek przeciwnika.

W związku z intensyfikacją operacji wojskowych prowadzonych w środowiskach zurbanizowanych, zaistniała potrzeba dostosowania systemów detekcji i lokalizacji do specyfiki współczesnych warunków pola walki. Wysoka precyzja pozycjonowania źródeł emisji w takich warunkach może znacząco wpłynąć na skuteczność realizowanych operacji oraz podnieść poziom bezpieczeństwa sił własnych. Jednym z perspektywicznych kierunków badań, umożliwiających skuteczną kompensację niekorzystnych efektów propagacyjnych, jest implementacja zaawansowanych algorytmów przetwarzania sygnałów wspomaganych technikami sztucznej inteligencji.

1.2. Algorytm MUSIC

W tym kontekście algorytm MUSIC (Multiple Signal Classification) stanowi zaawansowaną metodę estymacji parametrów, umożliwiającą jednoczesne wyznaczanie kierunków nadejścia wielu sygnałów nawet w warunkach znacznej wielodrogowości oraz braku bezpośredniej ścieżki propagacji (Line of Sight, LOS).

Podstawą działania algorytmu MUSIC jest analiza przestrzenna sygnałów odbieranych przez układ antenowy z wykorzystaniem macierzy korelacji, która podlega dekompozycji na podprzestrzeń sygnałową i szumową [3]. Dzięki temu możliwe jest precyzyjne oszacowanie zarówno azymutów nadejścia sygnałów (Direction of Arrival, DoA), jak i ich amplitud, co w dalszej kolejności pozwala na określenie lokalizacji źródeł emisji.

Z punktu widzenia formalnego, algorytm MUSIC bazuje na dekompozycji macierzy korelacji sygnału za pomocą analizy wartości osobliwych (Singular Value Decomposition, SVD), w wyniku której przestrzeń obserwacji rozdzielana jest na dwie ortogonalne składowe: podprzestrzeń sygnałową oraz podprzestrzeń szumową. Proces ten można opisać równaniem:

$$\boldsymbol{R} = \boldsymbol{A} \wedge \boldsymbol{A^H} + \boldsymbol{\sigma I} \qquad (1)$$

Gdzie R to macierz autokorelacji sygnału, A to macierz wektorów własnych podprzestrzeni sygnałowej, Λ to macierz wartości własnych związanych z mocą poszczególnych składowych sygnału, σ to moc szumu, a I to macierz jednostkowa.

Ponieważ algorytm MUSIC jest w stanie wskazać wiele DOA odpowiadających różnym źródłom promieniowania działających na tej samej częstotliwości autorzy niniejszego tekstu wysnuli hipotezę, że z punktu widzenia algorytmu nie ma różnicy czy sygnały pochodzą z różnych źródeł, czy są ta sygnały odbite na swojej drodze pochodzące z jednego źródła. Algorytm MUSIC może zatem zostać wykorzystany do estymacji kierunków nadejścia sygnałów odbitych przez przeszkody w terenie zurbanizowanym. [5]

1.3. Wsteczne śledzenie promieni

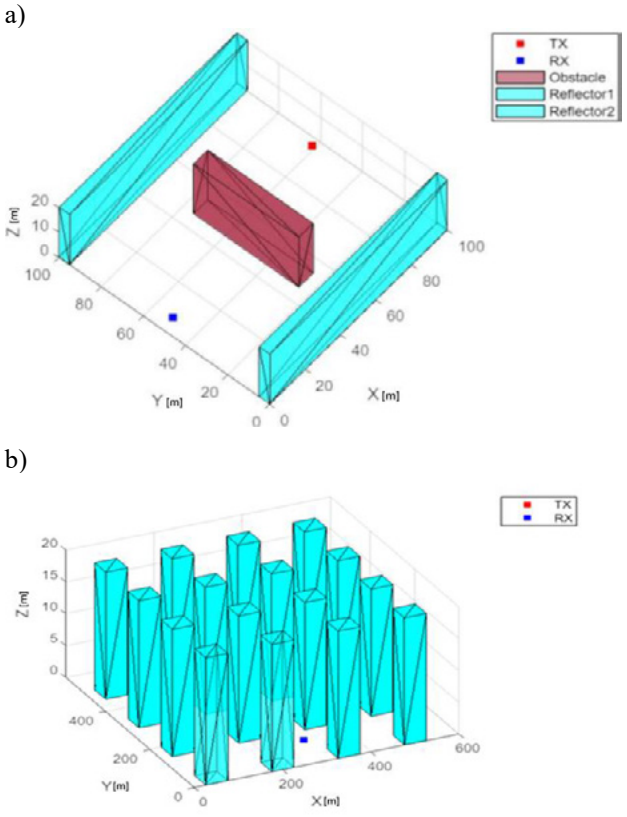
Informacja o azymutach poszczególnych ścieżek propagacyjnych może zostać wykorzystana do estymacji położenia źródła emisji z zastosowaniem metody wstecznego śledzenia promieni (ang. backward raytracing, BR), która polega na rekonstrukcji trajektorii odbitych fal radiowych. W podejściu tym symuluje się propagację fal w kierunku odwrotnym — od odbiornika w stronę nadajnika — uwzględniając interakcje fal elektromagnetycznych z przeszkodami terenowymi, takimi jak zabudowania, elementy infrastruktury czy ukształtowanie terenu (np. wzniesienia, góry). Dzięki temu możliwe jest odtworzenie rzeczywistego przebiegu torów propagacyjnych i precyzyjna lokalizacja źródła sygnału.

Pod względem formalnym metoda BR opiera się na odwzorowaniu trajektorii odbitych fal w przestrzeni trójwymiarowej i analizie ich punktów przecięcia. W procesie tym uwzględnia się zarówno kąty nadejścia sygnałów (DoA) wyznaczone wcześniej za pomocą algorytmu MUSIC, jak i topograficzne oraz strukturalne właściwości środowiska propagacyjnego.

Analiza przecięć odwzorowanych promieni umożliwia oszacowanie najbardziej prawdopodobnego położenia nadajnika, gdyż w środowiskach zurbanizowanych, ze względu na obecność licznych odbić i rozproszeń, ścieżki propagacyjne przyjmują złożone, nieliniowe trajektorie, co skutkuje wielokrotnym krzyżowaniem się torów fal w różnych punktach przestrzeni.

1.4. Metodologia symulacji.

W środowisku Matlab przeprowadzono dwie symulacje. Pierwszą w prostszym środowisku, gdzie łatwo można było z góry przewidzieć trajektorie sygnału od nadajnika do odbiornika i drugą w środowisku typu Manhattan.

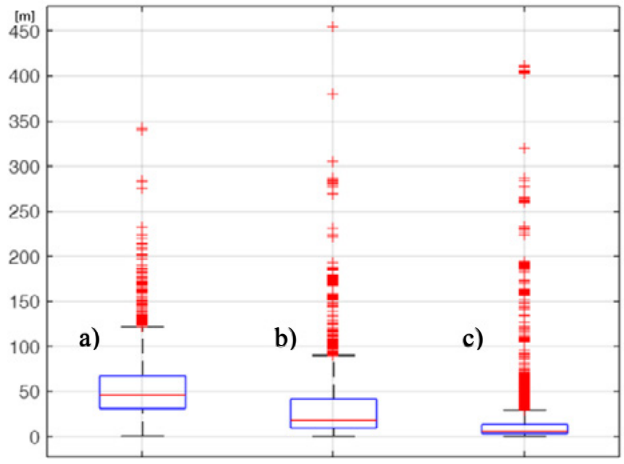


Rys. 1 - Scenariusz symulacji a) pierwszej b) drugiej

W pierwszym scenariuszu do lokalizacji źródła emisji wystarczające było uśrednienie wartości punktów przecięcia trajektorii sygnałów na ich drodze, ponieważ środowisko było proste, a większość punktów znajdowała się w obrębie lokalizacji nadajnika. Punkty przecięcia poza głównym skupiskiem można było odfiltrować algorytmem opartym na medianie wartości odległości punktów od siebie.

W drugim scenariuszu środowisko było już bardziej skomplikowane w związku z czym do odfiltrowania punktów przecięcia znajdujących się daleko od lokalizacji nadajnika wykorzystano sieci neuronowe, które na podstawie danych o namierzonych DOA określały sektor, w którym z określonym prawdopodobieństwem znajdowało się źródło sygnału. Punkty spoza wyznaczonego sektora nie były brane pod uwagę w trakcie uśredniania [7].

Oba scenariusze dały zadowalające wyniki z niskim poziomem błędu, choć w niektórych przypadkach w drugim scenariuszu błędy bywały nieakceptowalne. W związku z powyższym w trakcie opracowywania koncepcji stanowiska pomiarowego autorzy wzorowali się na pierwszym scenariuszu symulacyjnym.



Rys. 2 – Przykładowe wyniki dla drugiego scenariusza[8] a)bez NN, b)z NN dzielącą obszar na 6 sektorów c) z NN dzielącą obszar na 51 sektorów

2. KONCEPCJA STANOWISKA POMIAROWEGO

2.1. Wybór miejsca

Jako miejsce do przeprowadzenia pierwszych praktycznych pomiarów wybrano obszar Warszawy znajdujący się w bezpośredniej bliskości Wojskowej Akademii Technicznej, gdyż umożliwia ono odwzorowanie warunków z pierwszej symulacji w uproszczonym środowisku, gdzie z góry możemy przewidzieć trasę ścieżek sygnału i w łatwy z programistycznego punktu widzenia sposób wyliczyć je przy pomocy wstecznego śledzenia promieni.

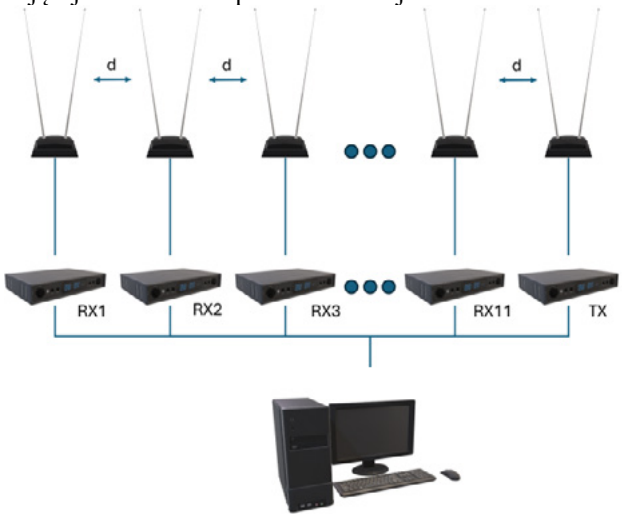
Rys. 3 – Wybrany teren umiejscowienia stanowiska

Taki dobór stanowiska pomiarowego umożliwia wyeliminowanie na wstępnym etapie badań błędów wynikających ze złożoności obliczeniowej wstecznego śledzenia promieni co umożliwia właściwą ocenę samej metody lokalizacyjnej.

2.2. Schemat stanowiska

Pierwszym elementem stanowiska pomiarowego jest źródło sygnału w postaci USRP X310 zaprogramowanego do pracy na stałej częstotliwości i emitowania jedynie fali nośnej sygnału. Umożliwia to uniknięcie uznania za błąd metody sytuacji, w której rodzaj modulacji sygnału mógłby wpłynąć na jakość namierzania. Za częstotliwość sygnału wybrano 868MHz, co jest odstępstwem od warunków symulacyjnych, lecz umożliwia znaczące zmniejszenie rozmiaru szyku anten odbiorczych i umieszczenie go na pojeździe.

Kolejnym elementem jest odbiornik w postaci liniowego szyku antenowego umieszczonego na pojeździe składającego się z 12 anten, gdzie odległość między antenami wynosi $d = \lambda / 2$. Każda z anten podłączona jest do USRP identycznego do nadajnika gdzie 11 z nich zaprogramowane jest na odbiór sygnału na zadanej częstotliwości, a 12 ustawiony jest na nadawanie i okresowo emituje sygnał kalibrujący umożliwiający pozostałym USRP uzyskanie stanu synchronizacji, koniecznego do poprawnej interpretacji wyników w części pracy systemu odpowiadającej za określenie położenia nadajnika.



Rys. 4 - Schemat odbiornika

Nadajnik i odbiornik umieszczone są na tej samej wysokości co umożliwia ograniczenie przestrzeni analizy wstecznego śledzenia promieni do jednej płaszczyzny, a położenie nadajnika i odbiornika jest dobrane w sposób uniemożliwiający drogę sygnału w ścieżce LOS aby wyeliminować możliwość maskowania sygnałów odbitych poprzez sygnał docierający do odbiornika ścieżką bezpośrednią [6].

Rys. 5 - Schemat stanowiska pomiarowego

2.3. Metodologia pomiarowa

Nadajnik zostaje umieszczony na miejscu wskazanym na schemacie i rozpoczyna nadawanie, a odbiornik przechodzi w stan odbioru.

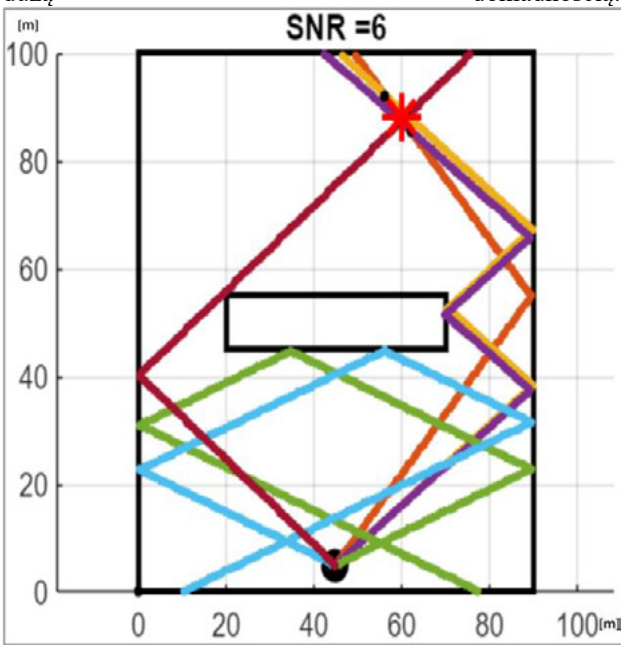
Odebrany sygnał zapisywany jest w macierzy, gdzie każda jej kolumna odpowiada próbkom sygnału odebranym przez pojedynczą antenę, a każdy wiersz oznacza kolejną próbkę czasową. Z powodu takiego zapisu próbek konieczna jest opisana wcześniej synchronizacja USRP, gdzie dodatkowe USRP emituje sygnał synchronizacji i kalibruje USRP ustawione na odbiór tak aby kolejne próbki pobierane były przez nie w tym samym czasie.

Pomiary powtarzane są dla kilku położeń nadajnika, a następnie program wykonany w środowisku Matlab poddaje je działaniu algorytmu MUSIC w celu ustalenia kierunku nadejścia poszczególnych ścieżek sygnału.

W dalszej części działania programu na podstawie określonych kierunków nadejścia sygnału wykonywany jest BR, który odwzorowuje trajektorie, po których poruszał się sygnał.

Dzięki odwzorowaniu niniejszych trajektorii możliwe jest wyznaczenie punktów przecięcia się sygnału na jego drodze oraz poddanie ich algorytmowi odfiltrowania punktów odległych co umożliwia niebranie pod uwagę w dalszych obliczeniach punktów znajdujących się poza głównym ich skupiskiem będącym w okolicy faktycznej lokalizacji nadajnika.

W ostatnim kroku wartość tych punktów jest uśredniania co (sądząc po wykonanej wcześniej symulacji) w takim środowisku umożliwia określenie lokalizacji nadajnika z dużą dokładnością.



Rys. 6 – Wynik z pierwszego scenariusza reprezentujący ideę stanowiska

3. WNIOSKI

Opisana w niniejszym artykule koncepcja wykonania stanowiska pomiarowego umożliwia skuteczną ocenę praktycznego aspektu opracowanej metody określania położenia w terenie zurbanizowanym przy pomocy pojedynczego urządzenia odbiorczego. Poprzez odpowiednie dobranie miejsca prowadzenia pomiarów oraz opracowaną metodologię możliwe stało się wyeliminowanie błędów możliwych do popełnienia w innych aspektach prowadzenia badania, które w wyniku ich zaistnienia mogłyby mylnie wpłynąć na ocenę samej metody. Przeprowadzenie pierwszych badań praktycznych konieczne jest do dalszego prowadzenia prac nad opracowanym algorytmem określania położenia emitera EM poprzez ocenę poprawności wykonania wcześniejszych symulacji oraz ocenę praktycznego aspektu zastosowania niniejszej metody. Jeżeli okaże się ona skuteczna możliwym stanie się nie tylko skuteczne namierzanie w terenie zurbanizowanym, gdzie DOA najsilniejszej ścieżki sygnału nie definiuje jednoznacznie azymutu na pracujące źródło emisji, ale również namierzanie z wykorzystaniem zaledwie jednego urządzenia przy wykorzystaniu wiedzy o umiejscowieniu przeszkód terenowych. KIERUNEK DALSZYCH BADAŃ

Dalsze badania w kierunku lokalizacji emiterów EM w terenie zurbanizowanym będą skupiać się na fizycznym wykonaniu stanowiska pomiarowego oraz praktycznych pomiarach i określeniu lokalizacji na ich podstawie.

W zależności od uzyskanych wyników należy przenieść stanowisko pomiarowe w bardziej złożone środowisko, jeżeli wskazania metody okażą się poprawne. W przeciwnym wypadku należałoby rozważyć wykorzystanie większej ilości urządzeń pomiarowych co umożliwiłoby uzyskanie większej liczby ścieżek sygnału i prawdopodobnie znacząco wpłynęło na wzrost dokładności metody.

LITERATURA

[1] N. Tayem, A. Hussain, V. Verameredy, A. Soliman, J. Alghazo, “Propagator Rooting Method Direction of Arrival Estimation Based on Real Data”, MILCOM 2021-2022 IEEE Military Communications Conference (MILCOM)

[2] A. Duarte, J Apolonairo, J Santos, “An efficient single receiver high resolution DOA estimation algoritm for modulated signals”, 2015 SBNO/IEEE MTT-S International Microwave Optoelectronics Conference (IMOC)

[3] M.M.Gunjal and A.A.B.Raj, “Improved Direction of Arrival Estimation Using Modified Music Algorithm”, 2020 5 the International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India 2020

[4] M Hawa, M Malek, M Sunar, „MUSIC Algorithm for Direction-of-Arrival Estimation: A comprehensive Review”, Sensors 2016

[5] Q Liu, H Bertoni, “MUSIC Algorithm and Its Application to Electromagnetic Wave Propagation in Urban Environments,” International Journal of Antennas and Propagation, 2012

[6] Jaromir Gajewski, Jerzy Łopatka, „Wyznaczenie wartości mocy sygnału maskującego sieć przed namierzeniem przy wykorzystaniu algorytmu MUSIC”, Przegląd Telekomunikacyjny 4/2023 zeszyt str. 369

[7] Jaromir Gajewski, Jerzy Łopatka, „Efektywna metoda lokalizacji urządzeń nadawczych w środowisku z propagacją wielodrożną”, Przegląd Telekomunikacyjny 4/2024 zeszyt str. 365

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

ZAKRES STOSOWALNOŚCI OPRACOWANEGO DETEKTORA SYGNATUR RADIOWYCH BSP

THE SCOPE OF APPLICABILITY OF THE DEVELOPED UAV RADIO SIGNATURE DETECTOR

Krzysztof K. Cwalina[1]; Piotr Rajchowski[1]; Jarosław Sadowski[1]; Paweł Skokowski[2]; Krzysztof Malon[2]; Michał Kryk[2]; Krzysztof Maślanka[2]; Jan M. Kelner[2]

[1] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk,
{kkcwalina, piorajch, jarsadow1}@pg.edu.pl
[2] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa
{pawel.skokowski, krzysztof.malon, michal.kryk, krzysztof.maslanka, jan.kelner}@wat.edu.pl

DOI: 10.15199/59.2025.4.76

Streszczenie: W niniejszym artykule przedstawiono badania i analizę zakresu stosowalności opracowanego detektora sygnatur radiowych. Przedstawiono opracowany detektor oparty na analizie charakterystyk cyklostacjonarności wraz z parametrami użytkowymi. Na podstawie przeprowadzonych badań pomiarowych wybranych BSP klasy mikro i mini wyznaczono wartości wyjściowe miękkodecyzyjnego detektora sygnatur radiowych w różnych warunkach operacyjnych.

Abstract: In the article research and analysis of the scope of applicability of the developed radio signature detector is presented. The developed detector based on the analysis of cyclostationarity characteristics together with operational parameters is described and based on the measurement tests of selected micro and mini UAVs, the output values of the soft-decision radio signature detector were determined in various operating conditions.

Słowa kluczowe: bezzałogowy statek powietrzny, cyklostacjonarność, detekcja, sygnatura radiowa

Keywords: cyclostationarity, detection, radio signature, unmanned aerial vehicle

1. WSTĘP

Niewątpliwie, ze względu na występujące różne urządzenia nadawcze, warunki propagacji, pasma częstotliwości, techniki formowania i odbioru sygnału, zastosowania i przypadki użycia, zarówno wojskowe, jak i cywilne, determinują realizację ciągłych badań i rozwój szerokopojętych technik wykrywania, identyfikacji i klasyfikacji emisji radiowych [1, 2, 3]. We wskazanych grupach urządzeń nadawczych i badań nad nimi można wskazać detekcję sygnału radiowego emitowanego przez bezzałogowe statki powietrzne (BSP) różnej klasy, które umożliwią zwiększenie bezpieczeństwa zarówno systemów do zastosowań cywilnych jak i wojskowych.

W obecnym stanie literatury proponowane są detektory, które można zaimplementować w warstwie fizycznej na podstawie wybranych parametrów statystycznych sygnałów BSP, w tym m.in. detektory energii lub mocy sygnałów odebranych [4, 5], właściwości cyklostacjonarności sygnałów [6, 7, 12, 8], czy detektory oparte na modelach uczenia maszynowego [9, 10].

Przegląd literatury wskazuje, że z oczywistych względów proponowane rozwiązania i zakres badań mają pewne ograniczenia, do których można zaliczyć m.in.: brak informacji o zachowaniu detektorów w przypadku ekspozycji na sygnały niebędące częścią zbioru treningowego, wpływ zmiennej struktury sygnałów bezzałogowych statków powietrznych nawet w trakcie ich pracy na skuteczność detekcji, brak lub ograniczone informacje o praktycznej implementacji wskazanych metod, badania wyłącznie symulacyjne lub badania z ograniczonymi zbiorami zarejestrowanych sygnałów. Wskazane ograniczenia uzasadniają prowadzone prace przedstawione w niniejszym artykule, które można uznać za innowacyjne w kontekście sposobu realizacji detektora sygnatur BSP i adaptacji metody wyznaczania charakterystyk cyklostacjonarności.

Niniejszy artykuł stanowi opis wyników prac będących kontynuacją badań opisanych wcześniej w [11], które oparto na realizacji badań pomiarowych w komorze bezechowej wybranych modeli BSP klasy mikro i mini, m.in. firmy DJI oraz Autel Evo. Sygnały rejestrowano z użyciem urządzenia radia programowalnego USRP X410 synchronicznie w dwóch pasmach 2,4 GHz oraz 5 GHz o szerokości 125 MHz. Szczegółowe informacje dotyczące przeprowadzonej kampanii pomiarowej oraz stanowiska pomiarowego przedstawiono w [11].

2. DETEKTOR SYGNATUR BSP

W ramach prac zrealizowano detektor sygnatur radiowych BSP bazujący na charakterystycznych parametrach cyklostacjonarności sygnałów poszczególnych BSP i estymacji ich unikatowych sygnatur radiowych. Na podstawie [12] parametry cyklostacjonarności można określić wyznaczając zespoloną cykliczną funkcję autokorelacji CAF jako funkcję częstotliwości cyklicznej α danego procesu stochastycznego oraz przesunięcia czasowego τ

$$\hat{R}_x^{\alpha}(\tau) = \frac{1}{M}\sum_{m=0}^{M-1} x[m]\cdot x^*[m+\tau]\cdot e^{-2j\pi\alpha m}, \qquad (1)$$

gdzie M to długość analizowanego okna sygnałów próbek. Uwzględnienie analizy częstotliwości cyklicznej oraz przesunięcia czasowego tworzy właściwie zespoloną przestrzeń sygnatur radiowych. W takim przypadku określenie odpowiedniego zakresu parametrów estymatora jest zasadniczo nieliniowym, trójwymiarowym problemem optymalizacyjnym opisanym jako funkcję $\hat{R}_x(\alpha, \tau, M)$. Jego efektywność definiowana jest prawdopodobieństwem poprawnej detekcji sygnatury oraz złożoności obliczeniowej (inercji całego układu decyzyjnego) w obliczu różnych scenariuszy jego użycia, tj. w przypadku braku zakłóceń, zmiennej wartości SNR (Signal-to-Noise Ratio), czy zmiennej wartości SINR (Signal-to-Interference-plus-Noise Ratio).

Jako że sygnały analizowanych BSP są emitowane w ogólnodostępnych pasmach nielicencjonowanych, i z oczywistych względów warunki propagacyjne mogą się zmieniać oraz mogą mieć istotny wpływ na prawdopodobieństwo poprawnej detekcji sygnałów, przeprowadzono badania trójwymiarową metodą siatkową (Grid Search) dla analizowanych sygnałów BSP.

Zgodnie z opisem zawartym w [11] znakowanie referencyjnych zbiorów sygnałów odbieranych zrealizowano na podstawie 40 dB progu względem maksymalnych wartości otrzymywanych z analizy czasowo-częstotliwościowej. Umożliwia to identyfikację występowania transmisji dla poszczególnych scenariuszy pomiarowych i określenie referencyjnych decyzji detektora co umożliwiło jego badanie w różnych symulowanych warunkach pracy.

Dodatkowo, należy wspomnieć, iż prezentowane wyniki dotyczyć będą analiz w podpasmach w obrębie całego monitorowanego pasma sygnałów projektowanego systemu. Na tym etapie prac przyjęto, iż monitorowane oba pasma 125 MHz podzielone będą na 10 podpasm o szerokości 12,5 MHz, co umożliwia ich współbieżną analizę i określenie nie tylko obecności danego BSP, ale także wskazanie występowania jego sygnatury w siatce czasowo-częstotliwościowej z odpowiednią dokładnością. Warto podkreślić, iż prezentowane dalsze rozważania są prawidłowe niezależnie od szerokości wybranych podpasm, ze względu na zrealizowaną normalizację uzyskiwanych wartości estymat cyklostacjonarnych, co powinno w ogólności uodpornić proponowaną metodę na zmienne wartości mocy sygnału odbieranego oraz szerokość monitorowanego podpasma.

3. PARAMETRYZACJA DETEKTORA SYGNATUR BSP

W celu określenia zakresu parametrów $\hat{R}_x(\alpha, \tau, M)$ zawężono zakres rozważanych parametrów na podstawie dotychczasowych wyników [11], np. w kontekście wykrytej powtarzalności pewnych składowych w różnych interwałach czasu (np. 20 ms, 1 ms i 500 μs). Determinowało to analizowane okno czasowe nie krótsze niż 21 ms (by objąć przynajmniej dwie emisje periodyczne odpowiadające potencjalnie czasowi trwania ramek zgodnie z emisją sygnałów w standardzie systemów komórkowych, lub popularnych organizacji transmisji wieloczęstotliwościowej techniki OFDM). Przy czym, wartości okna nie mogą być zbyt długie – ze względu na inercję całego detektora i jego dokładność wskazywania pozycji występowania emisji w siatce czasowo-częstotliwościowej. Z oczywistych względów – im dłuższe okno analizy – tym większa szansa na statystyczne uśrednianie uzyskiwanych wartości i wpływu kanału radiowego oraz szumu.

Na tej podstawie dla stałych zakresów wartości częstotliwości cyklicznej i przesunięcia (na tym etapie wystarczającym było wybranie wartości pozostałych parametrów odpowiadającym długości analizowanego okna czasowego) wybrano sygnały o czasie trwania 1 sekundy i z krokiem 50 ms zmniejszano wielkość okna analizy do momentu, w którym uzyskiwane sygnatury radiowe co do kształtu były skorelowane zgodnie z estymowaną wartością współczynnika Pearsona. Na podstawie uzyskanych wyników określono, iż długość analizowanego okna M powinno odpowiadać próbkom z przedziału 100 ms i taką wartość przyjęto w dalszych rozważaniach.

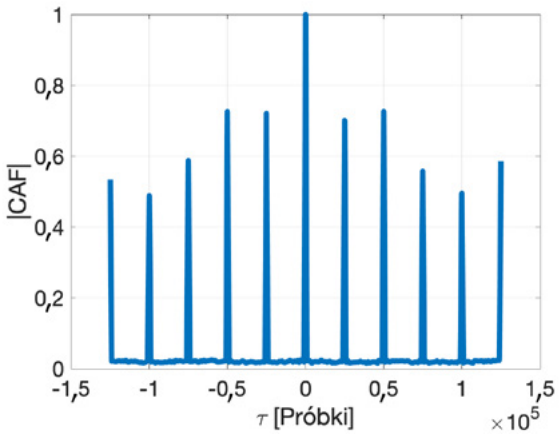
Dla przyjętego 100 ms okna analizy w poszczególnych podpasmach zdefiniowany problem optymalizacyjny estymacji parametrów CAF sprowadzono do problemu dwuwymiarowego. Jest on powiązany z periodycznością analizowanych sygnałów poszczególnych emisji BSP co znajduje także potwierdzenie w literaturze [13, 14] wskazując także różne zakresy parametrów pracy estymatora cyklostacjonarności. W takim przypadku, należało zweryfikować i próbować zdefiniować przestrzeń rozwiązań funkcji CAF by zmaksymalizować prawdopodobieństwo detekcji BSP. Przy czym, wartości te mogą się różnić w zależności od analizowanego BSP, ale zakres tych parametrów powinien być uniwersalny i umożliwiać detekcję każdego analizowanego BSP.

Zgodnie z tym podejściem analizowano amplitudę uzyskiwanych sygnatur cyklostacjonarnych jako funkcji częstotliwości cyklicznej i opóźnienia korelowanych funkcji. Przestrzeń rozwiązań została ograniczona do wartości $\alpha = \{100, 110, ..., 5000\}$ oraz $\tau = \{-0,001f_s, ..., 0,001f_s\}$, gdzie f_s to częstotliwość próbkowania użytego front-endu radiowego typu USRP, tj. 125 MHz dla analizowanych sygnatur dla dwóch BSP kolejno DJI Mavic 3 oraz Autel Evo Lite. Wyjaśnienia wymaga dolna oraz górna granica cyklicznej częstotliwości. Dolna wartość właściwie nie uwzględnia wskazanych wcześniej składowych dla częstotliwości 50 Hz (tj. 20 ms). Ze względu na fakt, iż analiza realizowana jest w podpasmach, co wymusza realizację filtracji dolnoprzepustowej sygnału oraz jego decymację, wpływa ona negatywnie na kształt sygnatur w dolnym zakresie częstotliwości cyklicznych oraz sprawia, iż dominujące w tym zakresie składowe są raczej produktami tejże filtracji aniżeli emisji radiowej nawet dla wysokich wartości SNR. Co więcej, w przypadku braku emisji i występowania wyłącznie nieskorelowanego szumu gaussowskiego widoczne są ich jednoznacznie skorelowane wartości właśnie ze względu na realizowaną filtrację.

Z tego względu, na podstawie analiz kształtów i istotnych składowych, które nie występują z oczywistych względów dla sygnałów nieperiodycznych np. szumu, wyznaczono dolną wartość częstotliwości cyklicznych $\alpha = 100$ Hz. Dodatkowo, określono, iż powyżej częstotliwości cyklicznych $\alpha = 5000$ Hz nie można wyróżnić istotnych składowych dla analizowanych sygnałów BSP [11], co jest odmiennym wynikiem, niż prezentowane w literaturze sygnatury z istotnymi składowymi nawet dla wartości powyżej 5000 Hz [14]. Przy czym były to inne warunki oraz inne modele BSP niż te analizowane w ramach niniejszych prac, a wskazywane składowe wysokoczęstotliwościowe nie były dominujące pod względem sumarycznej wartości mocy pozostałych składowych poniżej 5000 Hz.

Dla emisji BSP oprócz składowej dominującej widoczne są periodyczne składowe, które nie występują

w przypadku emisji WiFi oraz sygnałów bez emisji BSP (szum). Na rys. 1 przedstawiono wyniki znormalizowanej wartości modułu zespolonej CAF uśrednionej w funkcji częstotliwości cyklicznych dla emisji BSP. Widoczne są dominujące składowe co 200 μs (25 000 próbek) dla BSP DJI. Dla BSP firmy Autel dominujące składowe występują co 1 ms (125 000 próbek).



Rys.1. Przykładowy wynik znormalizowanej wartości modułu zespolonej cyklicznej funkcji autokorelacji, uśrednionej w funkcji częstotliwości cyklicznych dla DJI Mavic 3.

Na podstawie uzyskanych wyników (rys. 1) można znacząco zminimalizować liczbę realizowanych obliczeń dla poszczególnych opóźnień τ bowiem większość energii w przypadku występowania emisji BSP zawiera się w periodycznych realizacjach τ, które powinny być analizowane by nie zanieczyszczać wyznaczanych odbieranych sygnatur CAF. W takim przypadku określono, iż dalsze przetwarzanie uzyskiwanych wartości CAF powinno być realizowane wyłącznie dla 5-ciu największych (co do wartości i energii) estymowanych składowych cyklostacjonarności. Ogranicza to znacząco liczbę obliczeń i hipotez obliczając wyłącznie wybrane opóźnienia τ, a nie ich cały zbiór.

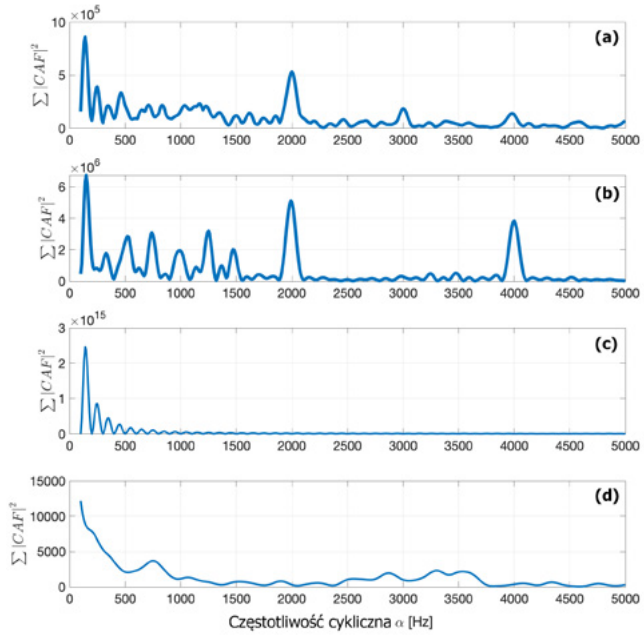
Z uwagi na fakt, że istotnym jest detekcja emisji dowolnego BSP, obliczono ujednoliconą sygnaturę emisji analizowanych BSP DJI oraz Autel, która stanowi kumulatywną, uśrednioną funkcję gęstości prawdopodobieństwa cyklicznej funkcji autokorelacji. Na tym etapie prac istotna jest wyłącznie detekcja emisji, a nie jej identyfikacja, czy klasyfikacja. W takim przypadku, detekcja emisji może być zrealizowana obliczając metrykę podobieństwa pomiędzy sygnaturą odebraną, a wspomnianą sygnaturą referencyjną, np. używając filtracji przestrzennej z użyciem sieci neuronowych, czy współczynnika korelacji Pearsona. Ze względu na złożoność obliczeniową i interpretowalność uzyskiwanych wyników wybrano realizację korelacji skrośnej. W takim przypadku wartość bezwzględna wynikowego współczynnika stanowi miarę prawdopodobieństwa występowania emisji BSP, gdzie:

- wartość 1 uzyskiwana jest dla teoretycznie idealnej korelacji wyników;
- wartość 0,5 uzyskiwana jest dla braku emisji sygnałów (szumu);
- wartość 0 uzyskiwana jest dla braku korelacji wyników, przy czym jest to informacja, iż w danym paśmie występują inne interferujące emisje radiowe, których charakter nie jest zbliżony do referencyjnej funkcji gęstości prawdopodobieństwa i nie mają charakterystyki szumu gaussowskiego.

W tym miejscu należy podkreślić ponownie, iż wartość 0,5 korelacji szumu i referencyjnej FGP należy interpretować jako poprawną ze względu na realizowaną filtrację pasmową przy wyodrębnianiu podpasm do analizy dla detektora. Stąd, szum nie uzyskuje wartości zbliżonych do 0, a raczej zbliżone są co do wartości do 0,5.

4. ZAKRES STOSOWALNOŚCI DETEKTORA SYGNATUR BSP

Na rys. 2 przedstawiono przykładowe wyniki kumulatywnej wartości kwadratu modułu zespolonej cyklicznej funkcji autokorelacji $\sum |CAF|^2$ kolejno dla emisji DJI Mavic 3 (a), Autel Evo Lite (b), bez emisji sygnałów (szum) (c) oraz emisji WiFi (d). Na przedstawionych wykresach wyraźnie widoczna jest inna dynamika zmian sygnatur dla poszczególnych typów sygnałów.

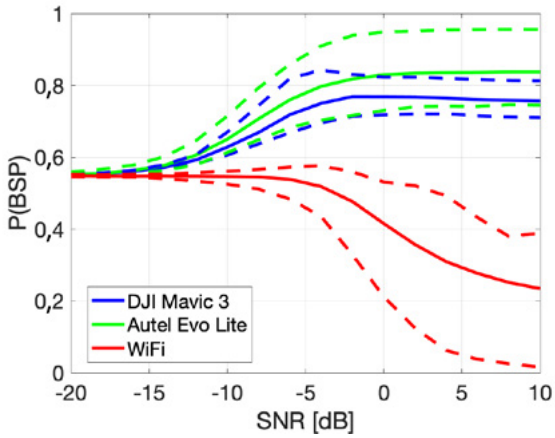


Rys.2. Przykładowe wyniki kumulatywnej wartości kwadratu modułu zespolonej cyklicznej funkcji autokorelacji kolejno dla emisji DJI Mavic 3 (a), Autel Evo Lite (b), bez emisji sygnałów (szum) (c) oraz emisji WiFi (d).

Można wyraźnie wskazać, że każdy komercyjny model BSP (nawet tego samego producenta) ma swoją własną unikatową sygnaturę w zakresie α 150-2000 Hz z dwiema dominującymi składowymi 140 Hz, 2000 Hz, a dla urządzeń DJI dodatkowo przy 3000 Hz, zaś urządzeń Autel Evo 4000 Hz. W wymienionym zakresie można wyróżnić unikatowe składowe cykliczne na podstawie BSP, np. dla DJI Mavic 3 (rys. 2a) komponenty {250, 350, 480, 2000} Hz, a dla Autel Evo Lite (rys. 2b) komponenty {330, 530, 740, 990, 1250, 1490, 2000} Hz. Wyniki te pokazują również możliwość znacznego zmniejszenia zakresu częstotliwości skanowania wartości α, co również wpłynie na minimalizację złożoności numerycznej prezentowanego podejścia w końcowej implementacji na urządzeniach docelowych [11].

Dla wyników zaprezentowanych na rys. 2 uzyskano kolejno wartości prawdopodobieństw występowania emisji $P(BSP)$: 0,91, 0,95, 0,55, 0,40 co potwierdza poprawność pracy zrealizowanego miękkodecyzyjnego detektora emisji BSP.

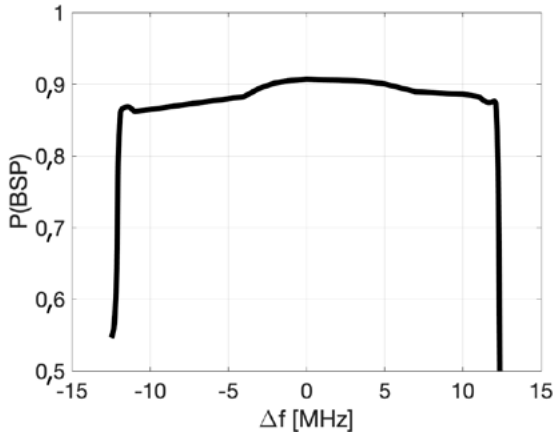
W celu usystematyzowania i weryfikacji poprawności pracy miękkodecyzyjnego detektora, zasymulowano zmienną wartość stosunku SNR dla emisji BSP DJI Mavic 3, Autel Evo Lite oraz sygnałów ogólnodostępnej sieci WiFi co przedstawiono na rys. 3 dla 9-ciu segmentów o czasie trwania 100 ms. Należy zaznaczyć, iż wskazane wyniki obliczano dla innych segmentów, niż te które były uwzględniane przy wyliczaniu referencyjnych sygnatur radiowych. Sygnały sieci WiFi zarejestrowano niezależnie, w wewnątrzbudynkowym środowisku na terenie Politechniki Gdańskiej podczas obecności wielu użytkowników korzystających z niniejszego ogólnodostępnego pasma.



Rys.3. Prawdopodobieństwo detekcji emisji BSP $P(BSP)$ z zaznaczonymi wartościami minimalną i maksymalną i wartością średnią przy użyciu opracowanego miękkodecyzyjnego detektora bazującego na cyklostacjonarności.

Przedstawione wyniki wskazują na poprawną pracę opracowanego detektora w funkcji zmiennej wartości SNR. Dla $SNR > -10$ dB przy przyjętym jednowartościowym progu 0,6 możliwa jest detekcja występowania emisji BSP. Dla wskazanego zakresu SNR możliwe jest również określenie występowania innych typów emisji w zadanym paśmie, które nie stanowią emisji BSP ani szumu.

Dodatkowo, w aktualnym stanie literatury często zakłada się, iż emisje radiowe, które mają być wykrywane umieszczone są w całości w oknie obserwacji detektora (tj. gdy szerokość pasma emisji jest mniejsza od szerokości pasma obserwacji), lub wypełniają je całkowicie (tj. gdy szerokość pasma emisji jest większa lub równa od szerokości pasma obserwacji). Biorąc pod uwagę rzeczywiste warunki pracy detektora, w ogólności nie można zapewnić, iż wspomniane warunki będą zawsze spełnione, tj. w oknie obserwacji może występować wyłącznie część emisji sygnatur radiowych. Mając na uwadze niniejsze, zweryfikowano efektywność pracy detektora jako funkcji przesunięcia emisji w dziedzinie częstotliwości Δf w stosunku do okna obserwacji. Na rys. 4 zaprezentowano uzyskane wyniki, przy czym wartość 0 oznacza brak przesunięcia emisji sygnału, tj. całe okno obserwacji jest wypełnione emisją BSP w dziedzinie częstotliwości, a graniczne przesunięcia $\pm 12,5$ MHz oznaczają brak emisji (szum) w oknie obserwowanego pasma.



Rys.4. Prawdopodobieństwo detekcji emisji BSP $P(BSP)$ w funkcji przesunięcia emisji w dziedzinie częstotliwości Δf.

Na podstawie uzyskanych wyników można określić, iż proponowany detektor umożliwia wykrycie emisji BSP w przypadku gdy w oknie obserwacji znajduje się część sygnatury radiowej o szerokości pasma 550 kHz, tj. dla przesunięcia maksymalnie ± 11950 kHz dla $P(BSP)$ większego od 0,85 oraz o szerokości pasma 300 kHz, tj. dla przesunięcia maksymalnie ± 12200 kHz dla $P(BSP)$ większego od 0,6. Zgodnie z oczekiwaniami, dla wskazanej granicznej wartości szerokości pasma emisji około 500 kHz efektywność detektora znacząco maleje wraz ze zmniejszaniem obserwowanego pasma emisji sygnału BSP. Warto nadmienić, iż główna emisja sygnałów wizyjnych zawierała się w całkowitym paśmie 18 MHz.

5. PODSUMOWANIE

Mając na uwadze przedstawione wyniki badań proponowany detektor sygnatur radiowych BSP umożliwia poprawną detekcję wybranych emisji dla wartości SNR powyżej -10 dB przy przyjętym progu $P(BSP) = 0,6$. Wskazano sposób przetwarzania danych wraz z metodyką i wynikowymi parametrami opracowanego miękkodecyzyjnego detektora. Zweryfikowano także, iż proponowana metoda umożliwia detekcję sygnału BSP nawet przy występowaniu w danym podpaśmie wyłącznie części emisji (nawet do 500 kHz) ich charakterystycznych sygnałów cyklicznych.

Przyszłe prace powinny obejmować dodatkowe sygnały zakłócające występujące w rzeczywistych środowiskach dla badanych pasm ISM, analizę współistnienia wielu (np. roju) BSP, rozszerzone badania nad wydajnością wykrywania lub możliwe wykorzystanie sieci neuronowych do wykrywania emisji BSP w kanale z zanikami, gdzie spodziewane jest obniżenie efektywności korelacji krzyżowej ze względu na nieliniowe zmiany amplitudy i fazy sygnału odbieranego.

6. FINANSOWANIE

Niniejsze badania zrealizowano w ramach projektu pt. “Mobilny system do zwalczania BSP klasy mikro i mini”, umowa nr DOB-SZAFIR/02/A/018/04/2021.

LITERATURA

[1] Seidaliyeva U., Ilipbayeva L., Taissariyeva K., Smailov N., Matson E.T., „Advances and Challenges in Drone Detection and Classification Techniques: A State-of-the-Art Review", Sensors, 2024.

[2] Jagannath A., Jagannath J., Kumar P. S. P. V., „A comprehensive survey on radio frequency (RF) fingerprinting: Traditional approaches, deep learning, and open challenges.", Computer Networks, 2022.

[3] Dobre, Octavia A., „Signal identification for emerging intelligent radios: Classical problems and new challenges", IEEE Instrumentation & Measurement Magazine, 2015.

[4] Sobron I., Diniz P. S. R., Martins W. A., Velez M., „Energy Detection Technique for Adaptive Spectrum Sensing, IEEE Transactions on Communications", wol. 63, nr 3, 2015.

[5] Lorincz J., Ramljak I., Begušić D., „A Survey on the Energy Detection of OFDM Signals with Dynamic Threshold Adaptation: Open Issues and Future Challenges", Sensors, 2021.

[6] Nafkha A., Naoues M., Cichon K., Kliks A., „Experimental spectrum sensing measurements using USRP Software Radio platform and GNU-radio", 9th International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM), 2014.

[7] Nafkha, A., Naoues, M., Cichony, K., Kliks, A., and Aziz, B., „Hybrid spectrum sensing experimental analysis using GNU radio and USRP for cognitive radio", International Symposium on Wireless Communication Systems (ISWCS), 2015.

[8] Antoni J., „Cyclostationarity by examples, Mechanical Systems and Signal Processing", wol. 23, 2009.

[9] Y. Zhang, D. Li, S. Hong, R. Lin, and H. Sun, „RF-Based UAV Detection and Identification Using Random Forest", IEEE 7th International Conference on Information Systems and Computer Aided Education (ICISCAE), 2024.

[10] C. Ge, S. Yang, W. Sun, Y. Luo, and C. Luo, „For RF Signal-Based UAV States Recognition, Is Preprocessing Still Important At The Era Of Deep Learning?", 7th International Conference on Computer and Communications (ICCC), 2021.

[11] Cwalina K. K., Rajchowski P., Sadowski J., Skokowski P., Malon K., Kryk M., Maślanka K., Kelner J.M., „Detection of Dual-Wideband UAV Radio Signatures in an Integrated C-UAS, 21st International Conference on Smart Technologies", IEEE Eurocon 2025, 2025 (przyjęty do publikacji).

[12] Gardner W. A., „Cyclostationarity in Communications and Signal Processing", IEEE Press, 1994.

[13] Kaplan B., Kahraman I., Ekti A. R., Yarkan S., Gorcin A., Ozdemir M. K., Cirpan H. A.„Detection, Identification, and Direction of Arrival Estimation of Drone FHSS Signals With Uniform Linear Antenna Array", IEEE Access 9, 2021.

[14] Yue X., i Zhang X., „Micro-UAV Detection and Identification Based on Radio Frequency Signature", 6th International Conference on Systems and Informatics (ICSAI), 2019.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

IMPLEMENTACJA AUTOENKODERA DO ETYKIETOWANIA DANYCH I DO DETEKCJI NOWYCH ATAKÓW SIECIOWYCH

AUTOENCODER-BASED APPROACH FOR DATA LABELING AND NOVEL NETWORK ATTACK DETECTION

Tomasz Rachwalik[1], Robert Białas[1], Przemysław Dąbrowski[1]

[1] Wojskowy Instytut Łączności – Państwowy Instytut Badawczy, Zegrze Płd.,
{t.rachwalik, r.bialas, p.dabrowski}@wil.waw.pl

DOI: 10.15199/59.2025.4.77

Streszczenie: **Detektory ataków sieciowych mogą być wspomagane przez techniki uczenia maszynowego. Do ich działania niezbędne są jednak dane wejściowe (datasety), które przygotowują model do klasyfikacji. W artykule przedstawiono framework i algorytm do generowania danych do nauki modelu klasyfikującego. Prezentowany model oparty o autoenkoder może być zaimplementowany i rozwijany w docelowym środowisku jako NIDS.**

Abstract: **Network attack detectors can be supported by machine learning techniques. However, their operation requires input data (datasets) that prepares the model for classification. The article presents a framework and an algorithm for generating data to train a classification model. The presented model, based on an autoencoder, can be implemented and developed in the target environment as a NIDS.**

Słowa kluczowe: **anomalie, ataki wolumetryczne, autoenkoder, dataset, uczenie głębokie.**

Keywords: **anomalies, volumetric attacks, autoencoder, dataset, deep learning.**

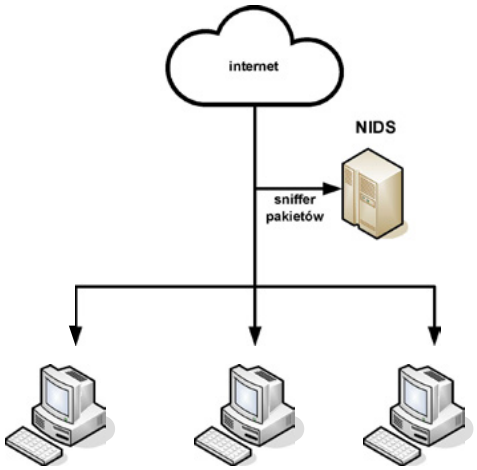
1. WSTĘP

Detektory wykrywania ataków sieciowych wykorzystujące uczenie maszynowe są coraz częściej stosowane w różnorodnych sieciach. Zastosowanie zaawansowanych technik analizy danych zapewniają skuteczniejsze wykrywanie anomalii i zagrożeń, które mogłyby zostać przeoczone przez klasyczne systemy detekcji. Szczególnie interesujące są detektory anomalii, ponieważ umożliwiają wykrywanie także tych ataków, które wcześniej nie były znane. Pojęciem anomalii określamy stany, które odbiegają od zachowań zaobserwowanych w trakcie niezaburzonej pracy sieci.

Zaprezentowana metoda i narzędzia do detekcji nowych ataków sieciowych umożliwiają wzbogacanie zbiorów danych wykorzystywanych przez mechanizmy klasyfikujące ataki, a ich precyzja może być zwiększona przez profilowanie danych uczących pod konkretną sieć.

1.1. NIDS

Systemy wykrywające ataki, nieautoryzowane próby dostępu lub inne podejrzane działania w sieci określane są pojęciem NIDS (*ang. Network Intrusion Detection System*). NIDS analizuje ruch sieciowy i ostrzega administratora o potencjalnych zagrożeniach. NIDS może być urządzeniem wykorzystującym reguły (np. SNORT, Suricata). Regułami określamy zbiór instrukcji, które definiują jakie pakiety powinny być analizowane i jak należy na nie reagować. Reguły są często stosowane do każdego pakietu osobno, a wykrywanie zagrożeń polega na porównywaniu zawartości pakietów z określonymi wzorcami i sygnaturami.



Rys. 1. NIDS w sieci.

NIDS może również być wspomagany przez model uzyskany w wyniku uczenia na odpowiednio spreparowanych danych (dataset). Przy czym „odpowiednio" jest tutaj słowem kluczowym i stanowi problem badawczy realizowanych prac.

1.2. Zbieranie danych do nauki

NIDS wykorzystujący uczenie maszynowe może analizować przepływy lub sesje. Przepływ to zestaw pakietów w określonym oknie czasowym. Sesja to sekwencja pakietów połączona na podstawie pięciu informacji sieciowych (IP klienta, port klienta, IP serwera, port serwera, protokół) [3].

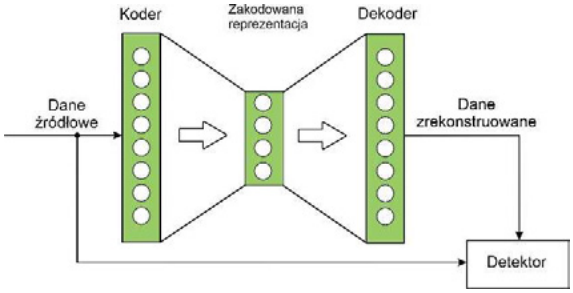
W niniejszej pracy skupiono się na przepływach jako formie gwarantującej zbiorczą informację o wymienianych danych. Odpowiednia struktura przepływów pozwoliła uniknąć problemów z przeciążaniem modelu, tym samym zapewniając dostępność wykorzystującego je detektora.

Proces oznakowywania (etykietowania) zbiorów danych można usprawnić poprzez wdrożenie mechanizmu

wykrywania anomalii, który wykorzystuje zarówno płytkie, nienadzorowane metody uczenia, takie jak algorytm Isolation Forest czy klastrowanie, jak i głębokie modele, na przykład autoenkodery. Autoenkodery realizują wykrywanie subtelnych i trudnych do wychwycenia odchyleń od normy, co przekłada się na wyższą jakość etykietowania danych i skuteczniejsze przygotowanie zbiorów do dalszego uczenia modeli klasyfikujących [3].

1.3. Autoenkoder

Autoenkoder jest rodzajem sieci neuronowej głębokiej, nienadzorowanej [1]. Autoenkoder jest w stanie znaleźć nieliniowe zależności między cechami, co zwiększa dokładność wykrywania [1]. Autoenkoder kompresuje dane, a następnie odtwarza ich postać na podstawie „zakodowanej" reprezentacji.



Rys.2. Autoenkoder – budowa i schemat działania.

Architektura autoenkodera może być dostosowana do charakterystyki danych – w przypadku danych sekwencyjnych (np. przepływy czasowe) stosuje się warstwy rekurencyjne – RNN (*ang. Recurrent Neutral Network*), LSTM (*ang. Long Short-Term Memory*) [3], a dla danych przestrzennych - konwolucyjne CNN (*ang. Convolutional Neutral Network*) [1]. Dzięki nienadzorowanemu charakterowi, model nie wymaga etykietowanych przykładów anomalii. Może też posłużyć do etykietowania zbiorów danych, co wykorzystano w pracy. Autoenkodery są efektywne obliczeniowo i mogą być wykorzystywane w systemach detekcji w czasie rzeczywistym [3].

1.4. Stan wiedzy i proponowane rozwiązanie

Wiele prac dotyczących zastosowania autoenkodera wykorzystuje do oceny gotowy zbiór danych NSL-KDD (*ang. Network Security Lab – Knowledge Discovery and Data Mining*) [1][2][3]. Przedstawiony Dataset jest zbalansowany, ale nie zawiera nowych ataków [3]. Autorzy wskazują, że dotychczasowe rozwiązania oparte na autoenkoderze różnią się skutecznością i nie analizują kompleksowo wpływu kluczowych czynników na jakość detekcji. W [5] autorzy koncentrują się na wdrożeniu systemu do wykrywania anomalii w sieci, który może działać przy strumieniowaniu dużej ilości danych przesyłanych w czasie rzeczywistym.

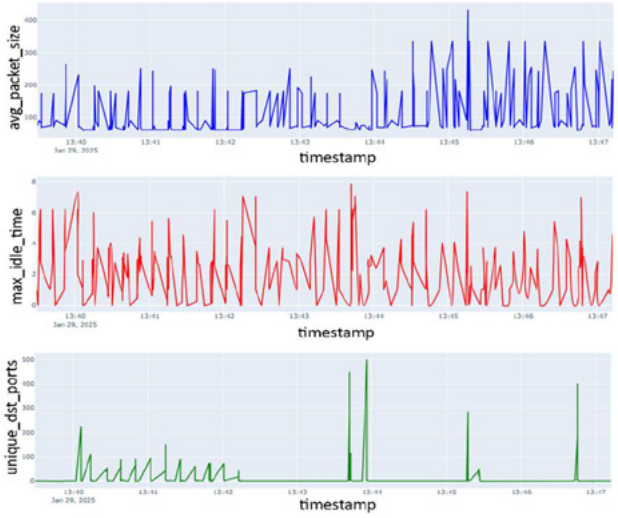
W pracy zaproponowano algorytm automatycznego etykietowania oraz przedstawiono własną implementację systemu NIDS. System wykorzystuje autoenkoder do wykrywania nowych ataków oraz model oparty na algorytmie Random Forest do klasyfikacji znanych ataków. Nowe ataki poddane analizie mogą rozszerzyć zakres klasyfikacji znanych ataków. Oprogramowanie zweryfikowano na dostępnych publicznie danych off-line, ale także przetestowano w wybranych, rzeczywistych sieciach o różnym zastosowaniu. Wyniki tych badań przedstawiono poniżej w p. 2.5.

2. IMPLEMENTACJA

2.1. Przetwarzanie danych

W środowisku open source dostępnych jest wiele narzędzi do agregacji i konwersji danych sieciowych, jednak często są one trudne w obsłudze, wymagają zaawansowanej konfiguracji i nie spełniały wszystkich stawianych w pracach wymagań. Brak jest także gotowych aplikacji napisanych w Pythonie, które umożliwiałyby bezpośrednią konwersję plików PCAP do IPFIX i eksport do CSV. W pierwszym etapie badań do tego celu wykorzystano narzędzie CICFlowMeter [6]. Jest to jedno z najpopularniejszych narzędzi open source służących do generowania i analizy przepływów sieciowych na podstawie plików PCAP. Jedną z kluczowych funkcji CICFlowMeter jest możliwość eksportu wyników bezpośrednio do pliku CSV, co bardzo ułatwia dalszą analizę, np. w narzędziach do uczenia maszynowego. W trakcie używania narzędzia okazało się jednak, że generuje ono dużą liczbę wpisów przy atakach wolumetrycznych. Można to przeanalizować np. na datasetach CICDDoS2019 [6].

W ramach prowadzonych badań zostało opracowane własne narzędzie pozwalające agregować, a tym samym znacznie ograniczać liczbę takich wpisów. W tym celu wprowadzono parametry *unique_dst_ports* i *unique_srt_ports* zdefiniowane jako liczba skanowanych portów na wybranym adresie IP. Wynik jego działania przedstawia rys. 3. Narzędzie to zostało napisane w języku Python z wykorzystaniem biblioteki *scapy*. Pozwoliło ono na uniknięcie przeładowania modelu, zatem na detekcję ataków i anomalii w czasie rzeczywistym.



Rys. 3. Zobrazowanie wybranych cech z wykorzystaniem biblioteki dash.

Proces tworzenia datasetu poprzedziła dokładna analiza cech danych przesyłanych w różnych środowiskach sieciowych. Posłużyła do tego graficzna reprezentacja cech w czasie - rys. 3, czy inne metody jak np. macierz korelacji - rys. 4.

Rys. 4. Macierz korelacji wybranych cech z przepływów.

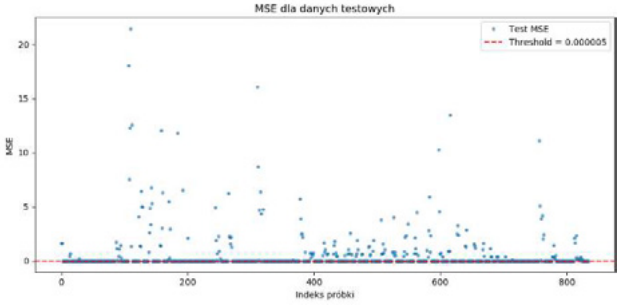
2.2. Detekcja anomalii

Model autoenkodera uczy się odwzorowania danych wejściowych $x \in \mathbb{R}^n$ na ich rekonstrukcję $\hat{x} = g_\phi(f_\theta(x))$, minimalizując błąd rekonstrukcji. Najczęściej stosowaną funkcją straty jest średni błąd kwadratowy (MSE), definiowany jako:

$$MSE(x, \hat{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2 \qquad (1)$$

Autoenkoder trenowany jest wyłącznie na danych normalnych, dzięki czemu uczy się ich wewnętrznej struktury i typowych wzorców. W przypadku wystąpienia anomalii, których model wcześniej nie identyfikował, błąd rekonstrukcji istotnie wzrasta, co umożliwia ich detekcję.

Próg detekcji może być ustalany empirycznie lub statystycznie. Najprostszym sposobem wyznaczenia progu detekcji dla przepływów jest zliczanie po kolejnych cechach. Wynik detekcji anomalii na podstawie tak obliczonego progu przedstawiono na rys. 5.

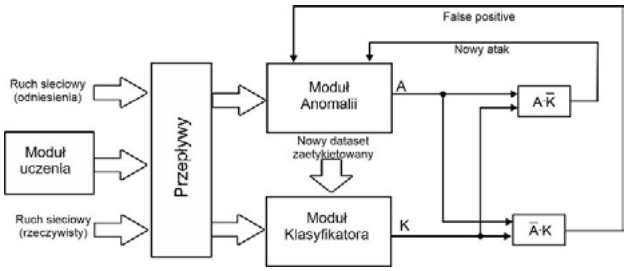


Rys. 5. Zobrazowanie progu detekcji i próbek przekraczających próg.

2.3. Proces etykietowania danych

Wynikiem działania większości narzędzi do zbierania przepływów są dane zawierające zestaw cech (atrybutów) bez etykiet, czyli bez informacji, czy dane są poprawne, czy wskazują na atak. Pełny zbiór danych (dataset), niezbędny do klasyfikacji, wymaga obecności etykiet. Etykietowanie danych jest możliwe przy zastosowaniu uczenia nienadzorowanego, które umożliwia identyfikację anomalii i ich grupowanie [3].

W ramach prac zbudowano autorski system NIDS (framework) o strukturze przedstawionej na rys. 6. Framework zawiera moduł uczenia, moduł anomalii oraz klasyfikator. Moduł uczenia obejmuje zestaw testów (ataków wolumetrycznych). Moduł anomalii służy do detekcji anomalii i formowania zbiorów danych. Klasyfikator wykorzystuje uczenie maszynowe (algorytm Random Forest) do klasyfikacji ataków.



Rys. 6. Schemat opisywanego framework

Model anomalii oparty na autoenkoderze jest utworzony na bazie znormalizowanego ruchu sieciowego. Wynikiem tego etapu jest wyznaczenie progu bazowego. Następnie z modułu uczenia wykonywane są wybrane ataki. Detekcja ataków jest przeprowadzana zgodnie z algorytmem przedstawionym w p. 2.4. Wynikiem tego procesu jest pierwotny zbiór danych dla klasyfikatora.

W dalszej fazie do modułów anomalii i klasyfikatora wprowadzane są przepływy z sieci w trybie operacyjnym. W przypadku niewykrycia ataku przez klasyfikator wykrycie anomalii jest archiwizowane. Moduł anomalii prowadzi statystyki częstości występowania ataków w celu dalszych analiz. Cechom związanym z atakiem przypisywana jest nowa etykieta. Na tej podstawie tworzony jest nowy zbiór danych - dataset.

Moduł anomalii analizuje również informacje o ataku wykrytym wyłącznie przez klasyfikator (False Positive). Dane takie są wykorzystywane do weryfikacji zbioru danych.

2.4. Algorytm etykietowania

Algorytm etykietowania danych na etapie wstępnego uczenia przedstawiono jako Alg. 1. Działanie algorytmu implementuje metodę Injection Timing [2], detekcję anomalii, stopniowe zwiększanie progu w celu uzyskania korelacji między czasem wykonania ataku, a rejestrowaną anomalią.

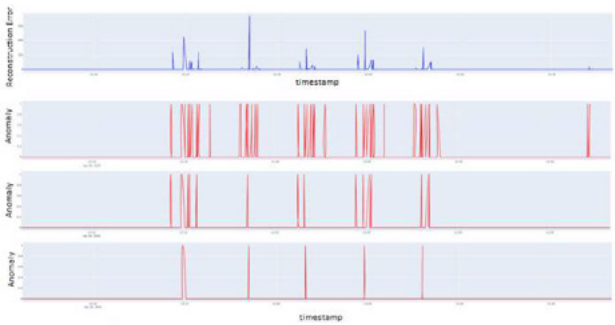
Algorytm 1 – Algorytm etykietowania danych

Wejście:
- Wytrenowany model na danych normalnych.
- Parametr τ: próg anomalii.
- Znormalizowany strumień ruchu sieciowego z możliwością iniekcji k-ataków w kontrolowanych odstępach czasowych T.

Kroki przetwarzania:
Powtórz k-krotnie:
1. Generacja ataków i ekstrakcja przepływów:
   - n-krotne wstrzykiwanie pakietów ataków z okresowością T.
   - Rejestracja znaczników czasowych ataków $\{t_1, t_2, \ldots, t_n\}$.
   - Grupowanie pakietów w przepływy
2. Test autoenkoderem.
3. Detekcja anomalii.
Obliczanie błędu rekonstrukcji (MSE) między wejściem a wyjściem.
4. Wykrywanie okresowości T z tolerancją ΔT.
Jeśli wykryto korelację z okresem $T \pm \Delta T \rightarrow$ oznacz jako atak. Wyjście.
5. Etykietowanie ataków:
   - Aktualizacja progu τ.
   - Powrót do kroku 3.

Wyjście:
Zaetykietowany dataset.

Zaetykietowany dataset jest niezbędny do klasyfikowania ataków z wykorzystaniem uczenia maszynowego nadzorowanego. Działanie algorytmu obrazuje rys. 7. Pierwszy wykres czasowy przedstawia błąd rekonstrukcji uzyskany w wyniku działania autoenkodera. Kolejne trzy wykresy ilustrują trzy kroki algorytmu, w których stopniowo zwiększany jest próg. Wyniki uzyskane z ostatniego wykresu posłużyły do oznaczenia cech jako atak oraz do wzbogacenia zbioru danych dla klasyfikatora.

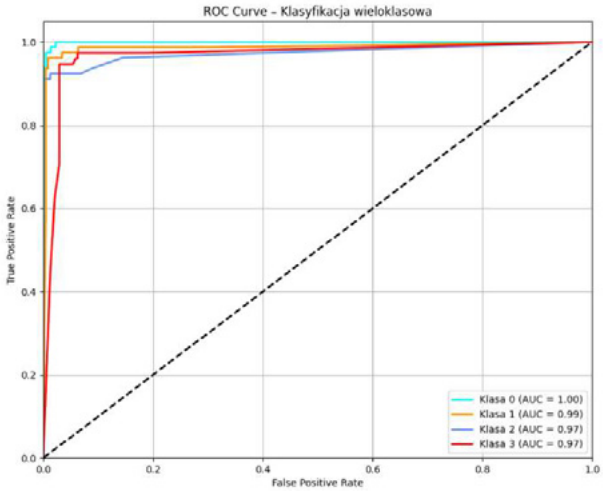


Rys. 7. Zobrazowanie działania algorytmu 1.

Algorytm wykrywa anomalie powtarzające się z zadanym okresem. Algorytm selekcjonuje najbardziej istotne wpisy z przepływu, co wspomaga prawidłowe szkolenie modelu. Wynikiem działania algorytmu 1 jest zaetykietowany dataset.

2.5. Wyniki

Opisywany system NIDS przetestowano w podsieciach IoT, SCADA i wewnętrznej sieci wybranej instytucji w ramach badań sieci szkieletowej 5G [4]. Ocenie poddano Klasyfikator pracujący w sieciach testowych. Do oszacowania systemu wykrywania intruzów opartego na metodach klasyfikacji wykorzystano dataset stworzony w trakcie wstępnej fazy uczenia rozwinięty o dodatkowe ataki. Rys. 8. przedstawia krzywą ROC (*ang. Receiver Operating Characteristic*) obrazującą wyniki działania klasyfikatora dla wybranych ataków.



Rys. 8. Krzywe ROC przedstawiające klasyfikację ataków w środowisku SCADA.

Klasa 0 to ruch zidentyfikowany przez Klasyfikator (a wcześniej zaetykietowany przez Moduł Anomalii) jako normalny. Klasy od 1 do 3 zostały zidentyfikowane jako trzy różne ataki wolumetryczne (którym na etapie wykrywania anomalii przypisano odpowiednio etykiety 1, 2 i 3). Uzyskane metryki, w tym AUC (*ang. Area Under Curve*) powyżej 0,95 potwierdzają dobrą jakość generowanych zbiorów danych.

3. WNIOSKI

Celem pracy było stworzenie narzędzi wspierających badania nad atakami w sieciach komputerowych. Kluczowym elementem był autoenkoder oparty na głębokim uczeniu, służący do wykrywania nietypowych zachowań w ruchu sieciowym. Zaproponowano także algorytm do etykietowania danych treningowych, co pozwoliło na lepsze przygotowanie modelu klasyfikującego. W ramach walidacji zbioru danych wdrożono dwustopniowy mechanizm obejmujący wykrywanie nowych anomalii poprzez analizę pojawiających się ataków oraz identyfikację przypadków błędnie sklasyfikowanych, co umożliwia dalszą korektę zbioru uczącego. Zauważono, że analizowane przepływy danych często układają się w krótkie sekwencje, co sugeruje istnienie lokalnych zależności czasowych. W odpowiedzi na tę obserwację rozpoczęto badania nad wykorzystaniem bardziej zaawansowanych architektur autoenkoderów, takich jak LSTM i CNN, które lepiej uchwycą te zależności.

LITERATURA

[1] Chen Z. et al. 2018. „ *Autoencoder-based Network Anomaly Detection*". *IEEE Xplore, Wireless Telecommunications Symposium*.

[2] Guerra Jorze Luis et al. 2022. "*Datasets are not enough: Challenges in labeling network traffic*". Computers & Security, Volume 120.

[3] Hongyu Liu, Bo Lang. 2019. "*Machine Learning and Deep Learning Methods for Intrusion Detection Systems: A Survey*". *Applied Science*. doi:10.3390/app9204396.

[4] Kosmowski Krzysztof, Bryś Rafał , Dudko Adam. 2024. „*Ataki na protokół PFCP sieci szkieletowej 5G oraz metody ich identyfikacji*". KRiT 2024.

[5] Rivera Javier Jose Diaz et al. 2021. *„An ML Based Anomaly Detection System in real time data streams"*. CSCI 2021.

[6] University of New Brunswick. „*CICFlowMeter"* CIC Research Applications. Dostęp 25 czerwca 2025, https://unb.ca/cic/research/applications.html

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

METODA PREDYKCJI RUCHU WYCHODZĄCEGO W SYSTEMACH KLASY CALL/CONTACT CENTER WYKORZYSTUJĄCA ALGORYTM UCZENIA ZE WZMOCNIENIEM

PREDICTION METHOD OF OUTGOING TRAFFIC IN CALL/CONTACT CENTER SYSTEMS USING REINFORCEMENT LEARNING ALGORITHM

Marcin Kozłowski; Mirosław Płaza[1]; Małgorzata Lucińska[1]; Marcin Kowalczyk[2]

[1] Politechnika Świętokrzyska, Kielce, mkozlowski@tu.kielce.pl
[2] Politechnika Warszawska, Instytut Telekomunikacji i Cyberbezpieczeństwa, Warszawa, marcin.kowalczyk@pw.edu.pl

DOI: 10.15199/59.2025.4.78

Streszczenie: **W artykule przedstawiono koncepcję metody predykcji ruchu wychodzącego dedykowaną dla systemów Call/Contact Center wykorzystującą algorytm uczenia przez wzmocnienie TD3. Rozwiązanie oparte na architekturze aktor-krytyk podejmuje decyzje w sposób ciągły, dostosowując tempo generowania połączeń bezpośrednio do aktualnego stanu środowiska. Uczenie oraz testowanie modelu przeprowadzono z użyciem zebranych rzeczywistych danych billingowych. Proponowane podejście pozwala na adaptacyjne sterowanie ruchem wychodzącym i znacznie zwiększa wydajność pracy agentów przy zachowaniu wysokiej jakości obsługi.**

Abstract: **The paper presents a method for predicting outbound traffic in Call/Contact Center systems using the TD3 algorithm. The model, based on an actor–critic architecture, makes continuous decisions by adjusting the call generation rate to the current state of the environment. Training and testing were carried out using real billing data. The proposed approach enables adaptive control of outbound traffic and improves agent work-time utilization while maintaining high service quality.**

Słowa kluczowe: **call/contact center, predykcja ruchu wychodzącego, TD3, uczenie ze wzmocnieniem.**

Keywords: **call/contact center, outbound traffic prediction, TD3, reinforcement learning.**

1. WSTĘP

Aktualnie obserwujemy bardzo dużą dynamikę w rozwoju systemów klasy CC (ang. *Call/Contact Center*) związaną z ich rosnącym znaczeniem w nowoczesnych usługach biznesowych [1-6]. Tym samym coraz większą uwagę poświęca się zagadnieniom optymalizacji ruchu wychodzącego, co samo w sobie stanowi podstawowe zagadnienie teorii kolejek [7]. Efektywne zarządzanie połączeniami inicjowanymi przez systemy CC ma bezpośredni wpływ na jakość obsługi klienta, poziom wykorzystania zasobów ludzkich oraz ostatecznie również na wyniki finansowe organizacji [8]. W szczególności, pożądane dostosowanie liczby generowanych połączeń wychodzących w stosunku do dostępności agentów oraz przewidywanych reakcji klientów jest definiowane poprzez dwa kluczowe wskaźniki KPI (ang. *Key Performance Indicators*) [9]. Pierwszy z nich to AR (ang. *Abandonment Rate*) definiujący odsetek połączeń porzucanych przez klientów przed połączeniem z agentem względem wszystkich połączeń odebranych. Drugim natomiast jest AO (ang. *Agent Occupancy*). Określa on procent czasu, który operatorzy poświęcają na pracę z klientami, względem całkowitego czasu ich obecności w systemie. Dodatkowo, w ramach prac uwzględniono również wskaźnik HR (ang. *Hit Rate*). Określa on odsetek połączeń odebranych przez klientów w stosunku do wszystkich połączeń wychodzących. Natomiast końcową skuteczność rozwiązania oceniano metryką SL (ang. *Service Level*), która definiowana jest jako procent połączeń zestawionych w czasie krótszym niż ustalony próg czasowy. W badaniach przyjęto 3 sekundy.

Wskaźniki AO oraz AR są najbardziej istotnymi metrykami stosowanymi w określaniu skuteczności predykcji ruchu wychodzącego [10]. Odzwierciedlają one dwa fundamentalne aspekty działania systemu, a mianowicie z jednej strony efektywność pracy agentów, a z drugiej jakość obsługi klienta. Wysoka zajętość agentów zwiększa produktywność, ale jednocześnie może prowadzić do wzrostu liczby porzuconych połączeń, jeśli system nie zapewni dostępnych agentów w odpowiednim czasie. Z kolei niski wskaźnik porzuceń świadczy o wysokim poziomie obsługi, ale może oznaczać niewykorzystany potencjał zasobów ludzkich. Dlatego też utrzymanie obu wskaźników na optymalnym poziomie jest kluczowe dla równowagi między efektywnością operacyjną, a satysfakcją klienta.

Dotychczasowe badania nad optymalizacją połączeń wychodzących w systemach CC skupiały się głównie na modelach statystycznych oraz algorytmach heurystycznych lub ewolucyjnych, które mimo niewątpliwych zalet wykazują też istotne ograniczenia w kontekście adaptacji w dynamicznych środowiskach rzeczywistych CC. Przykładowo, w pracy [11] zaprezentowano model symulacyjny mechanizmu regulacji tempa połączeń wychodzących oparty na klasycznym modelu kolejkowym typu M/G/m/m. Podejście to w miarę skutecznie odwzorowuje ogólną dynamikę systemu, ale nie wykorzystuje żadnych

mechanizmów optymalizacji, a tempo połączeń ustalane jest statycznie, na podstawie z góry określonych reguł. Brak w tym przypadku zdolności adaptacyjnej w czasie rzeczywistym ogranicza możliwość reagowania na nagłe zmiany w środowisku CC. Z kolei w pracy [12] zastosowano algorytmy genetyczne do optymalizacji działania oprogramowania nadzorującego procesy wybierania połączeń wychodzących o nazwie ASPD (ang. *Altitude Software Predictive Dialer*). Choć algorytmy te pozwalają na znalezienie lokalnie optymalnych wartości parametrów w oparciu o mechanizmy ewolucyjne, proces ich dopasowania jest zbyt czasochłonny, aby umożliwić bieżące dostosowywanie się do dynamicznych zmian środowiska. Ponadto, rozwiązanie zostało zaprojektowane dla konkretnego zastosowania komercyjnego, bez wskazania możliwości jego wdrożenia w innych tego typu systemach. W pracy [13] do modelowania procesu inicjowania połączeń wykorzystano łańcuchy Markowa w czasie ciągłym (CTMC, ang. *Continuous-Time Markov Chain*). Choć podejście to pozwala na określenie ogólnego przebiegu procesu to prognozy bazują tu wyłącznie na wartościach średnich bez uwzględniania danych historycznych. Co istotne, model zakłada stałą liczbę agentów i nie zawiera mechanizmów adaptacji, przez co jego zastosowanie w zmiennych warunkach rzeczywistych jest ograniczone. W pracy [14] zaproponowano natomiast heurystyczną optymalizację klasycznego algorytmu predykcji połączeń wychodzących, koncentrując się na doborze parametrów takich jak liczba jednoczesnych połączeń czy maksymalny czas oczekiwania klienta. Mimo uzyskania ciekawych wyników, metoda ta opiera się na manualnie dostrajanych regułach, bez wykorzystania algorytmów adaptacyjnych lub uczących się, co czyni ją mało elastyczną i nieprzystosowaną do dynamicznej natury środowisk klasy CC. W tabeli 1 zestawiono porównanie otrzymywanych wartości wskaźników AR oraz AO dla wyżej wymienionych metod.

Tab. 1. Porównanie rozwiązań znanych

Metoda	*AR*	*AO*	Lit.
M/G/m/m	5%	60-65%	[11]
Algorytmy genetyczne	3.5-5%	73-83%	[12]
CTMC	4%	70%	[13]
Heurystyka adaptacyjna	1-5%	92-93%	[14]

W analizowanych rozwiązaniach widoczna jest wyraźna zależność między wskaźnikami AR i AO. Tradycyjny model kolejkowy oparty na M/G/m/m osiąga akceptowalny poziom AR=5%, jednak kosztem niższego wykorzystania agentów, utrzymującego się w przedziale 60–65%. Zastosowanie algorytmów genetycznych w rozwiązaniu [12] pozwala na znaczne zwiększenie wskaźnika AO do 83% przy jednoczesnym utrzymaniu AR na umiarkowanym poziomie 3.5–5%. Model bazujący na łańcuchach Markowa zapewnia zrównoważone wartości: AR = 4% i AO = 70%, co wskazuje na stabilność systemu przy umiarkowanej efektywności operacyjnej. Wartości te zostały istotnie poprawione w rozwiązaniu [14], gdzie heurystyczna regulacja parametrów oprogramowania nadzorującego procesy wybierania połączeń wychodzących pozwoliła na uzyskanie bardzo wysokiego poziomu wykorzystania agentów (do 93%) przy jednoczesnym utrzymaniu AR w zakresie 1–5%. Przy czym najlepsze wyniki dla tego podejścia zaobserwowano przy liczbie agentów równej 10, gdzie AO przekroczyło 93%, a AR spadło poniżej 1%. Najsłabsze rezultaty uzyskano natomiast przy 6 agentach, gdzie AR wzrosło do około 5%. Warto jednak zaznaczyć, że rezultaty przedstawione w pracy [14] mogą być niewystarczające z punktu widzenia zastosowań praktycznych. Wynika to z faktu, że założono tu średni czas oczekiwania przez klienta na połączenie rzędu 20–30 sekund, co w rzeczywistych systemach klasy CC jest nieakceptowalne. Wyżej zdefiniowane ograniczenia były dla autorów motywacją do podjęcia dalszych prac w tym obszarze.

W niniejszej pracy zaproponowano koncepcję nowej metody predykcji ruchu wychodzącego dedykowanej na potrzeby systemów klasy CC wykorzystującej metody uczenia maszynowego. Podejście oparte zostało na modelu sieci neuronowej głębokiego uczenia ze wzmocnieniem TD3 (ang. *Twin Delayed Deep Deterministic Policy Gradient*). Według wiedzy autorów, jest to pierwsze zastosowanie tego algorytmu do sterowania predykcją połączeń wychodzących w systemie CC. Algorytm TD3 charakteryzuje się zdolnością do bieżącej adaptacji, samouczenia się na podstawie historycznych i aktualnych danych oraz podejmowania decyzji w czasie rzeczywistym – bez konieczności stosowania ręcznie definiowanych reguł [15]. Ponadto, zapewnia wysoką stabilność procesu uczenia w środowiskach podobnych do CC (o silnym szumie i nieliniowościach), jest zdolny do pracy w przestrzeni akcji rzeczywistych oraz odporny na przeszacowania funkcji Q (ang. *Action-Value Function*). Powyższe cechy wpłynęły na wybór TD3 do dalszych badań.

Wyniki badań potwierdzają, że model ten pozwala uzyskać istotną poprawę w zakresie kluczowych wskaźników operacyjnych AR oraz AO, przewyższając tym samym rozwiązania oparte na modelach statycznych, heurystykach oraz klasycznej optymalizacji.

2. METODOLOGIA

Głównym celem proponowanej metody była optymalizacja dwóch wspomnianych wskaźników KPI. W niniejszej sekcji opisano założenia modelu oraz jego strukturę.

2.1. Założenia modelu

Tab. 2. Stan środowiska dla modelu TD3

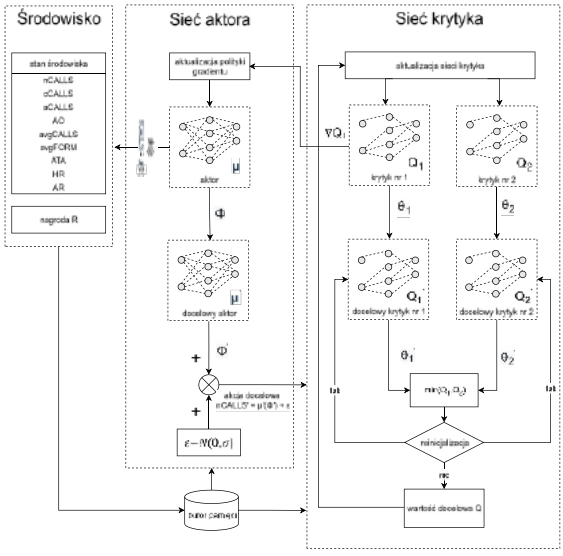
Parametr	*Opis*
nCALLS	Liczba wydzwonionych połączeń.
cCALLS	Liczba połączeń odebranych przez klientów.
aCALLS	Liczba połączeń zainicjowanych i podjętych przez agenta w czasie krótszym niż 3s.
AO	Stosunek liczby agentów zajętych rozmową do ilości wszystkich dostępnych agentów.
avgCALLS	Znormalizowana średnia długość rozmów.
avgFORM	Znormalizowana średnia długość czasu wypełniania skryptu po rozmowie.
ATA	Stosunek ilości rozmów znajdujących się w automacie do ilości dostępnych agentów.

***** wartości parametrów dotyczą aktualnego kroku czasowego

W tabeli 2 zestawiono parametry opisujące stan środowiska w danym kroku czasowym, stanowiące pierwszą część wektora danych wejściowych modelu. Ponadto, na wejście modelu TD3 podawane są również zdefiniowane wcześniej wartości wskaźników HR oraz AR dla ostatnich czterech kroków czasowych działania algorytmu. Omawiany model podejmuje decyzje w sposób ciągły, z krokiem co 5 sekund, określając liczbę połączeń, które należy zainicjować w następnym kroku czasowym.

2.2. **Struktura modelu**

Architektura zaproponowanego modelu została przedstawiona na rysunku 1. Rysunek odzwierciedla pełny cykl decyzyjny modelu TD3 począwszy od obserwacji stanu środowiska, przez wygenerowanie akcji, estymację jej wartości, aktualizację parametrów sieci, aż po wzmocnienie zwrotne ze środowiska. Poszczególne komponenty oznaczono zgodnie z przyjętą konwencją: Φ to parametry sieci aktora, θ_1 i θ_2 – parametry sieci krytyków, zaś ε – szum Gaussa używany do stabilizacji eksploracji.



Rys. 1. Architektura modelu TD3

Kluczowym elementem architektury jest agent uczący się w sposób ciągły na podstawie interakcji ze środowiskiem, którego stan modelowany jest przez realistyczny symulator systemu CC. W danym kroku czasowym agent otrzymuje wektor obserwacji, zawierający zestaw cech opisujących aktualną sytuację operacyjną systemu. Na podstawie tych danych, sieć aktora μ(Φ) generuje decyzję – akcję nCALLS, która reprezentuje liczbę połączeń wychodzących do zainicjowania w kolejnym kroku czasowym. Akcja ta trafia do środowiska, które zwraca zaktualizowany stan oraz wartość nagrody.

Model wykorzystuje architekturę typu aktor–krytyk z dwoma niezależnymi sieciami krytyków oznaczonymi jako $Q_1(\theta_1)$ oraz $Q_2(\theta_2)$. Obie sieci estymują wartość funkcji Q, czyli tzw. funkcję wartości akcji, która określa oczekiwaną sumę przyszłych nagród, jaką agent może uzyskać podejmując daną akcję w określonym stanie. Do aktualizacji parametrów wykorzystywana jest wartość minimalna spośród dwóch estymacji Q_1 i Q_2, co przeciwdziała przeszacowaniu wartości akcji, zgodnie z założeniami algorytmu TD3. Dla zwiększenia stabilności procesu uczenia zastosowano również docelowe kopie sieci: $Q_1'(\theta_1')$, $Q_2'(\theta_2')$ oraz $\mu'(\Phi')$. Sieć docelowego aktora $\mu'(\Phi')$ generuje akcję nCALLS′, do której dodawany jest szum Gaussowski $\varepsilon \sim \mathcal{N}(0, \sigma)$, wspomagający eksplorację przestrzeni akcji. Ostateczna akcja docelowa $nCALLS' = \mu'(\Phi') + \varepsilon$ przekazywana jest do docelowych krytyków Q_1' i Q_2' w celu oszacowania jej jakości w kolejnym stanie środowiska.

Uczenie aktora opiera się na sygnale gradientowym ∇Q_1, wyznaczanym względem akcji przez krytyka Q_1. Gradient ten informuje, jak zmienić parametry polityki, by generować akcje o wyższej oczekiwanej wartości Q. Funkcja nagrody R została zdefiniowana jako:

$$R = w_{AO} AO + w_{AR} AR \tag{1}$$

gdzie: w_{AO} – waga zajętości agentów CC; w_{AR} – waga współczynnika porzuceń.

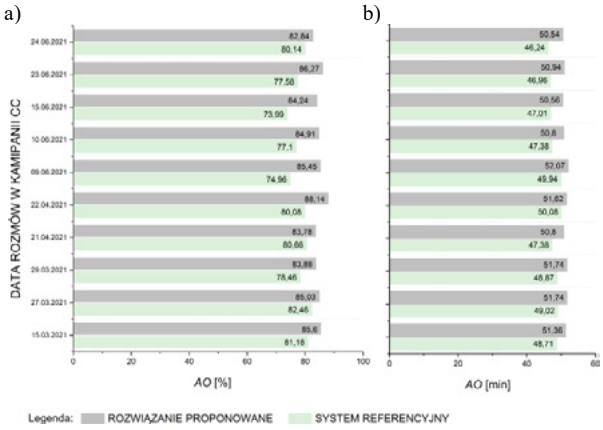
Jeśli którykolwiek ze wskaźników przekracza założone progi jakościowe (AR > 3% lub AO < 70%), agent otrzymuje ujemne wzmocnienie równe –1, co wpływa negatywnie na ocenę danego działania i zmienia jego dalszą strategię. Proces dostrajania hiperparametrów przeprowadzony został manualnie w ramach serii eksperymentów. Najlepsze wartości uzyskano dla: train_freq=(5, "step"), tau = 0.001, sigma = 0.15, batch_size = 64 i policy_delay = 3. Model trenowano w trybie offline na podstawie danych rzeczywistych, pozyskanych z dużego, komercyjnego systemu CC. Dzięki opracowaniu autorskiego, realistycznego symulatora możliwe było bezpieczne dostrajanie polityki agenta przed jej wdrożeniem do środowiska produkcyjnego.

3. **WYNIKI**

Rozwiązanie badano z wykorzystaniem danych historycznych ruchu wychodzącego z okresu czterech miesięcy działania rzeczywistego systemu CC (od marca do czerwca 2021 roku). W ocenie skuteczności przyjęto najczęściej stosowany wskaźnik SL, gdzie założono, że 98% rozmów powinno być zestawionych w ciągu 3 pierwszych sekund. Parametry te determinowały wartości wag, jak i graniczne wartości dla wskaźników AR oraz AO w ustalaniu wielkości wzmocnienia. Przyjęto wartości wag w_{AO} = 0,5 i w_{AR} = -1,5. Badania przeprowadzono dla ilości agentów CC równej 24.

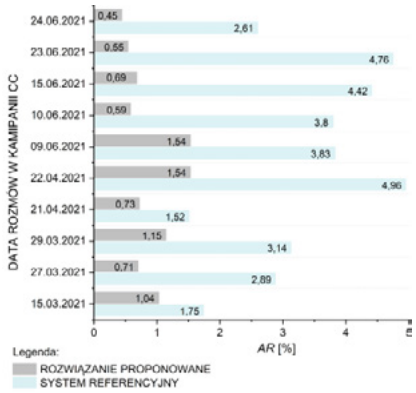
W celu porównania wyników uzyskanych przy wykorzystaniu opisywanego algorytmu z rzeczywistymi obciążeniami agentów pozyskano informacje dotyczące czasu zajętości i czasu rozmowy wszystkich agentów, zaangażowanych w daną usługę. Następnie, dla przedziałów godzinowych obliczono średni czas zajętości i średni czas rozmowy wszystkich agentów przypisanych do rozpatrywanej usługi w danej godzinie. Procentowy czas zajętości i procentowy czas rozmowy zostały następnie przeliczone na wartości definiowane w minutach na godzinę. Porównanie wyników rzeczywistych z proponowaną metodą zawiera rysunek 2.

Analiza prezentowanych wyników wskazuje na znaczną przewagę rozwiązania proponowanego w stosunku do rozwiązania systemu referencyjnego. Rozwiązanie referencyjne stanowią wartości AO oraz AR rzeczywiście uzyskane w dużym systemie CC dla tych samych danych, które przetwarzane były przez model TD3. W przypadku reprezentacji w ujęciu procentowym (Rys. 2a), model TD3 bardzo dobrze odwzorowuje zmienność zajętości agentów w czasie, co świadczy o skutecznym dostosowaniu się do rytmu operacyjnego systemu. Analogiczne obserwacje potwierdza rysunek 2b, na którym porównano rzeczywisty oraz przewidywany czas pracy agentów przeliczony na minuty.



Rys. 2. Porównanie wartości wskaźnika AO a) wartości procentowe, b) wartości w jednostce czasu

Widoczna zbieżność wskazuje, że agent był w stanie uczyć się na podstawie danych historycznych i generować strategię, która odpowiadała rzeczywistemu zapotrzebowaniu. We wszystkich analizowanych dniach pracy proponowany model wykazywał wyższe wartości AO niż otrzymane w oparciu o dane rzeczywiste, co przekłada się na lepsze wykorzystanie zasobów kadrowych oraz wyższą efektywność operacyjną systemu. Osiągnięty poziom AO przewyższa też wartości podawane w literaturze prezentującej wcześniejsze badania w tym obszarze. Dla przedstawionych wyników obliczano również wartość wskaźnika AR. Wyniki przedstawiono na rysunku 3.



Rys. 3. Porównanie wartości wskaźnika AR

Wartości AR utrzymują się na niskim i stabilnym poziomie, w zdecydowanej większości przypadków nie przekraczając progu 1% oraz są znacznie lepsze niż w rozwiązaniu referencyjnym. Potwierdza to skuteczność metody w ograniczeniu sytuacji, w których klient nie został połączony z dostępnym agentem. Jest to szczególnie istotne z perspektywy zapewnienia wysokiej jakości obsługi klienta. Proponowane rozwiązanie pozwala na pełne spełnienie wymagań SL.

4. WNIOSKI

W odróżnieniu od dotychczas stosowanych metod, takich jak modele statystyczne czy algorytmy genetyczne, zaproponowane rozwiązanie wykazuje się zdolnością do adaptacji w czasie rzeczywistym, a także umożliwia podejmowanie decyzji w sposób ciągły, co lepiej odpowiada rzeczywistym warunkom operacyjnym. Wysoka skuteczność opracowanego modelu została potwierdzona w eksperymentach opartych na rzeczywistych danych historycznych, gdzie osiągnięto bardzo dobre średnie wartości KPI równe AO=84,6% oraz AR=0,9%. Wartością dodaną jest autorski, realistyczny symulator ruchu wychodzącego, umożliwiający bezpieczne i efektywne trenowanie agenta przed jego wdrożeniem w środowisku rzeczywistym. Zaproponowane podejście jest skalowalne, uniwersalne i niezależne od konkretnej architektury systemu, tym samym w ocenie autorów może być postrzegane jako atrakcyjna propozycja dla implementacji w nowoczesnych środowiskach CC.

Praca powstała w ramach projektu EMTICA AI, współfinansowanego przez Narodowe Centrum Badań i Rozwoju z Programu Operacyjnego Inteligentny Rozwój w ramach poddziałania 4.1.4. Projekty aplikacyjne. Numer umowy: POIR.04.01.04-00-0079/19

LITERATURA

[1] Płaza, M., Pawlik, Ł., & Deniziak, R. S. 2021. „Call transcription methodology for Contact Center systems". *IEEE Access*.

[2] Płaza, M., Trusz, S., Kęczkowska, J., Boksa, E., Sadowski, S., & Koruba, Z. 2022. „Machine Learning Algorithms for Detection and Classification of Emotions in Contact Center Applications". *Sensors* , 22(14): 5311.

[3] Płaza, M., Kazała, R., Koruba, Z., Kozłowski, M., Lucińska, M., Sitek, K., & Spyrka, J. 2022. „Emotion Recognition Method for Call/Contact Centre Systems". *Applied Sciences*, 12(21): 10951.

[4] Pawlik, Ł., Płaza, M., Deniziak, R. S., & Boksa, E. 2022. „A method for improving bot effectiveness by recognising implicit customer intent in contact centre conversations". *Speech Communication*, 143: 33–45.

[5] Poczęta, K., Płaza, M., Michno, T., Krechowicz, M., & Zawadzki, M. 2023. „A multi-label text message classification method designed for applications in call/contact centre systems". *Applied Soft Computing*, 145.

[6] Poczęta, K., Płaza, M., Zawadzki, M., Michno, T., & Krechowicz, M. 2024. „Analysis of the retraining strategies for multi-label text message classification in call/contact center systems". *Scientific Reports*, 14: 1–16.

[7] Asanjarani, A., Nazarathy, Y., & Taylor, P. 2020. „A Survey of Parameter and State Estimation in Queues". arXiv:2012.14614.

[8] Escalante, F. 2008. „Effective Call Center Management: Evidence from Financial Services". *J. of Finan. Serv. Mark.*, 13(1): 35–50.

[9] Płaza, M., & Pawlik, Ł. 2021. „Influence of the Contact Center Systems Development on Key Performance Indicators". *IEEE Access*, 9.

[10] Gans, N., Koole, G., & Mandelbaum, A. 2003. „Telephone call centers: Tutorial, review, and research prospects". *Manufacturing & Service Operations Management*, 5(2): 79–141.

[11] Anisimov, N., Korolev, N., & Ristock, H. 2008. „Modeling and simulation of a pacing engine for proactive campaigns in contact center environment". *Proceedings of the 2008 Spring Simulation Multiconference*, San Diego.

[12] Amaral, P. M., & Vital, M. M. 2014. „Predictive dialer intensity optimization using genetic algorithms". *International Journal of Machine Learning and Computing*, 4(3): 286–291.

[13] Nadeeka Dilini, W. M., Karunaratne, A., & Ranasinghe, D. N. 2011. „Methods for controlling the call placement ratio for outbound dialing of a call center". *Statistical Concepts and Methods for the Modern World Conference*, Colombo, 1–12.

[14] Fourati, S., & Tabbane, S. 2010. „Optimization of a predictive dialing algorithm". *2010 Sixth Adv. Int. Conf. on Telecomm.*, Barcelona: 212–218.

[15] Fujimoto, S., Van Hoof, H., & Meger, D. 2018. „Addressing Function Approximation Error in Actor-Critic Methods". *Proc. 35th Int. Conf. on Mach. Lear. (ICML)*, Stockholm, 1587–1596.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

NOWY ALGORYTM PRZESTROJEŃ DLA PÓL KOMUTACYJNYCH TYPU SWS1

A NEW REARRANGEMENT ALGORITHM FOR SWS1-TYPE SWITCHING FABRICS

Grzegorz Danilewicz[1]; Remigiusz Rajewski[2]

[1] Politechnika Poznańska, Instytut Sieci Teleinformatycznych, Poznań, grzegorz.danilewicz@put.poznan.pl

[2] Politechnika Poznańska, Instytut Sieci Teleinformatycznych, Poznań, remigiusz.rajewski@put.poznan.pl

DOI: 10.15199/59.2025.4.79

Streszczenie: Niedawno Bey-Chi Lin opublikowała pracę poświęconą przestrajalnym i przepakowywalnym polom komutacyjnym, w których można realizować połączenia wieloszczelinowe, z zachowaniem wymogu występującego w elastycznych sieciach optycznych – połączenia wieloszczelinowe muszą być realizowane w sąsiednich szczelinach. Ten artykuł jest komentarzem do wspomnianej pracy dotyczącym jednego z przedstawionych w niej aspektów – przestrojeń połączeń jednoszczelinowych w polach typu SWS1 (ang. *Space-Wavelength-Space*). W niniejszym referacie wykazaliśmy, iż podejście do realizacji przestrojeń w polach SWS1 zaprezentowane w cytowanej pracy jest poprawne z matematycznego punktu widzenia, ale jest równocześnie technicznie nieużyteczne. Ponadto zaproponowaliśmy algorytm przestrojeń, który umożliwia wykorzystanie pól komutacyjnych SWS1 w praktycznych strukturach węzłów elastycznych sieci optycznych.

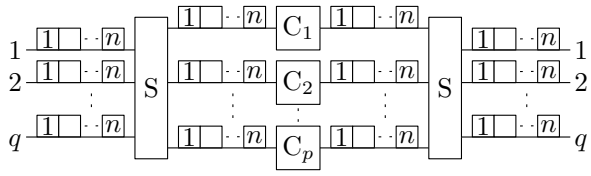
Abstract: Recently, Bey-Chi Lin published a paper on rearrangeable and repackagable switching fabrics capable of supporting multi-slot connections, addressing a key requirement in elastic optical networks where such connections must occupy contiguous slots. This article serves as a commentary on that work, focusing specifically on the rearrangement of single-slot connections within SWS1-type fabrics. We demonstrate that, while the rearrangement method proposed for SWS1 fabrics in the original paper is mathematically sound, it proves to be technically impractical. Furthermore, we introduce a new rearrangement algorithm that facilitates the practical deployment of SWS1 switching fabrics in the node architectures of elastic optical networks.

Słowa kluczowe: elastyczne sieci optyczne, kolorowanie grafów, przestrajalne optyczne pola komutacyjne, space-wavelength-space switching fabrics.

Keywords: elastic optical networks, graph coloring, repackable optical switching fabrics, space-wavelength-space switching fabrics

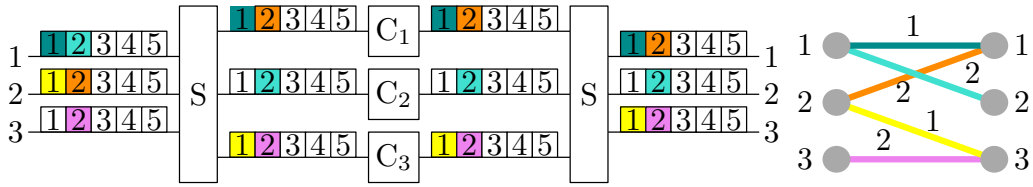
1. WSTĘP

W [10] zostały opublikowane warunki przestrajalności i przepakowalności pól komutacyjnych typu SWS1 dla połączeń wieloszczelinowych realizowanych w sąsiednich szczelinach. Architektura pola komutacyjnego SWS1 została zaproponowana w [4] jako struktura umożliwiająca kierowanie m-połączeń (połączeń wieloszczelinowych w m sąsiednich szczelinach) z dowolnego traktu wejściowego do dowolnego traktu wyjściowego, a dodatkowo z konwersją długości fali. Połączenia takie są realizowane w węzłach elastycznych sieci optycznych EON (ang. *Elastic Optical Networks*) [5, 8]. Architektura pola SWS1 została przedstawiona na rys. 1.



Rysunek 1: Architektura pola SWS1

Pole SWS1 składa się z komutatora przełączającego m-połączenie w przestrzeni, z dowolnego z q traktów wejściowych do dowolnego z p łączy międzysekcyjnych między pierwszą a drugą sekcją pola. Drugi komutator przestrzenny, umieszczony w trzeciej sekcji przełącza m-połączenie z dowolnego z p łączy międzysekcyjnych do dowolnego z q traktów wyjściowych. W sekcji środkowej znajduje się p komutatorów, które dokonują przełączenia m-połączenia z wejścia na wyjście z możliwością konwersji długości fali, to znaczy, że zestaw m szczelin w połączeniu na wejściu może różnić się od zestawu m szczelin na wyjściu komutatora. Warunkiem koniecznym poprawnego działania pola komutacyjnego SWS1 w strukturze węzła w sieci EON jest zachowanie integralności m-połączenia, to znaczy, że w każdym łączu pola m-połączenie musi być zrealizowane w m sąsiednich szczelinach. W każdym łączu pola SWS1 liczba szczelin jest równa n. Zestawienie nowego m-połączenia w polu SWS1 polega na znalezieniu komutatora środkowej sekcji, który ma równocześnie wolne m sąsiednich szczelin na wejściu, a szczeliny te mają te same indeksy jak szczeliny w trakcie wejściowym oraz wolne m sąsiednich szczelin na wyjściu, a ich indeksy są takie same jak indeksy szczelin żądanych w trakcie wyjściowym. W ogólności w literaturze rozważane są pola SWS1, w których można realizować m-połączenia, gdzie $1 \leqslant m \leqslant m_{max} \leqslant n$ [2, 3, 4, 10].



Rysunek 2: Graf dwudzielny używany w [10]

W teorii komutacji rozważane są struktury pól komutacyjnych, które są nieblokowalne w wąskim sensie (ang. *Strict-sense Nonblocking* – SNB) [1, 9]. W polach takich zawsze można zestawić połączenie między wolnym wejściem, a wolnym wyjściem niezależnie od używanego algorytmu wyboru drogi połączeniowej. Pola nieblokowalne w szerokim sensie (ang. *Wide-sense Nonblocking* – WNB) [6, 9] umożliwiają zestawienie połączenia między wolnym wejściem a wolnym wyjściem, ale pod warunkiem zastosowania odpowiedniego algorytmu wyboru drogi połączeniowej. W polach przestrajalnych (ang. *Rearrangeable Nonblocking* – RNB) może być wymagane rozłączenie istniejących w polu połączeń i ich ponowne zestawienie, tak aby znaleźć drogę połączeniową w polu dla nowego połączenia. Pola RNB zazwyczaj wymagają mniejszej liczby komutatorów niż pola SNB i WNB, ale ze względu na konieczność zastosowania algorytmu przestrojeń wymagają one bardziej skomplikowanego sterowania [6, 9]. Pola przepakowywalne (ang. *Repackable Nonblocking* – RPNB) działają na podobnej zasadzie jak pola przestrajalne, przy czym przepakowanie połączeń odbywa się po rozłączeniu istniejącego wcześniej w polu połączenia [7].

Warunki SNB dla pól SWS1 i dla m-połączeń, $1 \leqslant m \leqslant m_{max} \leqslant n$, zostały podane w [4]. Warunki SNB i WNB dla asymetrycznych pól SWS1 zostały zaprezentowane w [2]. W [10] został podany warunek przestrojeń dla pól SWS1, udowodniony za pomocą teorii kolorowania grafów dwudzielnych oraz algorytm zestawiania połączeń dla połączeń jednoszczelinowych ($m_{max} = 1$). Mimo podania warunku przestrajalności w [10] nie zdefiniowano algorytmu przestrojeń. W niniejszej pracy wykażemy, że podejście zaprezentowane w [10] jest błędne z punktu widzenia technicznego zastosowania przestrajalnych pól SWS1 w węzłach EON. Dodatkowo w referacie zostanie zaproponowany algorytm przestrojeń, który przy niezmienionym warunku przestrojeń umożliwi techniczne zastosowanie przestrajalnych pól SWS1 w węzłach EON.

W sekcji 2 zostaną przedstawione warunki przestrajalności pól SWS1 wprowadzone w [10], a w sekcji 3 skomentowano techniczną nieadekwatność podejścia przedstawionego w [10] do zestawiania połączeń w polach SWS1. W sekcjach 4 oraz 5 zaprezentowano nowe podejście do realizacji przestrojeń w polach SWS1. Na końcu umieszczono wnioski.

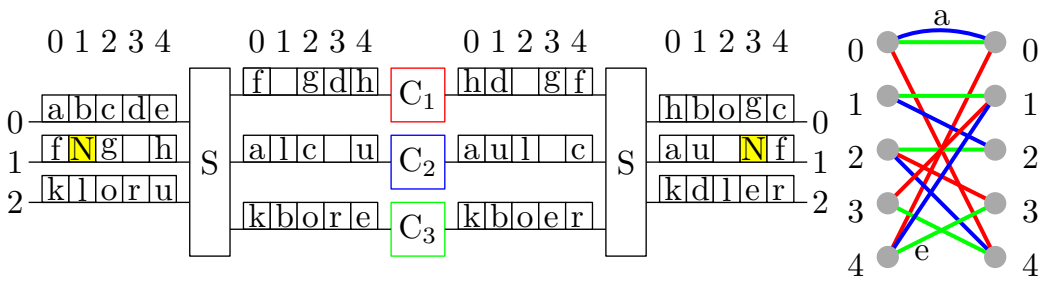
2. ISTNIEJĄCE WARUNKI PRZESTROJEŃ

W [10] zaprezentowano warunek przestrajalności pól SWS1 dla połączeń jednoszczelinowych ($m_{max} = 1$) jako: $p \geqslant q$. Zależność ta została udowodniona za pomocą teorii kolorowania krawędzi grafu dwudzielnego. Autorka stosuje graf dwudzielny, którego jeden zbiór q wierzchołków (oznaczanych przez i, gdzie $1 \leqslant i \leqslant q$) odpowiada q traktom wejściowym, a drugi zbiór q wierzchołków (oznaczanych przez j, gdzie $1 \leqslant j \leqslant q$) odpowiada q traktom wyjściowym. W oczywisty sposób zbiory tych wierzchołków są rozłączne. Na rysunku 2 przedstawiono przykład połączeń rozważany w [10]. Połączenia w polu są zestawiane za pomocą algorytmu rutingu, który wykorzystuje kolorowanie krawędzi grafu dwudzielnego. Przykład grafu dwudzielnego na podstawie którego nastąpiło zestawienie połączeń w polu także został pokazany na rysunku 2.

W przykładowym grafie kolorowanie krawędzi reprezentowane jest przez widoczne przy krawędziach numery (a nie przez kolory krawędzi, które służą tylko do graficznego powiązania krawędzi grafu i szczelin w polu z rys. 2). Numer każdej krawędzi wychodzącej z węzła ze zbioru reprezentującego łącza wejściowe oraz każdej krawędzi wchodzącej do węzła ze zbioru reprezentującego łącza wyjściowe musi być unikalny w ramach węzła. Każdy taki numer reprezentuje szczelinę w trakcie wejściowym, wyjściowym, i w łączach międzysekcyjnych. Oznacza to, że w każdym węźle zaproponowanego w [10] grafu można użyć maksymalnie n różnych numerów krawędzi. Stąd stopień grafu $\Delta = n$. Równocześnie oznacza to, że połączenie z traktu wejściowego i do traktu wyjściowego j zestawiane jest przez szczelinę o numerze c, $1 \leqslant c \leqslant n$ wskazaną w grafie. Według algorytmu zestawiania połączeń zaproponowanego w [10] o wyborze szczeliny decyduje dostępność wolnych szczelin w trakcie wyjściowym żądanym w połączeniu. W przykładzie z rys. 2 o wyborze szczeliny 2 w połączeniu między traktem wejściowym (węzłem) $i = 3$ i traktem wyjściowym (węzłem) $j = 3$ zadecydował fakt, że szczelina 1 była już używana w trakcie wyjściowym.

3. NIEADEKWATNOŚĆ ISTNIEJĄCEGO ROZWIĄZANIA

Ze względu na przyjęte w [10] założenie, że to dostępność wolnej szczeliny w trakcie wyjściowym decyduje o numerze szczeliny, która zostanie użyta do zestawienia nowego połączenia w całym polu komutacyjnym powoduje, że nowe połączenie jest zestawiane w szczelinie o tym samym numerze zarówno w trakcie wyjściowym jak i w łączach międzysekcyjnych, a ponadto także w trakcie wejściowym pola (patrz rys. 2). Przyjęcie tego założenia ma kilka znaczących konsekwencji. Po pierwsze doprowadziło ono Autorkę artykułu [10] do słusznego (z punktu widzenia przyjętego założenia) wniosku, że sekcja zawierająca komutatory z konwersją nie są w polu potrzebne (z przykładu z rys. 2 można zauważyć, że dla żadnego z połączeń oznaczonego tym samym kolorem konwersja numeru szczeliny nie ma miejsca). Dalszą konsekwencją jest konkluzja Autorki, że nie ma potrzeby używania pola SWS1, gdyż do zestawiania połączeń bez konwersji wystarczy tylko komutator przestrzenny o pojemności $q \times q$. Drugą konsekwencją niezauważoną przez Autorkę jest fakt, że gdy w sieci EON będzie konieczność realizacji połączenia w określonej relacji, które wymaga konwersji długości fali w węź-



Rysunek 3: Przykład połączeń w polu SWS1 (numery szczelin zaznaczono na górze rysunku; literami alfabetu zaznaczono zrealizowane połączenia; literą N na żółtym tle oznaczono nowe połączenie do zestawienia w polu) oraz przykład grafu dwudzielnego (kolory krawędzi odpowiadają kolorom komutatorów środkowej sekcji; literami alfabetu zaznaczono przykładowe połączenia w polu SWS1)

le EON, to przy proponowanym w [10] rozwiązaniu nie da się go zrealizować bez konieczności wprowadzenia dodatkowej konwersji sygnału w węźle. Kolejną konsekwencją wynikającą z podejścia zaprezentowanego w [10] jest odejście od paradygmatu realizacji połączeń w węzłach sieci, który jest powszechnie wykorzystywany w teorii komutacji, ale wynika z rozwiązań praktycznych. Działanie pól komutacyjnych jest powszechnie analizowane przy założeniu, że nowe połączenie pojawia się na konkretnym wejściu pola (na przykład wskutek działania algorytmu sterującego pracą węzła) i żąda konkretnego wyjścia pola (kierunku, w którym realizowane jest połączenie). W zamian tego rozwiązanie zaproponowane w [10] wymusza zestawianie nowego połączenia między wejściem i wyjściem, które są wymuszane przez algorytm wyboru drogi połączeniowej w polu. Takie rozwiązanie jest niestety nieużyteczne w praktycznych zastosowaniach w telekomunikacji.

4. GRAFOWE PODSTAWY NOWEGO ROZWIĄZANIA

Jednoszczelinowe połączenie między traktem wejściowym i, $0 \leqslant i \leqslant q-1$, pola SWS1 i szczeliną s, $0 \leqslant s \leqslant n-1$, tego traktu, a traktem wyjściowym j, $0 \leqslant j \leqslant q-1$, i szczeliną t, $0 \leqslant t \leqslant n-1$ tego traktu będzie oznaczane jako $\mathcal{C}(i[s], j[t])$. Zdefiniujmy graf dwudzielny G, który składa się ze zbioru n węzłów odpowiadających n szczelinom w traktach wejściowych pola SWS1 oraz z n węzłów odpowiadających n szczelinom w traktach wyjściowych pola SWS1. Węzły wejściowe będą oznaczane przez s, $0 \leqslant s \leqslant n-1$, a węzły wyjściowe przez t, $0 \leqslant t \leqslant n-1$. Do zestawiania i przestrajania połączeń w polu SWS1 zostanie wykorzystane kolorowanie krawędzi grafu dwudzielnego. Graf musi być tak pokolorowany, aby każda krawędź wychodząca z pojedynczego węzła ze zbioru węzłów wejściowych miała inny kolor, oraz żeby każda krawędź wchodząca do węzła ze zbioru węzłów wyjściowych także miała inny kolor. Każda z krawędzi grafu G reprezentuje wybór jednego z komutatorów sekcji środkowej (z konwersją), przez który zestawiane jest połączenie $\mathcal{C}(i[s], j[t])$. Szczelina o tym samym numerze w traktach wejściowych może być użyta tylko w q różnych połączeniach. Podobnie szczelina o tym samym numerze w traktach wyjściowych może być użyta także tylko q razy. Oznacza to, że z każdego węzła wejściowego może wychodzić q krawędzi, a każda z nich musi mieć inny kolor, oraz że do każdego węzła wyjściowego może wchodzić q krawędzi i każda z nich musi być w innym kolorze. W konsekwencji najwyższy stopień grafu G jest równy $\Delta(G) = q$. Z teorii grafów [11] wiadomo, że do kolorowania krawędzi grafu dwudzielnego o stopniu q wystarczy q kolorów. Ponieważ każda krawędź grafu G reprezentuje jeden komutator sekcji środkowej, a do pokolorowania krawędzi tego grafu wystarczy $\Delta(G) = q$ kolorów, to oznacza, że pole SWS1 jest przestrajalne, gdy $p \geqslant q$. Potwierdza to warunek przestrajalności podany w [10].

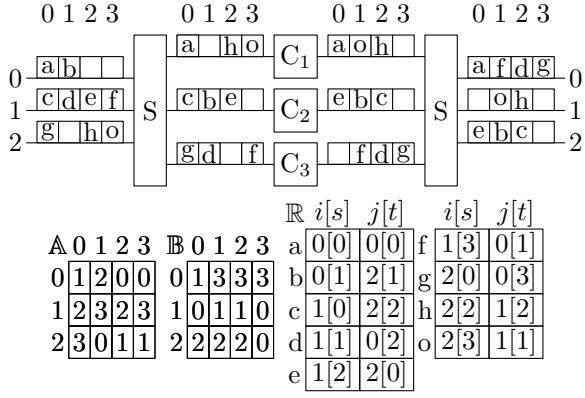
Na rysunku 3 przedstawiono przykład połączeń w polu SWS1 z $q = p = 3$ i $n = 5$. Numery szczelin w poszczególnych łączach oznaczono na górze rysunku. Małymi literami alfabetu zaznaczono połączenia zestawione już w polu. Wielką literą N na żółtym tle zaznaczono nowe połączenie $\mathcal{C}(1[1], 1[3])$, które ma być w polu zestawione. Na rysunku zaprezentowano także graf dwudzielny, który odzwierciedla bieżący stan w polu komutacyjnym z rysunku. Krawędzie grafu są pokolorowane kolorami odpowiadającymi kolorom komutatorów w sekcji środkowej pola z rys. 3. Na rysunku z grafem zaznaczono dodatkowo przykładowe połączenia (a i e), które przypisano do odpowiednich krawędzi. Połączenie a ($\mathcal{C}(0[0], 1[0])$) zestawione jest przez komutator C_2 oznaczony kolorem niebieskim. Z kolei połączenie e ($\mathcal{C}(0[4], 2[3])$) zestawione jest przez komutator C_3 oznaczony kolorem zielonym.

Konieczność przestrojeń wynika z pojawiających się stanów blokady w polu SWS1 z $p = q$, które należy rozwiązać przez przestrojenie istniejących w polu połączeń. Na rysunku 3 pokazano taki stan blokady dla nowego połączenia $\mathcal{C}(1[1], 1[3])$ (oznaczonego przez N). W łączach międzysekcyjnych między pierwszą a drugą sekcją wolna szczelina o numerze 1 jest tylko w komutatorze C_1, podczas gdy w łączach międzysekcyjnych między sekcją środkową a trzecią wolna szczelina o numerze 3 jest tylko w komutatorze C_2. Nie jest możliwe znalezienie wspólnego komutatora środkowej sekcji przez który można zestawić nowe połączenie N. Sytuację blokady można zaobserwować także w grafie dwudzielnym z rys. 3. Nowe połączenie wymaga znalezienia koloru dla krawędzi między węzłem wejściowym 1 (szczelina numer 1), a węzłem wyjściowym 3 (szczelina numer 3). Dla obecnego stanu pola nie można znaleźć koloru dla nowej krawędzi w taki sposób aby nie naruszyć zasady, że w każdym węźle grafu G mogą spotykać się tylko krawędzie o różnych kolorach. Z węzła wejściowego 1 wychodzą krawędzie o kolorach zielonym i niebieskim, a do węzła wyjściowego 3 dochodzą krawędzie w kolorach czerwonym i zielonym. W takiej sytuacji należy dokonać przekolorowania krawędzi grafu G, które doprowadzi do sytuacji, że warunek spotykania się krawędzi o różnych kolorach zostanie spełniony. W teorii do przestrojeń w polu SWS1 można wykorzystać dowolny algorytm kolorowania krawędzi grafu dwudzielnego.

5. NOWY ALGORYTM PRZESTROJEŃ

W sekcji zostanie zaproponowane rozwiązanie, które respektuje zasady zestawiania nowych połączeń między wskazanym wejściem i wskazanym wyjściem w przestrajalnych polach SWS1. Działanie algorytmu przestrojeń zostanie oparte na wykorzystaniu dwóch macierzy oraz rejestru połączeń w polu SWS1. Niech $\mathbb{A} = [a_{is}]_{q\times n}$ oznacza macierz wejść do pola SWS1 z q traktami wejściowymi i n szczelinami w trakcie. Numer wiersza i, $0 \leqslant i \leqslant q-1$, odpowiada numerowi traktu wejściowego, a numer kolumny s, $0 \leqslant s \leqslant n-1$, odpowiada indeksowi szczeliny w trakcie. Niech $\mathbb{B} = [b_{jt}]_{q\times n}$ oznacza macierz wyjść z pola SWS1 z q traktami wyjściowymi i n szczelinami w trakcie. Numer wiersza j, $0 \leqslant j \leqslant q-1$, odpowiada numerowi traktu wyjściowego, a numer kolumny t, $0 \leqslant t \leqslant n-1$, odpowiada indeksowi szczeliny w trakcie. Zestawienie nowego połączenia $\mathcal{C}(i[s], j[t])$ w polu SWS1 przez komutator sekcji środkowej k, $1 \leqslant k \leqslant p$, wymaga równoczesnych wpisów w macierzach $\mathbb{A}$ i $\mathbb{B}$ takich, że $a_{is} = k$ i $b_{jt} = k$. Można zauważyć, że różne połączenia zestawiane ze szczelin z tego samego traktu wejściowego oraz zestawiane do szczelin z tego samego traktu wyjściowego mogą być realizowane w ogólności przez ten sam komutator sekcji środkowej. Stąd numery komutatorów sekcji środkowej mogą się powtarzać w pojedynczych wierszach macierzy $\mathbb{A}$ i $\mathbb{B}$. Z kolei połączenia ze szczelin o tym samym indeksie, ale zestawiane z różnych traktów wejściowych muszą być realizowane przez różne komutatory sekcji środkowej. Podobnie połączenia kierowane do szczelin o tych samych indeksach w różnych traktach wyjściowych muszą być zestawiane przez różne komutatory sekcji środkowej. Oznacza to, że numery k, $1 \leqslant k \leqslant p$, w każdej kolumnie zarówno macierzy $\mathbb{A}$ jak i $\mathbb{B}$ muszą być różne.

W macierzach $\mathbb{A}$ i $\mathbb{B}$ elementy o wartościach 0 wskazują, że dana szczelina w trakcie nie jest obecnie używana do zestawiania połączenia. Do przechowywania informacji o zestawionych w polu połączeniach wykorzystywany jest rejestr $\mathbb{R}$, który przechowuje informacje o wejściu (w postaci $i[s]$), z którego jest zestawione połączenie do określonego wyjścia (w postaci $j[t]$). Na rysunku 4 przedstawiono stan połączeń w polu SWS1 z $q = p = 3$ i $n = 4$ i odpowiadające mu macierze $\mathbb{A}$ i $\mathbb{B}$ oraz rejestr $\mathbb{R}$.



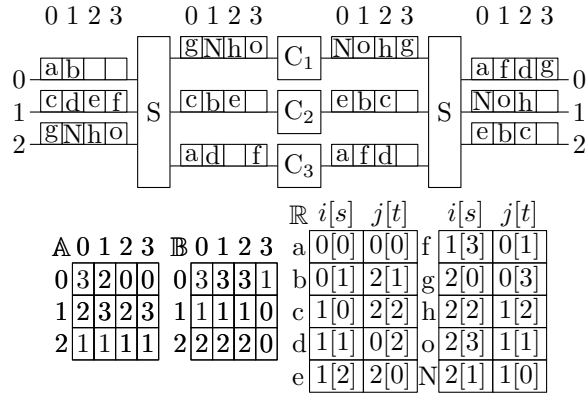
Rysunek 4: Przykład macierzy $\mathbb{A}$, $\mathbb{B}$ i rejestru $\mathbb{R}$ dla stanu połączeń w polu SWS1

Załóżmy, że w polu z rys. 4 ma być zestawione nowe połączenie $\mathcal{C}(2[1], 1[0])$. Oznaczmy przez $\mathbb{U}^s_{in}$ zbiór numerów komutatorów środkowej sekcji, które nie są wpisane w kolumnie s macierzy $\mathbb{A}$, a przez $\mathbb{U}^t_{out}$ zbiór numerów komutatorów środkowej sekcji, które nie są wpisane w kolumnie t macierzy $\mathbb{B}$. Dla połączenia $\mathcal{C}(2[1], 1[0])$ mamy $\mathbb{U}^1_{in} = \{1\}$ i $\mathbb{U}^0_{out} = \{3\}$. Ponieważ $\mathbb{U}^1_{in} \cap \mathbb{U}^0_{out} = \emptyset$, to nowe połączenie jest zablokowane, ponieważ nie można wybrać wspólnego komutatora środkowej sekcji do zestawienia tego połączenia. W takiej sytuacji należy przestroić istniejące w polu połączenia i uwolnić zasoby do zestawienia nowego połączenia. Algorytm przestrojeń został przedstawiony na rys. 5.

Algorithm: Set up and rearrangement connections
Input: $\mathbb{A}$, $\mathbb{B}$, $\mathbb{R}$, $\mathcal{C}(i[s], j[t])$
1: **if** $\mathbb{U}^s_{in} \cap \mathbb{U}^t_{out} = \emptyset$ **then**
set up $\mathcal{C}(i[s], j[t])$ through C_k, $k \in \mathbb{U}^s_{in}$
update_matrices($\mathbb{A}$, $\mathbb{B}$, $\mathbb{R}$, i, s, j, t, k) and **goto** 2
else goto 5
2: find $b_{j't} = k$, $0 \leqslant j' \leqslant q-1$, $j' \neq j$
find $\mathcal{C}'\left(i'[s'], j'[t]\right)$ from $\mathbb{R}$
$\mathcal{C}(i[s], j[t]) := \mathcal{C}'\left(i'[s'], j'[t]\right)$ and **goto** 3
3: **if** $\mathbb{U}^s_{in} \cap \mathbb{U}^t_{out} = \emptyset$ **then**
set up $\mathcal{C}(i[s], j[t])$ through C_k, $k \in \mathbb{U}^t_{out}$
update_matrices($\mathbb{A}$, $\mathbb{B}$, $\mathbb{R}$, i, s, j, t, k) and **goto** 4
else goto 5
4: find $a_{i's} = k$, $0 \leqslant i' \leqslant q-1$, $i' \neq i$
find $\mathcal{C}'\left(i'[s], j'[t']\right)$ from $\mathbb{R}$
$\mathcal{C}(i[s], j[t]) := \mathcal{C}'\left(i'[s], j'[t']\right)$ and **goto** 1
5: set up $\mathcal{C}(i[s], j[t])$ through C_k, $k \in \mathbb{U}^s_{in} \cap \mathbb{U}^t_{out}$
update_matrices($\mathbb{A}$, $\mathbb{B}$, $\mathbb{R}$, i, s, j, t, k) and **stop**
Procedure: update_matrices
Input: $\mathbb{A}$, $\mathbb{B}$, $\mathbb{R}$, i, s, j, t, k
$a_{is} := k$, $b_{jt} := k$
if absent **then** insert $i[s], j[t]$ into $\mathbb{R}$
return

Rysunek 5: Algorytm zestawiania i przestrajania połączeń w polu SWS1

Na rysunku 6 przedstawiono efekt działania zaproponowanego algorytmu przestrojeń w celu zestawienia pierwotnie zablokowanego połączenia $\mathcal{C}(2[1], 1[0])$ (oznaczonego przez N). Zestawienie nowego połączenia wymagało przestrojenia połączeń a z komutatora C_1 na C_3 oraz g z C_3 na C_1. Dało to możliwość zestawienia połączenia $\mathcal{C}(2[1], 1[0])$ przez komutator C_1. Należy podkreślić, że algorytm przestrojeń dokonuje nowej aranżacji połączeń jedynie w zakresie łączy międzysekcyjnych, a żądane w połączeniach wejścia i wyjścia pozostają niezmienione. Algorytm działa w sposób oczekiwany z technicznego punktu widzenia.



Rysunek 6: Stan pola SWS1 z rys. 4 po przestrojeniu połączeń i zestawieniu połączenia $\mathcal{C}(2[1], 1[0])$ oznaczonego przez N

6. WNIOSKI

W referacie zaproponowano nowy algorytm przestrojeń połączeń jednoszczelinowych w polach typu SWS1. Wskazano, że warunki RNB pól SWS1 podane w [10] są prawidłowe, ale zaproponowany w cytowanej pracy algorytm zestawiania połączeń nie ma zastosowania w praktycznych rozwiązaniach węzłów sieci EON. W referacie podano teoretyczne podstawy stosowania algorytmu przestrojeń oraz sam algorytm, który umożliwia na przestrajanie połączeń w polu SWS1 z zachowaniem żądanych w połączeniach wejść i wyjść pola, co jest zgodne z podejściem używanym zarówno w teorii komutacji jak i w praktyce budowy pól komutacyjnych pracujących w węzłach sieci.

Zaznaczyć należy, iż zaproponowany w niniejszym artykule algorytm jest w pełni skalowalny i nadaje się do wszystkich rodzajów pól, ale w niektórych przypadkach wymagany ciąg przestrojeń może być stosunkowo długi, bo zależy on od wartości jakie mają n oraz q. Może to skutkować dużą ilością przerwań i ponownych połączeń. Aby uniknąć dużej ilości przestrojeń alternatywnym rozwiązaniem może być używanie pola SWS1 spełniającego warunki SNB, które zostało zaproponowane w [4].

LITERATURA

[1] Clos C. 1953. „A study of non-blocking switching networks". Bell Syst. Tech. J. 32: 406–424.

[2] Danilewicz G. 2019. „Asymmetrical space-conversion-space SCS1 strict-sense and wide-sense nonblocking switching fabrics for continuous multislot connections". IEEE Access, 7: 107058-107072.

[3] Danilewicz G. 2019. „Supplement to "Asymmetrical Space-Conversion Space SCS1 Strict-Sense and Wide-Sense Nonblocking Switching Fabrics for Continuous Multislot Connections" - the SCS2 Switching Fabrics Case". IEEE Access, 7: 167577-167583.

[4] Danilewicz G., Kabaciński W., Rajewski R. 2016. „Strict-sense nonblocking space-wavelength-space switching fabrics for elastic optical network nodes". J. Opt. Commun. Netw., 8: 745–756.

[5] Gerstel O., Jinno M., Lord A., Yoo S.J.B. 2012. „Elastic optical networking: A new dawn for the optical layer?". IEEE Commun. Mag., 50: S12–S20.

[6] Hwang F. 2004. The Mathematical Theory of Nonblocking Switching Networks. World Scientific: Singapore.

[7] Jajszczyk A., Jekel G. 1993. „A new concept-repackable networks". IEEE Trans. Commun., 41: 1232–1237.

[8] Jinno M., Takara H., Kozicki B., Tsukishima Y., Sone Y., Matsuoka S. 2009. „Spectrum-efficient and scalable elastic optical path network: Architecture, benefits, and enabling technologies". IEEE Commun. Mag., 47: 66–73.

[9] Kabaciński W. 2005. Nonblocking Electronic and Photonic Switching Fabrics. Springer.

[10] Lin B.-C. 2020. „Rearrangeable and Repackable S-W-S Elastic Optical Networks for Connections with Limited Bandwidths". Appl. Sci., 10: 1251.

[11] West D.B. 2001. Introduction to Graph Theory. Pearson: New York City, NY, USA.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

TRANSMISJA DANYCH W SIECIACH O WYSOKIEJ STOPIE BŁĘDÓW Z WYKORZYSTANIEM PROTOKOŁÓW QUIC I HTTP/3

DATA TRANSMISSION IN HIGH ERROR RATE NETWORKS USING QUIC AND HTTP/3 PROTOCOLS

Wojciech Gumiński[1]; Robert Latoszewski[2];

[1] Politechnika Gdańska Wydział ETI, Gdańsk, wojciech.guminski@pg.edu.pl
[2] Politechnika Gdańska Wydział ETI, Gdańsk, s167866@student.pg.edu.pl

DOI: 10.15199/59.2025.4.80

Streszczenie**: W pracy omówiono potrzebę i uzasadnienie rezygnacji z protokołu TCP w sieciach o wysokiej stopie błędów. Pokazano wyniki testów zastosowania protokołów QUIC i HTTP/3 w transmisji o bardzo wysokiej stopie błędów oraz przydatność tych protokołów dla sieci o sporadycznej i przerywanej łączności.**

Abstract**: The paper discusses the need and justification for abandoning the TCP protocol in networks with a high error rate. It shows the test results of the QUIC and HTTP/3 protocols in transmission with a very high error rate and the usefulness of these protocols for networks with sporadic and disruption connectivity.**

Słowa kluczowe**: HTTP, protokoły, QUIC, transmisja danych.**

Keywords**: data transmission, HTTP, protocols, QUIC.**

1. WPROWADZENIE

Zmiany zachodzące w sposobie korzystania z sieci komputerowych, powszechność usług internetowych oraz wszechobecność mobilnych terminali bezprzewodowych powoduje konieczność zweryfikowania używanych protokołów transmisji danych. Projektowany dla sieci przewodowych i światłowodowych niezawodny protokół TCP (ang. *Transmission Control Protocol*) staje się kłopotliwym ograniczeniem w sieciach bezprzewodowych. Założenia o bitowej stopie błędów BER (ang. *Bit Error Rate*) nieprzekraczającej 1e-13 i pakietowej stopie błędów PER (ang. *Packet Error Rate*) nie większej niż 1e-8 w systemach wymiany danych [8] były słuszne w czasach, gdy dostęp do sieci zapewniany był łączem przewodowym. W ostatniej dekadzie najpowszechniej używanymi terminalami dostępowymi stały się smartfony, pojawiła się komunikacja pojazdów, rozpowszechniły się idee Internetu Rzeczy IoT (ang. *Internet of Things*), a większość urządzeń korzysta wyłącznie z łączności bezprzewodowej. Nieustannie zwiększający się wolumin danych przesyłanych do i z urządzeń mobilnych z jednej strony wymusił stosowanie coraz bardziej efektywnych metod modulacji, a z drugiej powoduje coraz większe narażenie transmisji na błędy [1], które skutkują często koniecznością pracy w środowiskach, w których PER łatwiej niż w skali wykładniczej wyrażać jest w skali procentowej.

Najpowszechniej stosowany protokół transportowy TCP, który zaprojektowano tak, aby zapewniał niezawodną i bezbłędną komunikację, zawiera szereg mechanizmów związanych z zapewnieniem tej niezawodności [3]. Niestety, przy stopie błędów znacznie przekraczającej założenia projektowe mechanizmy te są nieefektywne: prowadzą do chwilowego zatrzymywania transmisji danych i blokowania połączenia, resetowania mechanizmów zapewniających wydajność po wystąpieniu każdego błędu i w efekcie zmniejszają efektywną szybkość transmisji danych, często o rzędy wielkości, powodując równocześnie długotrwałe bezczynne blokowanie zasobów serwerów treści.

Najwięksi dostawcy treści w Internecie i zarazem najwięksi usługodawcy chmurowych usług hostingowych zauważyli tę nieefektywność wykorzystania zasobów serwerowych. Przy zwiększającym się wolumenie danych dostarczanych do urządzeń bezprzewodowych ta nieefektywność powodowała blokowanie kosztownych zasobów serwerowych w oczekiwaniu na sygnalizację błędów i retransmisje. Efektem były prace nad zwiększeniem efektywności wymiany informacji z wykorzystaniem TCP np. protokół SPDY [5] multipleksujący wiele strumieni w jednym połączeniu, a ostatecznie nad porzuceniem TCP na rzecz protokołów transmisji danych implementowanych w warstwie aplikacyjnej w przestrzeni użytkownika. Pozwala to na ich szybkie wdrożenie niezależne od cykli wydawniczych poszczególnych systemów operacyjnych oraz daje możliwość dostosowywania tak implementowanych protokołów do potrzeb konkretnych aplikacji [8]. Ostatecznie zdefiniowany został protokół QUIC [5], a następnie dla najpowszechniejszego ruchu aplikacji webowych ustandaryzowano go jako HTTP/3 [5].

Protokół QUIC wykorzystujący w warstwie transportowej bezpołączeniowy pozbawiony jakichkolwiek mechanizmów protokół UDP (ang. *User Datagram Protocol*) implementuje w warstwie aplikacyjnej niemal wszystkie mechanizmy optymalizacyjne TCP w sposób zapewniający detekcję błędów, retransmisje błędnych datagramów oraz multipleksację strumieni danych bez przerw w transmisji wywoływanych blokowaniem połączenia. Pozwala to na ograniczenie efektu zmniejszenia efektywnego wykorzystania przepustowości łącza do wynikającego jedynie z dodatkowego czasu niezbędnego na retransmisje błędnych datagramów.

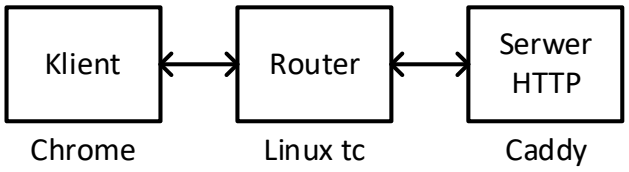
Dodatkowo warto zaznaczyć, że protokół QUIC rozwiązuje niezwykle istotny problem poufności transmisji. Obligatoryjne użycie TLS (ang. *Transport Layer Security*) w wersji 1.3 dostosowanej do transmisji UDP zapewnia szyfrowanie end-to-end i pozwala równocześnie na pominięcie czasochłonnego procesu *handshakingu* przy powtórnych połączeniach z tym samym serwerem usług.

W komunikacji bezprzewodowej często mamy do czynienia z pracą na granicy zasięgu radiowego albo w środowiskach o przerywanej, sporadycznej lub silnie zakłócanej transmisji [1, 4]. Stosowane zwykle w takich środowiskach protokoły przekazywania danych wykorzystują mechanizmy charakterystyczne dla sieci DTN (ang. *Delay/Disruption Tolerant Network*) takie jak: buforowanie, wieloskokową retransmisję czy stosowanie protokołu Bundle [2]. Na poziomie warstwy transportowej zwykle używany jest niezawodny protokół TCP, o którego wadach w sieciach o wysokiej stopie błędów wspomniano powyżej. Badania dotyczące wydajności protokołu QUIC zwykle przedstawiają jego wyższość nad HTTP/2 dla transmisji wielu niewielkich elementów lub w sieciach o wysokim opóźnieniu [7] i ich wyniki zwykle uwzględniają PER do poziomu 1-2% nie adresując problemów związanych z transmisją w gorszych warunkach np. gdy PER osiąga 10% albo 15%.

Motywacją tej pracy była potrzeba znalezienia odpowiedzi czy QUIC albo HTTP/3 będą nadal dobrym wyborem przy transmisji w środowiskach z PER rzędu 10% i więcej.

2. ŚRODOWISKO BADAWCZE

Dla potrzeb przeprowadzenia eksperymentów badawczych wykorzystano środowisko złożone z trzech elementów: serwera HTTP, routera pośredniczącego z możliwością kształtowania parametrów stratności i opóźnienia transmisji oraz odbiorcy wiadomości stanowiącego równocześnie stację pomiarową (rys. 1).



Rys. 1. Stanowisko badawcze

Jako serwer HTTP wykorzystywany był serwer Caddy z obsługą HTTP/2 i HTTP/3 oraz konfiguracją dodającą do każdego z zasobów nagłówki uniemożliwiające przechowywanie elementów w pamięci podręcznej i serwerach pośredniczących. Odbiornikiem była przeglądarka Google Chrome uruchamiana za pomocą aplikacji w języku Python z biblioteką Selenium automatyzującą użycie aplikacji Chrome i dostarczającej niezbędnych informacji o mierzonych czasach. Funkcję routera pełniła maszyna wirtualna z systemem operacyjnym Ubuntu Linux, a kształtowanie ruchu realizowane było z wykorzystaniem linuksowego mechanizmu *traffic control*. Całe stanowisko badawcze zostało zrealizowane w środowisku wirtualnym z wykorzystaniem oprogramowania Virtual-Box.

W badaniach przy zadanych parametrach opóźnienia transmisji i zadanej średniej pakietowej stopy błędów wielokrotnie powtarzano pomiar czasu załadowania lokalnej, statycznej kopii stron internetowych cnn.com i shopify.com złożonych z wielu elementów różnej wielkości o sumarycznej objętości ok 10MB. Pomiary wykonywano najpierw dla protokołu HTTP/2 wykorzystującego TCP, a następnie porównywano z wynikami dla protokołu HTTP/3 używającego protokołów QUIC i UDP.

3. WYNIKI

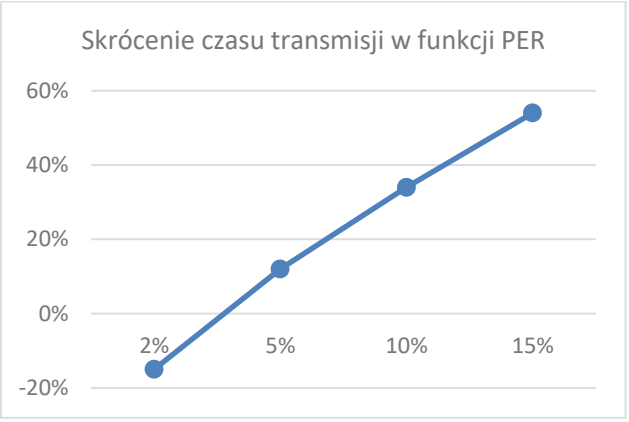
Zgodnie ze znanymi w literaturze [7] wynikami protokół QUIC nie wykazuje żadnej istotnej przewagi nad HTTP/2 dla transmisji w warunkach niskiej PER. W prezentacji wyników pominięto badania i analizy, których wyniki nie różnią się od znanych np. [7], skupiając się na analizie zmian łącznego czasu transmisji przy wysokiej pakietowej stopie błędów większej niż 5%.

Dla potrzeb prezentacji wyników zdefiniujmy skrócenie czasu transmisji jako procent czasu transmisji niezbędnej dla protokołu HTTP/2 (1).

$$skrócenie = \left(\frac{t_{H2}-t_{H3}}{t_{H2}}\right)100\% \qquad (1)$$

gdzie:
t_{H2} – to średni czas transmisji dla protokołu HTTP/2,
t_{H3} – to średni czas transmisji dla protokołu HTTP/3.

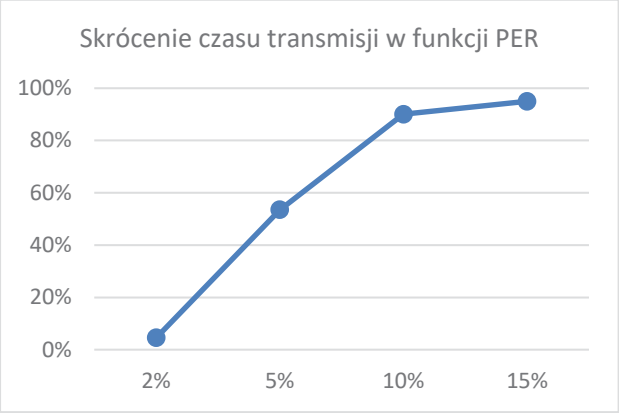
Wyniki prezentowane na rys. 2 pokazują, że dla niskiej stopy błędów i typowego dla aplikacji internetowych opóźnienia protokół HTTP/3 może być wolniejszy, ale dla wysokiej stopy błędów pozwala skrócić łączny czas transmisji mniej więcej o połowę.



Rys. 2. Skrócenie średniego łącznego czasu transmisji wielu elementów o łącznym rozmiarze 10MB przy wprowadzeniu dodatkowego 20ms opóźnienia w łączu.

Wyższość protokołu HTTP/3 nad HTTP/2 widoczna jest szczególnie w sieciach o wysokiej stopie błędów przekraczającej 2%. Dla połączeń bezpośrednich, charakterystycznych dla sieci wieloskokowych i DTN, o PER na poziomie większym niż 10% można zauważyć spektakularne kilkukrotne zmniejszenie średniego łącznego czasu transmisji (rys. 3). W warunkach tak wysokiej stopy błędów czasy transmisji dla protokołu TCP wzrastają o dwa rzędy wielkości praktycznie uniemożliwiając utrzymanie

stabilnego połączenia. Protokół QUIC w tych samych warunkach pracuje oczywiście również wolniej, ale transmisja jest ciągła i wielokrotnie szybsza.



Rys. 3. Skrócenie średniego łącznego czasu transmisji wielu elementów o łącznym rozmiarze 10MB przy połączeniu bez dodatkowego opóźnienia.

QUIC jako protokół zdefiniowany w warstwie aplikacyjnej posiada wiele odmiennych implementacji [6], które realizując ten sam standard, różnią się często założeniami projektowymi. Z jednej strony prowadzi to do różnic wydajności, a z drugiej daje możliwość dostrojenia protokołu do konkretnych wymagań aplikacji i konkretnego środowiska transmisji danych.

4. WNIOSKI

Zaprezentowane wyniki pokazują, że w sieciach o bardzo wysokiej stopie błędów zastąpienie klasycznych połączeń TCP protokołem QUIC albo HTTP/3 wykorzystujących UDP pozwala na uzyskanie nawet wielokrotnie krótszych łącznych czasów transmisji. Pozwala to na znacznie efektywniejsze wykorzystanie dostępnej przepustowości łącza. Jednocześnie nie wymaga żadnych kompromisów dotyczących poprawności, czy poufności transmisji. Projektowane do zwiększenia efektywności w aplikacjach internetowych protokoły QUIC i HTTP/3 wydają się być również doskonałym wyborem dla sieci o nieciągłej albo sporadycznej łączności w środowiskach o wysokiej stopie błędów.

LITERATURA

[1] Bai Haowei, Atiquzzaman Mohammed. 2003. "Error modeling schemes for fading channels in wireless communications: A survey". *IEEE Communications Surveys & Tutorials*, vol. 5, no. 2, pp. 2-9, Fourth Quarter 2003.

[2] Fall Kevin, Farrell Stephen. 2008. "DTN: an architectural retrospective". *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 5, pp. 828-836, June 2008.

[3] Grigorik Ilya. 2017. "HTTP Protocols". O'Reilly

[4] Gumiński Wojciech. 2018, "Integracja mechanizmów transportowych dla sieci o nieciągłej i sporadycznej łączności w systemie netBaltic". *Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne* nr 2-3/2018 pp. 57-59.

[5] Koch Jacob, Falowo Olufunsho, Elrod Nathan. 2024. "What We Know About HTTP/3 and Its Implementation: A Literature Review," *IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI)*, Mt Pleasant, MI, USA, 2024, pp. 1-7.

[6] Marx Robin, Herbots Joris, Lamotte Wim, Peter Quax. 2020. "Same Standards, Different Decisions: A Study of QUIC and HTTP/3 Implementation Diversity". *EPIQ'20: Proceedings of the Workshop on the Evolution, Performance, and Interoperability of QUIC*. Pp 14-20.

[7] Megyesi Péter, Krämer Zsolt, Molnár Sándor. 2016. "How quick is QUIC?". *IEEE International Conference on Communications (ICC)*, Kuala Lumpur, Malaysia, pp. 1-6.

[8] Tee Lai King. 2007. "Packet Error Rate and Latency Requirements for a Mobile Wireless Access System in an IP Network". *IEEE 66th Vehicular Technology Conference*, Baltimore, MD, USA, pp. 249-253.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

Weryfikacja możliwości opracowania skanera pasma PMR z wykorzystaniem radia definiowanego programowo

Verification of the possibility of developing a PMR band scanner using software-defined radio

Michał Maj[1]; Piotr Boryło[1];

[1]AGH w Krakowie, Instytut Telekomunikacji
mimaj@agh.edu.pl, piotr.borylo@agh.edu.pl

DOI: 10.15199/59.2025.4.81

Streszczenie: Niniejsza praca stanowi analizę możliwości opracowania skanera pasma PMR z wykorzystaniem radia definiowanego programowo RTL-SDR. Kluczowymi funcjonalnościami są jednoczesne monitorowanie 16 kanałów, blokada szumów oraz zapisywanie danych audio w czasie rzeczywistym. W ramach pracy przeprowadzono też testy geolokalizacji za pomocą trzech stacji odbiorczych. Wyniki wykazały, że maksymalny zasięg opracowanego skanera to ok. 1 km w otwartej przestrzeni. W odniesieniu do geolokalizacji uzyskano akceptowalne, ale nie w pełni satysfakcjonujące rezultaty. Podstawowym wyzwaniem była wielodrogowość.

Abstract: This paper presents an analysis of the feasibility of developing a PMR band scanner using RTL-SDR software-defined radio. The scanner's key functionalities include simultaneous monitoring of 16 channels, noise squelch control, and real-time audio recording. Geolocation trials were also conducted using three receiving stations. Tests showed that the scanner's maximum range is approximately 1 km in open terrain. For geolocation, acceptable but not entirely satisfactory results were obtained. In both cases, multipath propagation was the primary challenge.

Słowa kluczowe: demodulator FM, geolokalizacja, pasmo PMR, radio definiowane programowo, skaner radiowy

Keywords: FM demodulator, geolocation, PMR band, software-defined radio, radio scanner

1. WSTĘP

Postęp technologiczny sprawił, że radio definiowane programowo stało się powszechnie dostępnym narzędziem. Malejące ceny, a także liczne potencjalne zastosowania, stanowiły motywację do opracowania wielu aplikacji otwartoźródłowych, pozwalających na analizę sygnałów w czasie rzeczywistym. Należą do nich między innymi SDR++, Gqrx, SDR#, GNU Radio i SRDuno. Różnią się one znacząco oferowanym zakresem funkcjonalności. Skanowanie więcej niż jednego kanału jest dostępne jedynie w SDRuno. GnuRadio natomiast pozwala na stworzenie bardzo zaawansowanych ścieżek obróbki sygnału, ale dzieje się to kosztem wydajności i uniwersalności tworzonego projektu. Żadne z wymienionych narzędzi nie oferuje jednak możliwości zlokalizowania nadajnika w przestrzeni. Niniejsza praca stanowi analizę opracowania systemu, który nie tylko byłby w stanie skanować w czasie rzeczywistym wszystkie 16 kanałów pasma PMR (ang. Private Mobile Radio), ale także pozwoliłby na geolokalizację urządzenia nadającego komunikat. Proponowany system mógłby znaleźć zastosowanie w organach regulacyjnych, np. w Urzędzie Komunikacji Elektronicznej, ale także mógłby posłużyć jako podstawa dla amatorskiej sieci nasłuchowej.

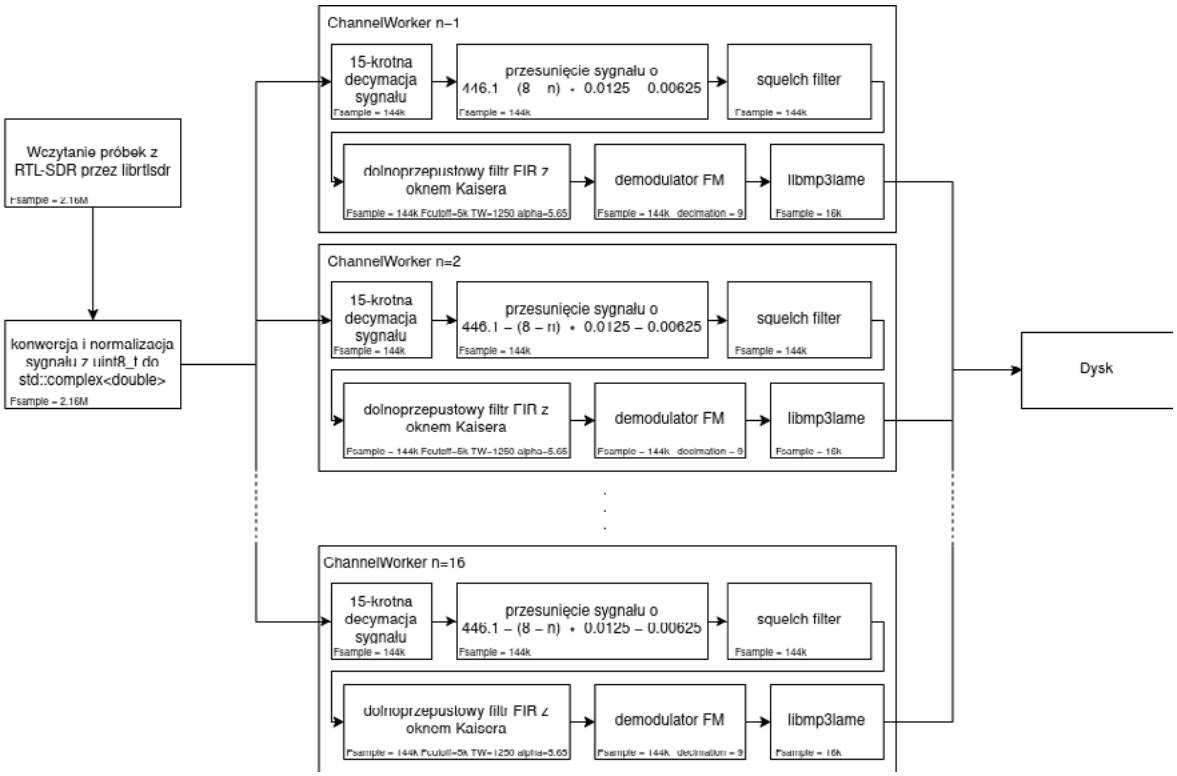
Proponowany system pracuje w nielicencjonowanym paśmie PMR przeznaczonym do krótkodystansowej komunikacji głosowej, operujące na częstotliwościach ok. 446 MHz. Pasmo to jest powszechnie wykorzystywane przez m.in. pracowników budowlanych, sportowców, harcerzy i inne osoby używające krótkofalówek typu walkie-talkie.

2. ARCHITEKTURA SYSTEMU

Architektura proponowanego systemu, przedstawiona na Rysunku 1, obejmuje aplikację podzieloną na logiczne moduły. Próbki I/Q pochodzące z RTL-SDR są konwertowane do formatu zmiennoprzecinkowego, a następnie przekazywane do 16 niezależnych wątków przetwarzających. Każdy wątek, oznaczony jako *ChannelWorker*, odpowiada za obsługę pojedynczego kanału PMR.

Każdy *ChannelWorker* działa w pętli kolejno przetwarzającej próbki. Pierwszym krokiem jest przesunięcie częstotliwości w taki sposób, aby 0 Hz (pasmo podstawowe) odpowiadało częstotliwości danego kanału. Następnie blokada szumu (ang. squelch filter) określa, czy odbierany jest sygnał FM, czy szum. W przypadku sygnału przetwarzanie jest kontynuowane poprzez odfiltrowanie zbędnych częstotliwości, co pozwola uniknąć zjawiska aliasingu w kolejnych krokach. Przedostatnim krokiem jest demodulacja FM, zaimplementowana jako demodulator różniczkująco-dzielący według [3], po której sygnał jest zapisywany na dysku jako plik WAVE lub MP3.

Do odbioru sygnału użyto radia RTL-SDR, które jest popularnym wśród hobbystów radiem definiowanym programowo. Radio zostało podłączone za pomocą kabla RG-174 do anteny typu J-pole [6] dostrojonej do częstotliwości 446,1 MHz. Do przetwarzania sygnałów użyto laptopa Framework 13 z procesorem Intel Core i5-1135G7 oraz 16 GB pamięci RAM. Jako radiotelefony wykorzystano Motorola T82 Extreme. Deklarowana moc nadawcza każdego z urządzeń wynosi 500 mW.



Rysunek 1: Logiczna architektura systemu

W RTL-SDR zastosowano oscylator niskiej jakości, co ma negatywny wpływ na zdolność wstrajania się radia. Jego częstotliwość pracy zależy od temperatury otoczenia, co powoduje odchyłki od wartości zadanej. W celu skompensowania tego efektu wykorzystano parametr *PPM offset*, którego wartość została wyznaczona empirycznie.

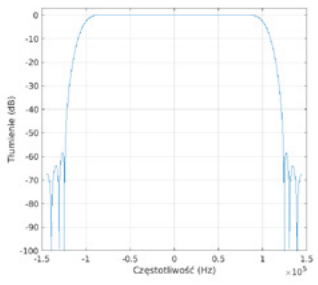
2.1 Dobór filtrów

Proponowana implementacja skanera pasma PMR wymaga filtrowania sygnału na trzech etapach: w decymatorze w celu zapobiegania aliasingowi, przy podziale sygnału na kanały w celu eliminacji szumu z sąsiednich kanałów, oraz w blokadzie szumów w celu zawężenia pomiar energii widma do obszaru w pobliżu nośnej.

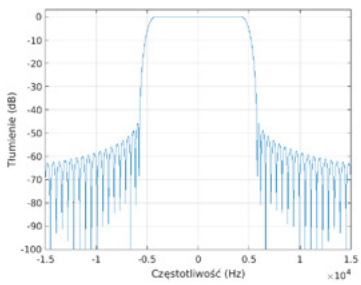
W decymatorze i mechaniźmie blokady szumów zastosowano filtr IIR z oknem Kaisera [7] [1], natomiast na potrzeby blokady szumów filtr Butterworth'a. Charakterystyki częstotliwościowe tych filtrów prezentują Rysunki 2-4.

2.2 Testy terenowe

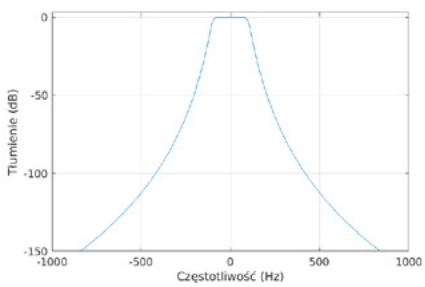
Na Rysunku 5 przedstawiono mapę z przybliżonymi lokalizacjami punktów pomiarowych. Punkty P1 i P16 są bezpośrednio widoczne z punktu B1. Punkty P2-P6 oraz P14-P15 znajdują się na otwartej przestrzeni, w której jedynymi przeszkodami są małe budynki i pojedyncze drzewa. Natomiast punkty P7-P13 znajdują się w obszarze gęstej zabudowy, gdzie wysokość przeszkód (głównie budynków) przekracza 10 metrów.



Rysunek 2: Odpowiedź częstotliwościowa filtru decymatora, $M = 29$, $f_c = 105$ kHz, $\alpha = 5.65$



Rysunek 3: Odpowiedź częstotliwościowa filtru blokady szumów, $M = 249$, $f_c = 5$ kHz, $\alpha = 3.97$



Rysunek 4: Odpowiedź częstotliwościowa filtru Butterworth'a 8-go stopnia, $f_c = 100$ Hz

W celu weryfikacji poprawności działania opracowanego skanera zaplanowano następujący eksperyment:

1. Odbiornik został tymczasowo zamontowany na 2,5-metrowym maszcie w lokalizacji oznaczonej jako B1 w Tabeli 1.

2. Operator wyposażony w nadajnik przemieścił się kolejno do lokalizacji P1 i P16, nadając z każdej z nich komunikat na wszystkich 16 kanałach.

3. Po potwierdzeniu poprawnego działania systemu na wszystkich kanałach wybrany został kanał o najmniejszym obciążeniu. Operator z nadajnikiem przemieścił się kolejno pomiędzy punktami P1-P16, nadając w każdym z nich komunikat głosowy zawierający numer punktu. System rejestrował komunikaty wraz z wartościami parametrów wykorzystywanymi przez mechanizm blokady szumów do podjęcia decyzji.

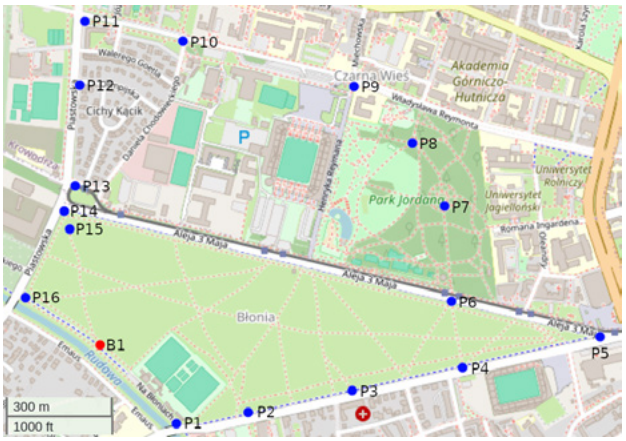
4. Pomiary zostały powtórzone trzykrotnie w celu redukcji niepewności pomiarowej.

Transmitowane wiadomości powinny być identyczne pod względem treści, różniąc się wyłącznie numerem punktu. Przykładem poprawnej wiadomości jest "Znajduję się w punkcie P1".

Tabela 1: Lokalizacje punktów pomiarowych

Nazwa	*Współrzędne GPS*
B1	50°03'31.7"N 19°54'14.4"E
P1	50°03'25.5"N 19°54'24.8"E
P2	50°03'26.6"N 19°54'36.1"E
P3	50°03'28.4"N 19°54'50.1"E
P4	50°03'30.6"N 19°55'05.8"E
P5	50°03'33.2"N 19°55'25.4"E
P6	50°03'36.5"N 19°55'03.9"E
P7	50°03'45.3"N 19°55'03.0"E
P8	50°03'51.0"N 19°54'58.9"E
P9	50°03'56.2"N 19°54'50.2"E
P10	50°04'00.3"N 19°54'26.0"E
P11	50°04'02.2"N 19°54'11.8"E
P12	50°03'56.2"N 19°54'11.3"E
P13	50°03'47.2"N 19°54'10.6"E
P14	50°03'44.8"N 19°54'09.6"E
P15	50°03'43.6"N 19°54'09.9"E
P16	50°03'36.9"N 19°54'03.9"E

Wyniki pomiarów przedstawiono w Tabeli 2. Skaner zarejestrował komunikaty z obszaru zabudowanego, jednak były one jedynie częściowo zrozumiałe. W punktach P10 i P11 odebrana została tylko jedna wiadomość, więc pomiary te zostały odrzucone jako statystycznie niewiarygodne. Najdalszym punktem, z którego udało się odebrać jakikolwiek komunikat, był punkt P6 oddalony o 993 metry od anteny odbiorczej. Nie wszystkie z odebranych wiadomości były w pełni zrozumiałe. Punkt P14, zlokalizowany w odległości 413 metrów od stacji odbiorczej, był najdalszym miejscem, z którego wszystkie wiadomości były w pełni zrozumiałe.



Rysunek 5: Mapa z przybliżonymi lokalizacjami punktów pomiarowych

2.3 Obliczanie wartości kompensujących dla modeli propagacyjnych

Obliczanie odległości na podstawie mocy odbieranego sygnału wymaga znajomości bezwzględnego poziomu mocy wyrażonego w dBm. W przedstawionych pomiarach wartości są przedstawione w dB względem wewnętrznego poziomu referencyjnego odbiornika RTL-SDR. Zależność pomiędzy wskazaniami RTL-SDR (w dB) a rzeczywistą mocą sygnału (w dBm) jest określona w następujący sposób:

$$P_{dBm} = P_{dB} + G_{offset} \tag{1}$$

Wartość G_{offset} można obliczyć przez porównanie zmierzonego poziomu sygnału w punktach P1 i P16 (znajdujących się w bezpośredniej linii widzenia z punktem B1) z teoretycznymi wartościami oczekiwanymi. Wartości te będą się różnić w zależności od przyjętego modelu propagacyjnego. Ta wartość może być następnie użyta do zamiany wszystkich pomiarów na dBm. Na potrzeby tych obliczeń założono, że nadajnik nadaje z maksymalną mocą 500 mW (26,98 dBm), zysk anteny odbiorczej to 2,15 dB natomiast zysk anteny nadawczej jest równy zyskowi dipola (2,15 dB). Założono także, że tłumienie kabla antenowego wraz ze złączami wynosi 1.5 dB. Obliczone w ten sposób wartości prezentuje Tabela 3.

Tabela 2: Wyniki pomiarów z odbiornikiem w punkcie B1 przy zysku tunera ustawionym na 48,0 dB i wykorzystaniu nadajnika Motorola T82 Extreme. Symbol X oznacza, że blokada szumów zablokowała sygnał ze względu na zbyt niski poziom mocy w pobliżu nośnej.

Punkt	*Odległość [m]*	*Średnia moc [dB]*
P1	280	-34,12
P2	459	-43,06
P3	714	-47,99
P4	1020	-42,72
P5	1410	X
P6	993	-37,93
P7	1050	X
P8	1060	X
P9	1030	X
P10	911	X
P11	940	X
P12	757	-41,59
P13	480	-28,11
P14	413	-27,12
P15	375	-25,01
P16	263	-16,77

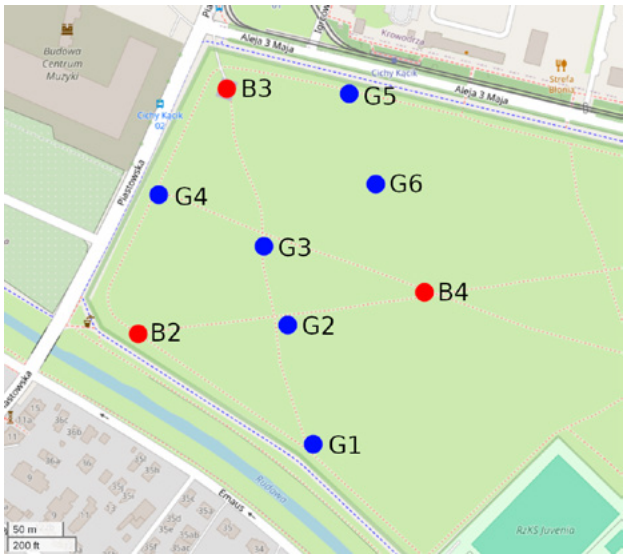
Tabela 3: Oszacowane wartości G_{offset}

Model propagacyjny G_{offset}	
Hata open space	-51.39 dBm
Hata suburban	-50.24 dBm
FSPL	14,13 dBm

3. MODUŁ GEOLOKALIZACJI

W celu realizacji modułu geolokalizacji zastosowano metodę multilateracji, będącą generalizacją trilaeracji. Wyniki pomiarów poddano obróbce po zakończeniu eksperymentu. Dla potrzeb eksperymentu opracowano dedykowany układ współrzędnych. Koordynaty są reprezentowane w dwuwymiarowym układzie kartezjańskim. Początek układu odpowiada punktowi 50°03'40.2"N 19°54'10.6"E, a jednostka układu odpowiada jednemu metrowi na ziemi. Pomiary zostały przeprowadzone z wykorzystaniem 3 stacji odbiorczych znajdujących się w punktach B2, B3 i B4. Mapę z przybliżoną lokalizacją tych punktów przedstawiono na Rysunku 6. Każda ze stacji składała się z identycznego zestawu jak poprzednio.

Analogicznie jak w poprzednim eksperymencie, oprator wyposażony w nadajnik Motorola T82 Extreme przemieszczał się pomiędzy punktami G1-G6, nadając w każdym z nich komunikat głosowy. Wszystkie anteny w punktach B2-B4 zostały zamontowane na masztach o wysokości 1,5 metra. Wyniki pomiarów przedstawiono w Tabeli 4. Widać w nich dużą wariancję między kolejnymi powtórzeniami pomiarów, którą można wyjaśnić zjawiskiem wielodrogowości.



Rysunek 6: Mapa z przybliżonym położeniem punktów pomiarowych i stacji odbiorczych

Tabela 4: Wyniki pomiarów w punktach G1-G6 przy zysku tunera ustawionym na 48,0 dB i wykorzystaniu nadajnika Motorola T82 Extreme. Symbol X oznacza, że blokada szumów odrzuciła sygnał ze względu na zbyt niski poziom mocy w pobliżu nośnej. Litera *p* jest skrótem od słowa *pomiar*.

Punkt	*Moc w B2 [dB]*			*Moc w B3 [dB]*			*Moc w B4 [dB]*		
	p. 1	*p. 2*	*p. 3*	*p. 1*	*p. 2*	*p. 3*	*p. 1*	*p. 2*	*p. 3*
G1	-25,26	-19,46	-26,73	-42,38	-48,48	-44,28	-36,05	-37,47	-47,13
G2	-27,08	-41,83	-42,17	-45,46	-38,56	-47,99	-20,76	-25,55	X
G3	-33,97	-37,45	-41,10	-37,12	-39,71	-42,65	-26,40	-28,26	X
G4	-33,63	-33,54	-36,83	-44,47	-40,48	-38,12	-31,06	-35,45	-32,06
G5	-44,56	-44,66	-47,81	-33,58	-43,25	-42,61	-37,63	-37,66	-36,90
G6	-49,06	-43,30	-38,52	-40,41	-41,59	-41,05	-25,25	-21,12	-20,70

3.1 Porównanie modeli propagacyjnych

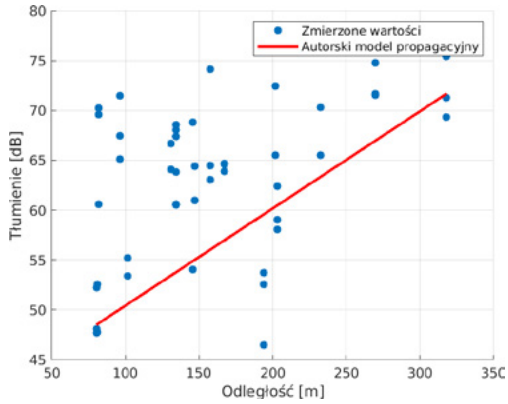
Dokładność systemu w znacznym stopniu zależy od wyboru odpowiedniego modelu propagacyjnego. W badaniach porównano cztery modele propagacyjne: model tłumienia swobodnej przestrzeni (ang. Free Space Path Loss, FSPL), model Hata dla otwartej przestrzeni (Hata Open Space [4]), model Hata dla terenów podmiejskich (Hata Suburban [4]), a także autorski model propagacyjny opisany poniżej. Modele Hata i ich pochodne wykazują dobrą skuteczność w zakresie pasm około 400 MHz, co potwierdzono w literaturze przedmiotu [2] [5].

Autorski model propagacyjny opracowano przez zastosowanie regresji liniowej do wszystkich zmierzonych próbek, a następnie dostrojenie współczynników A i B do momentu, gdy otrzymana formuła przestała zwracać wartości negatywne. Otrzymane wyniki przedstawia Rysunek 7. Otrzymany wzór ma postać:

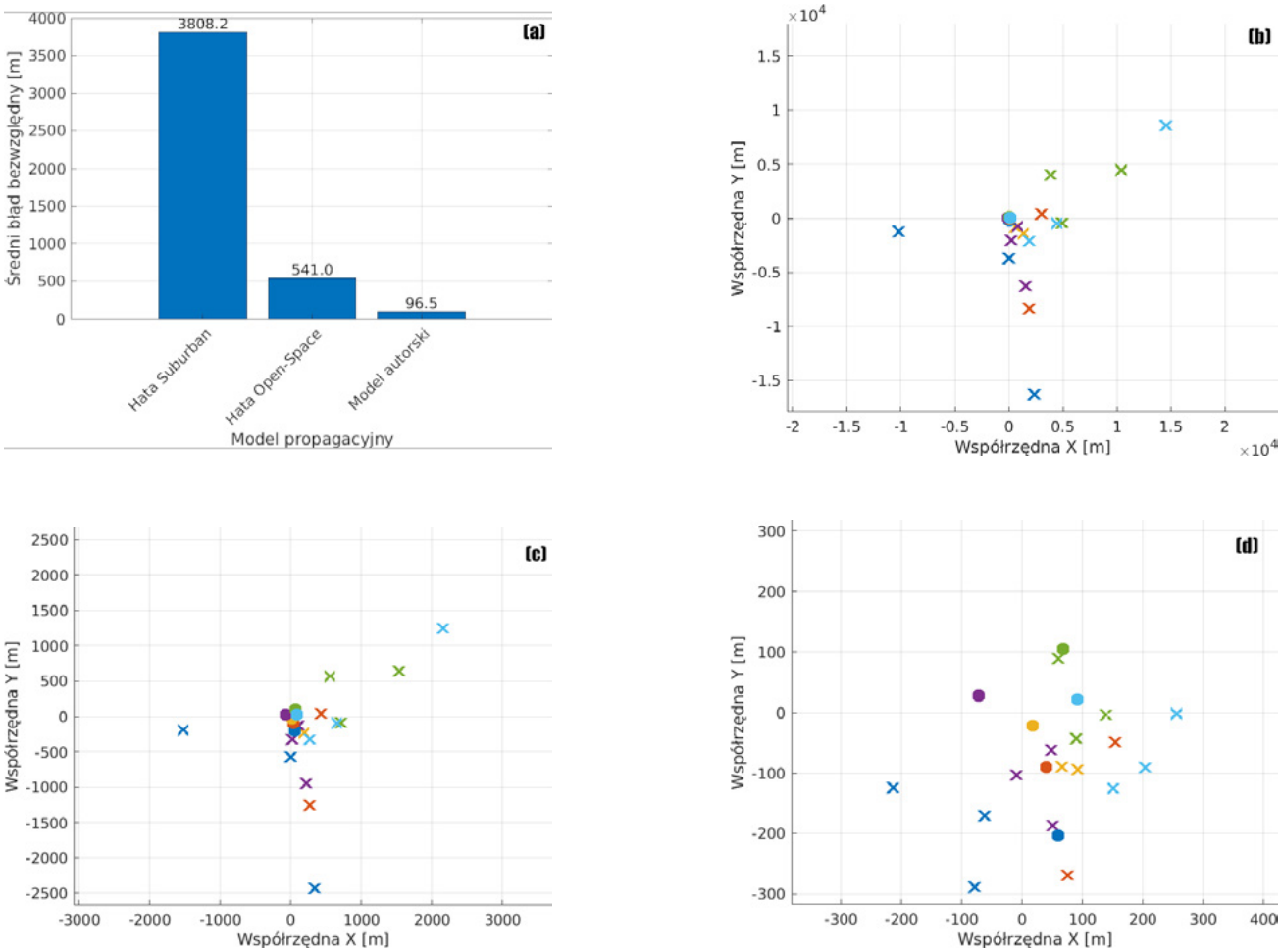
$$d_{CUSTOM} = \frac{L - 40.6704}{0.0975} \tag{2}$$

gdzie L oznacza tłumienie na danej odległości wyrażone w dB, a d_{CUSTOM} oznacza odległość pomiędzy nadajnikiem a odbiornikiem wyrażoną w metrach.

Rysunek 8 przedstawia porównanie średnich błędów bezwzględnych każdego z analizowanych modeli propagacyjnych oraz szczegółowe mapy ilustrujące dokładność poszczególnych modeli. Model FSPL wykluczono z porównania ze względu na uzyskiwane wartości błędów bezwzględnych rzędu 10^7 metrów.



Rysunek 7: Dopasowanie prostej regresji w autorskim modelu propagacyjnym



Rysunek 8: Lokalizacje przewidziane przy użyciu poszczególnych modeli przedstawione w układzie współrzędnych. Symbolem X oznaczono przewidzianą lokalizację, a kropką rzeczywistą lokalizację. **(a)** - podsumowanie błędów bezwzględnych, **(b)** - model Hata dla terenów podmiejskich, **(c)** - model Hata dla otwartej przestrzeni, **(d)** - model autorski

Model Hata dla terenów podmiejskich wykazuje lepszą dokładność niż model FSPL, jednak uzyskane wyniki wciąż nie osiągają poziomu akceptowalnego. Model ten nie jest dopasowany do charakterystyki obszaru pomiarów. Model Hata dla otwartej przestrzeni zapewnił lepsze rezultaty niż omówione wcześniej modele. Należy jednak zaznaczyć, że stacje odbiorcze były oddalone od siebie o 200-300 metrów, a największy odnotowany błąd wynosił prawie 2 km. Nie można więc uznać, że model ten daje satysfakcjonujące wyniki. Najlepsze rezultaty osiągnięto przy użyciu autorskiego modelu propagacyjnego. Średni błąd lokalizacji wynosił około 100 metrów, przy maksymalnym błędzie na poziomie 300 metrów. Taka dokładność, uzyskana przy wykorzystaniu zaledwie trzech stacji odbiorczych, może być uznana za akceptowalną, pozwalając na określenie przybliżonego obszaru występowania nadajnika, co stanowi informację użyteczną w praktyce. Należy jednak zaznaczyć, że jest to model dedykowany dla konkretnego obszaru.

4. WNIOSKI

Zbudowany system skanera działa w czasie rzeczywistym, równocześnie monitorując wszystkie 16 kanałów i zapisując sygnał audio na dysku. Po zamontowaniu na maszcie o wysokości 2,5 metra osiąga zasięg od 400 do 1000 metrów w otwartym terenie. Zaimplementowana blokada szumów funkcjonuje zgodnie z założeniami projektowymi. W obszarach zabudowanych obserwuje się jednak znaczący spadek zasięgu urządzenia.

Moduł geolokalizacji dał rezultaty, które można określić jako akceptowalne, choć niesatysfakcjonujące. Przyjmujemy bowiem, że za satysfakcjonujący należy uznać średni błąd bezwzględny nieprzekraczający 10% odległości między stacjami odbiorczymi. Model ma charakter lokalny, co przekłada się na jego małą uniwersalność.

Duży rozrzut zmierzonych wartości wynika prawdopodobnie z wielodrogowego charakteru propagacji sygnału, który zaburza liniową zależność pomiędzy odległością i tłumieniem. Innym powodem może być niestabilność oscylatora w odbiorniku RTL-SDR. Może ona powodować przesunięcia częstotliwości nośnej względem wartości oczekiwanych.

5. AFILIACJA

Prace były wspierane przez Ośrodek Przetwarzania Informacji - Państwowy Instytut Badawczy, w ramach grantu nr KPOD.01.18-IW.03-0009/24 pt. „Krajowe laboratorium sieci i usług 5G wraz z otoczeniem” w ramach Krajowego Planu Odbudowy i Zwiększania Odporności, Inwestycja A2.4.1: Inwestycje w rozbudowę potencjału badawczego. Praca została sfinansowana przez Ministerstwo Nauki i Szkolnictwa Wyższego w ramach subwencji dla Wydziału Informatyki, Elektroniki i Telekomunikacji AGH.

LITERATURA

[1] Al-Ani, M. S. (2017, Aug.). Study the characteristics of finite impulse response filter based on modified Kaiser window. *UHD Journal of Science and Technology 1*(2), 1–6.

[2] Alamoud, M. A. and W. Schütz (2012). Okumura-hata model tuning for TETRA mobile radio networks in Saudi Arabia. In *2012 2nd International Conference on Advances in Computational Tools for Engineering Applications (ACTEA)*, pp. 47–51.

[3] Chaudhari, Q. Frequency Modulation (FM) and demodulation using DSP techniques. Opublikowano na https://wirelesspi.com/ (Dostęp: 1.08.2024).

[4] Hata, M. (1980). Empirical formula for propagation loss in land mobile radio services. *IEEE Transactions on Vehicular Technology 29*(3), 317–325.

[5] Kasampalis, S., P. I. Lazaridis, Z. D. Zaharis, A. Bizopoulos, S. Zettas, and J. Cosmas (2014). Comparison of Longley-Rice, ITU-R P.1546 and Hata-Davidson propagation models for DVB-T coverage prediction. In *2014 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, pp. 1–4.

[6] Sekhar, B., P. Kavitha, B. Srinivas, and C. Mahesh (2022, 11). Design of J-pole antenna for receiving ADS-B signals. pp. 1–5.

[7] Soares Augusto, J. (2015, 11). The Io-sinh function, calculation of Kaiser windows and design of FIR filters.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

Autonomiczny lot bez GNSS i łączności: zastosowanie radaru z syntetyczną aperturą oraz AI w nawigacji bezzałogowców.

Autonomous flight without GNSS and communication: application of synthetic aperture radars and AI in UAV navigation.

Krzysztof Gromada[1]

[1] XY-Sensing Sp. z o.o., Warszawa, kgromada@xysensing.pl

DOI: 10.15199/59.2025.4.82

Streszczenie: W artykule przedstawiono algorytm lokalizacji bezzałogowych statków powietrznych (UAV) w warunkach braku dostępu do systemów GNSS. Rozwiązanie opiera się na dopasowywaniu wyniku segmentacji sematycznej zobrazowania radarów SAR lub kamer wizyjnych z mapami pokrycia terenu. Uzyskane wyniki eksperymentów symulacyjnych wskazują na wyjątkowo wysoką dokładność lokalizacji (błąd poniżej 13 metrów w 98,9% przypadków) oraz krótki czas obliczeń (poniżej 25 ms), umożliwiając lokalizację w czasie rzeczywistym.

Abstract: The paper presents a localization algorithm for unmanned aerial vehicles (UAVs) operating in GNSS-denied environments. The proposed method utilizes SAR or electrooptical imagery and semantic segmentation by deep neural networks (U-Net). Using described image-to-ground projection and mask-to-map alignment algorithm, high accuracy localization is possible. Simulation results demonstrate error below 13 meters in 98.9% of cases with low processing times of under 25ms, enabling real-time onboard implementation.

Słowa kluczowe: Bezzałogowe Statki Powietrzne (UAV), Głębokie Sieci Neuronowe (DNN), Lokalizacja bez sygnału GPS, Radar z Syntetyczną Aperturą (SAR), Widzenie Maszynowe (CV)

Keywords: Computer Vision (CV), Deep Neural Networks (DNN), Localisation in GPS-denied environments, Synthetic Aperture Radar (SAR), Unmanned Aerial Vehicles (UAV)

1. WSTEP

Współczesne platformy bezzałogowe (UAV od ang. Unmanned Aerial Vehicle) zyskują na znaczeniu w zastosowaniach cywilnych, służb mundurowych, oraz na polach bitwy. Trwające konflikty zbrojne ukazują przewagę małych bezzałogowców nad klasycznym uzbrojeniem ze względu na bezpieczeństwo ich stosowania i koszty budowy.

Najsprawniejszą metodą przeciwdziałania UAV jest zagłuszanie (ang. jamming), bądź spoofing sygnałów Globalnych Systemów Nawigacji Satelitarnej (GNSS od ang. Global Navigation Satelite System). Wynika to z nieprzygotowania większości UAV do pracy w przypadku braku sygnałów zewnętrznych (sygnałów lokalizacyjnych lokalizacji bądź łączności z infrastrukturą naziemną). Z tego powodu zagłuszane platformy na ogół automatycznie przerywają misję i podejmują próbę awaryjnego lądowania.

Zapewnienie niezawodnego mechanizmu lokalizacji dronów bez sygnałów zewnętrznych stanowi kluczowy element w kierunku wysokiej autonomii i niezawodności dronów. Platformy pełniące kluczowe role muszą być odporne na zagłuszanie lub modulację sygnału GNSS. Stąd, popularnym zagadnieniem rozwijanym przez naukowców i konsorcja przemysłowe jest lokalizowanie samolotu bez zewnętrznych sygnałów nawigacyjnych [3, 10].

2. PRZEGLĄD LITERATURY

W przypadku utraty sygnałów GNSS i komunikacji radiowej, większość UAV korzysta z lokalizacji opartej na pomiarach sensorów inercyjnych (IMU od ang. Inertial Measuring Unit), w większych jednostkach będącego kluczowym elementem AHRS (ang. Attitude and Heading Reference Systems). Jednak sensory IMU charakteryzują się dryftem, który uniemożliwia lokalizację na długich przelotach [2, 9].

W dziedzinie robotyki silnie rozwijane są natomiast alternatywne mechanizmy lokalizacji oparte o systemy obrazujące (najczęściej kamery wizyjne). Można je podzielić na 3 kategorie:
1. Visual Odometry – odometria optyczna – najprostsza w implementacji, pozwala na śledzenie zliczeniowe pozycji na podstawie obserwowanych punktów charakterystycznych bądź przepływów optycznych [9],
2. Markery/znaki nawigacyjne – na ogół optyczna lokalizacja na podstawie znanych punktów czy obiektów charakterystycznych na mapie znanej á priori [1],
3. Simultaneous Localization And Mapping (SLAM) – równoczesne mapowanie terenu i lokalizacja, może opierać się o pomiar z różnych sensorów (na przykład lidar, sonary), jednak popularnie prowadzony jest na podstawie obserwacji optycznej (co określane

jest często jako Visual SLAM lub VSLAM) [6].

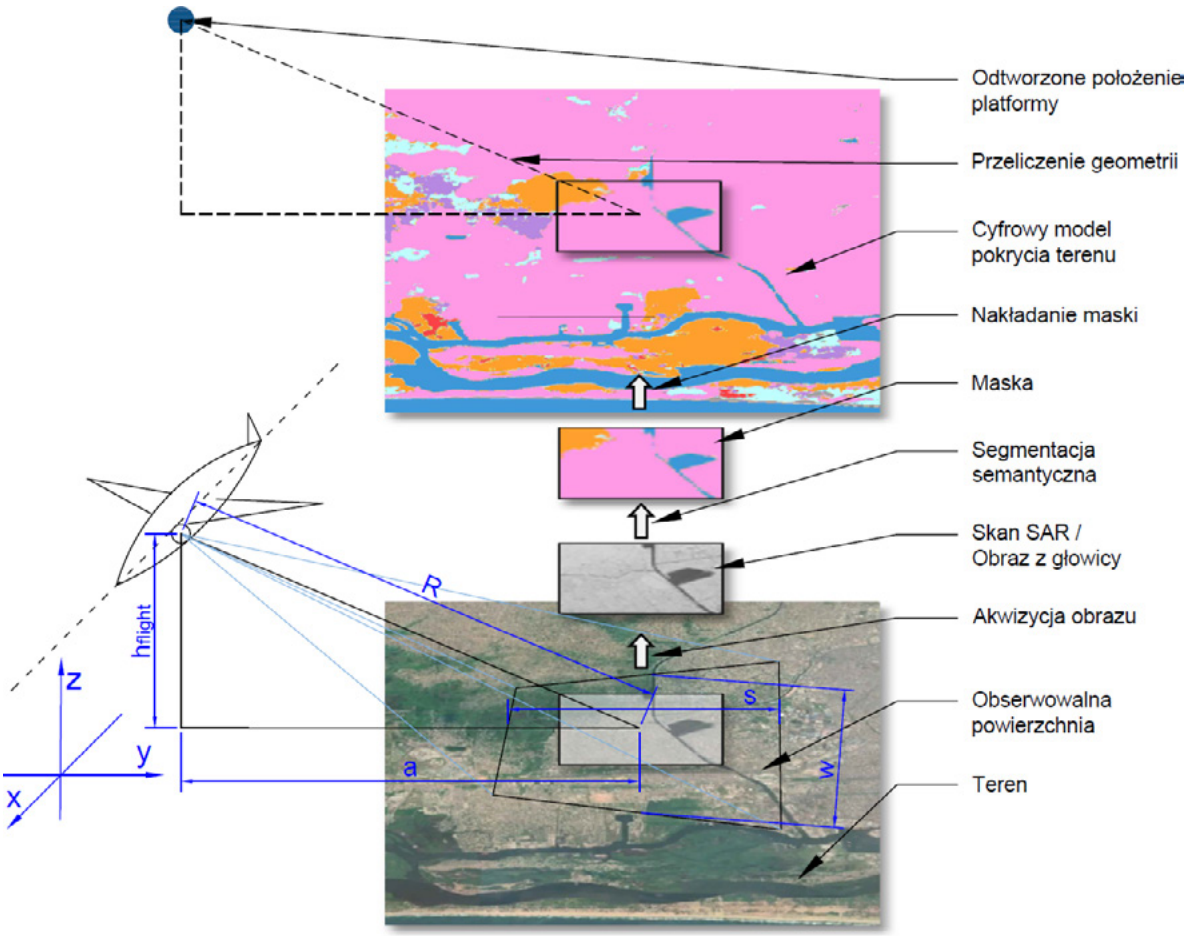
Omówione podejścia różnią się złożonością oraz zakresem zastosowania. Odometria optyczna jest najprostsza w implementacji, lecz jej główną wadą jest akumulacja błędów w czasie, co prowadzi do dryftu pozycji podobnego do tego w układach IMU. Metody wykorzystujące markery lub znaki nawigacyjne oferują wyższą dokładność, ale wymagają obecności wcześniej zmapowanych, rozpoznawalnych punktów w otoczeniu, co ogranicza ich użyteczność w środowiskach nieznanych. SLAM, a zwłaszcza jego wariant optyczny (VSLAM), umożliwia jednoczesne lokalizowanie i mapowanie otoczenia, co czyni go bardziej uniwersalnym i dokładniejszym, jednak kosztem wyższej złożoności obliczeniowej i podatności na błędy w przypadku ubogich danych sensorycznych lub braku punktów charakterystycznych. Istnieją inne algorytmy nawigacyjne robotów, jednak nie mają one zastosowania w przypadku statków powietrznych

W systemach rakietowych istniały alternatywne metody lokalizacji w oparciu o znane mapy terenu. Najdokładniejszymi z nich były systemy DSMAC (ang. Digital Scene Matching Area Correlator). Oparte są o proste przekształcenia amplitudowe obrazu (odszumianie, progowanie wartości) lub płytkie sieci neuronowe. Działają na zasadzie porównywania tak przetworzonego obrazu z kamery do podobnie przetworzonych zdjęć satelitarnych [5].

3. ALGORYTM LOKALIZACJI

Proponowany system lokalizacji przypomina zasadę działania DSMAC. Zgodnie ze schematem ideowym (rys. 1) pierwszym krokiem jest zebranie zobrazowania terenu w otoczeniu platformy. W tym celu wykorzystany może być dowolny sensor obrazujący: głowica elektro-optyczna, kamera termowizyjna, bądź radar z syntetyczną aperturą. Następnie, z użyciem sztucznej inteligencji, lub klasycznych metod przetwarzania, prowadzona jest segmentacja sematyczna zobrazowania (każdemu pikselowi przypisywana jest klasa). Co z kolei pozwala na lokalizację tak zsegmentowanego obrazu (zwanego dalej maską lub maską segmentacyjną) na wcześniej opracowanej mapie. Znajomość położenia zebranego obrazu względem mapy umożliwia określenie pozycji samolotu na podstawie znanych parametrów geometrycznych sensora.

Odnoszenie się do istniejącej mapy pozwala na prowadzenie lokalizacji platformy bez ograniczeń czasu pracy (jak w metodach zliczeniowych).



Rysunek 1: Schemat opracowanego algorytmu lokalizacji UAV w oparciu o segmentację obrazu

3.1 Segmentacja i dane wejściowe

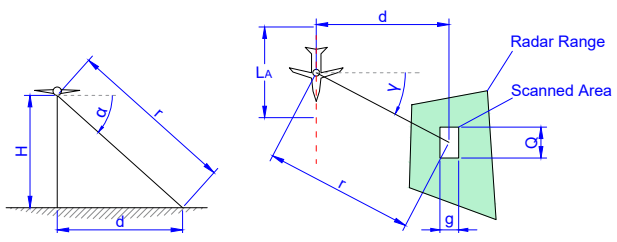
Zgodnie z poprzednim opisem system może korzystać z zobrazowania radarowego (SAR) lub optycznego/podczerwonego (EO/IR). Zastosowanie kamer zapewnia przewagę w częstotliwości zbierania zobrazowania (30–60 Hz). Zastosowanie SAR natomiast znacząco ułatwia rzutowanie zobrazowania na płaszczyznę terenu, co przekłada na możliwość zastosowania szybszych i dokładniejszych algorytmów dopasowania. W dalszym opisie przedstawiane będzie zastosowanie SAR do lokalizacji.

W celu segmentacji możliwe jest zastosowanie klasycznych metod przetwarzania obrazów [8]. Jednak ze względu na znaczące ograniczenia mocy komputerów pokładowych na UAV zastosowano metody sztucznej inteligencji, które charakteryzuje się znacznie niższą złożonością obliczeniową niż większość klasycznych metody przetwarzania zobrazowania SAR. W badaniach wykorzystano doskonale przebadaną sieć U-Net, ze względu na jej wysoką wydajność, dokładność segmentacji, a także architekturę. Zadaniem tej sieci jest określenie przynależności każdego piksela obrazu do 1 z 14 klas pokrycia terenu na obrazie (między innymi: budynki, woda, drogi, wysoka roślinność).

Warto podkreślić, że stosowanie AI do zobrazowania SAR jest silnie ograniczone przez brak dostępu do dużych zbiorów trenujących, które są powszechnie dostępny dla zobrazowania wizyjnego. Zastosowanie sieci U-Net przypominającej konstrukcją auto-enkodery, umożliwia przeprowadzenie wstępnego uczenia nienadzorowanego, które pozwala na podniesienie jakości całego treningu. Zobrazowanie SAR jest szczególnie podatne na metody rozszerzania zbioru (data augmentation) [4].

3.2 Dopasowanie maski do mapy

Aby dopasować maskę semantyczną do mapy, oblicza się projekcję maski na mapę na podstawie geometrii sensora i estymowanej pozycji UAV. Przykład geometrii rzutowania i zbieranego zobrazowania dla radaru SAR pokazano na rys. 2.



Rysunek 2: Geometria obserwacji SAR i jej parametry

Odpowiednio obrócona/przekształcona maska jest nakładana na mapę poprzez minimalizację różnicy między wartościami klas. Obliczana jest funkcja kosztu na podstawie wartości bezwzględnych różnic między maską a mapą. Szczegółowo opisana jest we wzorze 1, gdzie i, j - indeksy w macierzy maski, i_0, j_0 to indeksy mapy, dla których estymowana jest funkcja niedokładności nałożenia maski, o rozmiarze (n_{width}, n_{height}), $y_{mask}(i,j)$, $y_{map}(i,j)$ to wartości odpowiednio maski i mapy w indeksie (i, j).

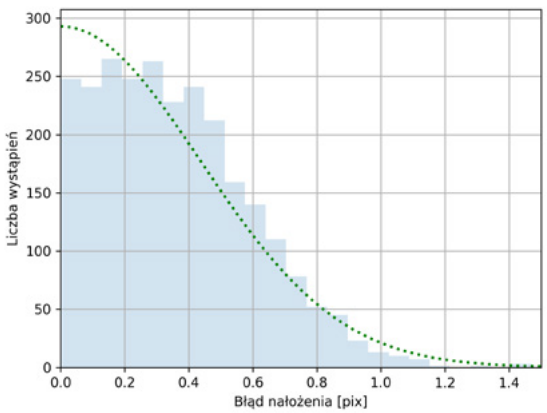
$$f(i_0, j_0) = \sum_{j=0}^{n_{height}} \sum_{i=0}^{n_{width}} |\boldsymbol{y}_{mask}(i, j) - \boldsymbol{y}_{map}(i_0+i, j_0+j)| \tag{1}$$

Warto podkreślić, że wartości segmentacji $\boldsymbol{y}$ to wektory, a zatem wynik funkcji może być przedstawiony jako suma bezwzględnych błędów dopasowania każdej z klas, uwzględniająca estymowaną przez sieć pewność klasyfikacji. Błąd opisywany takim wektorem można odpowiednio przystosować do odwzorowania pewności/jakości segmentacji. Przykładem może być podniesienie wagi niedopasowania budynków, które stanowią doskonałe cechy dla lokalizacji i są łatwo wykrywalne przez SAR.

4. WYNIKI LOKALIZACJI

W testach z rastrowymi mapami pokrycia terenu i zobrazowaniem satelitarnym Sentinel-2 osiągnięto dokładność lokalizacji lokalnej lepszą niż 13 m w 98,9% przypadków (rys. 3). Czas obliczeń (segmentacja i dopasowanie) nie przekracza 50 ms na procesorze AMD Ryzen Ryzen 5 3600X – 6 rdzeni, 3,80 GHz, albo 25ms z wykorzystaniem procesora graficznego Nvidia RTX 3080.

Dla porównania autorzy systemu nawigacji opartego o SAR osiągnęli dokładność nakładania na poziomie 25 m, przy czasie przetwarzania niewiele poniżej 2 minut [7], w najszybszej wersji.



Rysunek 3: Histogram błędów nakładania maski segmentacyjnej o rozmiarze 1024 px x 1024 px na mapę pokrycia terenu o rozmiarze 1424 px x 1424 px. Jeden piksel odpowiadał obszarowi 10 m x 10 m.

5. PODSUMOWANIE

Zaprezentowany algorytm umożliwia efektywną lokalizację UAV bez GNSS poprzez połączenie segmentacji semantycznej i dopasowania do map referencyjnych. Uzyskano bardzo wysoką dokładność i niskie czasy obliczeń, co umożliwia implementację na pokładzie UAV.

Warto podkreślić, że na błąd całkowity lokalizacji składa się więcej czynników niż sama dokładność dopasowania maski semantyczne na mapę, w tym odchylenia wynikające z niedokładności działania sensorów, czy zastosowanego filtru nawigacyjnego. Uwzględnienie niedokładności typowych sensorów nawigacyjnych pozwoliło na aproksymację niedokładności lokalizacji na podstawie jednego zobrazowania na poziomie 64 m.

Więcej informacji

Powyższy referat streszcza fragment rozprawy doktorskiej "Zwiększenie autonomii taktycznych bezzałogowych statków powietrznych w oparciu o radar z syntetyczną aperturą i kamerę elektrooptyczną," Krzysztof Gromada, 2024. Z dokładniejszym opisem powyższych algorytmów, można zapoznać się w dystertacji.

LITERATURA

[1] Benini Alessandro, Mancini Adriano oraz Longhi Sauro. “An IMU/UWB/visionbased extended kalman őlter for mini-UAV localization in indoor environment using 802.15.4a wireless sensor network": Journal of Intelligent and Robotic Systems: Theory and Applications 70.1-4 (Kwiecień 2013), pp. 461–476. issn: 09210296. doi: 10.1007/s10846-012-9742-1.

[2] Djuriatno Waru, Maulana Eka, Effendi Dodi Arisandi. “Velocity measurement based on inertial measuring unit": Telkomnika (Telecommunication Computing Electronics and Control) 17.4 (Sierpień 2019), pp. 1898–1906. issn: 2302929

[3] Greco Mario, Querry Stephane, Pinelli Gianpaolo, Kulpa Krzysztof, Samczynski Piotr, Gromek Damian, Gromek Artur, Malanowski Mateusz, Gal Bernard Querry oraz Aldo Bonsignore. “SAR-based augmented integrity navigation architecture - SARINA project results presentation": Proceedings International Radar Symposium (2012), pp. 225–229. issn: 21555753. doi: 10.1109/IRS.2012.6233320.

[4] Gromada Krzysztof. „Unsupervised SAR Imagery Feature Learning with Median Filter- Based Loss Value", Sensors, 22 (17), 6519 (2022) [13],

[5] Irani Geoffrey oraz Christ James. Image processing for Tomahawk scene matching. Tech. rep. Johns Hopkins APL Technical Digest, 1994, pp. 250–264. url: https://api.semanticscholar.org/CorpusID:221369275.

[6] Latcham Patrick oraz Taylor Clark. “Comparison of Visual Simultaneous Localization and Mapping Methods for Fixed-Wing Aircraft Using SLAMBench2": 2020 IEEE/ION Position, Location and Navigation Symposium, PLANS 2020. Institute of Electrical and Electronics Engineers Inc., Kwiecień 2020, pp. 1578–1586. isbn: 9781728102443. doi: 10.1109/PLANS46316.2020.9109945.

[7] Markiewicz Jakub, Abratkiewicz Karol, Gromek Artur, Ostrowski Wojciech, Samczyński Piotr oraz Gromek Damian. “Geometrical Matching of SAR and Optical Images Utilizing ASIFT Features for SAR-based Navigation Aided Systemsž: Sensors 2019, Vol. 19, Page 5500 19.24 (Grudzień 2019), p. 5500. issn: 1424-8220. doi: 10.3390/S19245500.

[8] Siemiątkowska Barbara, Gromada Krzysztof. „A New Approach to the Histogram-Based Segmentation of Synthetic Aperture Radar Images", Journal of Automation, Mobile Robotics and Intelligent Systems, 1 (2021) [10], 3. doi: 10.12928/telkomnika.v17i4.11826.

[9] Zhang He oraz Ye Cang. “Plane-Aided Visual-Inertial Odometry for 6-DOF Pose Estimation of a Robotic Navigation Aid.": IEEE access : practical innovations, open solutions 8 (2020), pp. 90042–90051. issn: 2169-3536. doi: 10.1109/access.2020.2994299.

[10] Katalog - Innowacje, Wdrożenia, Bezpieczeństwo - Obronność 2022 - Laureaci konkursu innowacje dla Bezpieczeństwa i Obronności. Tech. rep. Biuro Bezpieczeństwa Narodowego, Narodowe Centrum Badań i Rozwoju, 2022, p. 39. url: https://portalmundurowy.pl/images/wydawnictwa/katalog2022.pdf.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

ANALIZA PORÓWNAWCZA ASYNCHRONICZNYCH METOD RADIOLOKALIZACYJNYCH
COMPARATIVE ANALYSIS OF ASYNCHRONOUS RADIOLOCATION METHODS

Jacek Stefański[1]; Jarosław Sadowski[2]

[1] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, jacstefa1@pg.edu.pl
[2] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, jarsadow1@pg.edu.pl

DOI: 10.15199/59.2025.4.83

Streszczenie**: W artykule przedstawiono analizę porównawczą wybranych asynchronicznych metod radiolokalizacyjnych. Podejście asynchroniczne umożliwia m.in. integrację wielu niezależnych sieci radiolokalizacyjnych, zwiększając wydajność procesu estymacji położenia obiektów. Autorzy podjęli próbę usystematyzowania aktualnej wiedzy na temat asynchronicznych metod radiolokalizacyjnych, co może być przydatne podczas wyboru odpowiedniej metody szacowania położenia obiektu dla różnych zastosowań, biorąc pod uwagę ograniczenia spowodowane projektem sieci radiowej lub urządzenia mobilnego.**

Abstract**: The paper presents a comparative analysis of selected asynchronous radiolocation methods. The asynchronous approach enables, among others, the integration of many independent radiolocation networks, increasing the efficiency of the object position estimation process. The authors attempted to systematize the current knowledge on asynchronous radiolocation methods, which can be useful when selecting an appropriate method of object position estimation for various applications, taking into account the limitations caused by the design of the radio network or mobile device.**

Słowa kluczowe**: metody asynchroniczne, radiolokalizacja, radionawigacja, radiowe sieci sensorowe.**

Keywords**: asynchronous methods, radiolocation, radio navigation, wireless sensor networks.**

1. WSTĘP

Od ponad dwóch dekad można zaobserwować dynamiczny rozwój cyfrowych systemów radiokomunikacyjnych. Teoria cyfrowej radiokomunikacji jest obecnie rozległą dziedziną nauki, która ma liczne gałęzie i charakteryzuje się szeregiem powiązań z innymi obszarami nauki. Współczesna radiokomunikacja wspiera m.in. rozwój technologii informatycznych, w szczególności *ubiquitous* (*pervasive*) *computing* [1]. Termin ten odnosi się do wszelkiego rodzaju przetwarzania danych w komputerach i urządzeniach komputerowych, które automatycznie dostosowują się do potrzeb użytkowników. Często zdarza się, że urządzenia i użytkownicy, czyli obiekty, znajdują się w ruchu. Dlatego pełny rozwój *ubiquitous computing* nie jest możliwy bez dynamicznego rozwoju radiolokalizacji, która w naturalny sposób będzie prowadziła do przetwarzania danych o położeniu tych obiektów – *ubiquitous positioning* [2].

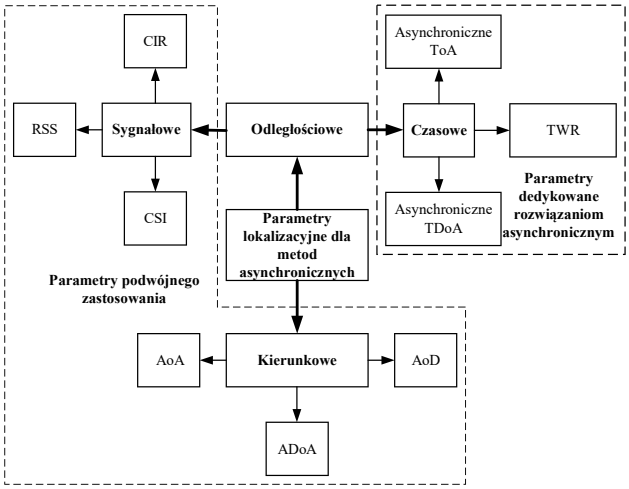
Rozwiązania radiolokalizacyjne opierają się na pomiarach parametrów fal radiowych przesyłanych pomiędzy stacjami referencyjnymi (bazowymi) o stałych i znanych współrzędnych a obiektem ruchomym, który ma zostać zlokalizowany. Transmisje te mogą być realizowane w kierunku od stacji bazowych do obiektu lub z obiektu do stacji bazowych. Większość nowoczesnych systemów radiolokalizacyjnych, które opierają się na skończonym czasie propagacji fal radiowych w środowisku, działa przy założeniu, że stacje referencyjne są w pełni zsynchronizowane. Z tego powodu powszechnie stosowane metody radiolokalizacyjne, które zapewniają stosunkowo wysoką dokładność określania współrzędnych obiektów, polegają na pomiarze odległości lub różnic odległości pomiędzy obiektem a stacjami bazowymi na podstawie opóźnienia propagacji sygnałów radiowych. Metody te są klasyfikowane jako rozwiązania synchroniczne. Kluczowym wyzwaniem dla tych systemów jest osiągnięcie i utrzymanie precyzyjnej synchronizacji pomiędzy stacjami bazowymi. Zapewnienie synchronizacji w sieciach radiowych wiąże się z dodatkowymi kosztami i wymaga rozbudowy infrastruktury sieciowej. Ponadto, niedokładna synchronizacja może prowadzić do znacznych błędów w szacowaniu położenia obiektów, a jeśli synchronizacja całkowicie zawiedzie, lokalizowanie obiektów staje się niemożliwe. Wady te można wyeliminować w asynchronicznych systemach pozycjonowania, które są projektowane przy założeniu braku synchronizacji pomiędzy elementami sieci. Pomimo, że dokładność pozycjonowania w metodach asynchronicznych jest zwykle niższa w porównaniu z rozwiązaniami synchronicznymi, brak synchronizacji sprawia, że asynchroniczne rozwiązania są bardziej odporne na awarie i celowe zakłócenia, dlatego mogą one działać jako rozwiązania zapasowe, jeśli systemy synchroniczne zawiodą. Ponadto, metody asynchroniczne stwarzają możliwości integracji wielu niezależnych sieci radiolokalizacyjnych w celu poprawy wydajności procesu lokalizacji. W artykule autorzy skupiają się na asynchronicznych metodach określania położenia obiektów w sieciach radiowych, w których synchronizacja pomiędzy stacjami bazowymi nie jest obowiązkowa lub nawet nie jest możliwa.

2. PARAMETRY SYGNAŁÓW RADIOWYCH

Na rys. 1 przedstawiono klasyfikację parametrów sygnałów radiowych wykorzystywanych w sieciach radiolokalizacyjnych, które można podzielić na dwie główne grupy:

– grupa parametrów podwójnego zastosowania, tzn. stosowanych zarówno w sieciach synchronicznych, jak i asynchronicznych,

– grupa parametrów dedykowanych rozwiązaniom asynchronicznym, które w szczególności znajdują się w obszarze zainteresowania autorów tego artykułu.



Rys. 1. Parametry sygnałów radiowych stosowane w asynchronicznych metodach radiolokalizacyjnych (na podstawie [3]).

Wymienione powyżej grupy parametrów mają uniwersalne zastosowanie, tj. w środowiskach wewnątrzbudynkowych oraz na zewnętrz budynków. Podstawowymi parametrami sygnałów radiowych, które są mierzone w sieciach radiowych na potrzeby estymacji położenia obiektu są: moc sygnałów odbieranych RSS (*Received Signal Strength*), odpowiedź impulsowa kanału CIR (*Channel Impulse Response*), informacja o stanie kanału CSI (*Channel State Information*), kąt nadejścia sygnałów AoA (*Angle of Arrival*), różnica kątów nadejścia sygnałów ADoA (*Angle Difference of Arrival*), kąt odejścia sygnałów AoD (*Angle of Departure*), czas propagacji sygnałów w dwóch kierunkach TWR (*Two-Way Ranging*), asynchroniczny czas nadejścia sygnałów ToA (*Time of Arrival*) i asynchroniczna różnica czasów nadejścia sygnałów TDoA (*Time Difference of Arrival*).

Metoda RSS została po raz pierwszy zaimplementowana w sieciach komórkowych 1G [4]. Polega ona na określeniu, bezpośrednio z budżetu łącza radiowego (znając moc odbieranego sygnału), tłumienia propagacyjnego pomiędzy stacją bazową (referencyjną) a stacją ruchomą i obliczeniu odległości stacji ruchomej od stacji bazowej na podstawie przyjętego modelu propagacyjnego. Metodę tę można stosować zarówno w stacji ruchomej, jak i w stacjach bazowych. Niestety, jednoznaczne powiązanie pomiędzy odległością a tłumieniem propagacyjnym za pomocą funkcji odwrotnej można uzyskać na ogół tylko w warunkach LoS (*Line of Sight*) lub w sytuacji bardzo zbliżonej do warunków LOS.

Odpowiedź impulsowa kanału (CIR), która jednoznacznie opisuje właściwości kanału radiowego pomiędzy nadajnikiem a odbiornikiem, może być używana do tworzenia map sygnatur obszaru objętego systemem radiolokalizacyjnym FP (*FingerPrinting*) [5]. Technika FP wykorzystuje RSS, CIR lub inne parametry sygnału radiowego do określania położenia stacji ruchomej w czasie rzeczywistym poprzez porównanie ich z danymi sygnałów (sygnaturami) przechowywanymi w bazie danych [6]. Metoda ta jest szczególnie skuteczna w środowiskach wewnątrzbudynkowych ze względu na wysoki poziom dokładności i brak potrzeby precyzyjnego modelowania środowiska za pomocą modeli deterministycznych.

W systemach bezprzewodowych informacje o stanie kanału (CSI) są powszechnie używane do opisu jakości kanału. CSI dostarcza szczegółów na temat tego, w jaki sposób sygnał radiowy jest modyfikowany przez otoczenie podczas propagacji z nadajnika do odbiornika, uwzględniając takie czynniki, jak tłumienie propagacyjne, rozpraszanie, zaniki i zakłócenia od innych sygnałów [7]. W rezultacie CSI oferuje dokładniejszy opis stanu łącza radiowego w porównaniu z RSS [8].

Metody kątowe, bazujące na pomiarze kątów pomiędzy liniami łączącymi stację ruchomą i zbiór stacji odniesienia, są zaliczane do metod triangulacyjnych, nawet jeśli liczba wykonanych pomiarów jest różna od trzech. Istnieją dwa główne warianty metod kątowych: pomiar kątów przybycia sygnałów radiowych (AoA) względem ustalonego kierunku odniesienia (np. kierunku północnego, orientacji określonej osi w układzie współrzędnych, itp.) [9] oraz pomiar różnic kątów nadejścia sygnałów radiowych (ADoA), gdy wykonanych pomiarów nie można wyrazić względem ustalonego kierunku odniesienia [10]. W dwuwymiarowym systemie pozycjonowania AoA wystarczy wykonać dwa niezależne pomiary kątów nadejścia sygnału względem ustalonego kierunku odniesienia, chyba że stacja ruchoma i para stacji odniesienia znajdują się na jednej linii prostej.

Pomiar kąta nadejścia sygnałów (AoA) ze stacji ruchomej w kierunku stacji bazowej wyposażonej w układ antenowy jest zwykle wykonywany poprzez znalezienie kierunku odbioru maksymalnej energii [11]. Z kolei w metodzie AoD układ antenowy w stacji bazowej określa optymalny kąt nadawania sygnałów w łączu w dół na podstawie raportów o poziomie sygnału odebranego uzyskanych ze stacji ruchomej. Z punktu widzenia podstaw matematycznych szacowanie położenia obiektów w obu metodach jest takie samo. Największy wysiłek przetwarzania leży po stronie stacji bazowych, ale złożoność implementacji jest większa w przypadku metody AoD, ponieważ stacja bazowa musi adaptacyjnie formować wiązkę anteny w kierunku każdej obsługiwanej stacji ruchomej oddzielnie.

Pomiar odległości pomiędzy węzłami w sieci radiolokalizacyjnej w podstawowej metodzie dwukierunkowego pomiaru odległości (TWR) lub metodzie pomiaru czasu w obu kierunkach RTT (*Round-Trip Time*) obejmuje dwustronną wymianę pakietów radiowych pomiędzy parą aktywnych urządzeń (nadajnik-odbiornik).

Klasyczna metoda pomiaru czasu nadejścia sygnałów (ToA) polega na pomiarze opóźnień propagacji pomiędzy lokalizowanym obiektem (stacją ruchomą) a co najmniej trzema stacjami odniesienia (bazowymi) [12]. Pomiary mogą być przeprowadzane zarówno w łączu *w dół*, jak i *w górę*. Systemy radiolokalizacyjne oparte na

metodzie ToA nazywane są systemami stadiametrycznymi. Zatem metoda ToA wykorzystuje zależność pomiędzy czasem propagacji sygnałów w łączu radiowym a odległością pomiędzy antenami nadawczą i odbiorczą.

Z kolei klasyczna metoda TDoA sprowadza się do określenia różnicy odległości pomiędzy stacją ruchomą a parami stacji odniesienia (bazowych) [13]. Stała różnica czasów propagacji sygnału ze stacji ruchomej do dwóch stacji bazowych w danej parze definiuje jedno ramię hiperboli, stąd nazwa systemy hiperboliczne. W matematyce hiperbola to krzywa na płaszczyźnie złożona z punktów, dla których wartość bezwzględna różnicy odległości od dwóch stałych punktów odniesienia, zwanych ogniskami hiperboli, jest stała. W systemach radiolokalizacyjnych stacje odniesienia znajdują się w ogniskach hiperboli. W przypadku lokalizacji trójwymiarowej (3D) dany wynik pomiaru odległości definiuje hiperboloidę, czyli powierzchnię utworzoną przez obrót hiperboli wokół osi symetrii przechodzącej przez jej ogniska.

3. STRUKTURY ASYNCHRONICZNYCH SIECI RADIOLOKALIZACYJNYCH

Systemy radiolokalizacyjne, zapewniające zakładaną dokładność szacowania położenia obiektu w danym obszarze działania, wymagają odpowiednio rozwiniętej infrastruktury sieciowej. Infrastruktura ta obejmuje przede wszystkim odpowiednią liczbę stacji referencyjnych (bazowych) oraz jednostek zarządzających i nadzorujących prawidłową pracę systemu. Dlatego systemy radiolokalizacyjne można podzielić na trzy główne grupy. Kryterium podziału stanowi miejsce, w którym wykonywane są pomiary i ich przetwarzanie w celu oszacowania położenia obiektu [14], tj.:

- pomiary i przetwarzanie wykonywane przez infrastrukturę sieci radiokomunikacyjnej (*network-based solution*),
- pomiary i przetwarzanie wykonywane przez urządzenia radiowe w lokalizowanym obiekcie (*mobile-based solution*),
- pomiary i przetwarzanie wykonywane przez urządzenia radiowe stacji ruchomej we współpracy z infrastrukturą sieci radiokomunikacyjnej lub odwrotnie (*hybrid solution)*.

Pierwszy przypadek występuje przede wszystkim w sieci komórkowej, gdzie sieć wykonuje pomiary i szacowanie położenia stacji ruchomej. Drugie rozwiązanie występuje również w sieciach mobilnych i większości systemów nawigacji satelitarnej, gdzie pomiar i przetwarzanie wykonuje stacja ruchoma. Uzyskane wyniki szacowania położenia mogą być wysyłane do sieci szkieletowej (rdzeniowej) na żądanie lub wykorzystywane lokalnie w stacji ruchomej. Rozwiązania hybrydowe występują zazwyczaj we wszystkich typach sieci sensorowych.

Oprócz autonomicznych asynchronicznych sieci radiolokalizacyjnych istnieją również rozwiązania, w których wszystkie urządzenia w sieci pracują asynchronicznie i są wspieranie z zewnętrz. W literaturze opisano kilka rozwiązań takich systemów, w których parametry przesunięć czasowych i częstotliwościowych zegarów w węzłach oraz momenty transmisji pakietów lokalizacyjnych są szacowane na podstawie pomiarów z niewspółpracujących nadajników odniesienia i wykorzystując redundancję związaną z odbiorem tego samego pakietu przez wielu odbiorców [15].

4. ASYNCHRONICZNE METODY RADIOLOKALIZACYJNE

W tab. 1 zestawiono asynchroniczne metody radiolokalizacyjne, które według autorów stanowią reprezentatywny zbiór metod dostępnych w literaturze przedmiotu, uwzględniając przynależność do danej grupy metod wraz z krótkim opisem każdej metody.

W metodach, w których pomiary i przetwarzanie odbywa się w sieci, położenie stacji ruchomej (MS) szacuje się przy użyciu pomiarów wykonywanych przez stacjonarne komponenty sieci radiowej pracujące asynchronicznie. Każda stacja bazowa mierzy parametry sygnałów radiowych wysyłanych przez MS. Po wstępnym etapie przetwarzania stacja bazowa wysyła te dane do centralnego serwera lokalizacyjnego w celu dalszej analizy i fuzji danych. Następnie położenie MS szacuje się przy użyciu zintegrowanych danych pomiarowych. Kluczową zaletą tych metod opartych na sieci jest to, że MS nie musi bezpośrednio uczestniczyć w procesie pozycjonowania. Dlatego w wielu przypadkach nie jest konieczna modyfikacja struktury stacji bazowej. Do metod tych można zaliczyć:

- asynchroniczną różnicę czasów nadejścia sygnałów ATDoA (*Asynchronous Time Difference of Arrival*),
- asynchroniczny WAM (*Wide Area Multilateration System*),
- asynchroniczną metodę z dwoma nadajnikami,
- asynchroniczną metodę lokalizacji nadajnika obsługiwaną przez odbiór znaczników czasu.

Metody radiolokalizacyjne, w których stacja ruchoma określa swoją położenie na podstawie sygnałów otrzymanych z sąsiednich stacji bazowych są najrzadziej opisywane w literaturze. Wymaga to wyposażenia MS w odpowiednie zasoby sprzętowe, aby wydajnie szacować swoją pozycję w czasie rzeczywistym na podstawie zebranych danych. Wytworzone informacje o położeniu MS mogą być następnie przesyłane do infrastruktury sieciowej w celu dalszego wykorzystania i budowania usług LBS (*Location Based Services*). Metody te obejmują:

- ATDoA w łączu *w dół*,
- asynchroniczna radiowa interferometryczna metoda lokalizacyjna ARIPS (*Asynchronous Radio Interferometric Position-ing System*),
- metoda ATDoA dla urządzeń pasywnego nasłuchu APLD (*ATDoA for Passive Listening Devices*).

Tab. 1. Zestawienie asynchronicznych metod radiolokalizacyjnych [16]

Metoda	Grupa metod	Krótka charakterystyka
ATDoA	Network-based	Wykorzystuje niezsynchronizowane stacje bazowe do określenia położenia stacji mobilnej poprzez obserwowane różnice czasu pomiędzy momentami odbioru sygnałów ze stacji bazowych.
Asynchroniczny WAM	Network-based	Szacuje położenie samolotu w 3D na podstawie sygnałów emitowanych przez nadajniki pokładowe, wykorzystując niezsynchronizowane stacje bazowe (referencyjne) rozproszone na znacznym obszarze działania.
Asynchroniczna metoda z dwoma nadajnikami	Network-based	Nadawanie zsynchronizowanych sygnałów lokalizacyjnych z dwóch nadajników umieszczonych na obiekcie umożliwia sieci niezsynchronizowanych stacji bazowych jednoczesne określenie położenia i orientacji obiektu, nawet gdy obiekt znajduje się w spoczynku.
Asynchroniczna metoda lokalizacji nadajnika obsługiwana przez odbiór znaczników czasu	Network-based	Wykorzystuje niezsynchronizowane czujniki o znanych lokalizacjach, aby określić położenie obiektu poprzez pomiar różnic czasowych odebranych sygnałów. Metoda nadaje się do śledzenia wielu obiektów jednocześnie.
ATDoA w łączu *w dół*	Mobile-based	Metoda, w której stacje mobilne określają swoje położenie na podstawie odbieranych sygnałów, które są nadawane z niezsynchronizowanych względem siebie stacji referencyjnych. Szczególne zastosowanie do śledzenia ruchomych obiektów, takich jak samoloty.
ARIPS	Mobile-based	Wykorzystuje różnice fazowe między nieznacznie różniącymi się sygnałami dwutonowymi przesyłanymi przez stacje referencyjne, aby oszacować wartości TDoA i określić położenie obiektu.
APLD	Mobile-based	Umożliwia pasywnym urządzeniom oszacowanie położenia poprzez odbieranie sygnałów ze stacji bazowych bez aktywnego uczestnictwa w transmisjach. Nadaje się do urządzeń o ograniczonych zasobach sprzętowych.
Asynchroniczna hiperboliczna	Hybrid	Asynchroniczna odmiana metody hiperbolicznej, w której stacje bazowe wykorzystują lokalne zegary do transmisji sygnałów lokalizacyjnych (brak synchronizacji w relacji stacja nadrzędna – stacja podrzędna), co obniża koszty systemu.
Asynchroniczna eliptyczna	Hybrid	Polega na określeniu położenia obiektu na podstawie asynchronicznego odbioru sygnałów z wielu stacji referencyjnych, przy czym obiekt leży na trajektorii elipsy utworzonej przez stacje referencyjne.
AToA	Hybrid	Wykorzystuje pomiary wartości ToA do pozycjonowania w sieciach asynchronicznych przy użyciu znaczników czasowych umieszczonych w pakietach, redukując potrzebę dodatkowych transmisji pakietów.
dTDoA	Hybrid	Wykorzystuje pomiary różnicowe w celu zwiększenia dokładności metody TDoA w systemach asynchronicznych, co jest przydatne przy szacowaniu położenia wielu obiektów w sieciach czujników.
2D-TDoA	Hybrid	Określa dwuwymiarowe położenie obiektu na podstawie pomiarów różnicy czasu przybycia (TDoA) pomiędzy niezsynchronizowanymi stacjami bazowymi.

W procesie szacowania położenia stacji ruchomej w metodach hybrydowych brane są pod uwagę wszystkie komponenty infrastruktury sieciowej, w tym sama MS, stacje bazowe i ich koordynacja poprzez sieć szkieletową z serwerem lokalizacyjnym. W zależności od metody, poszczególne elementy sieci są w różnym stopniu obciążane w procesie szacowania położenia MS. Na podstawie literatury metody hybrydowe można sklasyfikować według następujących wariantów:

- asynchroniczna hiperboliczna,
- asynchroniczna eliptyczna,
- asynchroniczna czasu nadejścia sygnałów AToA (*Asynchronous Time of Arrival*),
- asynchroniczna różnicowa różnicy czasów nadejścia sygnałów dTDoA (*differential Time Difference of Arrival*),
- asynchroniczna 2D-TDoA (*Two-Dimensional TDoA*).

5. OGRANICZENIA I KIERUNKI ROZWOJU

Asynchroniczne metody radiolokalizacyjne, oferując elastyczność i niezawodność w różnych zastosowaniach, napotykają na kilka wyzwań, które wpływają na ich wydajność i dokładność. Jednym z głównych ograniczeń jest złożoność związana z brakiem synchronizacji systemu, ponieważ brak zunifikowanego odniesienia czaso-

wego zwiększa trudność dokładnego przetwarzania informacji o sygnale. W wielu prezentowanych asynchronicznych systemach pozycjonowania nieznane relacje czasowe, które wpływają na wyniki kilku pomiarów w ten sam sposób, są usuwane z równań za pomocą równań różnicowych, co znacznie zwiększa podatność dokładności szacowania położenia na wszelkie błędy pomiarowe. Wyzwanie to jest spotęgowane przez efekty propagacji wielodrogowej, szczególnie w gęstych środowiskach miejskich lub wewnątrzbudynkowych, gdzie odbicia sygnału i przeszkody wprowadzają znaczne błędy w szacowaniu położenia. Obecność warunków NLOS (*Non-Line of Sight*) uwypukla te problemy, prowadząc do niewiarygodnych pomiarów, które nie zawsze można skompensować lub złagodzić za pomocą zaawansowanych technik przetwarzania sygnału. Złożoność obliczeniowa wielu asynchronicznych algorytmów estymacji położenia powoduje również zwiększone opóźnienie, co sprawia, że aplikacje czasu rzeczywistego, takie jak reagowanie awaryjne i autonomiczna nawigacja, stają się bardziej wymagające. Potrzeba zaawansowanego sprzętu i dużej mocy przetwarzania może prowadzić do znacznego zużycia energii, szczególnie w przypadku urządzeń zasilanych bateryjnie, ograniczając ich wdrażanie w środowiskach o ograniczonych zasobach. Skalowalność i integracja z istniejącymi sieciami bezprzewodowymi, takimi jak 5G i IoT (*Internet of Things*), również stanowią poważne przeszkody, ponieważ systemy te wymagają wysokiego poziomu kompatybilności, płynnej wymiany danych i wydajnego wykorzystania przepustowości. Ponadto rozwiązania sprzętowe stanowią ograniczenie, szczególnie w przypadku miniaturyzacji czujników lokalizacyjnych o niskim poborze mocy i wysokiej dokładności, często wymaganych podczas implementacji w aplikacjach mobilnych i IoT. Niezawodność systemów asynchronicznych w środowiskach o wysokiej dynamice to kolejny obszar budzący obawy, ponieważ szybkie zmiany warunków sygnału i mobilności urządzeń wprowadzają dodatkowe warstwy złożoności do szacowania położenia obiektów. Jednak obawy te pojawiają się również w synchronicznych systemach pozycjonowania. Na przykład synchronizacja urządzeń w systemach synchronicznych zwykle wymaga dodatkowego sprzętu i energii, a warunki propagacji NLOS obniżają jakość szacowania położenia w tych systemach.

Przyszłe badania nad asynchronicznymi metodami radiolokalizacyjnymi powinny skupić się na rozwijaniu technik przetwarzania sygnałów, w szczególności poprzez uczenie maszynowe i algorytmy oparte na sztucznej inteligencji, które mogą zmniejszyć błędy, przyczynić się do redukcji szumów, czyli efektywnie adaptować się do środowiska pracy. Wykorzystanie pasm wyższych częstotliwości, w tym EHF (*Extremely High Frequency*) i technologii fal milimetrowych, mogłoby dodatkowo poprawić dokładność lokalizacji poprzez wykorzystanie krótszych fal, szerszego pasma sygnałów i wyższych szybkości transmisji danych. Jednak ta zmiana wymagałaby starannej oceny charakterystyk propagacyjnych w różnych środowiskach w celu optymalizacji wydajności systemów radiolokalizacyjnych. Ponadto, hybrydowe metody lokalizacji, które integrują technologie GPS (*Global Positioning System*), Wi-Fi (*Wireless Fidelity*) i UWB (*Ultra WideBand*), oferują obiecujące możliwości zwiększenia niezawodności i dostępności usług pozycjonowania poprzez połączenie mocnych stron różnych technik lokalizacyjnych. Ulepszone techniki wizualizacji danych, w tym mapowanie 3D i 4D w czasie rzeczywistym, mogą zapewnić lepszą reprezentację mobilności obiektów, szczególnie w złożonych i dynamicznych scenariuszach, takich jak inteligentne systemy transportowe i automatyzacja przemysłowa. Zajęcie się kwestiami bezpieczeństwa i prywatności jest również niezbędne, szczególnie w celu łagodzenia luk w zabezpieczeniach, takich jak podszywanie się pod sygnał radiowy, nieautoryzowany dostęp do danych i potencjalne cyberataki na infrastrukturę lokalizacyjną. Należy rozwijać dedykowane protokoły szyfrowania i mechanizmy uwierzytelniania, aby zapewnić integralność usług opartych na lokalizacji. Ponadto integracja asynchronicznych metod radiolokalizacyjnych z sieciami nowej generacji, takimi jak 6G i systemy satelitarne, może utorować drogę dla bardziej zaawansowanych i globalnie dostępnych rozwiązań pozycjonowania. Wdrożenie samoorganizujących się i samonaprawiających się systemów lokalizacji byłoby szczególnie korzystne w przypadku zastosowań w odległych i podatnych na katastrofy obszarach, gdzie niezawodne pozycjonowanie ma kluczowe znaczenie dla operacji poszukiwawczo-ratowniczych. Podejmując te wyzwania i eksplorując nowe kierunki technologiczne, asynchroniczne metody radiolokalizacyjne mogą znacząco przyczynić się do ewolucji bezprzewodowych systemów pozycjonowania, wspierając innowacje w inteligentnych miastach, autonomicznej nawigacji i usługach opartych na lokalizacji, które wymagają wysokiej dokładności i niezawodności.

6. PODSUMOWANIE

W artykule autorzy skupili się na syntetycznym przedstawieniu wyników analiz metod lokalizacji obiektów ruchomych w asynchronicznych sieciach radiowych, które nazwano asynchronicznymi metodami radiolokalizacyjnymi. Asynchronizm w sieci radiowej rozumiany jest tutaj jako niezależna praca poszczególnych stacji odniesienia (bazowych). Wszystkie operacje przetwarzania sygnałów w takich stacjach odbywają się w rytmie lokalnych zegarów. Wybrane metody radiolokacyjne bazujące na zależnościach czasowych w asynchronicznych sieciach radiowych zostały zestawione w postaci tabelki (tab. 1).

Według autorów rozwój systemów radiolokalizacyjnych, w tym asynchronicznych, będzie się odbywał równolegle w trzech głównych obszarach: rozwiązań systemowych i warstwy fizycznej, przetwarzania i wizualizacji danych oraz wdrażania nowych usług. Aby osiągnąć te cele, należy poszukiwać nowych rozwiązań w warstwie fizycznej i systemowej. Planuje się wykorzystanie częstotliwości z zakresu EHF do zastosowań lokalnych, co wymaga przeprowadzenia kompleksowych badań propagacyjnych w tym obszarze w różnych warunkach pracy [17]. Równolegle z wykorzystaniem coraz wyższych pasm częstotliwości do celów lokalizacyjnych prowadzone są prace mające na celu opracowanie nowych metod zwiększających dokładność szacowania położenia obiektów w środowisku propagacji wielodrogowej. Choć nie należy

spodziewać się znaczących przełomów w metodach i algorytmach pozycjonowania, to jednak ciągły postęp technologiczny pozwoli na zwiększenie szerokości kanałów w systemach radiolokalizacyjnych, przesuwając granice dokładności szacowania parametrów sygnału radiowego związanych z lokalizacją stacji ruchomych. Proponuje się również jednoczesne stosowanie wielu metod, opartych na pomiarach czasu propagacji sygnałów radiowych (lub ich różnicy) oraz na analizie fazy i/lub polaryzacji odbieranej fali sygnału. Kolejnym istotnym zagadnieniem jest rozwój nowych technologii budowy miniaturowych urządzeń radiowych, które będą mogły współpracować z wieloma sieciami lokalizacyjnymi (lokalnymi i globalnymi) na różnych pasmach częstotliwości. Oczekuje się, że tym wyzwaniom sprosta technologia SDR (*Software Defined Radio*) dedykowana tym aplikacjom [18]. Rozwiązania systemowe stoją przed ogromnymi wyzwaniami, które docelowo powinny uwzględniać dużą liczbę urządzeń radiowych przeznaczonych do realizacji funkcji lokalizacyjnych, wraz z odpowiednimi protokołami komunikacyjnymi. Konieczne będzie zintegrowanie wielu technologii radiowych w jednej warstwie fizycznej, aby zapewnić kompleksową informację o otoczeniu użytkownika. Oczekuje się również, że urządzenia w aplikacjach typu *machine-to-machine* (M2M) będą coraz częściej klientami usługi lokalizacyjnej [19].

Przewiduje się również rewolucję technologiczną w obszarze wizualizacji danych w omawianych systemach radiolokalizacyjnych. Przede wszystkim udoskonalona zostanie technologia interakcji pomiędzy klientem a serwerem usług lokalizacyjnych. Prowadzone są również prace nad rozwojem wizualizacji 3D, a w niedalekiej przyszłości także czterowymiarowej (4D), w której czwartym wymiarem będzie czas rzeczywisty. Rozwój technologii systemów radiolokalizacyjnych oraz nowe możliwości przetwarzania i wizualizacji danych lokalizacyjnych stworzą potencjalne możliwości wdrożenia nowych usług. Warto wspomnieć, że część z wymienionych problemów badawczych jest już przedmiotem szczegółowych analiz teoretycznych i badań eksperymentalnych na całym świecie. W szczególności dotyczy to nowych rozwiązań warstwy fizycznej w systemach radiolokalizacyjnych, a także w obszarze zaawansowanych usług lokalizacyjnych w sieciach komórkowych. Mając na uwadze powyższe, w artykule przedstawiono istotne i aktualne zagadnienia związane z lokalizacją obiektów w sieciach radiowych, ze szczególnym uwzględnieniem rozwiązań asynchronicznych, które zdaniem autorów są niezwykle istotne, zwłaszcza w nietypowych sytuacjach, w których powszechnie stosowane metody synchroniczne nie zapewniają dostępu do usług lokalizacyjnych na skutek celowych działań osób trzecich lub na skutek awarii systemu.

LITERATURA

[1] Hightower Jeffrey, Borriello Gaetano. 2001. "Location Systems for Ubiquitous Computing". *Computer*, 34 (8): 57–66.

[2] Mannings Robin. 2008. *Ubiquitous Positioning*. Artech House.

[3] Geok Tan Kim, i in. 2021. "Review of Indoor Positioning: Radio Wave Technology". *Appl. Sci.*, 11 (1): 1–44.

[4] Hata Masaharu, Nagatsu Takayoshi. 1980. "Mobile location using signal strength measurements in a cellular system". *IEEE Trans. Veh. Technol.*, 29 (2): 245–252.

[5] Batres Amilcar Ernesto, i in. 2023. "Passive Radio Localization System Using Channel Impulse Response and Deep Learning". *Int. Electr. Eng. Congress (iEECON)*, Thailand, 429–433.

[6] Jin Yunye, i in. 2010. "Indoor localization with channel impulse response based fingerprint and nonparametric regression". *IEEE Trans. Wireless Commun.*, 9 (3): 1120–1127.

[7] Wu Guan-Sian, Tseng Po-Hsuan. 2018. "A Deep Neural Network-Based Indoor Positioning Method using Channel State Information". *Int. Conf. Comput., Netw. Commun. (ICNC)*, USA, 290–294.

[8] Yang Zheng, Zhou Zimu, Liu Yunhao. 2013. "From RSSI to CSI: Indoor localization via channel response". *ACM Comput. Surv.*, 46 (2): 1–32.

[9] Li Hang, Cheng Zhiqun. 2021. "Angle-of-Arrival Estimation Using Difference Beams in Localized Hybrid Arrays". *Sensors*, 21 (5): 1–11.

[10] Li Xiaonan, Bai Li, Li Jing. 2009. "ADOA / TDOA combined positioning with dual-antenna over fading channel". *IEEE Int. Conf. Intell. Comput. and Intell. Systems*, Shanghai, 9–12.

[11] Lee Hyeongtaek, i in. 2019. "Efficient Channel AoD/AoA Estimation Using Widebeams for Millimeter Wave MIMO Systems". *IEEE 20th Int. Workshop on Signal Process. Advances in Wireless Commun. (SPAWC)*, France, 1–5.

[12] Schiller Jochen, Voisard Agnes. 2004. *Location-Based Services*. Elsevier.

[13] LaMarca Anthony, de Lara Eyal. 2008. *Location Systems. An Introduction to the Technology Behind Location Awareness*. Morgan & Claypool.

[14] del Peral-Rosado José A., i in. 2018. "Survey of Cellular Mobile Radio Localization Methods: From 1G to 5G". *IEEE Commun. Survey. Tut.*, 20 (2): 1124-1148.

[15] Yan Chunpeng, Fan Howard. 2008. "Asynchronous differential TDOA for non-GPS navigation using signals of opportunity". *IEEE Int. Conf. on Acoustics, Speech and Signal Process.*, USA, 5312–5315.

[16] Stefański Jacek, Sadowski Jarosław. 2025. "Survey of Asynchronous Radiolocation Methods". *IEEE Access*, 13: 85613-85640.

[17] *FCC Online Table of Frequency Allocations*. 2002. Federal Communications Commission, USA.

[18] Mazurek Gustaw. 2019. "Lightweight SDR Platform for Passive Coherent Location Applications". *Signal Process. Symposium (SPSympo)*, Poland, 39–44.

[19] Das Sree Krishna, Hossain Md Farhad. 2018. "TDOA based Localization Architecture for M2M Communications over Cellular Networks". *10th Int. Conf. on Elect. and Comput. Eng. (ICECE)*, Bangladesh, 333–336.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

ANALIZA WPŁYWU ROZMIESZCZENIA MIKROFONÓW I DOBORU CZĘSTOTLIWOŚCI PRÓBKOWANIA SYGNAŁU AUDIO NA DOKŁADNOŚĆ LOKALIZACJI ŹRÓDŁA DŹWIĘKU

Analysis of the impact of microphone placement and selection od audio sampling rate on the accuracy of sound source localization

Gabriel Jekateryńczuk[1], Zbigniew Piotrowski[2]

[1] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, gabriel.jekaterynczuk@wat.edu.pl
[2] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, zbigniew.piotrowski@wat.edu.pl

DOI: 10.15199/59.2025.4.84

Streszczenie: **W artykule przeanalizowano wpływ rozmieszczenia mikrofonów oraz parametrów próbkowania sygnału audio na dokładność lokalizacji źródła dźwięku w przestrzeni 3D. Zastosowano metodę triangulacyjną opartą na różnicach czasów przybycia (TDOA), estymowanych z wykorzystaniem uogólnionej korelacji krzyżowej (GCC). Analizie poddano różne konfiguracje geometryczne układu mikrofonów oraz częstotliwości próbkowania. Wyniki symulacji potwierdzają istotny wpływ tych czynników na precyzję lokalizacji.**

Abstract: **This paper analyzes the impact of microphone placement and audio sampling parameters on the accuracy of 3D sound source localization. A triangulation method based on time differences of arrival (TDOA) is used, with TDOA values estimated through generalized cross-correlation (GCC). Various microphone array geometries and sampling rates are evaluated. Simulation results confirm a significant influence of these factors on localization accuracy.**

Słowa kluczowe: **lokalizacja źródła dźwięku, mikrofony, symulacja, różnica czasu przybycia, uogólniona korelacja krzyżowa**

Keywords: **sound source localization, microphones, simulation, time difference of arrival, generalized cross-correlation**

1. WSTĘP

W dobie rozwoju inteligentnych systemów monitorujących coraz większe znaczenie zyskują rozwiązania umożliwiające lokalizację źródeł dźwięku. Informacja o położeniu emitera może mieć kluczowe znaczenie zarówno w kontekście bezpieczeństwa, jak i środowiskowych systemów detekcji zdarzeń akustycznych. Współczesne podejścia w tej dziedzinie opierają się najczęściej na zestawach mikrofonów, które rejestrują sygnały docierające do nich z różnych kierunków i odległości, jednak coraz częściej wykorzystuje się także metody uczenia maszynowego – w tym sieci neuronowe analizujące spektrogramy lub cechy akustyczne.

Jednym z istotnych wyzwań pozostaje zapewnienie wysokiej precyzji lokalizacji, zwłaszcza w warunkach rzeczywistych, gdzie pojawiają się zakłócenia, odbicia i szum. W praktyce skuteczność lokalizacji zależy nie tylko od zastosowanego algorytmu, ale również od czynników technicznych i fizycznych, takich jak geometria rozmieszczenia mikrofonów oraz parametry akwizycji sygnału, w tym częstotliwość próbkowania i głębokość bitowa. Nieoptymalne rozmieszczenie mikrofonów może znacząco obniżyć dokładność oszacowania położenia źródła dźwięku, podobnie jak zbyt niska częstotliwość próbkowania może ograniczyć rozdzielczość czasową analizy.

Celem niniejszego artykułu jest identyfikacja oraz ocena wyzwań związanych z rozmieszczeniem mikrofonów i wyborem parametrów próbkowania sygnału w kontekście dokładności lokalizacji źródeł dźwięku. Przedstawiona analiza ma na celu lepsze zrozumienie wpływu wybranych czynników na dokładność lokalizacji w systemach akustycznych.

2. ŚRODOWISKO SYMULACYJNE

Symulacje przeprowadzono z wykorzystaniem biblioteki Pyroomacoustics [2], pozwalającej na realistyczne modelowanie propagacji dźwięku w przestrzeni trójwymiarowej. Środowisko symulacyjne odwzorowuje otwartą przestrzeń bez przeszkód terenowych i odbić, co umożliwia analizę lokalizacji źródła dźwięku w warunkach anechoicznych.

W każdej symulacji wykorzystano osiem mikrofonów, rozmieszczonych względem środka układu w jednej z trzech konfiguracji geometrycznych:

- układ sferyczny – mikrofony rozmieszczone równomiernie na powierzchni wyimaginowanej sfery,
- układ dwupoziomowy – mikrofony rozmieszczone na dwóch poziomach (górnym i dolnym),
- układ dwupoziomowy ortogonalny – mikrofony na dwóch poziomach z przesunięciem pozycji w poziomie.

W celu oceny wpływu skali układu mikrofonowego, analizowano dwa warianty odległości mikrofonów od środka macierzy: 5 metrów i 15 metrów. Dodatkowo badano wpływ częstotliwości próbkowania sygnału audio, testując pięć wartości: 16 kHz, 44,1 kHz, 48 kHz, 96 kHz oraz 192 kHz.

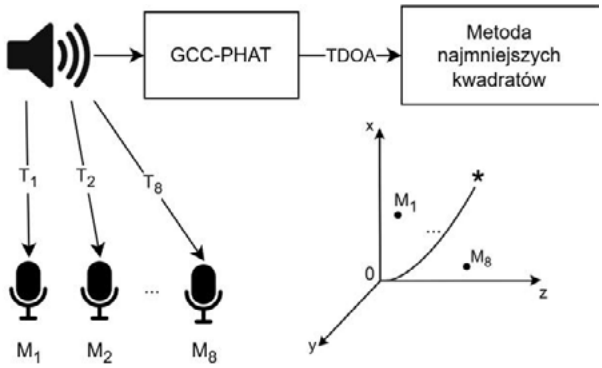
Jako źródło akustyczne wykorzystano nagranie impulsowego dźwięku wystrzału z pistoletu, co umożliwia precyzyjną analizę różnic czasów przybycia sygnału do mikrofonów. Dla każdej kombinacji parametrów testowych obliczono wartości średnie na podstawie 10 niezależnych prób, w których pozycja źródła dźwięku była losowo generowana w ustalonej odległości od środka macierzy mikrofonowej. Dzięki temu możliwe było uśrednienie wyników w obrębie danego scenariusza pomiarowego.

3. OPIS METODY

Metoda lokalizacji źródła dźwięku zastosowana w pracy opiera się na analizie różnic czasów przybycia sygnału (TDOA, ang. Time Difference of Arrival) do zestawu zsynchronizowanych mikrofonów rozmieszczonych w znanych pozycjach w przestrzeni. Różnice te wynikają z faktu, że fala dźwiękowa dociera do każdego mikrofonu w innym czasie, zależnie od odległości źródła dźwięku od danego sensora [3].

W pierwszym etapie procesu wykorzystywana jest uogólniona korelacja krzyżowa (GCC, ang. Generalized Cross-Correlation), która pozwala na estymację TDOA na podstawie par sygnałów zarejestrowanych przez mikrofony. W celu zwiększenia dokładności estymacji zastosowano filtr PHAT (Phase Transform), który normalizuje widmo amplitudowe sygnału, zachowując jedynie informację o fazie – co jest szczególnie korzystne w obecności zakłóceń lub pogłosu [1]. Po uzyskaniu estymowanych opóźnień czasowych pozycja źródła dźwięku wyznaczana jest na drodze rozwiązania układu równań, które opisują relację między opóźnieniami a geometrycznym położeniem źródła względem mikrofonów. Przed przystąpieniem do obliczeń, pozycje mikrofonów są przekształcane w taki sposób, aby ich środek geometryczny znajdował się w początku układu współrzędnych – co upraszcza obliczenia i zmniejsza błędy numeryczne.

Lokalizacja polega na znalezieniu takiego punktu w przestrzeni, którego odległości do mikrofonów – przeliczone na czasy propagacji sygnału – najlepiej odpowiadają zmierzonym opóźnieniom. Do rozwiązania tego zagadnienia stosowana jest metoda najmniejszych kwadratów (ang. Least Squares), która zwraca dwa potencjalne rozwiązania przestrzenne. Wybierane jest to, które minimalizuje błąd pseudoodległości.



Rys. 1. Schemat ilustrujący metodę lokalizacji źródła dźwięku z wykorzystaniem TDOA, algorytmu GCC-PHAT oraz metody najmniejszych kwadratów

Jeśli końcowa wartość błędu średniokwadratowego (RMS) przekracza określony próg, wynik lokalizacji jest odrzucany. W przeciwnym przypadku przyjmuje się go jako wiarygodny, a pozycja źródła zostaje zapisana w postaci współrzędnych kartezjańskich (x, y, z).

4. WYNIKI

W niniejszym rozdziale przedstawiono wyniki symulacji przeprowadzonych w ramach wcześniej opisanego systemu lokalizacji źródła dźwięku. Zaprezentowano porównanie dokładności trzech konfiguracji przestrzennych rozmieszczenia mikrofonów, z uwzględnieniem wpływu częstotliwości próbkowania oraz skali układu. Analizie poddano trzy kluczowe wskaźniki jakości lokalizacji: błąd azymutu, błąd elewacji oraz błąd dystansu, rozumiany jako odległość pomiędzy pozycją rzeczywistą a estymowaną.

4.1. Analiza średnich błędów azymutu, elewacji i lokalizacji

Tabela 1 przedstawia średnie wartości błędów lokalizacji źródła dźwięku uzyskane dla trzech konfiguracji przestrzennych rozmieszczenia mikrofonów: sferycznej, dwupoziomowej oraz dwupoziomowej z przesunięciem ortogonalnym. Dane zostały pogrupowane według dwóch wartości promienia macierzy mikrofonowej (5 m i 15 m) oraz pięciu częstotliwości próbkowania sygnału audio. Dla każdej kombinacji parametrów zestawiono średnie wartości błędu azymutu i elewacji (wyrażone w stopniach), a także błąd lokalizacji, rozumiany jako średnia odległość euklidesowa między pozycją rzeczywistą a estymowaną (w metrach). Tabela umożliwia bezpośrednie porównanie wpływu geometrii, skali układu oraz parametrów próbkowania na dokładność lokalizacji akustycznej w warunkach symulacyjnych.

Analiza uzyskanych wyników wskazuje jednoznacznie, że dokładność lokalizacji źródła dźwięku jest silnie zależna od częstotliwości próbkowania, promienia macierzy mikrofonowej oraz geometrii rozmieszczenia mikrofonów. Wzrost częstotliwości próbkowania przekłada się na systematyczne zmniejszanie się wszystkich trzech analizowanych błędów: azymutu, elewacji oraz dystansu. Najwyższe błędy lokalizacji występują przy najniższej częstotliwości (16 kHz), natomiast najlepsze rezultaty osiągane są przy 192 kHz. Przykładowo, dla konfiguracji sferycznej z promieniem 15 m błąd lokalizacji wynosi zaledwie 0,23 m. Zauważalną poprawę dokładności przynosi także zwiększenie promienia macierzy mikrofonowej. Konfiguracje o promieniu 15 m przewyższają pod względem precyzji te z promieniem 5 m, co wynika z większych różnic czasów dotarcia sygnału (TDOA) i lepszej rozdzielczości kierunkowej. Dla przykładu, w konfiguracji ortogonalnej przy 192 kHz błąd lokalizacji maleje z 9,98 m (dla promienia 5 m) do 0,80 m (dla promienia 15 m). Geometria rozmieszczenia mikrofonów również odgrywa istotną rolę. Konfiguracja sferyczna charakteryzuje się najniższymi błędami azymutu i elewacji, zwłaszcza przy większym promieniu macierzy. Z kolei konfiguracje dwupoziomowe z przesunięciem ortogonalnym uzyskują konkurencyjne, a często bardzo dobre wyniki w zakresie błędu lokalizacji, szczególnie przy wyższych częstotliwościach próbkowania.

Tab. 1. Średnie błędy kąta azymutu, elewacji oraz lokalizacji źródła dźwięku dla różnych konfiguracji przestrzennych mikrofonów, częstotliwości próbkowania i promieni macierzy.

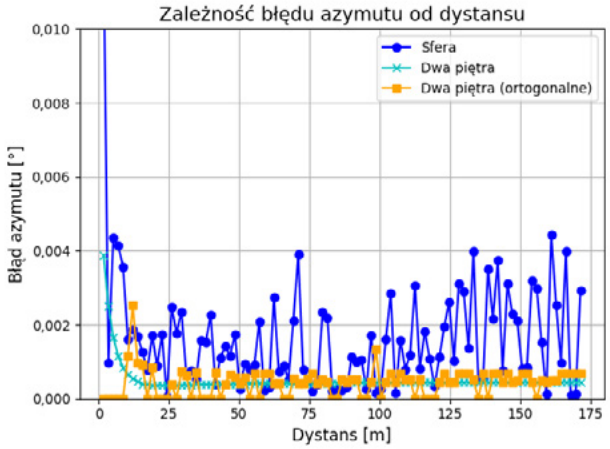
Promień [m]	*Konfiguracja*	*Częstotliwość [kHz]*	*Błąd azymutu [°]*	*Błąd elewacji [°]*	*Błąd lokalizacji [m]*
5	Sfera	16	0,06	2,55	22,86
		44,1	0,03	1,49	21,20
		48	0,02	1,11	12,83
		96	0,01	0,56	11,95
		192	0,01	0,28	12,13
	Dwa poziomy	16	0,01	19,01	35,11
		44,1	0,02	13,98	22,82
		48	0,02	8,28	35,55
		96	0,01	13,84	21,86
		192	0,00	2,92	12,61
	Dwa poziomy przesunięte ortogonalnie względem siebie	16	0,01	17,51	34,39
		44,1	0,02	13,98	22,82
		48	0,02	6,67	15,48
		96	0,01	3,32	9,68
		192	0,00	2,25	9,98
15	Sfera	16	0,02	0,84	1,50
		44,1	0,01	0,50	0,85
		48	0,01	0,51	0,86
		96	0,00	0,19	0,37
		192	0,00	0,07	0,23
	Dwa poziomy	16	0,01	11,95	9,77
		44,1	0,01	2,94	6,55
		48	0,01	5,52	5,52
		96	0,00	2,09	3,58
		192	0,00	0,97	1,85
	Dwa poziomy przesunięte ortogonalnie względem siebie	16	0,00	9,62	6,23
		44,1	0,01	3,16	3,65
		48	0,01	3,42	2,59
		96	0,00	1,38	1,45
		192	0,00	0,78	0,80

W wielu przypadkach błąd elewacji przewyższa błąd azymutu – szczególnie w konfiguracjach dwupoziomowych przy niskich częstotliwościach – co sugeruje większą wrażliwość estymacji wysokości na geometrię układu i rozdzielczość czasową. Najlepsze wyniki uzyskano dla konfiguracji sferycznej o promieniu 15 m i częstotliwości 192 kHz (błąd lokalizacji 0,23 m). Dobrze wypadła również konfiguracja ortogonalna (0,80 m). Dla mniejszego promienia (5 m) najniższy błąd również osiągnięto w konfiguracji sferycznej przy 192 kHz (12,13 m).

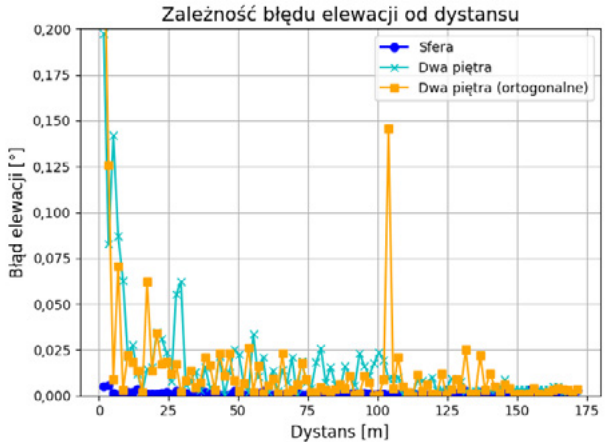
Uzupełniająco przeprowadzono analizę współczynnika GDOP (Geometric Dilution of Precision), która potwierdziła przewagę układów sferycznych o większym promieniu i ich niższą podatność na błędy pomiarowe.

4.2. Wizualizacja wyników symulacji

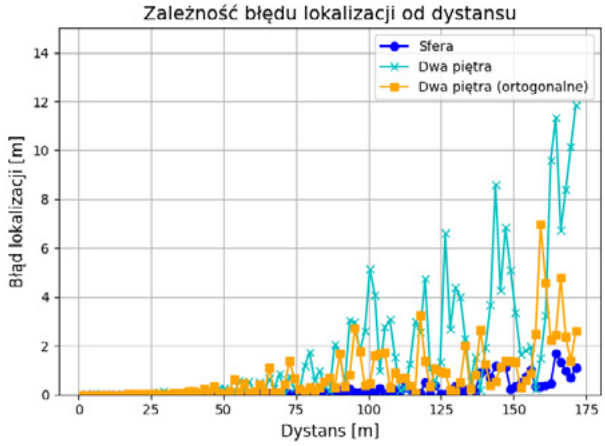
Na poniższych wykresach przedstawiono zależność trzech typów błędów lokalizacji – azymutu, elewacji oraz dystansu – od rzeczywistej odległości źródła dźwięku od środka macierzy mikrofonowej. Analiza przeprowadzona została dla wszystkich konfiguracji, przy promieniu równym 15 m oraz częstotliwości próbkowania 192 kHz. Wyniki ukazują wyraźnie, że konfiguracja sferyczna charakteryzuje się najniższymi błędami azymutu i elewacji w całym zakresie dystansu, zachowując jednocześnie stabilność wyników. Dla pozostałych konfiguracji zauważalne są wyższe odchylenia, zwłaszcza w zakresie elewacji, co wskazuje na większą wrażliwość tych układów na zmiany kierunku w płaszczyźnie pionowej.



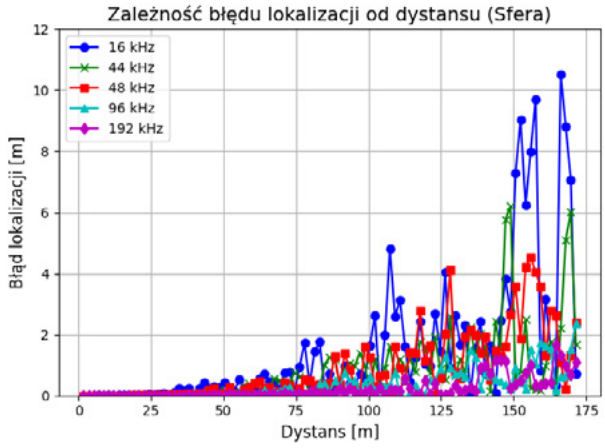
Rys. 2. Zależność błędu azymutu w funkcji dystansu



Rys. 3. Zależność błędu elewacji w funkcji dystansu



Rys. 4. Zależność błędu lokalizacji w funkcji dystansu



Rys. 5. Zależność błędu lokalizacji w funkcji dystansu dla różnych częstotliwości oraz konfiguracji sfery

Dodatkowo, rysunek 5 przedstawia wpływ częstotliwości próbkowania na dokładność lokalizacji w przypadku konfiguracji sferycznej. Można zaobserwować wyraźną zależność pomiędzy wzrostem częstotliwości a zmniejszeniem się błędów lokalizacji – im wyższa częstotliwość, tym mniejsze odchylenia estymowanej pozycji względem rzeczywistej. Efekt ten jest szczególnie widoczny przy większych odległościach źródła od macierzy mikrofonowej, gdzie różnice między badanymi wariantami są najbardziej wyraźne. Największe błędy pojawiają się przy najniższej częstotliwości próbkowania (16 kHz), natomiast najlepsze wyniki osiągane są przy 192 kHz, co jednoznacznie podkreśla znaczenie wysokiej rozdzielczości czasowej sygnału w kontekście precyzyjnej lokalizacji akustycznej.

5. PODSUMOWANIE

W pracy przeanalizowano wpływ wybranych parametrów systemu lokalizacji akustycznej na dokładność estymacji położenia źródła dźwięku. Badania przeprowadzono w środowisku symulacyjnym, wykorzystując osiem mikrofonów rozmieszczonych w trzech różnych konfiguracjach przestrzennych: sferycznej, dwupoziomowej oraz dwupoziomowej z przesunięciem ortogonalnym. Dodatkowymi zmiennymi były promień macierzy mikrofonowej (5 m i 15 m) oraz częstotliwość próbkowania sygnału (od 16 do 192 kHz).

Zarówno analiza numeryczna (tabelaryczna), jak i graficzna wykazały, że największy wpływ na jakość lokalizacji mają częstotliwość próbkowania oraz geometria układu. Wzrost częstotliwości próbkowania systematycznie obniżał błędy azymutu, elewacji i dystansu – najwyraźniej w przypadku większych odległości źródła od mikrofonów. Z kolei zwiększenie promienia macierzy mikrofonowej poprawiało rozdzielczość kierunkową, co przekładało się na znaczną redukcję błędów lokalizacji, szczególnie dla konfiguracji sferycznej.

Najlepsze rezultaty osiągnięto przy zastosowaniu konfiguracji sferycznej, promienia 15 m oraz częstotliwości 192 kHz – średni błąd lokalizacji wyniósł w tym przypadku zaledwie 0,23 m. Konfiguracja ortogonalna również uzyskała dobre wyniki, choć wykazywała większe odchylenia w zakresie elewacji. W przypadku mniejszych promieni i niższych częstotliwości zauważalne były wyraźnie większe błędy, szczególnie w konfiguracjach dwupoziomowych, co sugeruje ich ograniczoną skuteczność w bardziej wymagających warunkach.

Podsumowując, zaprezentowana analiza potwierdza, że optymalizacja rozmieszczenia mikrofonów oraz dobór parametrów systemowych – takich jak częstotliwość próbkowania i skala układu – mają kluczowe znaczenie dla skuteczności lokalizacji źródeł dźwięku. Symulacje uwzględniały obecność szumu, jednak zostały przeprowadzone w środowisku bez pogłosu i bez błędów kalibracji. W przyszłych pracach planowane jest rozszerzenie analizy o bardziej realistyczne scenariusze, w tym efekt pogłosu i niedoskonałości sprzętowych.

LITERATURA

[1] Grondin, Francois, and James Glass. 2018. *A Study of the Complexity and Accuracy of Direction of Arrival Estimation Methods Based on GCC-PHAT for a Pair of Close Microphones.* arXiv. https://arxiv.org/abs/1811.11787.

[2] Scheibler, Robin, Emanuele Bezzam, and Ivan Dokmanić. Pyroomacoustics: *A Python Package for Audio Room Simulations and Array Processing Algorithms*. Paper presented at the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Canada, 2018.

[3] Wu, Peng, Shaojing Su, Zhen Zuo, Xiaojun Guo, Bei Sun, and Xudong Wen. 2019. "*Time Difference of Arrival (TDoA) Localization Combining Weighted Least Squares and Firefly Algorithm*" *Sensors* 19, no. 11: 2554. https://doi.org/10.3390/s19112554

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

DYNAMICZNA ESTYMACJA BŁĘDU LOKALIZACJI W SYSTEMIE OPARTYM NA WI-FI

DYNAMIC ACCURACY ESTIMATION IN A WI-FI–BASED POSITIONING SYSTEM

Marcin Kołakowski[1]; Vitomir Djaja-Jośko[2]

[1] Politechnika Warszawska, IRiTM, Warszawa, marcin.kolakowski@pw.edu.pl
[2] Politechnika Warszawska, IRiTM, Warszawa, vitomir.djaja-josko@pw.edu.pl

DOI: 10.15199/59.2025.4.85

Streszczenie: W referacie przedstawiono koncepcję metody, w której błąd lokalizacji jest dynamicznie estymowany na podstawie wyników pomiarów wykorzystanych do określenia położenia obiektu. W ramach eksperymentów przeprowadzonych z wykorzystaniem systemu opartego na interfejsie Wi-Fi zbadano kilka algorytmów uczenia maszynowego. Dla najlepszego (lasu losowego), średni błąd estymacji błędu lokalizacji wyniósł 0.72 m.

Abstract: The paper presents a concept of a dynamic localization accuracy estimation method, which uses the same data as the positioning algorithm. In an experiment, in which a Wi-Fi–based system was used, several machine learning methods were tested. The highest positioning error estimation accuracy was achieved for random forest regression - mean absolute error of 0.72 m.

Słowa kluczowe: lokalizacja wewnątrzbudynkowa, dynamiczna estymacja dokładności, uczenie maszynowe, sieci neuronowe

Keywords: indoor positioning, dynamic accuracy estimation, machine learning, neural networks

1. WSTĘP

Niezależnie od zastosowanej technologii, jednym z najistotniejszych parametrów radiowych systemów lokalizacyjnych jest ich dokładność. To z jak małym błędem system jest w stanie określić lokalizację monitorowanych obiektów bezpośrednio wpływa na możliwość świadczenia określonych usług.

W przypadku większości systemów, dokładność jest podawana w nocie katalogowej jako uśredniona wartość błędu lokalizacji zmierzonego w określonych warunkach. Informacja ta daje ogólny obraz potencjalnych błędów lokalizacji, ale nie pozwala określić jakości pojedynczych wyników.

Informacja na temat możliwego błędu z jakim wyznaczona jest lokalizacja monitorowanego obiektu może być istotna w wielu zastosowaniach. Przykładowo, przekazanie jej użytkownikowi przy wizualizacji wyniku na planie budynku może pozwolić mu na jego łatwiejszą interpretację. W przypadku systemów autonomicznych, informacja o dokładności może być dodatkową zmienną wykorzystywaną przy podejmowaniu decyzji. W związku z tym, prowadzone są badania mające na celu opracowanie metod dynamicznej estymacji dokładności lokalizacji (ang. Dynamic Accuracy Estimation - DAE).

Celem metod DAE jest określenie potencjalnego błędu z jakim została wyznaczona lokalizacja monitorowanego obiektu. Typowo, wynik estymacji jest promieniem okręgu, w którym z założonym prawdopodobieństwem powinna znajdować się się rzeczywista lokalizacja użytkownika [2]. W przypadku metod rozróżniających błędy w poszczególnych osiach współrzędnych [8], obszar ten ma kształt elipsy. Błąd lokalizacji jest zazwyczaj estymowany bezpośrednio na podstawie wyników pomiarów wykorzystywanych do lokalizacji [7]. W niektórych metodach są wykorzystywane dodatkowe parametry odbieranego sygnału jak wariancja poziomu mocy [1].

Istnieją dwie klasy metod dynamicznej estymacji dokładności lokalizacji: metody oparte na modelowaniu matematycznym (ang. rule-based methods) oraz oparte na danych (ang. data-driven methods). W metodach opartych na modelowaniu poszukiwane są bezpośrednie zależności między zmierzonymi parametrami sygnału a dokładnością przeprowadzonej na ich podstawie lokalizacji. Przykładami są metoda oparta na analizie dolnej granicy błędu Cramer-Rao [6] oraz metoda propagacji błędu [8], w której niepewności wyników pomiarów są bezpośrednio przeliczane na wartości błędów wyznaczenia lokalizacji.

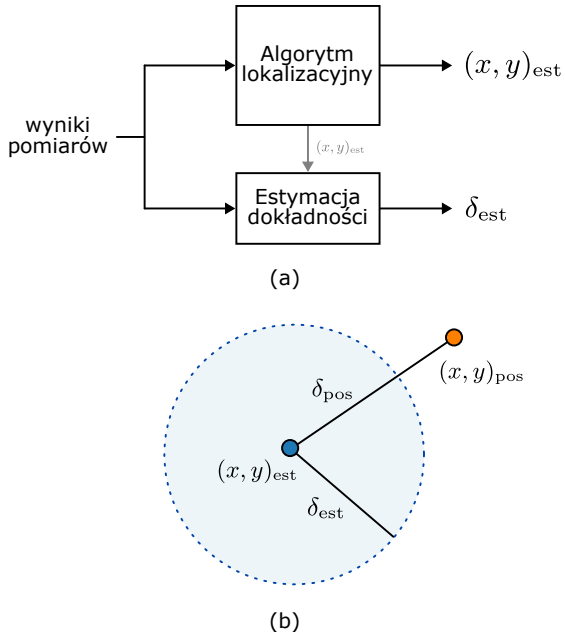
Postępujący w ostatnich latach rozwój metod uczenia maszynowego sprawił, że coraz większą popularnością zaczęły cieszyć się metody oparte na danych. Porównanie przeprowadzone w [4], dowiodło że zastosowanie tych metod pozwala uzyskać lepsze wyniki estymacji błędu niż metody pierwszej kategorii. Metody DAE oparte na danych stosują zazwyczaj algorytmy regresji takie jak lasy losowe czy maszyny wektorów nośnych [2, 7]. W przypadku większych zbiorów danych stosowane są również głębokie sieci neuronowe [3].

W referacie przedstawiono analizę możliwości zastosowania metod DAE opartych na danych do estymacji dokładności lokalizacji w systemie opartym na Wi-Fi, w którym mapa radiowa została w szybki sposób skonstruowana przy użyciu platformy jezdnej. W ramach badań eksperymentalnych przeprowadzono testy kilku metod regresji (algorytm k-najbliższych sąsiadów, las losowy, regresja liniowa, sieci neuronowe). Przeprowadzono również dodatkową weryfikację wytrenowanych modeli na danych zebranych przez użytkownika.

2. DYNAMICZNA ESTYMACJA DOKŁADNOŚCI

2.1 Koncepcja metody

Koncepcja dynamicznej estymacji dokładności lokalizacji zakłada implementację równoległej do algorytmu lokalizacyjnego metody pozwalającej na oszacowanie jakości otrzymywanych wyników. Schemat ideowy metody przedstawiono na Rysunku 1.a.



Rysunek 1: Ogólna idea dynamicznej estymacji dokładności lokalizacji; a) schemat algorytmu; b) ilustracja zależności między estymatą błędu a rzeczywistą lokalizacją.

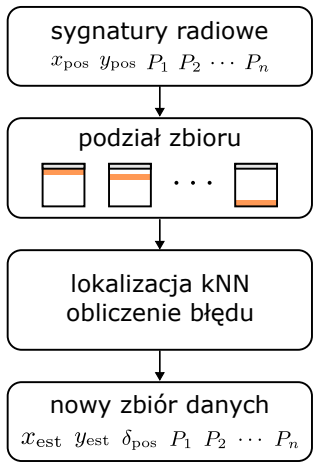
W zaproponowanej koncepcji, estymacja dokładności lokalizacji odbywa się na podstawie tych samych wyników pomiarów, które są przetwarzane przez algorytm lokalizacyjny. W badaniach eksperymentalnych przetestowano również wariant, w którym dodatkowo wykorzystywany jest wynik lokalizacji. Wynikiem działania metody jest estymowana wartość błędu δ_{est}. Zależność δ_{est} od rzeczywistego położenia obiektu i wyniku lokalizacji została przedstawiona na Rysunku 1.b.

Otrzymana estymata błędu δ_{est} może być zarówno większa jak i mniejsza od jego rzeczywistej wartości δ_{pos}. W przypadku systemów, w których na jej postawie podejmowane są istotne z punktu widzenia ich działania decyzję, parametry algorytmów DAE powinny zostać w odpowiedni sposób dostosowane.

2.2 Implementacja w systemie opartym na Wi-Fi

W ramach badań, w systemie lokalizacyjnym Wi-Fi zaimplementowano opartą na danych metodę DAE. W wykorzystanym systemie lokalizacja odbywa się na podstawie pomiarów mocy sygnałów odbieranych przez smartfon z punktów dostępowych Wi-Fi działających w pasmach 2,4 i 5 GHz. Do obliczenia lokalizacji wykorzystywana jest metoda mapy radiowej (ang. fingerprinting). Do lokalizacji wykorzystano algorytm k-najbliższych sąsiadów z k=4. Estymacja błędu odbywa się za pomocą metody regresji wytrenowanej na uprzednio przygotowanym zbiorze danych.

Zbiór danych, na podstawie którego trenowana jest metoda DAE powinien możliwie wiernie odzwierciedlać zależność między parametrami sygnału a rzeczywistymi błędami lokalizacji w systemie. Procedurę tworzenia zbioru danych przedstawiono na Rysunku 2.



Rysunek 2: Proces tworzenia zbioru danych do treningu metod DAE

Wstępny zbiór danych zawiera sygnatury radiowe - zmierzone poziomy sygnałów z punktów dostępowych wraz z lokalizacją punktu pomiarowego. Pierwszym krokiem jest dokonanie kilku podziałów zbiorów na treningowy i testowy zgodnie z zasadami znanymi z metody walidacji krzyżowej (za każdym razem zbiór testowy zawiera inne sygnatury). Dla każdego podziału zbiór treningowy jest traktowany jako mapa radiowa na podstawie której obliczana jest lokalizacja punktów ze zbioru testowego. Dla każdej z wyznaczonych lokalizacji $(x_{\text{pos}}, y_{\text{pos}})$ wyznaczany jest błąd lokalizacji δ_{pos}. Wartości te są łączone z poziomami sygnałów tworząc część składową zbioru danych, który zostanie wykorzystany do wytrenowania metody DAE.

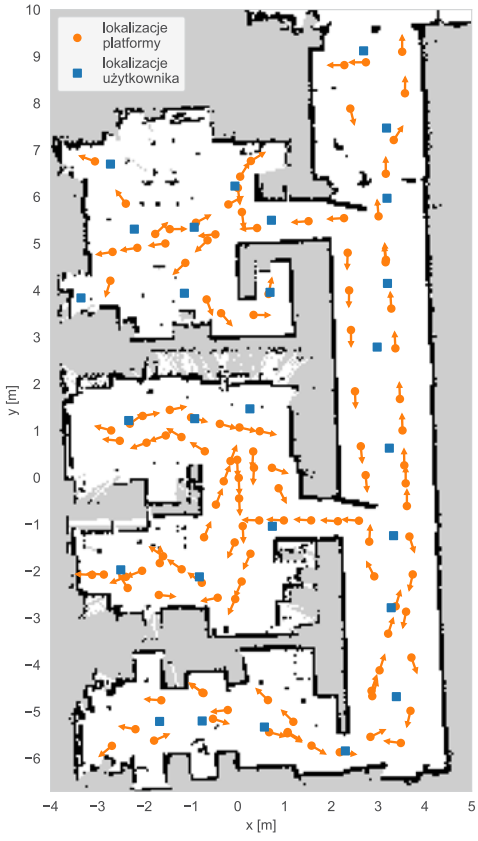
3. BADANIA EKSPERYMENTALNE

3.1 Środowisko eksperymentalne

Opisana koncepcja została zbadana eksperymentalnie w pomieszczeniach biurowych i laboratoryjnych Wydziału Elektroniki i Technik Informacyjnych Politechniki Warszawskiej. Ogólny plan pomieszczeń zaprezentowano na Rysunku 4. Lokalizacja badań cechuje się skomplikowanymi warunkami propagacji sygnału wynikającymi ze znaczącej ilości sprzętu zgromadzonego w pokojach oraz obecności elementów metalowych w ścianach.

Przeprowadzone badania polegały na akwizycji danych z wykorzystaniem platformy jezdnej oraz wytrenowaniu na ich podstawie metod i modeli pozwalających na ocenę dokładności wyznaczanych lokalizacji.

Rysunek 3: Wykorzystana do akwizycji danych platforma jezdna w środowisku eksperymentalnym



Rysunek 4: Lokalizacja punktów pomiarowych. Strzałki wskazują orientację platformy.

Podczas badań, smartfon (Samsung A41) był zamontowany na statywie przymocowanym do platformy, której zdjęcie przedstawiono na Rysunku 3. W trakcie badań, wysokość terminala nad podłogą wynosiła 0,74 m. Referencyjna lokalizacja platformy w punktach pomiarowych została określona z kilkucentymetrową dokładnością za pomocą lidaru RPLidar A2M12 o nominalnym zasięgu 12 m. Cała platforma sterowana była za pomocą bibliotek i narzędzi pakietu ROS (Robot Operating System).

W celu dodatkowej weryfikacji metody dynamicznej estymacji dokładności, przeprowadzono eksperyment, w którym dane zostały zebrane przez użytkownika trzymającego smartfon w dłoni.

3.2 Zbiory danych

Główny zbiór stanowiło 359 sygnatur radiowych (poziomy sygnałów Wi-Fi, referencyjna lokalizacja terminala) zebranych za pomocą platformy jezdnej w 117 punktach (minimum 3 skany na punkt). Lokalizacje punktów pomiarowych przedstawiono na Rysunku 4. Na badanym obszarze wykryto 78 różnych punktów dostępowych. Ze względu na niewielką dostępność części z nich, do dalszych prac wybrano 35. W związku z tym, że część testowanych metod regresji wymaga kompletnych danych, brakujące wyniki pomiarów zastąpiono stałą wartością -99 dBm (niższa, niż najniższa wartość mierzona przez wykorzystany smartfon). Za pomocą procedury opisanej w sekcji 2.2 utworzono zbiór do wytrenowania metody estymacji dokładności. Średni błąd lokalizacji robota w pomieszczeniu wyniósł 0,97 m.

W drugim zbiorze znajdowało się 108 sygnatur zebranych przez użytkownika w 27 punktach zaznaczonych na Rysunku 4. Średni błąd lokalizacji użytkownika na podstawie mapy utworzonej na podstawie danych zebranych za pomocą robota był stosunkowo wysoki i wyniósł 2,49 m. Wszystkie dane wykorzystane w badaniach są publicznie dostępne w serwisie Github [5].

3.3 Weryfikacja dokładności koncepcji

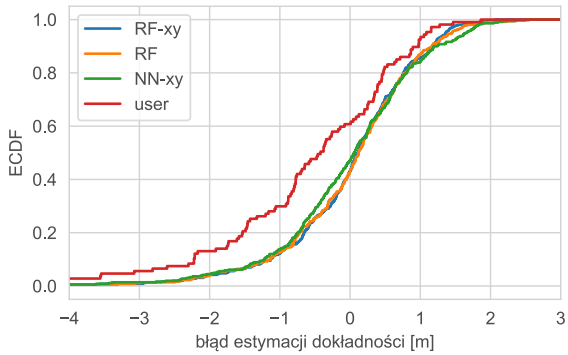
W celu weryfikacji koncepcji, za pomocą utworzonogo zbioru danych wytrenowano kilka różnych algorytmów regresji: regresji liniowej (LR), lasu losowego (RF), k-najbliższych sąsiadów (kNN) oraz gęstych sieci neuronowych (NN). Wszystkie metody z wyjątkiem sieci neuronowych zostały zaimplementowane z wykorzystaniem pakietów scikit-learn. Sieci neuronowe zaimplementowano korzystając z biblioteki pytorch. W ramach badań przetestowano kilkanaście architektur gęstych sieci neuronowych składających się z warstw gęstych po których następowała normalizacja wsadowa zapobiegająca przeuczeniu modelu. W referacie przedstawiono jedynie wyniki dla najlepszych modeli (liczba neuronów w poszczególnych warstwach została zamieszczona jako parametr w Tabeli 1).

W ramach badań oprócz metody przetwarzającej jedynie wyniki pomiarów przetestowano wariant wykorzystujący wynik lokalizacji (oznaczony sufiksem -xy). Jako miarę porównawczą przyjęto błąd estymacji dokładności lokalizacji zdefiniowany jako $\mathrm{DAE}_{\mathrm{error}} = \delta_{\mathrm{est}} - \delta_{\mathrm{pos}}$. Wartości średniego modułu błędu (MAE - Mean Absolute Error) oraz średniego błędu kwadratowego (MSE - Mean Squared Error) wraz z parametrami metod zebrano w Tabeli 1. Empiryczną dystrybuantę błędu estymacji dokładności oraz zależność estymowanych wartości od rzeczywistych przedstawiono na Rysunkach 5 i 6.

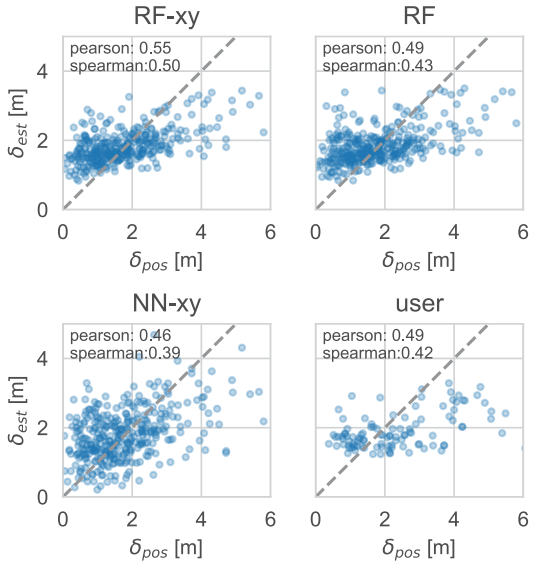
W przypadku wszystkich metod, uwzględnienie wyznaczonego położenia smartfona prowadzi do zwiększenia dokładności estymacji błędów lokalizacji. Największa dokładność została uzyskana dla lasu losowego (średni moduł błędu wynosił 0,726 m). Ta metoda została wykorzystana do określenia dokładości lokalizacji użytkownika. W tym przypadku błąd był większy co wynika z innych warunków w których została utworzona mapa radiowa (brak przysłaniania ciałem, smartfon na statywie).

Tabela 1: Średnie błędy estymacji dokładności lokalizacji platformy i użytkownika (user)

Metoda	*Parametry*	*MAE*	*MSE*
LR	-	0,926	1,529
LR-xy	-	0,921	1,511
RF	trees=100	0,753	1,046
RF-xy	trees=300	0,726	0,980
kNN	k=4	0,831	1,295
kNN-xy	k=4	0,813	1,261
NN	[128, 128, 128]	0,859	1,283
NN-xy	[256, 512, 256]	0,806	1,135
user	RF-xy (300)	1,082	2,112



Rysunek 5: Empiryczna dystrybuanta błędu estymacji dokładności



Rysunek 6: Zależność między estymowaną dokładnością a rzeczywistym błędem

W wynikach estymacji błędu platformy ruchomej nie występuje widoczna tendencja do zaniżania lub zawyżania otrzymanych błędów. Tendencja ta występuje dla lokalizacji użytkownika gdzie estymacje są bardziej optymistyczne niż w rzeczywistości. W przypadku analizowanych, najdokładniejszych metod, estymowane błędy wykazują umiarkowaną korelację z wynikami rzeczywistymi. Współczynnik korelacji Pearsona wynosi około 0.5 co jest wynikiem zbliżonym do osiąganego przez inne metody [4].

4. PODSUMOWANIE

W referacie przedstawiono metodę estymacji błędu lokalizacji w systemie lokalizacyjnym wykorzystującym interfejs Wi-Fi. Metoda polega na wytrenowaniu modelu uczenia maszynowego, który na podstawie wyników pomiarów ocenia dokładność wyznaczanej lokalizacji. Eksperymentalna weryfikacja metody wykazała, że może zostać ona wykorzystana do estymacji błędów lokalizacji. Jakość predykcji błędu spada w przypadku gdy lokalizowany jest użytkownik i pomiary przeprowadzane są w innych warunkach. w takim przypadku konieczna może być adaptacja modelu.

Przedstawiona koncepcja ma potencjał do dalszego rozwoju. Możliwym kierunkiem jest na przykład uwzględnienie innych parametrów bądź zastosowanie funkcji straty penalizujących optymistyczne prognozy.

W referacie przedstawiono wyniki prac częściowo sfinansowanych w ramach projektu finansowanego przez Narodowe Centrum Badań i Rozwoju w ramach programu THCS JTC1, nr umowy THCS/I/24/RENEW/2025.

LITERATURA

[1] Chen, Wei i inni. 2020. „The Analysis of Positioning Accuracy Based on Range and Angle Measurements" In 2020 3rd International Conference on Unmanned Systems (ICUS), 1030–34.

[2] Anagnostopoulos Grigorios G. i Kalousis Alexandros. 2021. „Analysing the Data-Driven Approach of Dynamically Estimating Positioning Accuracy" In ICC 2021 - IEEE International Conference on Communications, 1–7.

[3] Lemic Filip, i Famaey Jeroen . 2020. „Artificial Neural Network-Based Estimation of Individual Localization Errors in Fingerprinting" In 2020 IEEE 17th Annual Consumer Communications & Networking Conference (CCNC), 1–6. Las Vegas, NV, USA: IEEE.

[4] Anagnostopoulos Grigorios G. i Kalousis Alexandros. 2022. „Can I Trust This Location Estimate? Reproducibly Benchmarking the Methods of Dynamic Accuracy Estimation of Localization" Sensors 22 (3): 1088.

[5] Kołakowski Marcin, 2025, „Dataset: DAE fingerprints 2025", https://github.com/marckolak/dae-fingerprints-2025 (dostęp 28.05.2025)

[6] Nikitin Artyom, i inni. 2017. „Indoor Localization Accuracy Estimation from Fingerprint Data." In 2017 18th IEEE International Conference on Mobile Data Management (MDM), 196–205.

[7] Lemic Filip i inni. 2019. „Toward Regression-Based Estimation of Localization Errors in Fingerprinting-Based Localization" In 2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring), 1–5. Kuala Lumpur, Malaysia: IEEE.

[8] Valero-Abundio Cristian i inni. 2024. „Using Ellipses to Improve the Estimation of the Error Bound on Indoor Positioning Problems." In 2024 International Symposium on Sensing and Instrumentation in 5G and IoT Era (ISSI), 1–6. Lagoa, Portugal: IEEE.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

NAWIGACJA INERCYJNA PRZY POMOCY SEKWENCYJNEJ SIECI NEURONOWEJ Z WYKORZYSTANIEM NIEWYSPECJALIZOWANYCH CZUJNIKÓW IMU

INERTIAL NAVIGATION USING A SEQUENTIAL NEURAL NETWORK WITH CONSUMER-GRADE IMU SENSORS

Maciej Kawka; Jakub Kołodziej; Paweł Skokowski

Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa
{maciej.kawka, jakub.kolodziej, pawel.skokowski }@wat.edu.pl

DOI: 10.15199/59.2025.4.86

Streszczenie**: Niniejszy artykuł przedstawia propozycję nawigacji inercyjnej przy pomocy sekwencyjnej sieci neuronowej z wykorzystaniem niewyspecjalizowanych czujników IMU. Głównym celem jest uzyskanie odpowiedzi na pytanie, czy możliwe jest wykorzystanie czujników IMU klasy konsumenckiej do estymowania położenia obiektów z wykorzystaniem sieci neuronowej jako środka do fuzji i estymacji danych. W ramach prac przeprowadzono badania w warunkach rzeczywistych z wykorzystaniem proponowanego rozwiązania.**

Abstract**: This paper proposes an inertial navigation using a sequential neural network with non-specialized IMU sensors. The primary objective is to determine whether consumer-grade IMU sensors can be utilized to estimate the position of objects using a neural network as a means of data fusion and estimation. Tests were conducted using the proposed solution in a real-world scenario as part of the work.**

Słowa kluczowe**: fuzja danych, INS, LSTM, IMU, nawigacja inercyjna, sieci neuronowe**

Keywords**: data fusion, INS, inertial navigation, LSTM IMU, neural networks**

1. WSTĘP

Współczesne systemy nawigacyjne i śledzące pozycję obiektów coraz częściej wykorzystują tanie, konsumenckie czujniki IMU (*eng. Inertial Measurement Unit*), które charakteryzują się niewielkim rozmiarem oraz szeroką dostępnością. Z drugiej strony, czujniki te cechują się znaczną niestabilnością pomiarów, błędami dryfu oraz wysokim poziomem szumów. Zwykle estymacja położenia obiektów szczególnie jeśli są to obiekty latające (bezzałogowce) odbywa się w połączniu z danymi z odbiornika GNSS, przez wgląd na ich precyzję. Stąd lot czy nawigacja w warunkach o ograniczonym dostępie do danych z globalnych systemów pozycjonujących stanowi duże wyzwanie. Nawigacja inercyjna z klasycznej definicji jest możliwa do uzyskania, lecz metoda ta oczekuje bardzo dokładnych danych oraz dużej ilości obliczeń. Szczególnie przy wysokiej precyzji wymaga uwzględnienia panujących warunków pogodowych oraz charakterystyki lotu poszczególnych platform.

W literaturze coraz wyraźniej zaznacza się potencjał sieci neuronowych jako elastycznego narzędzia do fuzji danych pomiarowych oraz korekcji błędów tradycyjnych metod[1][2]. Widać, iż odpowiednio skonstruowane modele uczenia maszynowego mogą nie tylko kompensować dryf, czy adaptować się do zmiennych charakterystyk czujników, lecz także uwzględniać zmienne warunki środowiska. W przypadku proponowanej implementacji pozwala to na uzyskiwanie wyższej dokładności estymacji trajektorii obiektów. Niniejszy artykuł ma na celu zbadanie możliwości wykorzystania sieci neuronowej do estymowania położenia obiektów z wykorzystaniem niewyspecjalizowanych czujników IMU[3]. Szczegółowe cele pracy obejmują: implementację urządzenia zdolnego do pozyskiwania dokładnych, niezależnych, lecz skorelowanych danych pomiarowych potrzebnych do dalszych części badań; analizie pozyskanych danych; zaproponowaniu architektury sieci neuronowej zdolnej do fuzji danych oraz estymacji położenia przy wykorzystaniu jedynie danych z IMU oraz ocenie jakości estymacji na podstawie eksperymentów przeprowadzonych w środowisku rzeczywistym.

2. KLASYCZNA METODA PRZYROSTOWA

Klasyczna nawigacja inercyjna, polega na ciągłej integracji pomiarów sensorów, zwykle żyroskopu i akcelerometru w celu wyznaczenia orientacji, prędkości, a ostatecznie pozycji obiektu w czasie. W przypadku metody przyrostowej podstawowe zależności można opisać w kilku krokach. Początkowym etapem dla wielu urządzeń INS jest etap kalibracji i inicjalizacji. W tym punkcie realizowana jest orientacja sensorów (czyli finalne zorientowanie obiektu tak jak na Rys. 1) względem układu nawigacyjnego czy geograficznego, do czego wykorzystać można wiedzę a priori lub inne układy nawigacyjne jak magnetometr, czy odbiornik GNSS. W tym momencie przyjmowane są początkowe założenia odnośnie do błędów sensorów (ich offsetu, dryfów oraz szumów) oraz przyjęcie prędkości początkowych.

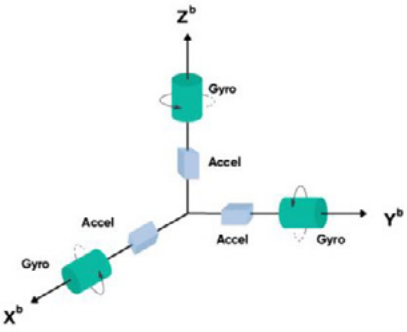
Kolejnym etapem jest estymacja orientacji przestrzennej w przyjętym położeniu nawigacyjnym. Następuje ona poprzez wyznaczanie przyrostu kąta obrotu czujnika - obliczenie przesunięcia kątowego w zdefiniowanym kroku czasowym i znanych prędkościach kątowych z żyroskopu. Całość uproszczonego modelu nawigacyjnego można zapisać jako klasyczne równanie ruchu obiektów (1):

$$\begin{cases} v_k = v_{k-1} + a_{k-1}\Delta t, \\ p_k = p_{k-1} + v_{k-1}\Delta t + \frac{1}{2}a_{k-1}(\Delta t)^2 \end{cases} \quad (1)$$

Gdzie:
p_k – wektor pozycji w ukł. nawigacyjnym w chwili k
v_k– wektor prędkości w ukł. nawigacyjnym w chwili k
a_{k-1}– wektor przysp. w ukł. nawigacyjnym w chwili k-1
Δt – odstęp czasu pomiędzy pomiarami [4].

Rys. 1 Quadcopter z zaznaczonymi osiami obrotu

Rozłożenie przestrzennej estymacji wektorów magnetometru i żyroskopu przedstawia Rys. 2.



Rys. 2 Akcelerometry i żyroskopy w 3 osiach ruchu (ortagonalnie) [5]

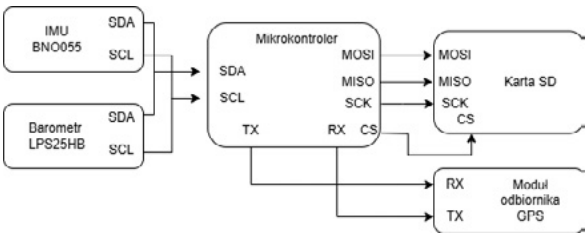
Następnie wyznacza się poszczególne przyspieszenia, a następnie prędkości poprzez wiedzę o położeniu przestrzennym (koniecznym do wyznaczenia różnicy pomiędzy wektorem grawitacji a przyspieszeń wyznaczanych z czujnika). Przyspieszenia w kolejnych krokach czasowych dają prędkości, które przy znanych wartościach położenia przestrzennego stanowią o kierunku i drodze obiektu. Przykładowe działanie INS przestawione jest na Rys.3.

Rys. 3 Blokowe działanie INS w celu wyznaczenia danych do estymacji położenia

W klasycznej metodzie przyrostowej każdy etap przetwarzania danych z IMU niesie ze sobą potencjalne źródła błędu, które z czasem kumulują się, prowadząc do coraz większych odchyleń w estymowanej pozycji. Żyroskopy w konsumenckich IMU mają tendencję do dryfu – czyli w miarę upływu czasu ich wskazania kątowych prędkości własnych powoli odbiegają od rzeczywistych wartości. Oznacza to, iż nawet przy pozornie niedużym błędzie offsetu (niekoniecznie stałym i zdefiniowanym) po pewnym czasie działania wypracowany wynik orientacji nie będzie zgodny z rzeczywistością. Akcelerometr także sam z siebie wprowadza nieduży, lecz ciągły szum, który dodatkowo wprowadza znaczne błędy obliczeniowe. Dynamiczna praca tego typu układów w rzeczywistym środowisku operacyjnym jest źródłem dodatkowych błędów. Na przykład oddziaływanie wiatru, turbulencji czy innych czynników atmosferycznych może być trudne do odróżnienia od chwilowych skoków wartości szczytowych przyspieszeń, wynikających z nagłych zmian położenia, czy dynamicznych manewrów obiektu [6].

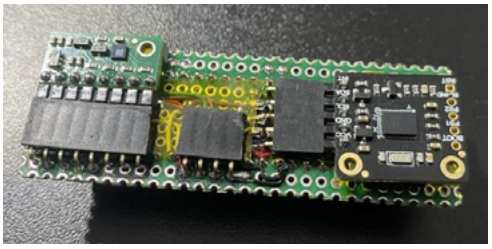
3. PROPONOWANE ROZWIĄZANIE

Zbudowany układ pomiarowy (Rys. 4) opiera się na mikrokontrolerze Teensy 3.6, do którego podłączono główne moduły MEMS: czujnik IMU (BNO055), barometr (LPS25HB) oraz odbiornik GNSS, a wszystkie odczyty są rejestrowane na karcie SD. Mikrokontroler Teensy został wybrany ze względu na jego dużą moc obliczeniową (ARM Cortex-M4) w połączeniu z rozbudowanymi peryferiami i prostotą implementacji.



Rys. 4 Schematu układu pomiarowego

Zastosowano moduł IMU BNO055, który integruje akcelerometr, żyroskop i magnetometr oraz posiada wbudowane algorytmy fuzji. Pomimo niskiego kosztu, urządzenie to zapewnia dokładność pomiarów wystarczającą dla celów przeprowadzonych badań. Barometr LPS dopełnia zestaw o pomiar ciśnienia, co pozwala uzyskać niezależną od GNSS i IMU referencję wysokości (do dalszych badań), a GNSS służy zarówno jako źródło współrzędnych i prędkości (do walidacji i trenowania sieci), jak i punkt odniesienia do oceny błędów dryfu IMU. Karta SD umożliwia prosty, ciągły zapis surowych pomiarów w formacie CSV, bez konieczności stałej komunikacji z komputerem.



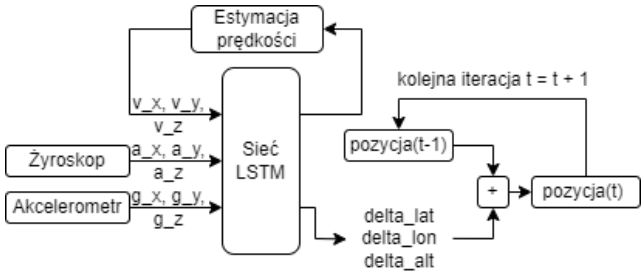
Rys. 5 Fizyczna implementacja układu pomiarowego

Po włączeniu zasilania mikrokontroler inicjalizuje wszystkie podłączone moduły: uruchamiany jest interfejs karty SD, wczytywane są i konfigurowane do pracy czujniki IMU (BNO055) oraz barometr (LPS), a następnie konfiguracja modułu GNSS, ustawiana jest komunikacja i tryb pracy. Gdy wszystkie elementy są gotowe, program przechodzi do głównej pętli, w której zapisywane są próbki danych. W jednym cyklu pobierane są najnowsze wartości z IMU (orientacja, przyspieszenie, żyroskop, magnetometr), barometru (ciśnienie, wysokość, temperatura) oraz z GNSS (współrzędne, prędkość, kurs, liczba satelitów). Cały zestaw danych jest zapisywany w kolejnej linii pliku CSV na karcie SD. W rezultacie na karcie

SD rejestrowany jest zbiór danych pomiarowych (dataset), wykorzystywany na późniejszym etapie do trenowania i walidacji sieci neuronowych.

4. NAWIGACJA INERCYJNA PRZY WYKORZYSTANIU SIECI LSTM

W proponowanym podejściu zastąpiono klasyczną estymację pozycji siecią LSTM. W trakcie uczenia model analizuje kolejne sekwencje surowych pomiarów z akcelerometru i żyroskopu wraz z odpowiadającym im wektorem prędkość, a następnie bezpośrednio przewiduje przyrosty współrzędnych [7]. Architektura sieci LSTM umożliwia wychwytywanie charakterystycznych wzorców ruchu i kompensować powtarzalne odchylenia wynikające zarówno z dryfu, szumów, niedokładności czujników jak i z wykorzystanych metod pomiarowych (np. drgania generowane przez silniki oraz niedoskonałości montażu). Dodatkowo identyfikuje czynniki atmosferyczne takie jak wiatr, temperaturę, wilgotność, czy ciśnienie.



Rys. 6 Schemat blokowy estymacji przyrostów pozycji przez sieć LSTM na podstawie pomiarów z IMU oraz bieżącej prędkości z akumulacją pozycji

Struktura sieci została zaprojektowana w sposób przedstawiony na Rys. 6. Na wejście modelu trafiają trójwymiarowe surowe sygnały z akcelerometru i żyroskopu oraz autoregresyjnie przekazywana estymowana prędkość w trzech osiach. Dzięki uwzględnieniu wektorów prędkości sieć może adaptować sposób analizy danych z czujników do różnych manewrów, które zależą od prędkości przelotowej bezzałogowca[8]. Na wyjściu sieć przewiduje chwilowe przyrosty współrzędnych (szerokość i długość geograficzna oraz wysokość) oraz aktualną prędkość, która następnie służy jako dane wejściowe w kolejnych krokach. W ten sposób, przez skumulowanie prognozowanych przesunięć, obliczana jest łącznie zmiana pozycji od początku lotu, co umożliwia ciągłe określanie bieżącego położenia platformy bezzałogowej. Architektura modelu opiera się na LSTM składającym się z trzech warstw. Pierwsza warstwa to LSTM z 30 komórkami, przyjmująca sekwencje o ustalonej długości (w eksperymentach 15 kolejnych pomiarów). Drugą warstwę stanowi gęsty (perceptronowy) blok o rozmiarze 15 neuronów, a trzecią – kolejny blok perceptronowy z 12 neuronami[9][10][11].

5. SCENARIUSZE BADAWCZE ORAZ WYNIKI

Celem badań było zweryfikowanie:

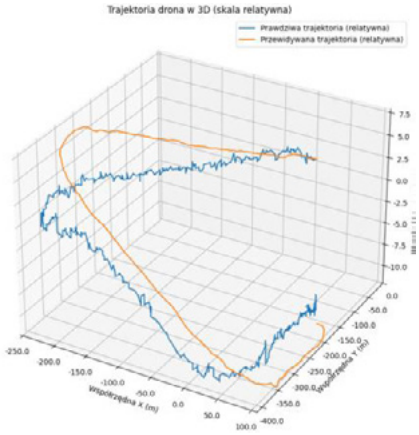
- poprawności akwizycji oraz zbierania próbek z czujników pomiarowych;
- skuteczności proponowanej metody nawigacji inercyjnej opartej na zaproponowanej sieci LSTM;
- odporności modelu na różnorodne warunki atmosferyczne;
- zdolności sieci trenowanej w określonych warunkach do generalizacji i pracy w środowisku o odmiennych parametrach.

Badania zostały przeprowadzona z wykorzystaniem opisanego autorskiego układu oraz DJI Mavic 3 jako platformą nośną (Rys. 7).

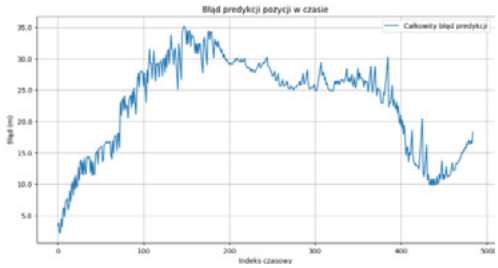
Rys. 7 Widok platformy nośnej (DJI Mavic 3) wraz z zamocowanym układem pomiarowym

W eksperymentach wykorzystano sieć o stałej architekturze i niezmiennych hiperparametrach, aby ocenić jej rzeczywiste możliwości. Model trenowano na 12 000 pomiarów, podzielonych na sekwencje długości 15 próbek, przez 255 epok z batchem 128. Przeprowadzone badania (wyniki poniżej) pozwoliły odpowiedzieć na postawione pytania.

W pierwszym eksperymencie sieć uczono na danych z lotów w stabilnym wietrze (≤ 8 węzłów), a następnie testowano na zbiorze walidacyjnym z tych samych warunków. Na Rys. 8 i Rys. 9 przedstawiono odpowiednio trajektorię rzeczywistą i przewidywaną na zbiorze walidacyjnym oraz błąd predykcji w czasie.



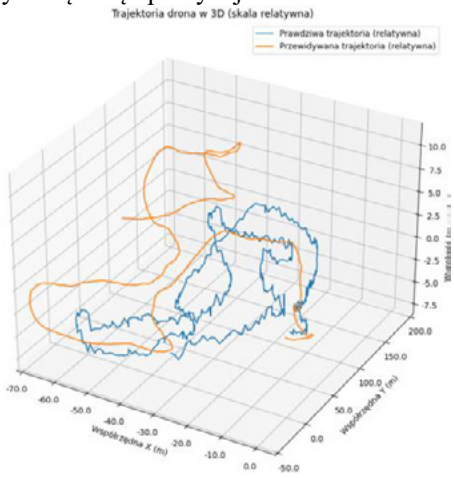
Rys. 8. Wyniki walidacji sieci wytrenowanej w wietrze < 8 węzłów testowanej w tych samych warunkach.



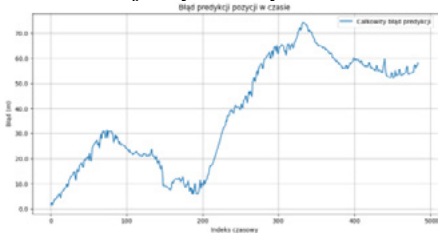
Rys. 9. Błędy predykcji sieci wytrenowanej w wietrze < 8 węzłów testowanej w tych samych warunkach.

Uzyskane wyniki wskazują, że błąd predykcji nie kumuluje się w sposób liniowy. Świadczy to o poprawnym uczeniu się sieci, pomimo występowania obustronnego odchylenia – raz przewidywane przesunięcia są za duże, innym razem za małe. Dzięki temu w ciągu 50 s lotu błąd pozostaje względnie stabilny i nie przekracza 30 m, co umożliwia wykorzystanie sieci do powrotu drona na miejsce startu w razie utraty sygnału GNSS.

W drugim eksperymencie sieć trenowano na danych z lotów przy wietrze powyżej 14 węzłów i testowano na analogicznym zbiorze walidacyjnym. Na Rys. 10 i Rys. 11 pokazano odpowiednio trajektorię rzeczywistą oraz przewidywaną i błąd predykcji w czasie.



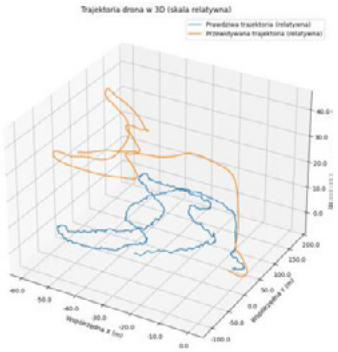
Rys. 10 Wyniki walidacji sieci wytrenowanej w wietrze > 14 węzłów testowanej w tych samych warunkach.



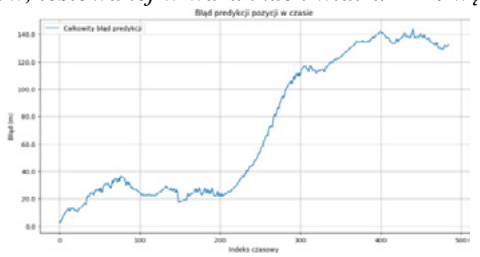
Rys. 11 Błędy predykcji sieci wytrenowanej w wietrze < 8 węzłów testowanej w tych samych warunkach.

W trzecim eksperymencie, mimo tych samych hiperparametrów i liczby próbek, błąd rośnie do ok. 70 m w 50 s i wykazuje tendencję wzrostową – szczególnie boczny wiatr, działając raz z jednej, raz z drugiej strony, jest trudny do uwzględnienia. Mimo to model nadal nadaje się do nawigacji powrotnej w razie utraty GNSS.

W czwartym eksperymencie sieć trenowano na danych z wiatrem ≤ 8 węzłów, a testowano przy wietrze > 14 węzłów. Na Rys. 12 i Rys. 13 pokazano trajektorię rzeczywistą vs. przewidywaną oraz błąd w czasie.

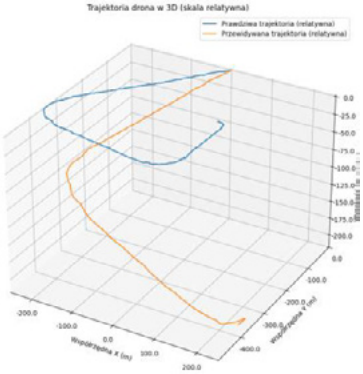


Rys. 12 Wyniki walidacji sieci wytrenowanej w wietrze < 8 węzłów, testowanej w warunkach wiatru > 14 węzłów.

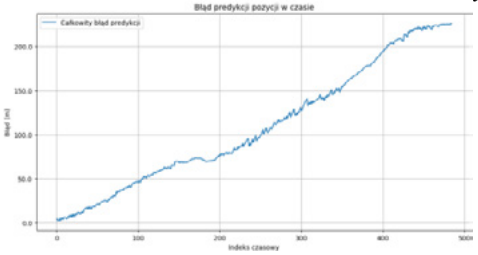


Rys. 13 Błędy predykcji sieci wytrenowanej w wietrze < 8 węzłów, testowanej w warunkach wiatru > 14 węzłów.

W tym eksperymencie błąd sięga około 140 m w ciągu 50 s i nie maleje zauważalnie, co wskazuje na niewystarczające uwzględnianie wpływu wiatru podczas predykcji. Ostatnie badanie polegało na trenowaniu sieci na danych z wiatrem > 14 węzłów i testowaniu jej na danych z wiatrem ≤ 8 węzłów. Na Rys. 14 i Rys. 15 przedstawiono odpowiednio trajektorię rzeczywistą vs. przewidywaną oraz błąd predykcji w czasie.



Rys. 14. Wyniki walidacji sieci wytrenowanej w wietrze > 14 węzłów, testowanej w warunkach wiatru > 8 węzłów.



Rys. 15 Błędy predykcji sieci wytrenowanej w wietrze > 14 węzłów, testowanej w warunkach wiatru < 8 węzłów.

W ostatnim eksperymencie błąd był jeszcze większy niż w poprzednich badaniach i wciąż narastał. Sieć nieprawidłowo uwzględniała wpływ wiatru i innych zakłóceń, co skutkowało stopniowo pogarszającą się dokładnością lokalizacji.

6. WNIOSKI I MOŻLIWOŚCI ROZWOJU

Na podstawie uzyskanych wyników można stwierdzić, że zastosowanie sieci neuronowej typu LSTM stanowi obiecującą alternatywę dla klasycznych metod nawigacji inercyjnej, jednak jej skuteczność w dużym stopniu zależy od warunków treningu. Najdokładniejsze wyniki osiągane są, gdy model uczy się w środowisku z minimalnym wiatrem – w miarę wzrostu siły wiatru w danych treningowych pojawiają się nieprzewidywalne zakłócenia, które pogarszają adaptację. Co więcej, sieć wytrenowana w silnym wietrze gorzej działa w warunkach łagodnych, ponieważ obecność dużych szumów utrudnia późniejsze przystosowanie do spokojniejszych warunków atmosferycznych. Z drugiej strony, model uczony przy niewielkim wietrze radzi sobie lepiej w środowisku ekstremalnym niż odwrotnie, co wskazuje na potrzebę stopniowania trudności scenariuszy szkoleniowych. W związku z powyższym, proponuje się dwa podejścia: Trenować sieć na danych z różnych warunków pogodowych (wymaga wielu testów); Wykorzystać dane z bezwietrznego środowiska i analizować odchylenia błędu przy innym wietrze — może to dać lepsze efekty, ale wymaga dodatkowej empirycznej weryfikacji.

Otrzymane wyniki dowodzą, że LSTM uwzględnia zarówno czynniki środowiskowe (np. siłę i turbulencje wiatru), jak i specyfikę konkretnej platformy (np. przechyły drona w zakrętach). Świadczą o tym obserwowane błędy wysokości w testach krzyżowych – sieć uczy się nie tylko uniwersalnych parametrów fizycznych (prędkości czy przyspieszenia), lecz także specyficznych wzorców zachowania każdego egzemplarza platformy i warunków. Stanowi to zaletę jak również i wadę tego rozwiązania. Predykcje sieci mają charakter wygładzony, przez co model nie nadąża za nagłymi, krótkotrwałymi odchyleniami czujników. Celem zwiększenia dokładności, planuje się optymalizację hiperparametrów oraz wstępne przygotowanie danych (np. filtr Kalmana, FIR) w celu ograniczenia wpływu losowych zakłóceń. W przyszłości niezbędne będzie opracowanie metody kompresji lub selekcji kluczowych informacji pogodowych, aby model mógł działać efektywnie przy mniejszym zestawie danych treningowych w różnych warunkach.

7. PODSUMOWANIE

W przeprowadzonych badaniach wykazano, że sieć LSTM, zasilana bezpośrednio surowymi sygnałami z konsumenckiego IMU oraz uzupełniona o wektory prędkości, może skutecznie zastąpić tradycyjną, przyrostową integrację przyspieszeń i prędkości kątowych w zadaniu nawigacji inercyjnej. Klasyczna metoda jest również podatna na wszystkie wykryte problemy jednakże analityczny sposób nie uwzględnienia warunków środowiska. W przypadku uwzględnienia warunków środowiskowych w dalszym ciągu może być obarczony nieprzewidywalnym środowiskiem. Stworzony na mikrokontrolerze Teensy 3.6 system pomiarowy, łączący IMU BNO055, barometr LPS25HB oraz odbiornik GNSS, umożliwił zebranie obszernych, skorelowanych danych treningowych, zawierających zarówno parametry orientacji i przyspieszeń, jak i referencyjne współrzędne z odbiornika GNSS.

Na podstawie wyników badań można stwierdzić, że do stworzenia uniwersalnego INS przy wykorzystaniu uczenia maszynowego istotny jest początkowy trening danymi pozyskanych w stabilnej atmosferze (możliwie niskiej wartości prędkości wiatru obecnego podczas analizy), a następnie stopniowe wzbogacanie zbioru o trudniejsze scenariusze (zarówno środowiskowe jak i parametrów oraz trajektorii lotu). Najlepsze efekty osiągnięto przy treningu na danych z umiarkowanym wiatrem – model wykazał zdolność generalizacji do trudniejszych warunków. Równoczesne wdrożenie technik wstępnej filtracji sygnałów pochodzących z czujników oraz optymalizacji struktury sieci pozwoli na minimalizację wpływu zakłóceń i poprawę precyzji śledzenia trajektorii, a jednocześnie umożliwi lepszą implementację na architekturach o mniejszej mocy obliczeniowej.

LITERATURA

[1] El-Sheimy, N., Youssef, A. Inertial sensors technologies for navigation applications: state of the art and future trends. *Satell Navig* 1, 2 (2020). https://doi.org/10.1186/s43020-019-0001-5

[2] Chen, Changhao & Pan, Xianfei. (2023). Deep Learning for Inertial Positioning: A Survey. 10.48550/arXiv.2303.03757.

[3] Zhenshu, Zhu & Bo, Yuming & Jiang, Changhui. (2019). A MEMS Gyroscope Noise Suppressing Method Using Neural Architecture Search Neural Network. 2019. 10.1155/2019/5491243.

[4] en.wikipedia.org/wiki/Inertial_navigation_system

[5] www.advancednavigation.com/tech-articles/inertial-navigation-systems-ins-an-introduction/

[6] Jiang, Changhui, Shuai Chen, Yuwei Chen, Boya Zhang, Ziyi Feng, Hui Zhou, and Yuming Bo. 2018. "A MEMS IMU De-Noising Method Using Long Short Term Memory Recurrent Neural Networks (LSTM-RNN)" https://doi.org/10.3390/s18103470

[7] Chen, Changhao & Pan, Xianfei. (2023). Deep Learning for Inertial Positioning: A Survey. 10.48550/arXiv.2303.03757.

[8] Li, Ruihao & Fu, Chunlian & Yi, Wei & Yi, Xiaodong. (2022). Calib-Net: Calibrating the Low-Cost IMU via Deep Convolutional Neural Network. 8. 10.3389/frobt.2021.772583.

[9] L. Tejmlova, J. Sebesta and P. Zelina, "Artificial neural networks in an inertial measurement unit," 2016, doi:10.1109/RADIOELEK.2016.7477431.

[10] Zhang, Ming & zhang, mingming & Chen, Yiming & Li, Mingyang. (2021). IMU Data Processing For Inertial Aided Navigation: A Recurrent Neural Network Based Approach.10.1109/ICRA48506.2021.9561172.

[11] Zeinali, Behnam & Zandizari, Hadi & Chang, J.. (2022). IMUNet: Efficient Regression Architecture for IMU Navigation and Positioning. 10.48550/arXiv.2208.00068.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

Porównanie algorytmów pozycjonowania dla systemu lokalizacji wewnątrzbudynkowej czasu rzeczywistego z wykorzystaniem sieci Wi-Fi

Comparison of the positioning algorithms for a real-time indoor locating system using Wi-Fi network

Piotr Kocik[1]; Grzegorz Cichocki[1]; Stanisław Nowak[1]; Łukasz Jezierski[1]; Krzysztof Cichoń[1];

[1] Politechnika Poznańska, Poznań, {piotr.kocik, grzegorz.cichocki, stanislaw.nowak, lukasz.jezierski}@student.put.poznan.pl, krzysztof.cichon@put.poznan.pl

DOI: 10.15199/59.2025.4.87

Streszczenie: W pracy przeanalizowano dokładność klasycznego pozycjonowania wewnątrzbudynkowego z wykorzystaniem sygnału RSSI w sieciach Wi-Fi. Oceniono działanie multilateracji, metody centroidowej oraz metody opartej o NLS (Nonlinear least squares). Wyniki potwierdzają wyraźne różnice w skuteczności rozważanych podejść oraz potencjał dalszych usprawnień.

Abstract: This paper analyzes the baseline accuracy of indoor positioning based on RSSI signals in Wi-Fi networks. We evaluate the performance of multilateration, the centroid method, and the method using a nonlinear least squares (NLS) algorithm. The results highlight significant differences in the effectiveness of the considered methods and indicate potential directions for further improvement.

Słowa kluczowe: lokalizacja wewnątrzbudynkowa, RTLS, pozycjonowanie RSSI, multilateracja.

Keywords: indoor localization, RTLS, RSSI positioning, multilateration.

1. WSTĘP

Wnioskowanie na temat otoczenia na podstawie sygnałów radiowych obecnie budzi szczególne zainteresowanie społeczności naukowej w kontekście prac nad systemami szóstej generacji (6G) [1]. Choć pojawiają się już ciekawe prace określające cechy elementów infrastruktury w pomieszczeniach wyłącznie na podstawie odebranych sygnałów radiowych [2], to niezmiennie jednym z kluczowych wyzwań jest wyznaczanie pozycji urządzeń w przestrzeni 3D oraz określenie orientacji urządzeń w przestrzeni z co najmniej centymetrową dokładnością w środowisku wewnątrzbudynkowym. Pozycjonowanie wewnątrz budynków zdobyło już spore zainteresowanie społeczności naukowej, a lokalizacja wewnątrzbudynkowa oparta na algorytmie tzw. fingerprintingu jest jedną z kluczowych jej metod [3].

W niniejszej pracy zbadano dokładność estymacji położenia użytkownika dla trzech wybranych metod. Skupiono się na klasycznej metodzie lokalizacji wykorzystującej wskaźnik mocy sygnału odbieranego (RSSI) w standardowych sieciach Wi-Fi. Choć technika ta jest znana od wielu lat i charakteryzuje się ograniczoną dokładnością, jej główną zaletą jest powszechna dostępność — możliwa do realizacji bez dodatkowego sprzętu, jedynie przy użyciu istniejącej infrastruktury i urządzenia mobilnego.

Celem pracy jest przeanalizowanie dokładności bazowej klasycznej metody, metody centroidowej a także przetestowanie metody opartej o NLS, które wykorzystuje znane pozycje ruterów oraz rzeczywiste wartości RSSI.

Plan artykułu jest następujący. W rozdziale drugim przedstawiono rozważane algorytmy lokalizacji wewnątrzbudynkowej. Szczegóły implementacji algorytmów obejmujące m. in. opis scenariusza pomiarowego i sposobu przetwarzania danych pomiarowych opisano w rozdziale trzecim. Wyniki pomiarów zaprezentowano w rozdziale czwartym, a całość podsumowano wnioskami w rozdziale piątym.

2. ALGORYTMY LOKALIZACJI WEWNĄTRZBUDYNKOWEJ

2.1 Multilateracja

Metoda wyznaczania pozycji urządzenia polega na wykorzystaniu pomiarów odległości d_i od co najmniej trzech punktów odniesienia o znanych współrzędnych (x_i, y_i). Pozycja urządzenia (x, y) spełnia równania okręgu:

$$(x - x_i)^2 + (y - y_i)^2 = d_i^2, \quad i = 1, \ldots, N, \tag{1}$$

gdzie $N \geq 3$.

Aby uprościć rozwiązanie, odejmujemy od kolejnych równań równanie dla pierwszego punktu odniesienia (x_1, y_1, d_1), co daje liniowy układ [4]:

$$2(x_i - x_1)x + 2(y_i - y_1)y = d_1^2 - d_i^2 + x_i^2 - x_1^2 + y_i^2 - y_1^2, \quad i = 2, \ldots, N. \tag{2}$$

W postaci macierzowej:

$$Ap = b, \tag{3}$$

gdzie

$$A = \begin{bmatrix} 2(x_2 - x_1) & 2(y_2 - y_1) \\ \vdots & \vdots \\ 2(x_N - x_1) & 2(y_N - y_1) \end{bmatrix}, \tag{4}$$

$$p = \begin{bmatrix} x \\ y \end{bmatrix}, \tag{5}$$

$$b = \begin{bmatrix} d_1^2 - d_2^2 + x_2^2 - x_1^2 + y_2^2 - y_1^2 \\ \vdots \\ d_1^2 - d_N^2 + x_N^2 - x_1^2 + y_N^2 - y_1^2 \end{bmatrix}. \tag{6}$$

Rozwiązanie $\hat{p}$ wyznaczane jest metodą najmniejszych kwadratów [5]:

$$\hat{p} = \arg\min_p ||Ap - b||_2^2 = \arg\min_p \sum_i ((Ap)_i - b_i)^2, \tag{7}$$

co pozwala oszacować pozycję urządzenia.

2.2 Metoda centroidowa

Metoda Weighted Centroid Localization (WCL) stanowi rozwinięcie klasycznego podejścia centroidowego, w którym pozycja nieznanego węzła estymowana jest jako średnia współrzędnych otaczających go punktów odniesienia. W przeciwieństwie do zwykłego centroidu, WCL przypisuje każdemu nadajnikowi referencyjnemu wagę w zależności od jego odległości od lokalizowanego węzła — im bliższy nadajnik, tym większy jego wpływ na końcową estymację pozycji.

Szacowana pozycja $P_i' = (x_i', y_i')$ nieznanego węzła i wyznaczana jest według wzoru [6]:

$$P_i' = \frac{\sum_{j=1}^{n} w_{ij}(x_j, y_j)}{\sum_{j=1}^{n} w_{ij}}, \tag{8}$$

gdzie n to liczba dostępnych nadajników odniesienia, (x_j, y_j) – współrzędne nadajnika j, w_{ij} – waga przypisana nadajnikowi j względem węzła i, zależna od oszacowanej odległości d_{ij} między nimi. Wagi widoczne w wzorze (8) obliczane są, zgodnie z równaniem:

$$w_{ij} = \frac{1}{d_{ij}^g}, \quad g \geq 1, \tag{9}$$

gdzie g jest parametrem czułości, regulującym wpływ odległości (typowo w zakresie od 1 do 4). Większe wartości g powodują silniejsze faworyzowanie bliższych nadajników.

WCL zachowuje prostotę obliczeniową klasycznego centroidu, a jednocześnie poprawia dokładność lokalizacji, uwzględniając geometryczne relacje przestrzenne bez konieczności stosowania złożonych metod optymalizacyjnych, takich jak multilateracja czy algorytmy nieliniowej estymacji.

2.3 Metoda oparta na NLS

Aby oszacować pozycję użytkownika w przestrzeni 2D, definiuje się funkcję błędu (10), która mierzy niedopasowanie pomiędzy odległościami szacowanymi na podstawie przybliżonej pozycji użytkownika a odległościami wyznaczonymi z modelu propagacji sygnału RSSI. Formalnie, niech (x, y) oznacza szacowaną pozycję użytkownika, a (x_i, y_i) – współrzędne i-tego punktu odniesienia. Wartość d_i odpowiada odległości wyliczonej na podstawie zmierzonej wartości wskaźnika RSSI i zastosowanego modelu propagacji fal radiowych. Funkcja błędu przyjmuje postać sumy kwadratów różnic pomiędzy kwadratami odległości geometrycznych od użytkownika do punktów odniesienia a odległościami wyliczonymi z RSSI [7]:

$$\epsilon(x, y) = \sum_{i=1}^{N} \left((x - x_i)^2 + (y - y_i)^2 - d_i^2 \right)^2, \tag{10}$$

gdzie N to liczba punktów odniesienia.

Proces estymacji pozycji sprowadza się do znalezienia takiego punktu (x, y), który minimalizuje tę funkcję błędu. W praktyce zrealizowano to za pomocą algorytmu SLSQP (Sequential Least SQuares Programming). Optymalizacja rozpoczyna się od pozycji startowej w środku badanego pomieszczenia i iteracyjnie dopasowuje współrzędne użytkownika tak, aby najlepiej odwzorować obserwowane wartości RSSI dla wszystkich punktów odniesienia. W efekcie uzyskujemy estymatę pozycji użytkownika, która minimalizuje błąd pomiarowy i najlepiej zgadza się z modelowanymi parametrami propagacji sygnału.

3. SCENARIUSZ POMIAROWY

Pomieszczenie, w którym prowadzono pomiary, to sala laboratoryjna wyposażona w typowe elementy infrastruktury technicznej, takie jak oscyloskopy, generatory sygnałów, anteny oraz meble. W sali wyznaczono regularną siatkę punktów pomiarowych obejmującą obszar o wymiarach 3×7 metrów. Punkty rozmieszczono co 1 metr, tworząc jednorodną siatkę pomiarową. Łącznie zarejestrowano dane w 32 punktach.

Do lokalizacji wykorzystywano pięć punktów dostępowych (ruterów), z których każdy nadawał w paśmie 5 GHz. Miejsca ruterów wyznaczono tak, aby jak najmniej kolidowały z infrastrukturą sali tzn. znajdowały się w punktach (0,0), (0,5), (0,7), (3,0), (3,5) przy założeniu, że lewy dolny róg siatki to punkt (0,0). Użyte rutery to TP-Link Archer C60. Odbiór sygnału realizowany był za pomocą smartfona Samsung Galaxy S10+, który zbierał dane RSSI od wszystkich dostępnych ruterów. Dla każdego punktu pomiarowego zarejestrowano około 70 próbek RSSI na każde widoczne urządzenie, przy interwale czasowym wynoszącym 4 sekundy, co odpowiada łącznemu czasowi pomiaru około 4 minut i 40 sekund na punkt.

W celu rejestracji danych pomiarowych opracowano aplikację na system Android, przygotowaną na potrzeby niniejszego badania. Rozwiązanie zostało zaprojektowane z wykorzystaniem języka Java oraz środowiska Android SDK. Program pozwala na automatyczne skanowanie wartości RSSI emitowanego przez pobliskie punkty dostępowe, wykorzystując w tym celu standardowe API systemowe (WifiManager).

Aplikacja zawiera również wbudowany moduł serwera lokalnego, obsługującego protokoły HTTP i WebSocket. Te funkcje pozwalają na jednoczesną wizualizację i monitoring pomiarów w czasie rzeczywistym oraz zdalną konfigurację parametrów eksperymentu. Dodatkowo, oprogramowanie udostępnia interfejs API, zapewniający zdalny dostęp do pełnego zbioru zgromadzonych danych w formacie JSON.

3.1 Model propagacji

Używany jest klasyczny logarytmiczny model propagacji sygnału radiowego [8], [9]:

$$RSSI = A - 10n \log_{10}(d), \tag{11}$$

gdzie d to odległość od rutera, A – stała, która uwzględnia straty systemowe, n – wykładnik propagacji (tłumienia sygnału), zależny od środowiska. Do obliczeń wzór (11) został przekształcony do postaci:

$$d = 10^{\frac{A-RSSI}{10n}}. \tag{12}$$

3.2 Przetwarzanie danych i metryki błędu

W celu zwiększenia stabilności i wiarygodności pomiarów RSSI zastosowano uproszczoną metodę filtrowania wartości sygnału. Dla każdego rutera w danym punkcie pomiarowym wyznaczano średnią z trzech najczęściej występujących wartości RSSI. Zabieg ten miał na celu ograniczenie wpływu wartości odstających oraz poprawę jakości danych wejściowych wykorzystywanych do estymacji pozycji.

Ocena dokładności estymowanych pozycji oparta została na analizie różnicy między pozycją rzeczywistą a wyznaczoną. Jako podstawową miarę błędu przyjęto odległość euklidesową:

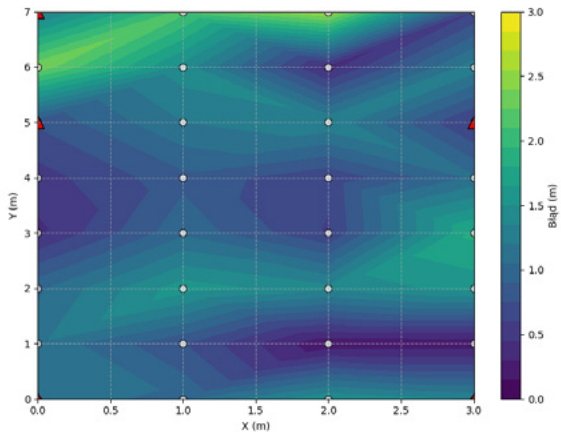
$$\varepsilon = \sqrt{(x_{\text{true}} - x_{\text{est}})^2 + (y_{\text{true}} - y_{\text{est}})^2}, \tag{13}$$

gdzie $(x_{\text{true}}, y_{\text{true}})$ — rzeczywista pozycja punktu pomiarowego, $(x_{\text{est}}, y_{\text{est}})$ — pozycja estymowana przez algorytm lokalizacyjny, ε — wartość błędu lokalizacji w metrach.

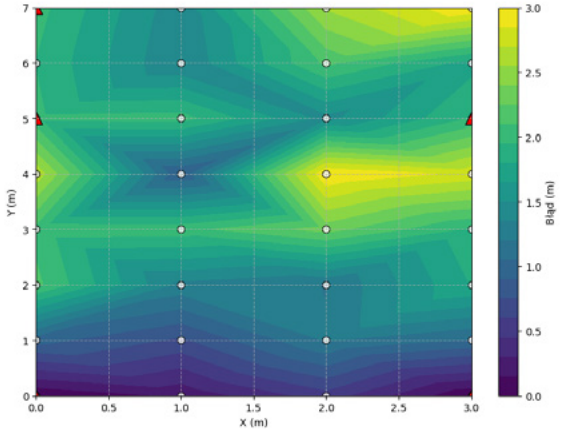
4. WYNIKI

4.1 Mapy ciepła dla pełnego zestawu pięciu punktów dostępowych

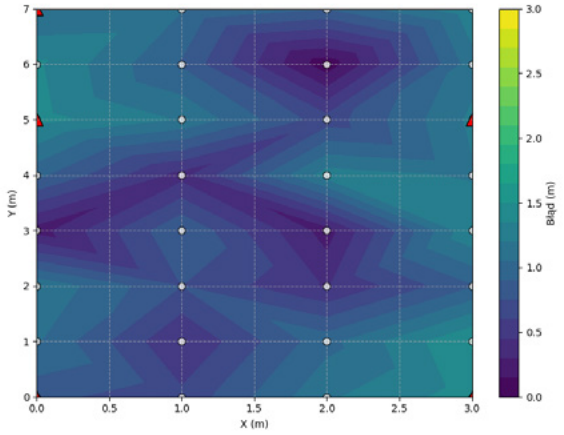
Poniżej przedstawiono przestrzenny rozkład średnich błędów lokalizacji dla poszczególnych metod, zobrazowany w formie map ciepła. Analiza została przeprowadzona na podstawie wyników uzyskanych przy wykorzystaniu pełnego zestawu pięciu ruterów.



Rysunek 1: Przestrzeń pomiarowa z zaznaczonymi średnimi błędami lokalizacji dla multilateracji



Rysunek 2: Przestrzeń pomiarowa z zaznaczonymi średnimi błędami lokalizacji dla metody centroidowej



Rysunek 3: Przestrzeń pomiarowa z zaznaczonymi średnimi błędami lokalizacji dla metody opartej o NLS

Analiza przestrzennego rozkładu błędów (rys. 1–3) wskazuje wyraźne różnice w dokładności i stabilności działania poszczególnych metod lokalizacji.

Najlepsze rezultaty uzyskano dla metody opartej na NLS (rys. 3), gdzie błędy estymacji były najmniejsze i najbardziej równomiernie rozłożone, szczególnie w centralnej części pomieszczenia. Świadczy to o wysokiej odporności tej metody na niesymetryczne rozłożenie punktów odniesienia oraz zakłócenia propagacyjne.

W przypadku multilateracji (rys. 1) zauważalne są większe błędy w obszarach brzegowych i narożnych. Widzimy, że metoda ta nie radzi sobie tak dobrze jak NLS.

Największe błędy odnotowano przy metodzie centroidowej (rys. 2). Metoda ta nie uwzględnia informacji o odległości oraz jest wrażliwa na niesymetryczne rozłożenie nadajników, co skutkuje niską dokładnością lokalizacji.

4.2 Porównanie dokładności lokalizacji przy różnej liczbie punktów dostępowych

W tabeli 1 przedstawiono porównanie skuteczności działania algorytmów lokalizacji przy wykorzystaniu 3, 4 oraz 5 ruterów. W przypadku konfiguracji z 3 (bez routerów (0,7) oraz (3,5)) i 4 (bez (0,7)) ruterami, wybrane zostały te punkty dostępowe, które zapewniały najniższy średni błąd estymacji pozycji w ujęciu globalnym — tj. wybrano jeden zestaw ruterów dla wszystkich punktów pomiarowych oraz dla wszystkich metod.

Tabela 1: Podsumowanie wyników dokładności uzyskanych dla wykorzystanych metod lokalizacji

Metoda	Średni błąd [m]	Mediana [m]	Minimum [m]	Maximum [m]	Odchylenie standardowe [m]
Dla 3 ruterów					
Multilateracja	2.02	1.60	0.22	6.11	1.51
Centroidowa	2.52	2.53	0.01	5.69	1.46
NLS	1.10	0.99	0.15	2.44	0.58
Dla 4 ruterów					
Multilateracja	1.14	1.21	0.17	2.32	0.62
Centroidowa	1.99	2.17	0.01	4.02	1.04
NLS	1.07	1.16	0.19	1.76	0.40
Dla 5 ruterów					
Multilateracja	1.11	1.08	0.16	2.59	0.58
Centroidowa	1.64	1.86	0.01	3.00	0.80
NLS	0.94	1.02	0.04	1.51	0.38

5. WNIOSKI

Na podstawie wyników przedstawionych w Tabeli 1, odnotowano wyraźną poprawę dokładności wraz ze wzrostem liczby ruterów dla metod multilateracji i centroidowej. Natomiast metoda NLS uzyskała najwyższą dokładność lokalizacji we wszystkich konfiguracjach. Charakteryzowała się ona najniższymi wartościami średniego błędu oraz odchylenia standardowego, co jest szczególnie użyteczne w środowisku pomiarowym o dużej zmienności.

Metoda multilateracji osiąga wyniki porównywalne z NLS przy większej liczbie ruterów. Sugeruje to, że jej efektywność rośnie wraz z dostępnością większej liczby danych. Przy ograniczonej liczbie punktów odniesienia metoda ta wykazuje trudności, co znajduje odzwierciedlenie w znacząco większym błędzie maksymalnym. Metoda centroidowa przedstawia największy błąd średni lokalizacji, szczególnie w konfiguracji z 3 punktami odniesienia, co wskazuje na jej ograniczoną dokładność w środowisku o mniejszej liczbie punktów dostępowych.

Podsumowując, wyniki te potwierdzają zasadność stosowania metody NLS, konieczność zapewnienia odpowiedniej liczby punktów odniesienia wraz z odpowiednim ich umiejscowieniem przy zastosowaniu konkretnych metod dla uzyskania satysfakcjonującej dokładności lokalizacji w systemach opartych na RSSI.

LITERATURA

[1] N. González-Prelcic et al., "The Integrated Sensing and Communication Revolution for 6G: Vision, Techniques, and Applications" in: Proceedings of the IEEE, tom 112, nr 7, str. 676-723, 2024.

[2] M. F. De Guzman and K. Haneda, "Analysis of Wave-Interacting Objects in Indoor and Outdoor Environments at 142 GHz," in IEEE Transactions on Antennas and Propagation, vol. 71, no. 12, pp. 9838-9848, Dec. 2023.

[3] K. Cichon, Z. Buksch, B. Gregorczyk, "Adaptive Indoor Positioning System Based on Fingerprinting Principle", 2024 20th IEEE International Conference on Wireless and Mobile Computing, Networking and Communications, Oct. 21-23, 2024, Paris.

[4] M. Pelka, "Position Calculation with Least Squares based on Distance Measurements." in Lübeck University of Applied Sciences: Technical Report (2015): 1-3.

[5] S. Mazuelas et al., "Robust Indoor Positioning Provided by Real-Time RSSI Values in Unmodified WLAN Networks," in IEEE Journal of Selected Topics in Signal Processing, vol. 3, no. 5, pp. 821-831, Oct. 2009,

[6] J. Blumenthal, R.Grossmann, F.Golatowski and D.Timmermann, "Weighted Centroid Localization in Zigbee-based Sensor Networks". Proceedings of the IEEE International Symposium on Intelligent Signal Processing. 2007,

[7] Du Jinze, Chao Yuan, Min Yue, and Tao Ma. 2022. "A Novel Localization Algorithm Based on RSSI and Multilateration for Indoor Environments" Electronics 11, no. 2: 289.

[8] F. Zafari, A. Gkelias and K. K. Leung, "A Survey of Indoor Localization Systems and Technologies," in IEEE Communications Surveys & Tutorials, vol. 21, no. 3, pp. 2568-2599, thirdquarter 2019,

[9] S. Sadowski and P. Spachos, "RSSI-Based Indoor Localization With the Internet of Things," in IEEE Access, vol. 6, pp. 30149-30161, 2018.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

SAL - lokalizacja z syntetyczną kotwicą

SAL - synthetic anchor localization

Bartosz Czaja[1]; Krzysztof Maślanka[2]

[1] Wojskowa Akademia Techniczna, Warszawa, bartosz.czaja@wat.edu.pl
[2] Wojskowa Akademia Techniczna, Warszawa, krzysztof.maslanka@wat.edu.pl

DOI: 10.15199/59.2025.4.88

Streszczenie: W referacie przedstawione zostały wyniki badań symulacyjnych systemu lokalizacji z syntetyczną kotwicą. Na podstawie pomiarów odległości estymuje się położenie nieruchomego źródła promieniowania względem przelatującego nad nim bezzałogowego statku powietrznego (BSP). Przeprowadzono testy w środowisku symulacyjnym w dwóch scenariuszach: trajektorii złożonej i prostej oraz porównano tradycyjne i optymalizacyjne metody estymacji położenia.

Abstract: The results of simulation studies of synthetic anchor localization (SAL) system were presented in this paper. The position of stationary emitter is estimated based on distance measurements between itself and the unmanned aerial vehicle (UAV). Two scenarios were simulated: the complex trajectory and simple trajectory. The traditional and optimization based position estimation methods were compared.

Słowa kluczowe: estymacja położenia, dynamiczny węzeł kotwicowy, lokalizacja z syntetyczną kotwicą

Keywords: dynamic anchor node, position estimation, synthetic anchor localization

1. WSTĘP

Metoda SAL polega na wyznaczeniu położenia nieruchomego źródła promieniowania (zwanego dalej terminalem naziemnym) na podstawie pomiarów odległości między nim, a przelatującym nad nim bezzałogowym statkiem powietrznym (BSP). Na podstawie jedynie pomiarów odległości można wyznaczyć położenie terminala względem BSP. W celu estymacji bezwzględnego położenia terminala w przestrzeni wykorzystywany jest odbiornik GNSS określający współrzędne BSP w każdym momencie lotu.

W pracy zbadano jakość estymacji położenia terminala nieruchomego z użyciem tradycyjnych algorytmów oraz metod optymalizacyjnych dla dwóch scenariuszy. Pierwszy scenariusz zakładał lot BSP po złożonej trajektorii dobranej specjalnie w celu zminimalizowania komplanarności punktów na trajektorii (syntetyczne kotwice), z których wykonywane były symulowane pomiary odległości. Przy stosowaniu tradycyjnych algorytmów estymacji wymagany jest warunek braku współpłaszczyznowych punktów kotwicowych. W niniejszej pracy termin "punkt kotwicowy" oznacza punkt w przestrzeni, o znanym położeniu, który pełni rolę węzła odniesienia. Względem punktów kotwicowych określa się położenie lokalizowanych obiektów. Następnie, zdecydowano zrealizować badania jakości estymacji dla trajektorii o kształcie linii prostej, aby zweryfikować hipotezę o konieczności zachowania warunku dotyczącego komplanarności punktów kotwicowych. W punkcie 2 przedstawiono opis środowiska symulacyjnego użytego do badań. Punkt 3 zawiera otrzymane rezultaty i ich dyskusję wraz z wnioskami. W punkcie 4 zamieszczono podsumowanie pracy.

2. ŚRODOWISKO BADAWCZE

Środowisko badawcze opracowano w postaci programu symulacyjnego zaimplementowanego w środowisku MATLAB. Symulacja polegała na generowaniu pomiarów położenia BSP wraz z pomiarami odległości BSP-terminal. Aby zachować zwięzłość tekstu w dalszej części słowo "pomiar" każdorazowo odnosi się do symulowanego pomiaru czyli wartości odniesienia z addytywnym szumem symulującym pomiar. Wyjątkiem jest jedynie pomiar czasu obliczeń badanych estymatorów - w tym przypadku mamy do czynienia z prawdziwymi pomiarami wielkości fizycznej jaką jest czas.

W ramach badań zdecydowano się przeprowadzić 100 symulacji generując nowy szum pomiarowy w każdej iteracji. W każdej iteracji generowano nowe pomiary trajektorii lotu BSP i odległości między BSP, a terminalem nieruchomym. Na podstawie tych pomiarów obliczano estymaty położenia terminala naziemnego z użyciem 6 różnych algorytmów. Zbadano dokładność estymacji algorytmów klasycznych (Foy'a, Chan'a) [4, 5] i opartych na metodach optymalizacyjnych: SA (Simulated Annealing) [2], GA (Genetic Algorithm) [7], GWO (Grey Wolf Optimizer) [9], SPSA (Simultaneous Perturbation Stochastic Approximation) [12].

Do badań jakości estymacji położenia terminala zaproponowano dwie różne trajektorie. Każda trajektoria została spróbkowana w 30 punktach. W każdym punkcie trajektorii idealnej obliczono odległość między tym punktem i terminalem nieruchomym umieszczonym w (0,0,1). Następnie, do każdej próbki idealnej trajektorii dodano szum gaussowski o zerowej wartości oczekiwanej i odchyleniu standardowym równym 0,1 [m] symulującym szum odbiornika GNSS z usługą precyzyjnej lokalizacji (np. GPS+RTK - Global Positioning System + Real Time Ki-

nematics) zamontowanego na BSP. Według autorów [3] tanie odbiorniki GPS+RTK pozwalają uzyskać błąd lokalizacji wzdłuż wszystkich 3 osi mniejszy niż przyjęte 0,1 metra. W następnym kroku do idealnych pomiarów odległości dodano szum gaussowski o zerowej wartości oczekiwanej i odchyleniu standardowym równym 0,3 [m] w celu zasymulowania szumu pomiaru odległości realizowanego z użyciem technologii UWB (ang. Ultra Wide Band). Autorzy [11] uzyskali błędy pomiaru metodą TWR (Two-Way Ranging) odległości nie większe niż wspomniane 0,3 metra dla modułów UWB znajdujących się w ruchu. W ten sposób otrzymano dwa zestawy danych (trajektoria i pomiary odległości): idealny i mierzony. Z danych pochodzących z symulowanych pomiarów, w sposób losowy, wybierane były 4 punkty na zaszumionej trajektorii i odpowiadające im pomiary odległości. Taka metoda zmniejsza ryzyko wybrania 4 punktów, które byłyby współpłaszczyznowe. W przypadku wyboru 4 punktów występujących kolejno takie ryzyko jest duże i zależy od odległości między próbkami, która jest związana z okresem próbkowania. Pomiary odległości metodą UWB charakteryzują się różnym czasem trwania więc zapewnienie równomiernego okresu próbkowania tej wielkości byłoby kłopotliwe.

W kolejnym kroku te 4 pomiary położenia i odległości stawały się zmiennymi wejściowymi do 6 różnych estymatorów położenia terminala nieruchomego. Algorytmy optymalizacji poszukiwały, takiej estymaty położenia terminala nieruchomego, która minimalizowałaby sumę kwadratów różnic odległości zmierzonych i estymowanych. Estymaty położenia terminala dla każdego z algorytmów były użyte do obliczenia dwóch miar jakości.

Miara pierwiastka z błędu średniokwadratowego (RMSE - ang. Root Mean Squared Error) została obliczona dla każdej współrzędnej z osobna (x, y, z) z użyciem równania (1), gdzie $\hat{x}_i$ jest estymatą współrzędnej x terminala w i-tej iteracji, x_p jest prawdziwym położeniem terminala, a N liczbą symulacji (w tym przypadku N=100).

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{x}_i - x_p)^2} \tag{1}$$

Kolejną miarą błędu estymacji, którą zastosowano w badaniach, jest odległość euklidesowa między prawdziwym położeniem terminala w przestrzeni, a tym estymowanym (DE - ang. distance error). Miara DE została obliczona dla każdej i-tej symulacji i wyraża się wzorem (2).

$$\mathrm{DE}_i = \sqrt{(\hat{x}_i - x_p)^2 + (\hat{y}_i - y_p)^2 + (\hat{z}_i - z_p)^2} \tag{2}$$

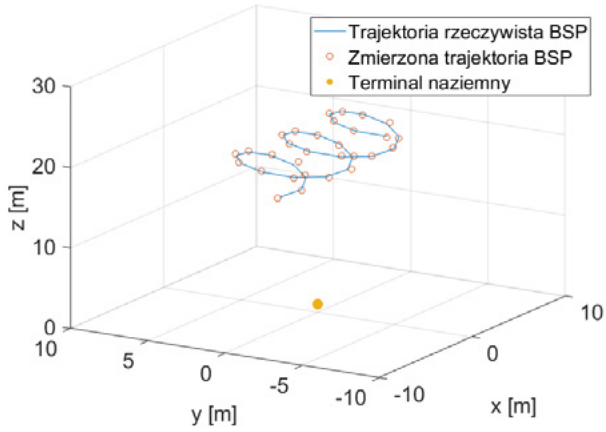
Dana próbka DE jest klasyfikowana jako wartość odstająca (ang. outlier) gdy znajduje się poza przedziałem o szerokości 6 bezwzględnych dewiacji medianowych (MAD - ang. Median Absolute Deviation) względem mediany. Wartość MAD jest obliczana zgodnie ze wzorem (3).

$$\mathrm{MAD} = mediana(|\mathrm{DE}_i - mediana(\mathrm{DE})|) \tag{3}$$

2.1 Trajektoria nieliniowa

Pierwsza trajektoria, nazwana nieliniową, została zaprojektowana z myślą o minimalizacji liczby przypadków, w których wylosowane punkty kotwicowe są współpłaszczyznowe [1, 6, 8]. Takie warunki są niezbędne ze względu na stabilność klasycznych estymatorów tj. algorytmów Foy'a i Chan'a, które w przypadku komplanarnych punktów kotwicowych charakteryzują się dużymi błędami (Chan) lub brakiem rozwiązania (Foy). Przebieg współrzędnych tej trajektorii wyraża równanie (4), a rys. 1 przedstawia środowisko symulacyjne z tego scenariuszu.

$$S_{nl} = \begin{cases} x \in [-7; 7, 5] \\ y = 3sin(0, 4\pi x) \\ z = 20 + 2cos(0, 4\pi x) \end{cases} \tag{4}$$

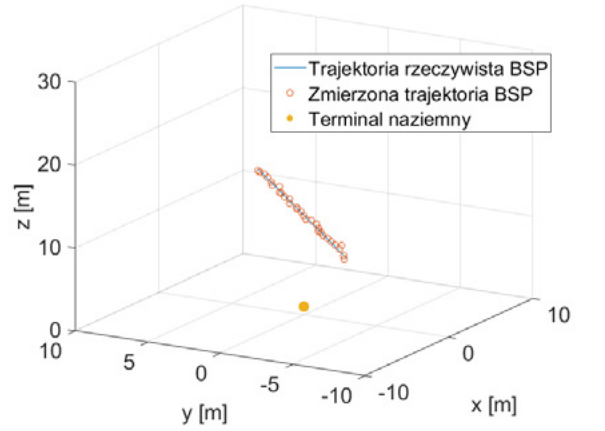


Rysunek 1: Środowisko symulacyjne - trajektoria nieliniowa

2.2 Trajektoria Liniowa

Drugi scenariusz obejmował uproszczoną trajektorię o kształcie linii prostej. Zastosowanie trajektorii o niskiej złożoności umożliwi użycie innych rodzajów BSP, których konstrukcja uniemożliwia wykonanie akrobatycznej trajektorii nieliniowej oraz pozwoli zaoszczędzić energię, co przełoży się bezpośrednio na wydłużenie czasu trwania lotu. Równanie (5) opisuje przebieg współrzędnych w uproszczonej trajektorii, a na rys. 2 zobrazowano środowisko symulacyjne.

$$S_l = \begin{cases} x \in [-7; 7, 5] \\ y = x \\ z = 12 \end{cases} \tag{5}$$



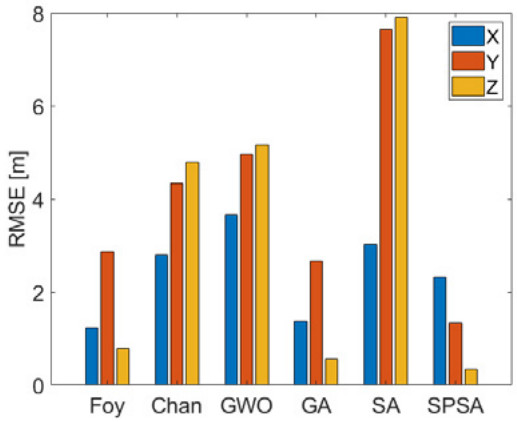
Rysunek 2: Środowisko symulacyjne - trajektoria liniowa

3. WYNIKI

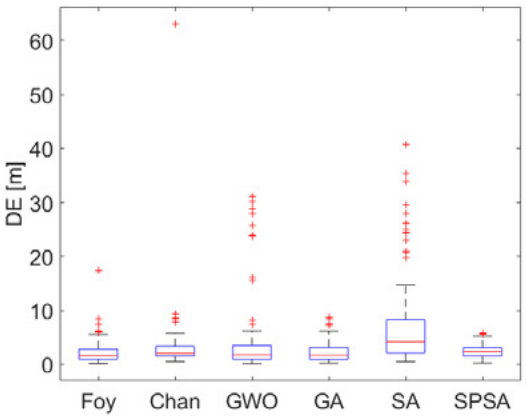
Po wykonaniu 100 symulacji dla każdego scenariusza obliczone zostały miary jakości estymacji położenia terminala naziemnego wspomniane w rozdziale 2 (RMSE i DE). Na podstawie miar sporządzono wykresy umożliwiające szybką ocenę badanych algorytmów. Dodatkowo, zmierzono średni czas wykonania wszystkich algorytmów, aby ocenić zasadność ich stosowania w czasie rzeczywistym.

3.1 Trajektoria nieliniowa

W przypadku trajektorii nieliniowej, wartości RMSE dla każdego algorytmu z podziałem na współrzędne (X, Y, Z) przedstawiono na rys. 3. Miarę DE zaprezentowano na wykresie pudełkowym zamieszczonym na rys. 4, a tabela 1 zawiera parametry statystyczne takie jak odchylenie standardowe σ, wartość średnią μ i medianę Med dla tej miary. W tabeli 2 zamieszczono zmierzony czas wykonania poszczególnych algorytmów w tym scenariuszu.



Rysunek 3: RMSE estymacji poszczególnych współrzędnych - trajektoria nieliniowa



Rysunek 4: DE estymacji poszczególnych współrzędnych - trajektoria nieliniowa

Tabela 1: Parametry statystyczne błędu estymacji DE - trajektoria nieliniowa

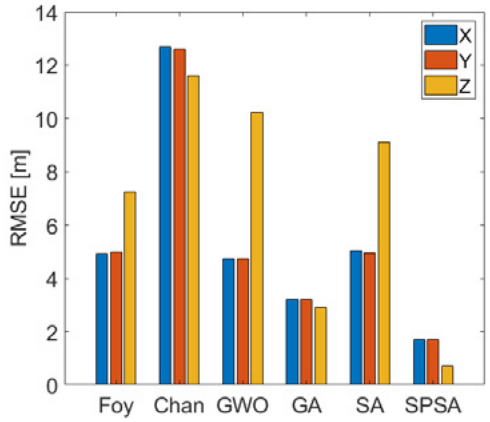
Algorytm	σ *[m]*	μ *[m]*	*Med [m]*
Foy	2,289	2,291	1,657
Chan	6,288	3,209	2,059
GWO	6,919	4,155	1,786
GA	1,949	2,352	1,715
SA	8,601	7,548	4,180
SPSA	1,213	2,418	2,350

Tabela 2: Średni czas obliczeń estymatorów - trajektoria nieliniowa

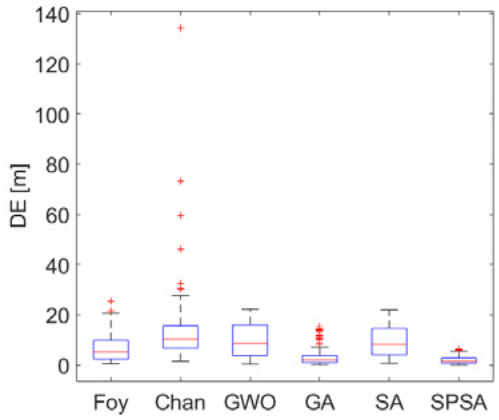
Algorytm	*Średni czas obliczeń [ms]*
Foy	0,654
Chan	0,178
GWO	7,983
GA	60,640
SA	11,057
SPSA	0,271

3.2 Trajektoria liniowa

Scenariusz pozwalający na ocenę trajektorii liniowej poskutkował obliczonymi wartościami RMSE dla wszystkich algorytmów z rozdzieleniem współrzędnych (X, Y, Z). Na rys. 5 przedstawiono wykres słupkowy tych wartości. Natomiast rys. 6 zawiera wykres pudełkowy miary DE, a w tabeli 3 zamieszczono podstawowe parametry statystyczne (odchylenie standardowe σ, wartość średnią μ i medianę Med) dla tego wskaźnika. Tabeli 4 zawiera średnie czasy wykonania każdego zbadanego algorytmu estymacji położenia.



Rysunek 5: RMSE estymacji poszczególnych współrzędnych - trajektoria liniowa



Rysunek 6: DE estymacji poszczególnych współrzędnych - trajektoria liniowa

Tabela 3: Parametry statystyczne błędu estymacji DE - trajektoria liniowa

Algorytm	σ *[m]*	μ *[m]*	*Med [m]*
Foy	6,796	7,547	5,107
Chan	16,151	13,989	10,336
GWO	7,272	9,858	8,677
GA	4,078	3,525	1,948
SA	6,411	9,593	8,131
SPSA	1,534	1,989	1,587

W przypadku metody Foy'a algorytm 72 ze 100 razy nie uzyskał zbieżności i nie zwrócił rozwiązania. Obliczone statystyki i miary jakości dla tej metody pochodzą z próby pomniejszonej o te przypadki tj. 28 próbek.

Tabela 4: Średni czas obliczeń estymatorów - trajektoria liniowa

Algorytm	*Średni czas obliczeń [ms]*
Foy	0,788
Chan	0,161
GWO	8,201
GA	64,267
SA	13,499
SPSA	0,302

3.3 Dyskusja Wyników

Zgodnie z oczekiwaniem, klasyczne metody (Foy, Chan) wykazały się niestabilnością estymacji położenia, co objawiło się dużymi błędami w metodzie Chan'a oraz brakiem 72 estymat w metodzie Foy'a. Dla obu trajektorii najlepszym estymatorem pod względem miary RMSE okazał się algorytm SPSA z wyjątkiem współrzędnej X dla trajektorii nieliniowej - w tej osi najlepszy rezultat uzyskał algorytm Foy'a. W przypadku DE metoda bazująca na SPSA, dla obu trajektorii, uzyskała najmniejsze odchylenie standardowe błędów. W przypadku trajektorii nieliniowej wartość średnia i mediana były najniższe dla algorytmu Foy'a, a w drugiej kolejności algorytmu genetycznego. Natomiast dla trajektorii liniowej algorytm SPSA uzyskał znacznie niższą wartość średnią błędu DE niż pozostałe algorytmy. Również dla tej trajektorii, biorąc pod uwagę medianę, algorytm SPSA okazał się być najlepszy wraz z algorytmem genetycznym - pozostałe estymatory uzyskały znacznie gorsze wyniki.

Uproszczenie trajektorii spowodowało nieznaczne wydłużenie czasów estymacji dla wszystkich algorytmów. Dla obu trajektorii najszybszy okazał się algorytm Chana, a algorytm SPSA również pozwalał na uzyskanie wyników w nieznacznie dłuższym czasie. Algorytm genetyczny oferuje wysoką dokładność estymacji natomiast, w porównaniu z pozostałymi estymatorami, jego czas wykonania może okazać się zbyt długi aby móc zastosować go w systemach czasu rzeczywistego.

3.4 Wnioski

Negatywny wpływ stosowania współpłaszczyznowych punktów kotwicowych na dokładność estymacji położenia terminala naziemnego z użyciem klasycznych algorytmów został potwierdzony. Algorytmy Foy'a i Chan'a zakładają dokładną znajomość położenia kotwic więc błędy w ich symulowanym pomiarze bezpośrednio wpływają na jakość estymacji. Rozważając metody optymalizacyjne, algorytm SPSA okazał się najbardziej dokładny z powodu uwzględnienia obecności szumu w funkcji kosztu. Pozostałe metody nie zakładają zaszumionej funkcji kosztu i przeszukują przestrzeń rozwiązań bazując na przekłamanych wartościach kosztu. Algorytm genetyczny (GA) również wykazał się wysoką odpornością na szum obecny w wartości funkcji kosztu, natomiast długi czas obliczeń wynoszący około 60 [ms] wyklucza go z zastosowań w systemach czasu rzeczywistego. Zastosowanie algorytmu SPSA wraz z uproszczoną trajektorią lotu pozwoliłoby wydłużyć czas lotu BSP (równoważny z czasem trwania usług lokalizacyjnych) oraz umożliwiłoby lokalizować nieruchome obiekty ze stosunkowo dużą dokładnością. Czas obliczeń dla algorytmu SPSA sugeruje zasadność stosowania go w rozważanym systemie czasu rzeczywistego.

4. PODZIĘKOWANIA

Publikacja została sfinansowana przez Wojskową Akademię Techniczną w ramach projektu numer UGB/22-059/2025/WAT.

LITERATURA

[1] Cho, Jungmin, Seungjin Jeong, i Byoungkil Lee. 2024. „A Study on Anchor Placement and 3D Positioning Algorithm for UWB Application in Small Sites". KSCE Journal of Civil Engineering 28 (10): 4575–87. https://doi.org/10.1007/s12205-024-2107-z.

[2] Guilmeau, Thomas, Emilie Chouzenoux, i Víctor Elvira. 2021. „Simulated Annealing: A Review and a New Scheme". W . https://inria.hal.science/hal-03275401.

[3] Janos, Daniel, Przemysław Kuras, i Łukasz Ortyl. 2022. „Evaluation of low-cost RTK GNSS receiver in motion under demanding conditions". Measurement 201 (wrzesień):111647. https://doi.org/10.1016/j.measurement.2022.111647.

[4] Jarosław Sadowski. 2020. „Badanie i analiza systemów radiolokalizacyjnych do zastosowań specjalnych". Gdańsk: Wydawnictwo Politechniki Gdańskiej.

[5] Jian-wu, Zhang, Yu Cheng-lei, Tang Bin, i Ji Yingying. 2008. „Chan Location Algorithm Application in 3-Dimension Space Location". W 2008 ISECS International Colloquium on Computing, Communication, Control, and Management, 2:622–25. https://doi.org/10.1109/CCCM.2008.76.

[6] Kumar, Sudhir, i Rajesh M. Hegde. 2016. „Multi-sensor data fusion methods for indoor localization under collinear ambiguity". Pervasive and Mobile Computing 30 (sierpień):18–31. https://doi.org/10.1016/j.pmcj.2015.09.001.

[7] Lambora, Annu, Kunal Gupta, i Kriti Chopra. 2019. „Genetic Algorithm- A Literature Review". W 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 380–84. https://doi.org/10.1109/COMITCon.2019.8862255.

[8] Liu, Wenyan, Guo Wei, i Ma Zhu. 2024. „A Survey on Multi-Dimensional Path Planning Method for Mobile Anchor Node Localization in Wireless Sensor Networks". Ad Hoc Networks 156 (kwiecień):103416. https://doi.org/10.1016/j.adhoc.2024.103416.

[9] Mirjalili, S., S.M. Mirjalili, i A. Lewis. 2014. „Grey Wolf Optimizer". Advances in Engineering Software 69:46–61. https://doi.org/10.1016/j.advengsoft.2013.12.007.

[10] Piotr Rajchowski. 2017. „Badanie i analiza dokładności estymacji położenia obiektów ruchomych w hybrydowym systemie lokalizacyjnym". Gdańsk.

[11] Risset, Tanguy, Claire Goursaud, Xavier Brun, Kevin Marquet, i Fabrice Meyer. 2018. „UWB Ranging for Rapid Movements". W 2018 International Conference on Indoor Positioning and Indoor Navigation (IPIN), 1–8. Nantes: IEEE. https://doi.org/10.1109/IPIN.2018.8533820.

[12] Spall, J.C. 1998. „Implementation of the simultaneous perturbation algorithm for stochastic optimization". IEEE Transactions on Aerospace and Electronic Systems 34 (3): 817–23. https://doi.org/10.1109/7.705889.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

Wykorzystanie praktycznie zweryfikowanych modeli predykcji zasięgowej radaru pasywnego do optymalizacji wyboru nadajników okazjonalnych

Utilization of Practically Verified Coverage Prediction Models of Passive Radar for Optimization of Transmitters of Opportunity Selection

Marcin Żywek

Politechnika Warszawska, Instytut Systemów Elektronicznych, Warszawa, marcin.zywek@pw.edu.pl

DOI: 10.15199/59.2025.4.89

Streszczenie: **Radar pasywny opiera swoje działanie na zewnętrznych nadajnikach okazjonalnych, takich jak radio FM czy telewizja DVB-T. Ponieważ system ten jest całkowicie uzależniony od dostępności i parametrów tych sygnałów, kluczowe znaczenie mają zarówno wiarygodna predykcja zasięgu detekcji, jak i właściwy dobór nadajników wykorzystywanych przez radar. Dodatkowo, z powodu ograniczonych zasobów obliczeniowych oraz często dużej liczby dostępnych nadajników, system może przetwarzać jednocześnie tylko ograniczoną liczbę kanałów częstotliwościowych, co wymusza wybór sygnałów zapewniających najlepszą skuteczność detekcji w danych warunkach.**

Abstract: **A passive radar relies on external transmitters of opportunity, such as FM radio stations or DVB-T television broadcasts. Since the system is entirely dependent on the availability and parameters of these signals, both reliable detection range prediction and the appropriate selection of transmitters used by the radar are of key importance. Additionally, due to limited computational resources and the often large number of available transmitters, the system can process only a limited number of frequency channels simultaneously. This necessitates the selection of signals that ensure the highest detection performance under given propagation conditions.**

Słowa kluczowe: **radar, radar pasywny, predykcja zasięgowa, propagacja fal radiowych**

Keywords: **radar, passive bistatic radar, coverage prediction, RF propagation and modeling**

1. WSTĘP

W ostatnich latach obserwuje się rosnące zainteresowanie zastosowaniami radarów pasywnych, które – w przeciwieństwie do systemów aktywnych – opierają swoje działanie na odbiorze sygnałów emitowanych przez zewnętrzne nadajniki okazjonalne, takie jak stacje radiowe FM, DAB czy nadajniki DVB-T. Istotnym ograniczeniem tych systemów jest możliwość jednoczesnego wykorzystania tylko wybranych kanałów częstotliwościowych, co sprawia, że kluczowe staje się taki dobór nadajników, aby zapewnić możliwie największą skuteczność detekcji oraz lokalizacji obiektów. Zadanie to komplikuje dodatkowo trudność wiarygodnej predykcji zasięgu detekcji, wynikająca m. in. z multistatycznego charakteru działania systemu, bistatycznej skutecznej powierzchni odbicia oraz dużej liczby czynników wpływających na propagację sygnału, takich jak np. charakterystyka anten, wielodrogowość i dyfrakcja [1, 11, 12].

W niniejszym artykule przedstawiono podejście do optymalizacji wyboru nadajników okazjonalnych na podstawie praktycznie zweryfikowanych modeli predykcji zasięgowej dla radaru pasywnego. Połączenie analiz teoretycznych, symulacji komputerowych oraz rzeczywistych pomiarów pozwala na opracowanie rozszerzonych modeli zasięgowych, które uwzględniają zarówno parametry nadajników, radaru, jak i wpływ środowiska propagacyjnego. Wyniki te stanowią podstawę do efektywnego planowania pracy radaru pasywnego, określania jego parametrów i zwiększenia jego skuteczności w rzeczywistych warunkach.

2. PREDYKCJA ZASIĘGOWA

2.1. Rozszerzone równanie zasięgu

Podstawą teoretyczną do predykcji zasięgu radaru pasywnego jest rozszerzone równanie zasięgu bistatycznego, które jest punktem wyjścia do analizy zasięgowej [2, 11, 16]:

$$P_r = \frac{\mathrm{EIRP}}{L_t(\phi_t)} \frac{G_{r\max}}{L_r(\phi_r)} \frac{\sigma\lambda^2}{(4\pi)^3 R_1^2 R_2^2 L F_t^2 F_r^2} \tag{1}$$

gdzie:

- EIRP - zastępcza izotropowa moc promieniowania (ang. Effective Isotropical Radiated Power) nadajnika,
- ϕ_t - kąt azymutalny obiektu względem nadajnika,
- $G_{r\max}$ - maksymalny zysk anteny odbiornika,
- $L_t(\phi_t)$ i $L_r(\phi_r)$ - charakterystyki promieniowania nadajnika oraz anten odbiornika (radaru),
- F_t i F_r - współczynniki propagacji odpowiadających odpowiednio drodze nadajnik-obiekt i odbiornik-obiekt ,
- σ - bistatyczna powierzchnia skuteczna (RCS),
- λ – długość fali,
- R_1 – odległość między nadajnikiem a obiektem,

- R_2 - odległość między odbiornikiem a obiektem,
- L - całkowite tłumienie.

Stosunek sygnał-szum SNR (ang. Signal-to-Noise Ratio) można obliczyć w następujący sposób [6, 10]:

$$\mathrm{SNR} = \frac{P_r}{P_n} \cdot B \cdot T = \frac{P_r}{P_n} \cdot G_i \qquad (2)$$

gdzie:

- P_n to całkowita moc szumu w systemie (szum termiczny i zewnętrzny szum z otoczenia),
- B to szerokość pasma sygnału,
- T to czas integracji,
- G_i to zysk przetwarzania.

Iloczyn $B \cdot T$ jest nazywany zyskiem przetwarzania G_i i osiagany jest poprzez koherentną integrację sygnału o szerokości pasma B w czasie T. Typowe wartości iloczynu $B \cdot T$ dla radarów pasywnych wynoszą 40-60 dB [9]. Szum termiczny można obliczyć jako $k_B T_0 B$, gdzie k_B to stała Boltzmanna, T_0 to temperatura szumu odbiornika, a B to szerokość pasma odbiornika . Poziom szumu jest zwiększany przez współczynnik szumu odbiornika F, co powoduje, że efektywna moc szumu wynosi $F k_B T_0 B$.

2.2. Detekcja

Aby wykryć obiekt w radarze pasywnym, wykorzystywana jest funkcja nieoznaczoności wzajemnej sygnału referencyjnego i echa (CAF, ang. Cross-Ambiguity Function). Pozwala ona na odnalezienie echa obiektu na tle szumu poprzez korelację sygnału echa $x_e(t)$ z odpowiednio opóźnionym i przemnożonym sygnałem referencyjnym $x_r(t)$ [10]:

$$\psi(R,V) = \int_{-T/2}^{T/2} x_e(t) x_r^*\left(t - \frac{R}{c}\right) \exp\left(-j\frac{2\pi}{\lambda} Vt\right) dt \qquad (3)$$

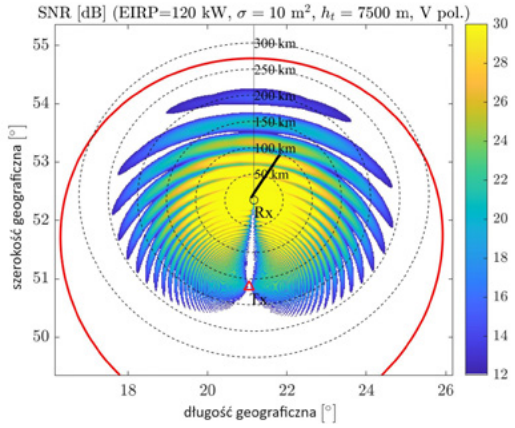
gdzie:

- R - bistatyczna odległość,
- V - bistatyczna prędkość,
- c - prędkość światła.

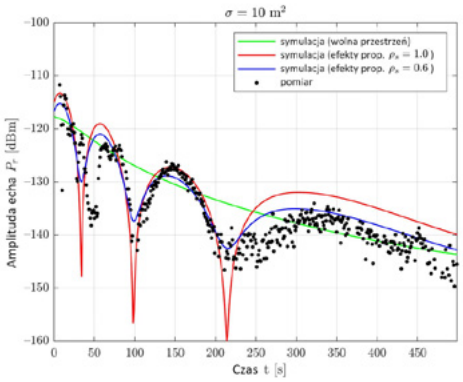
W praktyce sygnał echa zawiera nie tylko odbicie od obiektu, ale także silny sygnał bezpośredni oraz szumy i inne zakłócenia. Dlatego przed obliczeniem funkcji niejednoznaczności wzajemnej stosuje się filtry adaptacyjne lub algorytmy np. CLEAN, które pozwalają usunąć sygnał bezpośredni i zredukować zakłócenia [4]. Detekcja w radarze pasywnym polega na porównaniu modułu funkcji $\psi(R,V)$ z ustalonym progiem. W praktyce, dla zapewnienia niskiego poziomu fałszywych alarmów ($10^{-3} - 10^{-6}$) stosuje się próg detekcji ok. 13 dB [.

2.3. Porównanie wyników predykcji zasięgowej z pomiarami rzeczywistymi

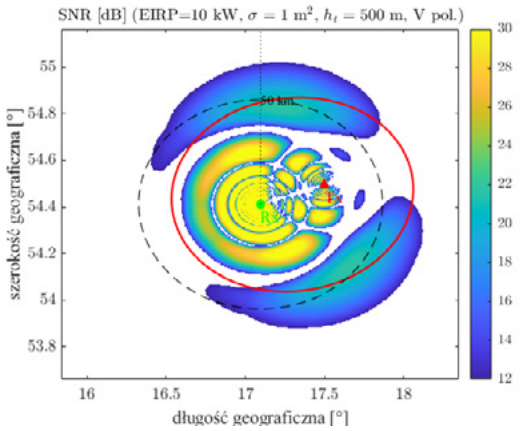
Rys. 1 przedstawia wyniki symulacji SNR echa samolotu Boeing 737-400 z wykorzystaniem sygnału radia FM (88.2 MHz) [5]. Na wyniki symulacji czarną linią nałożono trajektorię lotu samolotu. Widać, że w czasie pomiaru samolot przemieszcza się przez kilka minimów



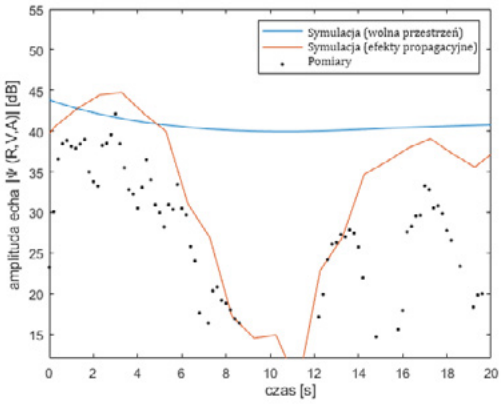
Rys. 2. Symulacja SNR echa samolotu (sygnał FM)



Rys. 3. Symulowana i zmierzona moc sygnału echa dla samolotu Boeing 737-400 (sygnał FM)



Rys. 4. Symulacja SNR echa rakiety (sygnał DVB-T)



Rys. 4. Symulowana i zmierzona moc sygnału echa dla rakiety doświadczalnej (sygnał DVB-T)

spowodowanych efektem wielodrogowości na ścieżce nadajnik-obiekt. Rys. 2. przedstawia symulowaną oraz zmierzoną moc sygnału echa. Dla współczynnika rozpraszania $\rho_s = 0.6$, który związany jest z chropowatością powierzchni, symulacja (niebieska linia) dobrze odpowiada pomiarom (czarne kropki).

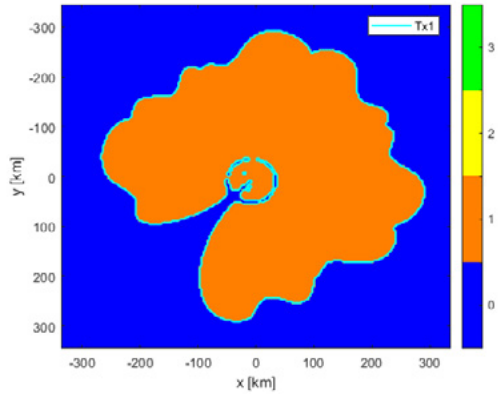
Rys. 3. Przedstawia wyniki symulacji SNR dla rakiety doświadczalnej z wykorzystaniem sygnału DVB-T w paśmie VHF (184.5 MHz) [3, 15]. Rys. 4. Przedstawia symulowaną i zmierzoną moc sygnału echa. Widać, że pomiary (czarne kropki) dobrze pokrywają się z symulacją uwzględniającą efekty propagacyjne (czerwona linia).

3. OPTYMALIZACJA WYBORU NADAJNIKÓW OKAZJONALNYCH

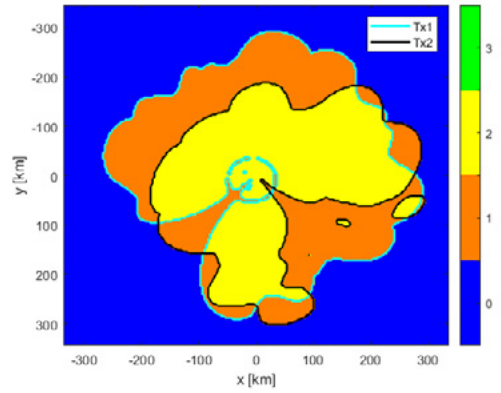
Ze względu na brak własnego nadajnika, zdolności detekcyjne radaru pasywnego PCL są uzależnione od wykorzystywanych nadajników okazjonalnych. W idealnych warunkach radar pasywny mógłby korzystać z wszystkich dostępnych sygnałów do detekcji obiektów. Niestety, w praktyce nie jest to możliwe ze względu na wysoką złożoność obliczeniową algorytmów przetwarzania sygnałów, ograniczone zasoby sprzętowe oraz dużą liczbę dostępnych transmisji [14, 16]. Pojawia się zatem problem wyboru odpowiednich nadajników okazjonalnych, a dokładniej wyboru konkretnych transmisji z wielu masztów radiowych.

Rys. 5-8 ilustrują cztery kroki w przykładowym procesie optymalizacji wyboru nadajników (Tx) dla radaru pasywnego (Rx). W najprostszym przypadku procedura ta polega na: przeliczeniu obszarów detekcji dla wszystkich dostępnych nadajników na zadanym obszarze, posortowaniu ich według wielkości tych obszarów, a następnie wyborze nadajników oferujących największy zasięg wykrywania obiektów. Zielony obszar wskazuje region, w którym co najmniej trzy unikalne pary nadajnik–odbiornik mogą wykryć i zlokalizować obiekt (na danej wysokości). Kontury w kolorze jasnoniebieskim, czarnym, czerwonym i różowym odpowiadają odpowiednio pierwszemu, drugiemu, trzeciemu i czwartemu wybranemu nadajnikowi. Widoczny jest złożony kształt obszaru detekcji, który wyraźnie odbiega od tych dla radarów aktywnych. Warto zaznaczyć, że przedstawione wyniki dotyczą sytuacji 2D (przy stałej wysokości lotu obiektu oraz stałej skutecznej powierzchni odbicia); w praktyce należałoby rozważyć przypadek 3D, obejmujący m. in. różne wysokości lotu oraz zmianę bistatycznej skutecznej powierzchni odbicia.

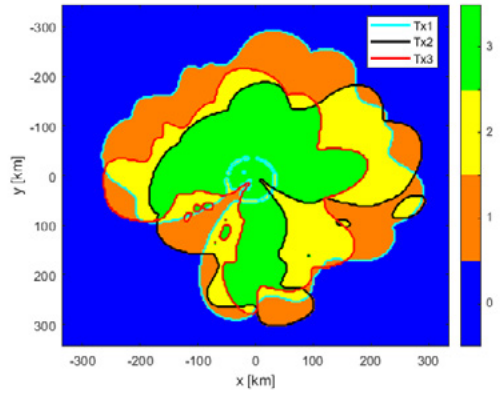
Optymalizacja wyboru nadajników może przebiegać na różne sposoby [7, 8, 13]. Jednym z podejść może być również maksymalizacja zdolności detekcyjnych w wybranych kierunkach lub na określonych obszarach. Inną strategią może być uwzględnienie pracy wielu współpracujących stacji odbiorczych, co może znacząco poprawić możliwości detekcyjne systemu. Ważnym kryterium może być również taki dobór nadajników okazjonalnych, który pozwoli zwiększyć dokładność uzyskiwanych danych, ponieważ ze względu na geometrię bistatyczną precyzja pomiarów i lokalizacji obiektów zależy



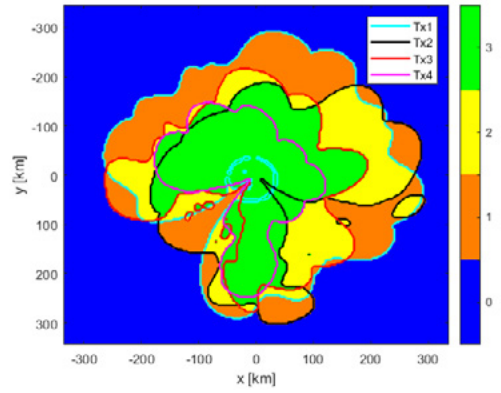
Rys. 5. Przewidywany obszar detekcji z 1 parą Tx-Rx



Rys. 6. Przewidywany obszar detekcji z 2 parami Tx-Rx



Rys. 7. Przewidywany obszar detekcji z 3 parami Tx-Rx



Rys. 8. Przewidywany obszar detekcji z 4 parami Tx-Rx

od wzajemnego położenia nadajnika, radaru oraz obiektu [5].

Wszystkie te aspekty sprawiają, że właściwy wybór nadajników okazjonalnych w radarze pasywnym jest zadaniem złożonym i zależy od wielu czynników, takich jak: dostępność i konfiguracja nadajników (rodzaj sygnału, moc, charakterystyka antenowa, wysokość masztu), lokalizacja i konfiguracja radaru pasywnego, typ obserwowanego obiektu, ukształtowanie terenu i parametry gruntu, zakładana dokładność oraz inne wymagania systemowe. Jednakże, wykorzystując praktycznie zweryfikowane modele predykcji zasięgu radaru pasywnego, możliwy jest zarówno właściwy wybór nadajników okazjonalnych, jak i określenie wiarygodnych charakterystyk zasięgowych takiego systemu, co jest kluczowe dla jego praktycznego zastosowania, np. w kontroli ruchu lotniczego.

4. PODSUMOWANIE

W niniejszym artykule przedstawiono wykorzystanie praktycznie zweryfikowanych modeli predykcji zasięgowej radaru pasywnego, wykorzystującego sygnały radia FM i telewizji cyfrowej DVB-T, do optymalizacji wyboru nadajników okazjonalnych. Zaprezentowano rozszerzone równania zasięgowe oraz metody detekcji obiektów w radarze pasywnym. Przedstawiono również porównanie wyników symulacyjnych z rzeczywistymi wynikami detekcji samolotu pasażerskiego oraz rakiety doświadczalnej dla różnych typów sygnałów, co potwierdziło poprawność zastosowanych modeli. Dzięki możliwości wyznaczania obszarów detekcji obiektów zaproponowano metody wyboru nadajników okazjonalnych. Przedstawione wyniki mogą znaleźć praktyczne zastosowanie w radarach pasywnych.

LITERATURA

[1] Barott William C., T. Dabrowski, Braham Himed. 2015. „Fidelity and Complexity in Passive Radar Simulations". *2015 IEEE 16th International Symposium on High Assurance Systems Engineering* : 277–278.

[2] Barton David K. 2012. Radar Equations for Modern Radar. Boston, MA: Artech House.

[3] Brzozowski Marek, Mariusz Pakowski, Zbigniew Jakielaszek, Mirosław Michalczewski, Mirosław Myszka. 2021. „Challenges of Preparation and Realization of Combined Field Tests of Passive and Active Radar Sensors on an Example APART-GAS 2019 Trials". *2021 IEEE 8th International Workshop on Metrology for AeroSpace (MetroAeroSpace)* : 188–192.

[4] Colone Fabiola, David W. O'Hagan, Pierfrancesco Lombardo, Christopher J. Baker. 2009. „A Multistage Processing Algorithm for Disturbance Removal and Target Detection in Passive Bistatic Radar". *IEEE Transactions on Aerospace and Electronic Systems* 45 (2) : 698–722.

[5] Greco Maria, Pietro Stinco, Fulvio Gini, Alfonso Farina, Muralidhar Rangaswamy. 2010. „Cramér-Rao Bounds and TX-RX Selection in a Multistatic Radar Scenario". *2010 IEEE Radar Conference* : 1371–1376.

[6] Griffiths Hugh D., Christopher J. Baker. 2017. An Introduction to Passive Radar. Norwood, MA: Artech House.

[7] Gumiero Francesca, Cinzia Nucciarone, Valeria Anastasio, Pierfrancesco Lombardo, Fabiola Colone. 2010. „Multistatic Passive Radar Geometry Optimization for Target 3D Positioning Accuracy". *The 7th European Radar Conference* : 467–470.

[8] Inggs Martin, Yoann Paichard, Gunther Lange. 2009. „Passive Coherent Location System Planning Tool". *Proc. Int. Radar Conf. 'Surveill. Safer World' (RADAR)* : 1–5.

[9] Malanowski Mateusz. 2009. Optymalizacja przetwarzania sygnałów w radarach z pasywną koherentną lokalizacją obiektów. Rozprawa doktorska, Politechnika Warszawska.

[10] Malanowski Mateusz. 2019. Signal Processing for Passive Bistatic Radar. Norwood, MA: Artech House.

[11] Malanowski Mateusz, Marcin Żywek, Marek Płotka, Krzysztof Kulpa. 2022. „Passive Bistatic Radar Detection Performance Prediction Considering Antenna Patterns and Propagation Effects*". IEEE Transactions on Geoscience and Remote Sensing* 60 : 1–16.

[12] O'Hagan Daniel W., Hugh D. Griffiths, Stefan M. Ummenhofer, Stephen T. Paine. 2017. „Elevation Pattern Analysis of Common Passive Bistatic Radar Illuminators of Opportunity". *IEEE Transactions on Aerospace and Electronic Systems 53 (6)* : 3008–3019.

[13] Winkler Volker, Sebastian Lutz, Michael Brandfass. 2021. „Expert Systems for Passive Radar Configuration". *2021 21st International Radar Symposium (IRS)* : 1–10.

[14] Żywek Marcin, Malanowski Mateusz, "Performance Optimization for Passive Bistatic Radar Based on Detection Range Predictions," *2021 Signal Processing Symposium (SPSympo)*, LODZ, Poland, 2021, pp. 322-326, doi: 10.1109/SPSympo51155.2020.9593804.

[15] Żywek Marcin. 2024. „Rocket Launch Detection Using VHF DVB-T Passive Bistatic Radar". 2024 *International Radar Symposium (IRS),* Wrocław : 293–297.

[16] Żywek Marcin. 2024. Optymalizacja metod wykrywania obiektów szybkich w radarze pasywnym. Rozprawa doktorska, Politechnika Warszawska.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

ANTENA INTELIGENTNEGO ZEGARKA PRACUJĄCA W PAŚMIE 2,4 GHz

SMARTWATCH ANTENNA FOR 2.4 GHz BAND

Łukasz Januszkiewicz[1]; Remigiusz Danych[2]; Maciej Łaski [3]

[1] Instytut Elektroniki, Politechnika Łódzka, Łódź, lukasz.januszkiewicz@p.lodz.pl
[2] GlobalRnD sp. z o. o., Łódź, danych@globalrnd.pl
[3] GlobalRnD sp. z o. o., Łódź, laski.maciej@globalrnd.pl

DOI: 10.15199/59.2025.4.90

Streszczenie: **W artykule przedstawiono projekt anteny typu odwrócone F przeznaczonej do zastosowania w inteligentnym zegarku pracującym w paśmie 2,4 GHz. Projekt anteny uwzględnia ograniczenia przestrzenne wynikające z małych wymiarów urządzenia oraz bliskości komponentów elektronicznych, takich jak obwód drukowany, wyświetlacz i akumulator. W projekcie anteny uwzględniono wpływ ciała użytkownika na jej charakterystykę wykorzystując symulacje numeryczne metodą różnic skończonych w dziedzinie czasu (FDTD) z wykorzystaniem heterogenicznego modelu ciała. Zaprezentowano wyniki symulacji oraz pomiarów prototypowej anteny, które wykazały zadowalające parametry w zakresie dopasowania impedancyjnego i zysku energetycznego.**

Abstract: **This paper presents the design of an inverted-F antenna intended for integration into a smartwatch operating within the 2.4 GHz band. The antenna design addresses spatial constraints imposed by the device's miniaturized form factor and the proximity of electronic components, including the printed circuit board, display, and battery. The influence of the user's body on the antenna's performance characteristics was considered during the design phase through numerical simulations employing the Finite-Difference Time-Domain (FDTD) method with a heterogeneous human body model. Simulation results and measurements of a fabricated prototype antenna are presented, demonstrating satisfactory performance in terms of impedance matching and gain.**

Słowa kluczowe: **antena, systemy noszone, FDTD, sieci WBAN.**

Keywords: **antenna, FDTD, wearable electronics, WBAN.**

1. WSTĘP

Wraz z dynamicznym rozwojem elektroniki obserwuje się ciągłą miniaturyzację komponentów i integrację coraz bardziej złożonych funkcji w urządzeniach o niewielkich rozmiarach. Trend ten znajduje odzwierciedlenie w rozwoju inteligentnych urządzeń noszonych (ang. wearables), które zyskują na popularności dzięki możliwości monitorowania różnorodnych parametrów użytkownika i interakcji z otoczeniem.

W odpowiedzi na wyzwania demograficzne związane ze starzeniem się społeczeństwa i rosnącym zapotrzebowaniem na zdalną opiekę, powstał projekt Wirtualny Dom Opieki AnnA realizowany przez firmę Global R&D przy współpracy z Instytutem Elektroniki Politechniki Łódzkiej, w ramach projektu POIR.01.01.01-00-0836/20, p.t. Opracowanie w drodze prac b+r infrastruktury do innowacyjnego systemu 24h opieki osób z niedołęstwem w formie „OSOBISTEGO OPIEKUNA" i „WIRTUALNEGO DOMU OPIEKI" [1]. Jego celem jest stworzenie innowacyjnego systemu, wykorzystującego elektroniczne urządzenia noszone, który umożliwia zdalny nadzór nad zdrowiem i samopoczuciem osób wymagających wsparcia, przy jednoczesnym zachowaniu ich niezależności w domowym środowisku. System przedstawiono schematycznie na rysunku 1.

Rys. 1. Widok zaprojektowanej anteny przeznaczonej do montażu w zegarku

Kluczowym elementem tego systemu jest zaawansowany zegarek inteligentny AnnA band. To urządzenie, wyposażone w szereg sensorów, w sposób ciągły monitoruje podstawowe parametry życiowe, takie jak temperatura ciała, puls, poziom aktywności fizycznej (liczbę kroków i ruch) oraz saturację krwi tlenem. Wbudowany alarm stanowi istotną funkcję zegarka, umożliwiając natychmiastowe powiadomienie wskazanych osób w sytuacji nagłego zdarzenia, takiego jak upadek czy nagłe pogorszenie samopoczucia. Kluczowe dla sprawnego działania wszystkich funkcji monitorujących i alarmowych

zegarka AnnA band jest zastosowanie łączności bezprzewodowej, która umożliwia transmisję danych w czasie rzeczywistym do dedykowanej aplikacji AnnA. Ze względu na dostępność nielicencjonowanego pasma częstotliwości 2,4 GHz oraz relatywnie niski koszt modułów bezprzewodowych, zdecydowano się na wykorzystanie go w projektowanym systemie. Zegarek komunikuje się z modułem centralnym umieszczonym wewnątrz budynku za pośrednictwem sieci modułów pośrednich, tworzących niezawodną infrastrukturę do przesyłania danych. Dzięki temu opiekunowie i bliscy mają stały wgląd w aktualny stan zdrowia i aktywność podopiecznego, co znacząco zwiększa poziom bezpieczeństwa i efektywność opieki na odległość.

Niniejszy artykuł poświęcony jest szczegółowemu opisowi procesu projektowania anteny zintegrowanej z zegarkiem AnnA band. W dalszej części pracy przedstawione zostaną założenia projektowe, projekt anteny oraz jej model numeryczny. Przedstawiony zostanie również prototyp anteny oraz wyniki jego weryfikacji pomiarowej.

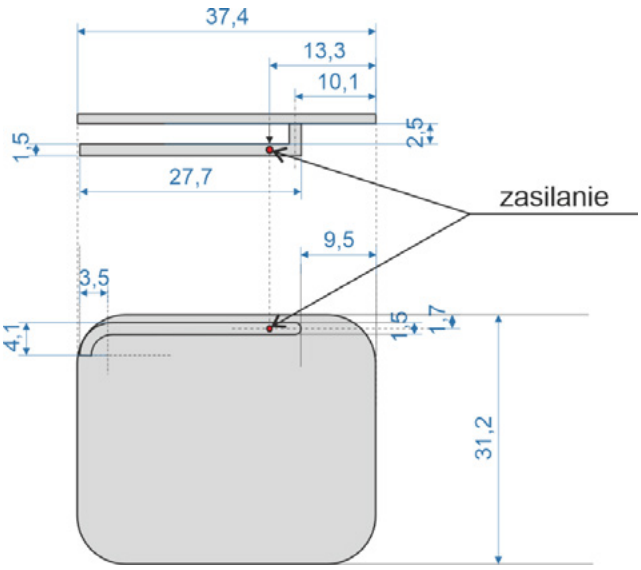
2. PROJEKT ANTENY

W ramach projektu opracowano antenę dla inteligentnego zegarka pracującego w popularnym paśmie ISM 2,4 GHz, wykorzystywanym m.in. przez technologie Bluetooth i Wi-Fi. Ze względu na niewielkie rozmiary urządzenia oraz wymóg integracji z istniejącym projektem podzespołów elektronicznych, wybrano antenę typu odwrócone F (IFA). Anteny tego typu charakteryzują się kompaktowymi wymiarami, prostą konstrukcją oraz możliwością łatwego dopasowania impedancyjnego do 50 Ω oraz dzięki czemu były stosowane w zegarkach inteligentnych [2-4].

Na potrzeby opracowania projektu anteny przyjęto następujące założenia: pasmo impedancyjne anteny pokrywa się z pasmem ISM 2,4 GHz, impedancja wejściowa anteny powinna być zbliżona do 50 Ω, przy czym maksymalne niedopasowanie w obrębie pasma nie powinno przekroczyć wartości 2 współczynnika fali stojącej (WFS). Kolejne założenie dotyczy przeznaczenia anteny do montażu w prototypowym zegarku, na prototypowym obwodzie drukowanym o wymiarach 31,2 mm x 37,4 mm. Ponadto, wymiary anteny umożliwią jej implementację w opracowanym urządzeniu, które zawiera akumulator i wyświetlacz LCD.

Na rysunku 2 przedstawiono widok ogólny zaprojektowanej anteny typu F. Promiennik anteny został wykonany z drutu miedzianego o średnicy 1,5 mm i jest podłączony do płaszczyzny masy obwodu drukowanego. Wymiary anteny zostały szczegółowo zaprezentowane na rysunku 3. Zostały one tak dobrane, aby wykorzystać wolną przestrzeń we wnętrzu zegarka. Ze względu na niewielką odległość promiennika anteny od obudowy zegarka, akumulatora i ciała człowieka konieczne było uwzględnienie w projekcie efektu odstrojenia impedancyjnego. Został on skompensowany poprzez skorygowanie długości elementu promieniującego względem początkowo wyznaczonej wielkości typowej dla anten pracujących w wolnej przestrzeni. W tym celu zastosowano program komputerowy XFdtd wykorzystujący metodę różnic skończonych w dziedzinie czasu oraz numeryczny model ciała człowieka [5,6]. Model ciała składał się z wokseli, których długość krawędzi wynosi 2 mm. Model odwzorowuje 39 tkanek ciała. Na rysunku 4 przedstawiono model numeryczny anteny umieszczonej w prototypowym urządzeniu usytuowanym na nadgarstku prawej ręki. W modelu pominięto pasek mocujący zegarek do nadgarstka ponieważ jest on wykonany z cienkiego, syntetycznego materiału tekstylnego o małej przenikalności elektrycznej (około 1,5) i niewielkiej stratności. Ponadto pasek jest odseparowany od promiennika anteny obudową, której wpływ na parametry anteny jest dominujący.

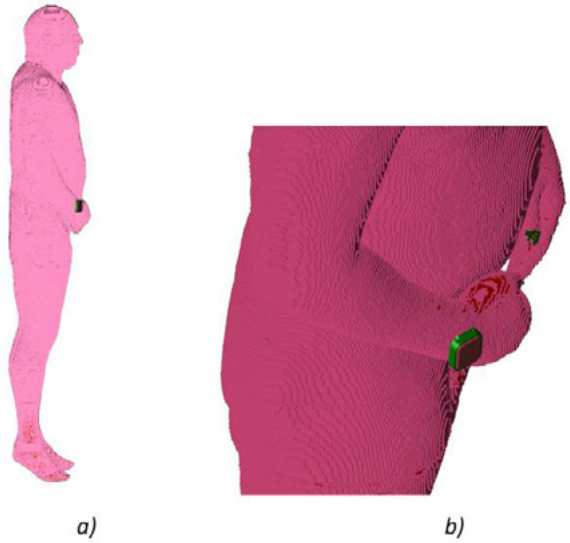
Rys. 2. Widok zaprojektowanej anteny przeznaczonej do montażu w zegarku



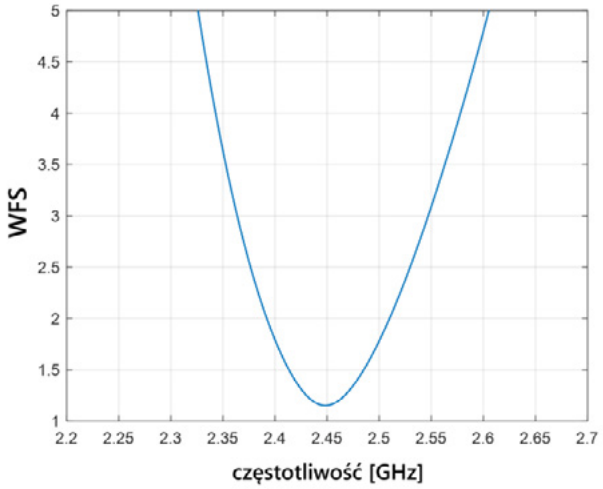
Rys. 3. Wymiary anteny

Model numeryczny wykorzystany w symulacjach obejmował promiennik anteny, obwód drukowany, obudowę wykonaną z polimeru (PLA) oraz wyświetlacz LCD i akumulator. Wyświetlacz LCD oraz akumulator od strony wnętrza urządzenia otoczone są warstwą folii metalowej co również zostało odwzorowane w modelu numerycznym. Przeprowadzono symulacje komputerowe dopasowania impedancyjnego anteny umieszczonej w urządzaniu znajdującym się na ręce człowieka. Wyniki uzyskane dla modelu numerycznego przedstawione na

rys. 5 wskazują na dobre dopasowanie impedancyjne anteny do założonego pasma częstotliwości obejmującego zakres od 2,4 GHz do 2,5 GHz.

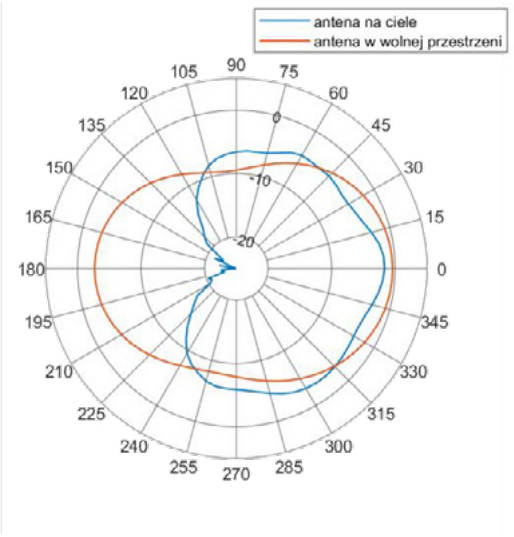


Rys. 4. Model numeryczny anteny umieszczonej w prototypowym urządzeniu usytuowanym na nadgarstku; a – widok całej sylwetki, b – widok obszaru ręki wraz z urządzeniem



Rys. 5. Wyniki symulacji dopasowania impedancyjnego anteny umieszczonej w urządzeniu

Korzystając z opracowanego modelu numerycznego przeprowadzono symulacje charakterystyki promieniowania anteny umieszczonej w urządzeniu znajdującym się na prawym nadgarstku. Na potrzeby symulacji przyjęto warunek brzegowy odpowiadający wolnej przestrzeni tak aby wyeliminować wpływ ograniczonej powierzchni podłoża na wyniki. Odpowiada to warunkom komory bezodbiciowej, w której dokonano później weryfikacji pomiarowej. Wyniki symulacji charakterystyki promieniowania w płaszczyźnie poziomej przedstawione zostały na rysunku 6. Jest to charakterystyka zysku energetycznego którego wartość maksymalna dla anteny usytuowanej na ciele wynosi -2 dBi. W przypadku anteny usytuowanej w wolnej przestrzeni wartość maksymalna zysku wyniosła -0.5 dBi.



Rys. 6. Wyniki symulacji charakterystyki promieniowania anteny w płaszczyźnie poziomej dla anteny umieszczonej w wolnej przestrzeni i na ciele człowieka

Na podstawie wyników symulacji przedstawionych na rys. 6 widoczny jest wpływ ciała człowieka na charakterystykę promieniowania anteny. Wytłumienie energii wypromieniowanej przez antenę w ciele człowieka znacząco ogranicza zysk anteny dla kierunków przysłoniętych.

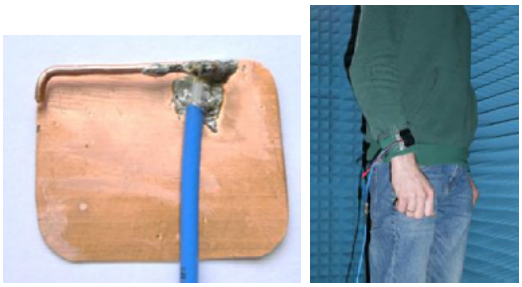
3. WERYFIKACJA POMIAROWA PROJEKTU

Opracowany projekt anteny zweryfikowano za pomocą pomiarów przeprowadzonych z wykorzystaniem prototypowej anteny przedstawionej na rysunku 6. Płaszczyzna masy obwodu drukowanego, do którego podłączony jest promiennik została odwzorowana w prototypie za pomocą kształtki wykonanej z blachy miedzianej o grubości 0,5 mm. Promiennik anteny wykonano z drutu miedzianego o średnicy 1,5 mm. Antena zasilana jest kablem koncentrycznym typu Huber + Suhner Multiflex 86 zakończonym złączem typu SMA.

Weryfikację pomiarową parametrów anteny przeprowadzono w warunkach komory bezodbiciowej. Antena została umieszczona w urządzeniu prototypowym, które przymocowano do nadgarstka ochotnika, co prezentuje rysunek 7. Pomiary dopasowania impedancyjnego przeprowadzono za pomocą wektorowego analizatora obwodów Anritsu MS 4647B. Płaszczyzna kalibracji przesunięta została do punktu otwarcia kabla koncentrycznego. Wyniki pomiarów dopasowania impedancyjnego do impedancji Z_0=50 Ω przedstawiono za pomocą współczynnika fali stojącej (WFS) na rysunku 8.

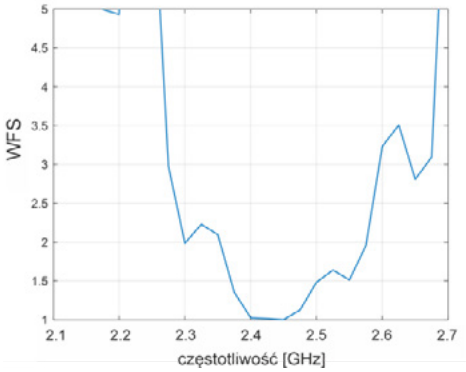
Przeprowadzono również pomiary charakterystyki promieniowania anteny prototypowej umieszczonej w urządzeniu przymocowanym do nadgarstka. Do pomiarów wykorzystano generator sygnałowy Rohde & Schwarz SMA 100 oraz analizator widma Rohde & Schwarz FSC 6 i anteny pomiarowe Rohde & Schwarz HF-907. Wyniki przedstawiono na rysunku 9. Zysk maksymalny anteny prototypowej oszacowano za pomocą porównania z antena wzorcową, jaką stanowił 2 egzemplarz

anteny Rohde & Schwarz HF-907. Szacowana wartość maksymalna zysku energetycznego wyniosła -3.5 dBi.

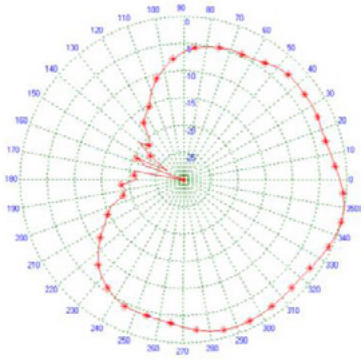


a) *b)*

Rys. 7. Prototyp anteny: a – widok promiennika, b – usytuowanie na ciele człowieka podczas pomiarów



Rys. 8. Wyniki pomiaru dopasowania impedancyjnego anteny prototypowej umieszczonej w urządzeniu przymocowanym do nadgarstka.



Rys. 9. Wyniki pomiarów charakterystyki promieniowania anteny prototypowej umieszczonej na ciele, w płaszczyźnie poziomej

4. PODSUMOWANIE I WNIOSKI

W artykule przedstawiono projekt i analizę kompaktowej anteny typu odwrócone F przeznaczonej do inteligentnego zegarka pracującego w paśmie 2,4 GHz. Uwzględnienie ograniczeń projektowych oraz wpływu ciała użytkownika wymagało przeprowadzenia symulacji komputerowych anteny z wykorzystaniem heterogenicznego modelu ciała.

Wyniki pomiarów dopasowania impedancyjnego prototypowej anteny wykazały dobrą zgodność z wynikami symulacji. Zmierzony współczynnik fali stojącej w paśmie 2.4 GHz nie przekraczał wartości 2, co świadczy o dobrym dopasowaniu impedancyjnym anteny do systemu transmisyjnego. Charakterystyka promieniowania anteny uzyskana na podstawie pomiarów ma bardzo zbliżony kształt do charakterystyki wyznaczonej na podstawie symulacji komputerowych. Wartość maksymalna zysku energetycznego uzyskana za pomocą pomiarów jest jednak mniejsza o około 1.5 dB od wartości wyznaczonej za pomocą modelu numerycznego. Może to wynikać z różnic pomiędzy wartościami parametrów materiałowych obudowy zegarka przyjętych w modelu i występującymi w rzeczywistości. Należy zauważyć, że wartość zysku maksymalnego na poziomie -3.5 dBi jest akceptowalna przy uwzględnieniu niewielkich wymiarów anteny. Również bliskość innych komponentów elektronicznych wykonanych z materiałów przewodzących znacząco pogarsza warunki pracy anteny.

Dalsze prace w ramach tego projektu będą miały na celu zweryfikowanie możliwości zastosowania technologii elastycznego laminatu cienkowarstwowego do wykonania anteny. Będzie się to wiązało z koniecznością zmodyfikowania projektu obudowy, ale może ułatwić wielkoseryjną produkcję inteligentnego zegarka Anna band.

Badania opisane w artykule były finansowane z projektu POIR.01.01.01-00-0836/20, p.t. Opracowanie w drodze prac b+r infrastruktury do innowacyjnego systemu 24h opieki osób z niedołęstwem w formie „OSOBISTEGO OPIEKUNA" i „WIRTUALNEGO DOMU OPIEKI".

LITERATURA

[1] https://globalrnd.pl/anna/
[2] Chen, W.-S., C.-K. Yang, and W.-S. Sin. "MIMO antenna with Wi-Fi and Blue-Tooth for smart watch applications." In *2015 IEEE MTT-S 2015 International Microwave Workshop Series on RF and Wireless Technologies for Biomedical and Healthcare Applications (IMWS-BIO)*, 212-213. Taipei, Taiwan: IEEE, 2015.
[3] Hong, S.-e., and J.-H. Kwon. "Numerical calculation of specific absorption rate for smart-watch with planar inverted F antenna." In *2016 International Conference on Electromagnetics in Advanced Applications (ICEAA)*, 646-647. Cairns, QLD, Australia: IEEE, 2016.
[4] Baloch, B. A., Z. Zahid, and A. A. Khan. "Self-Decoupled Dual Band PIFA for Wi-Fi 6E Smartwatch MIMO Applications." In *2022 19th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, 1036-1040. Islamabad, Pakistan: IEEE, 2022.
[5] Luebbers, R. "XFDTD and beyond—from classroom to corporation." In *2006 IEEE Antennas and Propagation Society International Symposium, 119–122. San Diego, CA, USA*: IEEE, 2006.
[6] Homsup, N., and J. Breakall. "Application of XFDTD and FEKO program to the analysis of planar antennas." In *2010 10th International Symposium on Communications and Information Technologies, 646–650. Tokyo, Japan*: IEEE, 2010.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

ANTENA NA PASMO 2,4 GHz ZINTEGROWANA Z HEŁMEM

2.4 GHZ BAND ANTENNA INTEGRATED WITH A HELMET

Kornelia Bendzel[1]; Łukasz Januszkiewicz[2];

[1] Wydział Elektrotechniki, Elektroniki, Informatyki i Automatyki, Politechnika Łódzka, Łódź, 246953@edu.p.lodz.pl
[2] Instytut Elektroniki, Politechnika Łódzka, Łódź, lukasz.januszkiewicz@p.lodz.pl

DOI: 10.15199/59.2025.4.91

Streszczenie: **W artykule opisano projekt anteny zintegrowanej z hełmem taktycznym, działającej w paśmie 2,4 GHz, przeznaczonej do transmisji obrazu z miniaturowej kamery nasobnej. Opracowano zmodyfikowaną antenę typu odwrócone F, którą można wykonać z elastycznego laminatu i zintegrować z szyną montażową hełmu. Wyniki symulacji komputerowych oraz pomiary prototypu potwierdziły dobre dopasowanie impedancyjne i zadowalający zysk energetyczny anteny.**

Abstract: **The article describes the design of a helmet-integrated antenna operating in the 2.4 GHz band, intended for transmitting images from a miniature wearable camera. A modified inverted-F type antenna was developed, which can be made from flexible laminate and integrated with the helmet's mounting rail. Computer simulation results and prototype measurements confirmed good impedance matching and satisfactory antenna gain.**

Słowa kluczowe: **antena, systemy nasobne, sieci WBAN**

Keywords: **antenna, wearable electronics, WBAN.**

1. WSTĘP

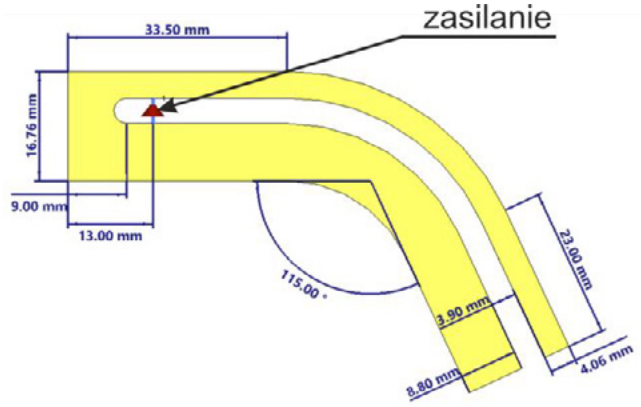
Bezprzewodowa transmisja danych odgrywa kluczową rolę w wielu dziedzinach życia codziennego, od łączności osobistej po systemy monitoringu. W zastosowaniach specjalistycznych, takich jak systemy nasobne, projektowanie anten staje się szczególnie wymagającym zadaniem. Kluczowe wyzwania obejmują dopasowanie projektu do ograniczeń przestrzennych, wynikających z konieczności integracji anteny z elementami ubioru lub wyposażenia, oraz zapewnienie dobrych parametrów elektrycznych w bliskim sąsiedztwie ciała ludzkiego. Bliskość złożonego środowiska, jakim jest ciało, może prowadzić do znacznego pogorszenia dopasowania impedancyjnego i obniżenia zysku energetycznego anteny.

W artykule przedstawiono projekt anteny nasobnej zintegrowanej z hełmem taktycznym, przeznaczonej do bezprzewodowej transmisji obrazu z miniaturowej kamery nasobnej wyposażonej w nadajnik pracujący w paśmie w paśmie 2,4 GHz. Opisano opracowaną przez autorów modyfikację anteny typu odwrócone F, którą można wykonać z elastycznego laminatu, co umożliwia jej integrację z szyną montażową hełmu. Przyjęto, że antena umieszczona zostanie w szynie „GFC Tactical" do hełmów typu Future Assault Shell Technology - FAST. Prezentowane badania obejmują zarówno symulacje komputerowe, jak i pomiary z wykorzystaniem prototypu, które potwierdziły satysfakcjonujące parametry anteny w docelowym środowisku pracy.

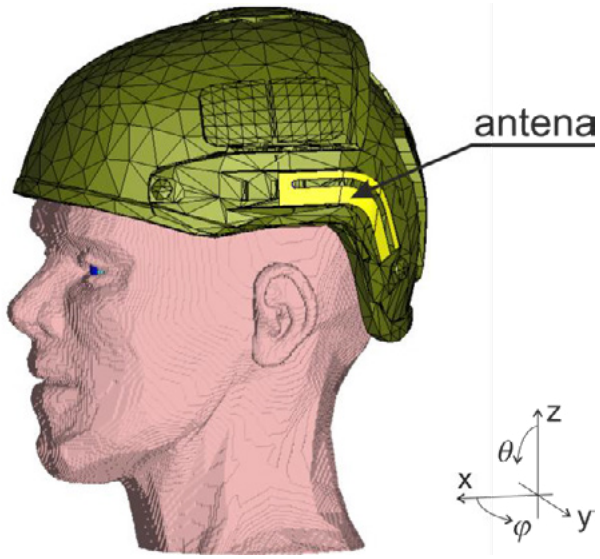
2. PROJEKT ANTENY

Opracowana antena jest zmodyfikowaną anteną typu odwrócone F. Anteny tego typu, dzięki swojej prostej konstrukcji, mogą być stosunkowo łatwo dostosowane do montażu w niewielkich przestrzeniach [1,2] oraz systemach nasobnych [3] i Internetu rzeczy [4]. Są zasilane niesymetrycznie, a ich impedancja wejściowa może zostać dobrze dopasowana do 50 Ω. Ponadto nadają się do wykonania na podłożu dielektrycznym z jednostronną metalizacją co znacząco ułatwia ich wykonanie. Antena tego rodzaju, pracująca w pasmie 150 MHz, była z powodzeniem zintegrowana z kaskiem jednak w tym przypadku cała powierzchnia kasku przewodzącego stanowiła przeciwwagę elektryczną promiennika [5]. W projekcie przedstawionym w tym artykule, również zastosowano antenę typu F, ale ze względu na to, że hełm wykonany jest z materiału dielektrycznego opracowano zmodyfikowaną antenę, która nie wymaga połącznia galwanicznego z podłożem. Ponadto, opracowana antena typu F, posiada kształt i wymiary umożliwiające zintegrowanie jej z szyną montażową hełmu. W przypadku tej anteny możliwe jest zastosowanie cienkowarstwowego laminatu elastycznego, który pozwala na ułożenie anteny na zakrzywionej powierzchni podłoża.

Ze względu na przyjęte założenia projektowe, obrys anteny nie może przekraczać wymiarów szczeliny w szynie montażowej hełmu. Antena jest również ograniczona kształtem tej szczeliny, która jest zagięta pod kątem około 115°. Aby wykorzystać dostępną przestrzeń wykorzystano część fragmentu poziomego szczeliny oraz jej fragment zagięty w dół. Dzięki temu możliwe jest umieszczenie miniaturowej kamery lub innych akcesoriów w przedniej części szyny. Ugięcie promiennika umożliwia antenie sprzęganie się z polaryzacją pionową oraz poziomą, co jest pożądane w przypadku warunków propagacji wielodrogowej we wnętrzach budynków. Wymiary opracowanej anteny typu F przedstawione zostały na Rys. 1.



Rys. 1. Wymiary anteny



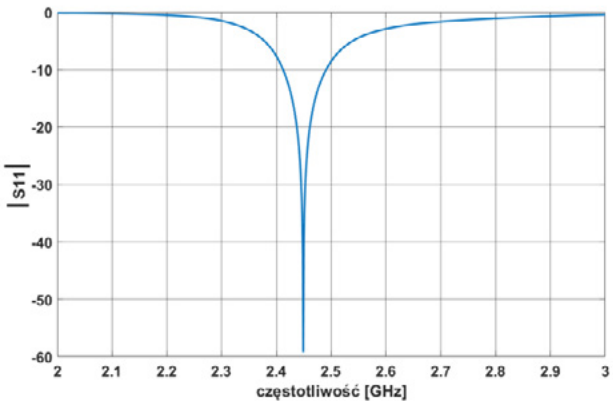
Rys. 2. Model anteny w programie CST Studio Suite

3. WYNIKI SYMULACJI

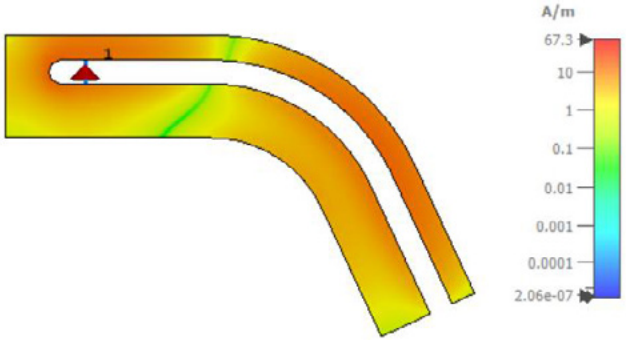
Prototyp anteny został zaprojektowany w programie CST Studio Suite. Opracowany model przedstawiony na Rys. 2 zawiera promiennik anteny, hełm oraz głowę człowieka. Wykorzystano w tym celu heterogeniczny model ciała "Tom" dostępny w bibliotece programu CST Studio Suite 2022, odwzorowujący ciało mężczyzny o wzroście 183 cm i wadze 77 kg. Jest on zbudowany z prostopadłościennych komórek elementarnych (wokseli) o długości krawędzi równej 1 mm i odwzorowuje 33 tkanki. Ze względu na to, że największy wpływ na antenę usytuowaną na hełmie ma głowa, to model ciała ograniczono do tego obszaru, dzięki czemu do przeprowadzenia symulacji wykorzystano jedynie 15 600 000 komórek elementarnych i można było wykorzystać do obliczeń kartę graficzną Nvidia Quadro GV100 dysponującą pamięcią 32 GB. Model hełmu wykorzystany do symulacji zaczerpnięto z dostępnej w Internecie bazy modeli obiektów trójwymiarowych [6]. Ponieważ hełmy taktyczne dostępne na rynku, które nie posiadają ochron balistycznych, wykonywane są z tworzyw sztucznych takich jak ABS to przyjęto, że przenikalność elektryczna względna hełmu wynosi ε_r=3, a stratność dielektryczna tg(δ)=0.015 [7]. Symulacje przeprowadzono w dziedzinie czasu za pomocą metody FIT (Finite Integration Method) [8].

Wyniki symulacji dopasowania impedancyjnego anteny przedstawiono na Rys. 3. Częstotliwość, dla której występuje najlepsze dopasowanie pokrywa się z założoną częstotliwością środkową pasma ISM i wynosi 2,45 GHz. Pasmo impedancyjne anteny dla którego moduł współczynnika S_{11} nie przekracza –10 dB obejmuje zakres od 2,41 GHz do 2,49 GHz. Jest to wystarczający zakres dla systemu Wi-Fi pracującego w paśmie 2,4 GHz.

Na Rys. 4 przedstawiony został rozkład gęstości prądu na powierzchni anteny obliczony dla częstotliwości 2,45 GHz. Widoczne są na nim względnie duże wartości prądu wzdłuż krawędzi wewnętrznej w zagiętej części anteny, dzięki czemu możliwe będzie uzyskanie komponentów polaryzacji pionowej jak i poziomej.

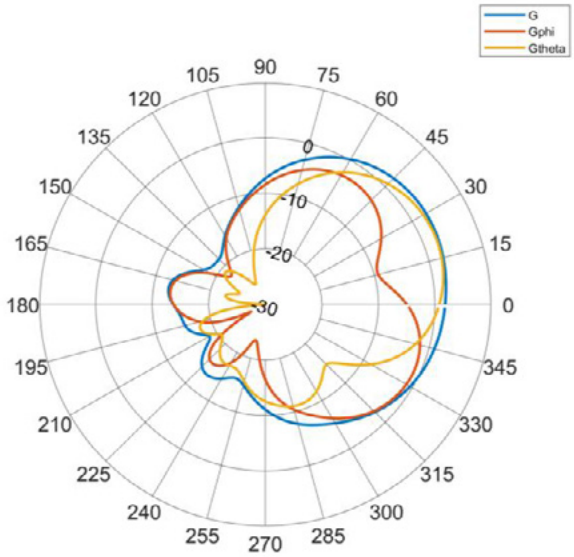


Rys. 3. Dopasowanie impedancyjne anteny - wynik symulacji komputerowej



Rys. 4. Rozkład gęstości prądu na powierzchni anteny dla częstotliwości 2,45 GHz

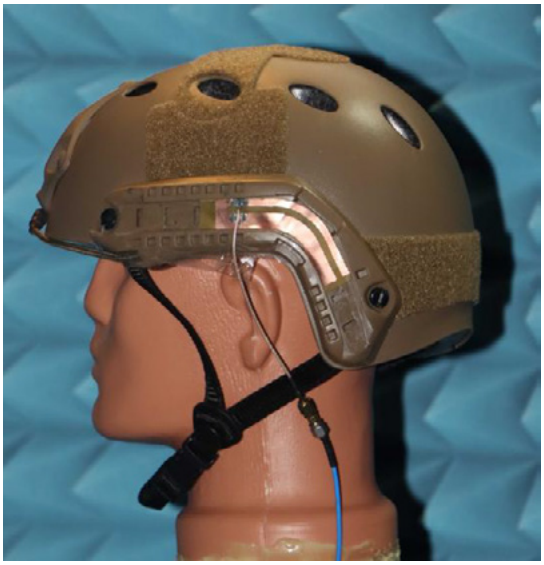
Charakterystyka promieniowania anteny obliczona została w płaszczyźnie poziomej. Na Rys. 5 przedstawiono zysk wypadkowy (G) oraz zysk dla polaryzacji liniowej poziomej (G_φ) oraz pionowej (G_θ). Wartość maksymalna zysku wypadkowego wynosi 2,8 dBi. Wartość maksymalna dla polaryzacji pionowej wynosi 2,4 dBi, a dla polaryzacji poziomej –1 dBi.



Rys. 5. Charakterystyka promieniowania w płaszczyźnie x-y uzyskana za pomocą symulacji komputerowych

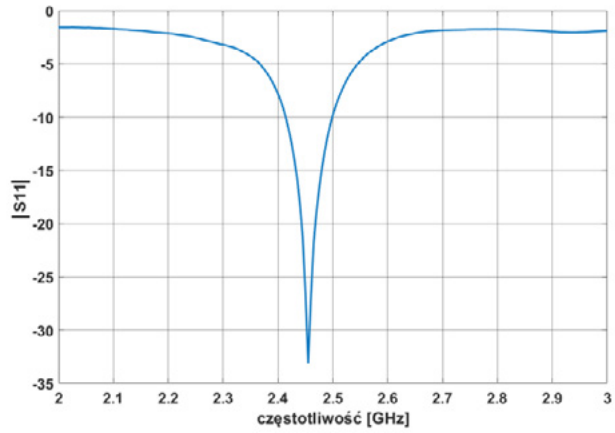
4. WERYFIKACJA POMIAROWA

Prototyp anteny został wykonany na laminacie elastycznym z jednostronną metalizacją typu Pyralux firmy DuPont za pomocą techniki wytrawiania. W celu przeprowadzenia pomiarów antenę podłączono do kabla koncentrycznego. Rys. 6 prezentuje prototyp anteny przymocowanej do hełmu taktycznego umieszczonego na fantomie głowy. Fantom imitujący głowę został wypełniony roztworem imitującym tkanki ciała człowieka [9].



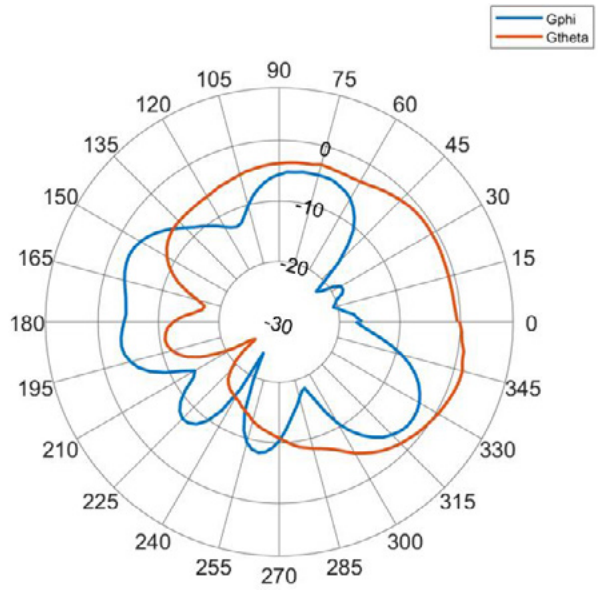
Rys. 6. Prototyp anteny przymocowanej do hełmu umieszczonego na fantomie głowy.

Dopasowanie impedancyjne prototypowej anteny przymocowanej do hełmu i umieszczonej na fantomie zmierzono wykorzystując wektorowy analizator obwodów Rohde & Schwarz ZVB-14. W procesie kalibracji uwzględniono kabel koncentryczny doprowadzający sygnał do anteny. Wyniki pomiarów przedstawione na Rys. 7. są bardzo zbliżone do wyników symulacji. Najlepsze dopasowanie występuje dla częstotliwości 2,45 GHz, a pasmo, dla którego S_{11}<-10 mieści się w przedziale od 2,41 GHz do 2,5 GHz.



Rys. 7. Dopasowanie impedancyjne anteny - wynik pomiarów

Pomiary charakterystyki promieniowania przeprowadzono w warunkach komory quasi-bezodbiciowej w płaszczyźnie poziomej. Wyniki uzyskane dla polaryzacji pionowej i poziomej przedstawiono na Rys. 8.



Rys. 8. Wyniki pomiarów charakterystyki promieniowania prototypowej anteny.

Wartość maksymalna zysku energetycznego anteny została oszacowana metodą anteny wzorcowej. Zastosowano w tym celu antenę Rohde & Schwarz HF907. Wartość maksymalna zysku dla polaryzacji pionowej wyniosła 1,4 dBi, a dla polaryzacji poziomej – 2,6 dBi. Są to wartości zbliżone do wyników symulacji. Największe rozbieżności pomiędzy wynikami symulacji i pomiarów zaobserwowano w przypadku charakterystyki promieniowania dla polaryzacji poziomej. Może to wynikać z wpływu przewodu doprowadzającego sygnał

do anteny na rozkład pola elektromagnetycznego. Ponadto, występowały niewielkie różnice w wymiarach pomiędzy modelem numerycznym i fizycznym zarówno w przypadku hełmu jak i fantomu głowy. Mogą one być przyczyna rozbieżności pomiędzy wynikami pomiarów i symulacji

5. PODSUMOWANIE I WNIOSKI

Przedstawiona w artykule antena została zaprojektowana w celu jej integracji z hełmem taktycznym. Przeprowadzone symulacje komputerowe oraz pomiary charakterystyk anteny prototypowej potwierdziły dobre dopasowanie impedancyjne anteny. Najlepsze dopasowanie uzyskano dla częstotliwości 2,45 GHz, a pasmo pracy (dla S11 < -10 dB) mieści się w zakresie od 2,41 GHz do 2,5 GHz. Pomiary dopasowania anteny prototypowej przymocowanej do hełmu i umieszczonej na fantomie, wykazały dużą zbieżność z wynikami symulacji.

Wyniki symulacji opracowanej anteny świadczą o tym, że posiada maksymalny zysk energetyczny o typowej wartości dla konstrukcji tego typu (2,4 dBi).

Na podstawie pomiarów charakterystyki promieniowania anteny prototypowej oszacowano, że wartość maksymalnego zysku energetycznego anteny dla polaryzacji pionowej wynosi około 1,4 dBi, a dla polaryzacji poziomej -2,6 dBi. Wartości te są mniejsze niż wartości uzyskane na podstawie pomiarów, co może wynikać z dodatkowego tłumienia wprowadzonego przez sondę koncentryczną, którą zasilano antenę prototypową. Rozbieżności pomiędzy wynikami pomiarów i symulacji widoczne w rozkładzie zysku anteny mogą wynikać z różnic pomiędzy modelem numerycznym i fizycznym głowy i hełmu. Ze względu na brak dostępu do modelu 3D zastosowanego w pomiarach hełmu, konieczne było wykorzystanie przykładowego modelu numerycznego, który miał charakter modelu wstępnego. Różnice pomiędzy modelami dotyczą również zastosowanych materiałów.

Opracowana antena może zostać wykorzystana w systemach komunikacji bezprzewodowej zintegrowanych z wyposażeniem osobistym, takich jak hełmy taktyczne. Jej elastyczna konstrukcja i możliwość integracji z istniejącymi elementami hełmu (szyną montażową GFC) czyni ją przydatną do wykorzystania w systemach bezprzewodowego monitorowania uczestników ćwiczeń taktycznych. Zauważone rozbieżności między wynikami symulacji a pomiarów, szczególnie w przypadku charakterystyki promieniowania dla polaryzacji poziomej, wskazują na potencjalne obszary do dalszych badań i udoskonalania modelu numerycznego anteny.

LITERATURA

[1] Stutzman, Warren L., and Gary A. Thiele. 2013. *Antenna Theory and Design*. 3rd ed. River Street, NJ: John Wiley & Sons.

[2] Cheung, Cheuk Yin, Joseph S. M. Yuen, and Steve W. Y. Mung. 2018. "Miniaturized Printed Inverted-F Antenna for Internet of Things: A Design on PCB with a Meandering Line and Shorting Strip." *International Journal of Antennas and Propagation*, Article ID 5172960, 5 pages.

[3] Michishita, Naobumi, Yusuke Saita, Hisashi Morishita, Michitaka Ameya, Masanobu Hirose, and Satoru Kurokawa. 2020. "Helmet-Mounted Inverted-F Antenna at VHF Band." *Journal of Advanced Simulation in Science and Engineering* 7, no. 2: 291-299.

[4] Arya A. K. et al., "Compact BLE Antenna With a Modified PIFA Configuration for Wearable EMG Monitor," *IEEE Sensors Journal*, vol. 23, no. 15, pp. 17648-17655, 1 Aug.1, 2023

[5] Bengloan G., Felicio J. M., Fernandes C. A., Chousseaud A., Froppier B., Cruz E. M., "A Compact and Flexible Dual-Band Antenna for Near-Body Applications," *2021 International Symposium on Antennas and Propagation (ISAP)*, Taipei, Taiwan, 2021, pp. 1-2

[6] https://grabcad.com

[7] Ali Al Takach et al., 2019. "Two-Line Technique for Dielectric Material Characterization with Application in 3D-Printing Filament Electrical Parameters Extraction," *Progress In Electromagnetics Research M* 85: 195-207.

[8] Schuhmann, R., and T. Weiland. 2004. "Recent Advances in Finite Integration Technique for High Frequency Applications." In *Scientific Computing in Electrical Engineering*, edited by W. H. A. Schilders, E. J. W. ter Maten, and S. H. M. J. Houben. Mathematics in Industry 4. Berlin/Heidelberg, Germany: Springer.

[9] Oleksy, P. Januszkiewicz. 2016. “Stanowisko do pomiaru odstrojenia impedancyjnego anten nasobnych” *Przegląd telekomunikacyjny - wiadomości telekomunikacyjne,*, nr 6, str. 479-482.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

MIKROPASKOWE UKŁADY ANTENOWE NA PASMO HIGH-BAND SYSTEMU 5G

MICROSTRIP ANTENNA ARRAYS FOR 5G HIGH-BAND FREQUENCY

Rafał Przesmycki[1]; Marek Bugaj[2] Klaudia Januszczak[3];

[1] Wojskowa Akademia Techniczna, ul. Gen. Sylwestra Kaliskiego 2, 00-908 Warszawa, rafal.przesmycki@wat.edu.pl
[2] Wojskowa Akademia Techniczna, ul. Gen. Sylwestra Kaliskiego 2, 00-908 Warszawa, marek.bugaj@wat.edu.pl [3] Wojskowa Akademia Techniczna, ul. Gen. Sylwestra Kaliskiego 2, 00-908 Warszawa, klaudia.januszczak@student.wat.edu.pl

DOI: 10.15199/59.2025.4.92

Streszczenie: **W artykule przedstawiono mikropaskowe układy antenowe o dużym zysku energetycznym, które wykonane zostały na podłożu dielektrycznym i przeznaczona są do pracy w systemie 5G w paśmie high-band. Opracowane anteny posiadają zwartą konstrukcję o wymiarach nie przekraczających (25,0 x 10,0 x 0,254) mm. Jako podłoże do budowy anteny wykorzystano laminat ROGERS RT5880 o przenikalności elektrycznej 2,2 i grubości h=0,254 mm. W artykule przedstawiono analizę wyników symulacji i pomiarów parametrów elektrycznych i charakterystyk promieniowania zaprojektowanych anten. Opisane w artykule mikropaskowe układy antenowe o dużym zysku energetycznym, pracujące w systemach 5G o częstotliwości środkowej 38 GHz charakteryzują się niskim współczynnikiem odbicia, dużym zyskiem energetycznym i pasmem pracy na poziomie 3,55%.**

Abstract: **The article presents microstrip antenna systems with high energy gain, which were made on a dielectric substrate and are designed to operate in the 5G system in the high-band band. The developed antennas have a compact structure with dimensions not exceeding (25.0 x 10.0 x 0.254) mm. The substrate for the antenna construction was used ROGERS RT5880 laminate with dielectric permittivity of 2.2 and thickness h=0.254 mm. The article presents an analysis of the simulation and measurement results of electrical parameters and radiation characteristics of the designed antennas. The microstrip antenna systems with high energy gain described in the article, operating in 5G systems with a center frequency of 38 GHz, are characterized by a low reflection coefficient, high energy gain and an operating band of 3.55%.**

Słowa kluczowe: **antena, antena mikropaskowa, układy antenowe, duży zysk energetyczny, system 5G.**

Keywords: **antenna, microstrip antenna, antenna arrays, high gain, 5G wireless system.**

1. WSTĘP

System piątej generacji (5G) stanowi kontynuację rozwoju lądowych systemów radiokomunikacyjnych ruchomych, powszechnie znanych jako systemy telefonii komórkowej, które funkcjonują globalnie od wielu lat. Infrastruktura sieci 5G oparta jest na połączeniach światłowodowych, które łączą poszczególne elementy systemu, tworząc zintegrowaną i ultraszybką sieć telekomunikacyjną nowej generacji. Standard 5G zakłada znaczący wzrost przepustowości – nawet do 50–100 razy w porównaniu z technologią 4G – umożliwiając rozwój zaawansowanych usług, takich jak Internet rzeczy (IoT), telemedycyna, pojazdy autonomiczne czy inteligentne miasta [1]. Efektywne wdrożenie systemu 5G wymaga zapewnienia odpowiednich zasobów widma częstotliwościowego oraz implementacji tzw. „gęstej sieci" (dense network), charakteryzującej się rozmieszczeniem dużej liczby stacji bazowych o niewielkich zasięgach. Rozwiązanie to pozwala na zwiększenie efektywności widmowej oraz przepustowości dostępnej dla użytkownika końcowego. W odróżnieniu od sieci wcześniejszych generacji, architektura 5G oparta jest na licznych komórkach małej mocy, rezygnując z tradycyjnych makrokomórek o dużym zasięgu [2].

Technologia sieci 5G umożliwia wykorzystanie częstotliwości z różnych pasm. Zgodnie z aktualnym stanem normalizacji sieci 5G oraz dostępnością częstotliwości dedykowanych do jej funkcjonowania, przewiduje się wykorzystanie trzech pasm częstotliwości. Pasma te są umownie określane jako: low-band 700 MHz (694 – 790 MHz); mid-band 3,6 GHz (3,4 – 3,8 GHz) oraz high-band 26 GHz (24,25 – 27,5 GHz) i 38 GHz (37 GHz – 40 GHz). Maksymalne zasoby częstotliwości w tych pasmach wynoszą odpowiednio: 2 × 30 MHz (część pasma), 400 MHz i 3250 MHz. Należy podkreślić, że podlegają one harmonizacji na terenie Europy, a także w wielu krajach na świecie, co oznacza, że użytkownicy terminali ruchomych będą mogli mieć dostęp do usług sieci 5G na terenie różnych krajów. Sieci 5G będą wykorzystywać wiele pasm jednocześnie, również włączając pasma, które aktualnie wykorzystywane są przez systemy 4G (m.in. 1800 MHz, 2100 MHz, 2600 MHz) [2].

Częstotliwości pasm 26 GHz (Europa) i 38 GHz (Chiny i USA) obejmują odpowiednio zakres od 24,25 GHz do 27,5 GHz oraz od 37 GHz do 40 GHz. Udostępnienie częstotliwości w tym zakresie wymaga uwzględnienia tego faktu w Krajowej Tablicy Przeznaczeń Częstotliwości. Powinno się umożliwić dystrybucję maksymalnie czterech kanałów po 200 MHz z zakresu 26,5 – 27,3 GHz oraz maksymalnie dwóch kanałów po 400 MHz z zakresu 37,5 – 37,8 GHz. Częstotliwości z tego zakresu umożliwią uzyskanie bardzo dużych przepustowości przy bardzo małych opóźnieniach transmisji. Jednak należy się spodziewać, że zasięgi łączności dla terminali ruchomych w terenie otwartym nie przekroczą 500 m. Zatem planowane jest ich wykorzystanie głównie

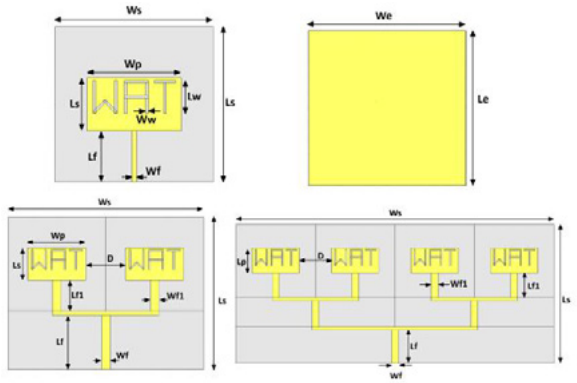
na potrzeby Internetu rzeczy i łączy stałych, określanych jako „radiowy światłowód" [3][4].

W niniejszym artykule przedstawiono propozycję mikropaskowych układów antenowych przeznaczonych do pracy w systemie 5G dla pasma high-band 38 GHz oraz wyniki przeprowadzonych symulacji i pomiarów parametrów elektrycznych tych anten. Oprócz pasma częstotliwości innym ważnym wymaganiem stawianym dla projektowanych anten są wymiary, które powinny być nie większe niż 25 mm x 10 mm, oraz jak największy zysk energetyczny.

2. MIKROPASKOWE UKŁADY ANTENOWE

Ze względu na planowane zakresy częstotliwości pracy systemu 5G, głównym założeniem dla projektowanych anten przeznaczonych do pracy w tym systemie, jest częstotliwość pracy w zakresie 38 GHz (high-band). Oprócz pasma częstotliwości innym ważnym wymaganiem stawianym projektowanym antenom są ich wymiary, które powinny być nie większe niż 25 x 10 mm, jak największy zysk energetyczny, liniowa polaryzacja oraz rodzaj materiału, z którego wykonana ma być antena. Ostatnim założeniem stawianym przed projektowaniem anten było uzyskanie anteny jednoelementowej (1x1) układu dwuelementowego (2x1) oraz układu czteroelementowego (4x1). Do projektów mikropaskowych anten wybrano profesjonalny laminat ROGERS RT5880. Jest to materiał o grubości h = 0,254 mm, przenikalności elektrycznej równej ε_r = 2,2 oraz niskich stratach tan(δ) = 0,0009. Składa się on z 3 warstw, gdzie środkowa część wykonana jest z tlenku glinu (AL203) a warstwy zewnętrzne wypełnione są szkłem i epoksydem. Taki układ warstw zapewnia sporą trwałość, ochronę przed wilgocią, oraz łatwość w obróbce, co czyni go popularnym wyborem do produkcji anten mikropaskowych.

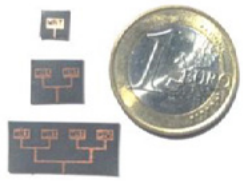
Dla założonych parametrów w oprogramowaniu CST Microwave Studio opracowano jednoelementowy model anteny, którego szczegółowy opis oraz wyniki symulacji i pomiarów przedstawiono w 2024 roku w publikacji [5]. W oprogramowaniu CST Microwave Studio można wyznaczyć parametry elektryczne i charakterystyki promieniowania anteny. Przy projektowaniu anten na początku pod uwagę brano głównie wartość współczynnika S_{11} oraz wartość zysku energetycznego G. W oprogramowaniu CST Microwave Studio dokonano również ostatecznego procesu optymalizacji uzyskanych modeli anten pod kątem stawianych im wymagań modyfikując jej wymiary oraz sposób zasilania. W procesie optymalizacji zmieniane były głównie długość i szerokość całkowita anten, długość i szerokość linii zasilających, odległość pomiędzy promiennikami oraz wymiary dzielników mocy w przypadku układów antenowych. Pozostałe elementy struktury anteny jednoelementowej, takie jak wymiary promiennika, grubość i przenikalność elektryczna podłoża, wielkość liter „WAT" zostały bez zmian. W wyniku przeprowadzonego procesu optymalizacji ostatecznie otrzymano modele anten przedstawione na rys. 1, których wymiary przedstawiono w Tab. 1. Dla tak uzyskanych modeli antenowych wykonano modele fizyczne anteny, których wygląd pokazano na rys. 2. Dla wykonanych fizycznych modeli proponowanych anten przeprowadzono proces symulacji z wykorzystaniem oprogramowania CST Microwave Studio oraz wykonano pomiary z wykorzystaniem analizatora sieci ZVA67 firmy Rohde&Schwarz wybranych parametrów elektrycznych. Wygląd zaprojektowanych anten podczas pomiarów przedstawiono na rys. 3.



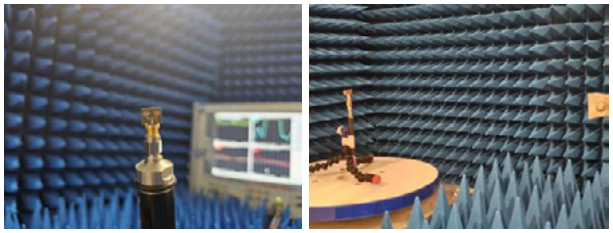
Rys. 1. Projekt mikropaskowych układów antenowych wraz z oznaczeniami wymiarów

Tab. 1. Wymiary projektów anten mikropaskowych

Nazwa komponentu anteny	Symbol	Wymiar (mm)		
		Antena 1x1	Antena 2x1	Antena 4x1
Szerokość podłoża	W_s	5,16	10,32	20,64
Długość podłoża	L_s	5,05	8,20	9,60
Szerokość ekranu	W_e	5,16	10,32	20,64
Długość ekranu	L_e	5,05	8,20	9,60
Szerokość promiennika	W_p	3,10	3,10	3,10
Długość promiennika	L_p	2,06	2,06	2,06
Grubość podłoża	h	0,254	0,254	0,254
Przenikalność podłoża	ε_r	2,20	2,20	2,20
Szerokość linii zasilającej	W_f	0,22	0,44	0,44
Długość linii zasilającej	L_f	2,93	2,93	2,31
Wysokość liter WAT	L_w	1,00	1,00	1,00
Grubość liter WAT	W_w	0,10	0,10	0,10
Odległość między promiennikami	D	-	2,10	2,10
Długość linii w dzielnikach mocy	L_{fl}	-	1,66	1,66
Szerokość linii w dzielnikach mocy	W_{fl}	-	0,44	0,44



Rys. 2. Wygląd modeli fizycznych proponowanych anten mikropaskowych



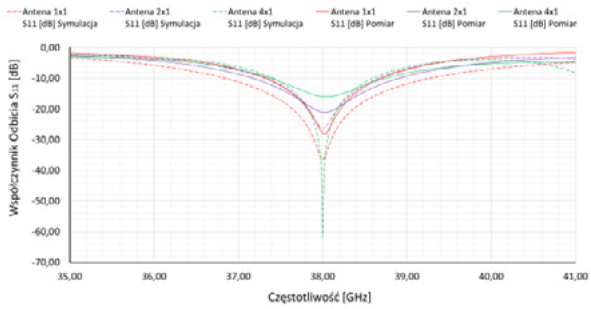
Rys. 3. Wygląd modeli fizycznych proponowanych anten mikropaskowych podczas realizacji pomiarów: a)współczynnika odbicia S_{11}, b) charakterystyk promieniowania

3. WYNIKI POMIARÓW DLA MIKROPASKOWYCH UKŁADÓW ANTENOWYCH

Zaprojektowane konstrukcje anten o wysokim zysku energetycznym składają się z trzech komponentów: ekranu, promiennika i podłoża, w przypadku układów 2x1 i 4x1 również z dzielników mocy. Dla tak zaprojektowanych anten przeprowadzono proces symulacji z wykorzystaniem oprogramowania CST Microwave Studio oraz wykonano pomiary z wykorzystaniem analizatora sieci ZVA67 firmy Rohde&Schwarz. W rezultacie czego otrzymano wyniki wartości parametrów elektrycznych takich jak: współczynnik odbicia, współczynnik fali stojącej, impedancja wejściowa, zysk energetyczny oraz charakterystyki promieniowania. Wszystkie wyżej wymienione parametry porównano ze sobą w celu analizy właściwości wykonanych mikropaskowych układów antenowych.

3.1. Współczynnik odbicia S_{11}

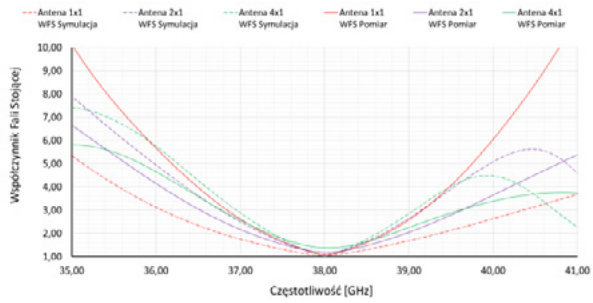
Pierwszym współczynnikiem, który pozwala stwierdzić jaki jest zakres pracy proponowanej anteny jest współczynnik odbicia S_{11}. Na rys. 4 przedstawiono jego wyniki symulacji oraz wyniki pomiarów dla proponowanych mikropaskowych układów antenowych „WAT". Proponowane anteny posiadają rezonans na częstotliwości 38 GHz ze stratami odbiciowymi wynoszącymi dla anteny 1x1, anteny 2x1, anteny 4x1 odpowiednio -28,81 dB, -21,07 dB, -16,01 dB. Proponowane anteny mają szerokość pasma pracy wynoszącą odpowiednio 1,35 GHz, 1,74 GHz, 1,37 GHz co daje szerokość względną pasma pracy na poziomie 3,55%. Na podstawie tego możemy powiedzieć, że są to anteny rezonansowe stosowana tylko w wybranym paśmie systemu 5G.



Rys. 4. Wartość współczynnika odbicia S_{11} w funkcji częstotliwości dla finalnych modeli proponowanych anten

3.2. Współczynnik fali stojącej WFS

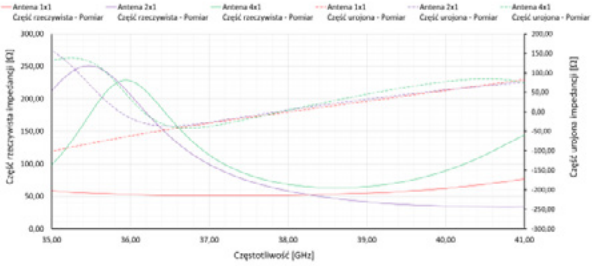
Kolejnym parametrem, który określa właściwości impedancyjne anten jest współczynnik fali stojącej (WFS). W przypadku anten mikropaskowych napięciowy współczynnik fali stojącej nie powinien być większy niż 2 w całej szerokości pasma częstotliwości. Na rys. 5 przedstawiono napięciowy współczynnik fali stojącej w funkcji częstotliwości. Jak można zaobserwować wartość WFS uzyskana przy częstotliwości rezonansowej 38 GHz wynosi 1,08, 1,19, 1,38 odpowiednio dla anteny 1x1, anteny 2x1, anteny 4x1. Wartość WFS wynosząca 2 dla wszystkich anten wyznaczona została przy częstotliwościach około 37,3 GHz i 38,7 GHz. Z powyższych wartości wynika, że proponowane anteny pracują w zakładanym wysokim paśmie częstotliwości 5G tj. 38 GHz, w którym wymagane szerokości kanału wynoszą 50 MHz, 100 MHz, 200 MHz, 400 MHz. Pasma pracy proponowanych anten pokrywają wszystkie szerokości kanałów stosowanych w paśmie 38 GHz, co potwierdza możliwość ich zastosowania w systemie 5G.



Rys. 5. Wartość współczynnika fali stojącej dla finalnych modeli proponowanych anten

3.3. Impedancja wejściowa

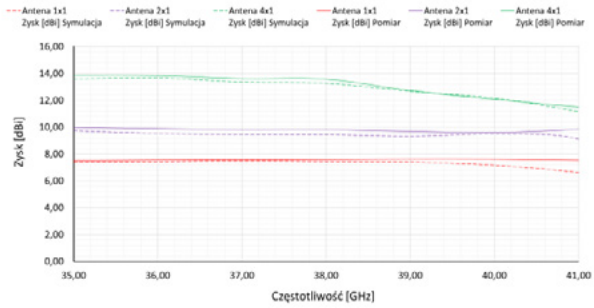
Impedancja wejściowa anteny składa się z części rzeczywistej i urojonej, których wartości zmieniają się wraz z częstotliwością. W projekcie anten przyjęto, że impedancja linii zasilającej powinna wynosić 50 Ω. Należy tą wartość uwzględnić podczas projektowania anten, ponieważ jest to łatwy sposób na pogorszenie właściwości takich anten. Zaprojektowane szerokości i długości linii zasilających w antenach pozwoliły we wszystkich trzech przypadkach na uzyskanie impedancji wejściowej o wartości zbliżonej do wartości zakładanej tj. $Z = 50{,}00 - j0{,}00\ \Omega$ przy częstotliwości rezonansowej 38,0 GHz. Szczegółowe wartości impedancji wejściowej dla proponowanej anteny w funkcji częstotliwości przedstawiono na rys. 6.



Rys. 6. Wartość impedancji wejściowej dla finalnych modeli proponowanych anten

3.4. Zysk energetyczny

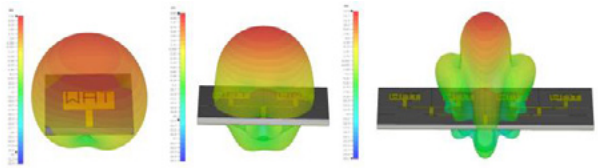
Jednym z ważniejszych parametrów branych pod uwagę podczas projektowania proponowanych mikropaskowych układów antenowych jest zysk energetyczny, który powinien być jak największy. Zysk energetyczny anteny podawany jest w odniesieniu do anteny izotropowej i wyrażany jest w jednostkach dBi. Proponowane anteny posiadają zysk równy 7,57 dBi, 9,82 dBi, 13,57 dBi (odpowiednio dla anteny 1x1, anteny 2x1, anteny 4x1) przy częstotliwości rezonansowej 38 GHz. Uzyskane wartości są uważane za bardzo dobre dla takiego rodzaju anten mikropaskowych, co jest dużym osiągnięciem dla takich konstrukcji anteny. Wartość ta została uzyskana dzięki odpowiedniemu kształtowi promiennika, w którym zastosowano wycięcie w postaci napisu „WAT". Przebieg wartości zysku energetycznego w funkcji częstotliwości dla zaprojektowanych anten „WAT" przedstawiono na rys. 7.



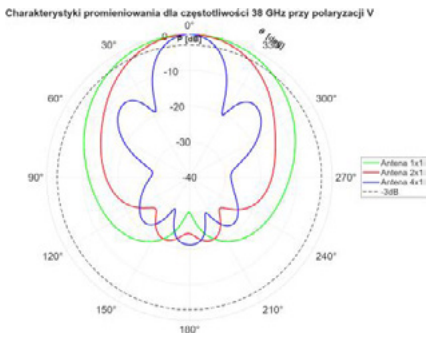
Rys. 7. Wartość zysku energetycznego dla finalnego modelu proponowanej anteny

3.5. Charakterystyki promieniowania

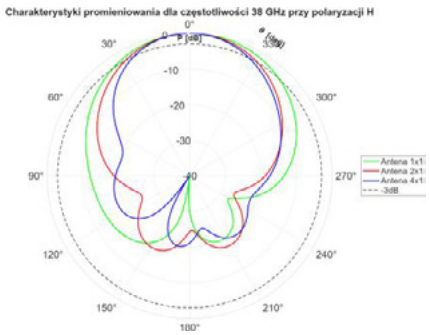
Charakterystyka promieniowania obrazuje, w jaki sposób antena promieniuje energię w zależności od kierunku. Przedstawia ona unormowany rozkład pola elektrycznego lub też względny rozkład powierzchniowej gęstości mocy. Charakterystyki są wyznaczane w dwóch płaszczyznach, poziomej i pionowej i mogą być także przedstawiane w postaci trójwymiarowej. Na rys. 8, rys. 9 oraz rys. 10 przedstawiono wygląd charakterystyk promieniowania proponowanych anten „WAT" dla częstotliwości 38 GHz uzyskany podczas przeprowadzonych symulacji w CST Microwave Studio i pomiarów.



Rys. 8. Charakterystyka promieniowania 3D anten „WAT" dla częstotliwości 38 GHz. Od lewej antena 1x1, antena 2x1, antena 4x1



Rys. 9. Unormowane pomierzone charakterystyki promieniowania anten „WAT" dla częstotliwości 38 GHz w polaryzacji pionowej V



Rys. 10. Unormowane pomierzone charakterystyki promieniowania anten „WAT" dla częstotliwości 38 GHz w polaryzacji poziomej H

4. WNIOSKI

W ramach artykułu przedstawiono projekt mikropaskowych anten „WAT" przeznaczonych do odbioru sygnałów w technologii 5G w paśmie częstotliwości „high-band". Anteny zostały zaprojektowane specjalnie do odbioru sygnałów 5G, które dają możliwość transmisji bardzo dużej ilości danych w krótkim czasie. Opracowane anteny to bardzo kompaktowe urządzenia, które umożliwiają odbiór sygnałów 5G w sposób bardzo efektywny i niezawodny. Dzięki swoim konstrukcjom anteny mogą być z łatwością zintegrowane z różnymi urządzeniami mobilnymi, takimi jak telefony, tablety czy laptopy. Wyniki przeprowadzonych symulacji i pomiarów dla proponowanych anten są bardzo obiecujące i dają nadzieję, że zaprojektowane anteny mogą służyć do komunikacji w gęsto zaludnionych miastach, gdzie znajduje się wiele przeszkód terenowych, które w przypadku sieci 4G utrudniają przesyłanie informacji.

Opisane w artykule anteny „WAT" wykonane została na podłożu dielektrycznym i przeznaczone są do pracy w systemie 5G w wysokim paśmie 38 GHz. Opracowane anteny posiadają zwartą konstrukcję o wymiarach nie przekraczających (25 x 10 x 0,254) mm. W artykule przedstawiono analizę wyników symulacji i pomiarów parametrów elektrycznych i charakterystyk promieniowania zaprojektowanych anten. Opisane w artykule anteny mikropaskowe o dużym zysku energetycznym, pracujące w systemach 5G o częstotliwości środkowej 38 GHz charakteryzują się niskim współczynnikiem odbicia, bardzo dużym zyskiem energetycznym i paśmie pracy na poziomie 3,55 %.

Praca została sfinansowana przez Wojskową Akademię Techniczną w ramach projektu nr UGB/22-055/2025/WAT pt. "Badanie odporności wybranych smartfonów na promieniowane pole elektromagnetyczne w aspekcie wymagań normy PN-EN 17025"

LITERATURA

[1] Korzeniowska E., Krawczyk A., Łada-Tondyra E., Plewako J., „Technologia 5G jako etap rozwoju komunikacji bezprzewodowej", *Przegląd Elektrotechniczny* R.95 nr 12/2019, str. 144-147.

[2] J. Senic et al., "Challenges for 5G and Beyond," 2022 16th European Conference on Antennas and Propagation (EuCAP), Madrid, Spain, 2022, pp. 1-5.

[3] J. Senic et al., "Challenges for 5G and Beyond," 2022 16th European Conference on Antennas and Propagation (EuCAP), Madrid, Spain, 2022, pp. 1-5.

[4] C. Hausl, J. Emmert, M. Mielke, B. Mehlhorn and C. Rowell, "Mobile Network Testing of 5G NR FR1 and FR2 Networks: Challenges and Solutions," 2022 16th European Conference on Antennas and Propagation (EuCAP), Madrid, Spain, 2022, pp. 1-5.

[5] Rafał Przesmycki, Marek Bugaj, Leszek Nowosielski, Bartosz Dudziński – Antena mikropaskowa na pasmo high-band systemu 5G, Przegląd Telekomunikacyjny – Wiadomości Telekomunikacyjne, ISSN 1230-3496, NR 4/2024, str: 260-263, doi: 10.15199/59.2024.4.56

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

SZEROKOPASMOWA ANTENA MIKROPASKOWA DO ZASTOSOWAŃ W SYSTEMACH 5G mmWAVE I 6G

BROADBAND MICROSTRIP ANTENNA FOR 5G mmWAVE AND 6G APPLICATIONS

Rafał Przesmycki[1]; Marek Bugaj[2]; Klaudia Januszczak[3]

[1] Wojskowa Akademia Techniczna, ul. Gen. Sylwestra Kaliskiego 2, 00-908 Warszawa, rafal.przesmycki@wat.edu.pl
[2] Wojskowa Akademia Techniczna, ul. Gen. Sylwestra Kaliskiego 2, 00-908 Warszawa, marek.bugaj@wat.edu.pl
[3] Wojskowa Akademia Techniczna, ul. Gen. Sylwestra Kaliskiego 2, 00-908 Warszawa, klaudia.januszczak@student.wat.edu.pl

DOI: 10.15199/59.2025.4.93

Streszczenie: **W artykule przedstawiono szerokopasmową antenę mikropaskową do zastosowań w 5G mmWave i 6G. Opracowana antena posiada zwartą konstrukcję o wymiarach nie przekraczających (5,0 x 5,0 x 0,102) mm. Jako podłoże do budowy anteny wykorzystano laminat ROGERS RO4450F o grubości h = 0,102 mm i przenikalności elektrycznej równej ε_r = 3,7 oraz niskich stratach tan(δ) = 0,004. W artykule przedstawiono analizę wyników symulacji wybranych parametrów elektrycznych i charakterystyk promieniowania zaprojektowanej anteny. Opisana w artykule antena mikropaskowa posiada główne pasmo pracy w zakresie 5G mmWave i dwa dodatkowe rezonanse z zakresu 6G. Charakteryzuje się niskim współczynnikiem odbicia, dużym zyskiem energetycznym (jak dla anten dookólnych) i szerokim pasmem pracy na poziomie 97,63 % (49,05 GHz).**

Abstract: **The paper presents a broadband microstrip antenna for 5G mmWave and 6G applications. The developed antenna has a compact structure with dimensions not exceeding (5.0 x 5.0 x 0.102) mm. The ROGERS RO4450F laminate with thickness h = 0.102 mm and electric permittivity equal to ε_r = 3.7 and low losses tan(δ) = 0.004 was used as a substrate for the antenna construction. The paper presents an analysis of simulation results of selected electrical parameters and radiation characteristics of the designed antenna. The microstrip antenna described in the paper has a main operating band in the 5G mmWave range and two additional resonances in the 6G range. It is characterized by a low reflection coefficient, high energy gain (as for omnidirectional antennas) and a wide operating band of 97.63% (49.05 GHz).**

Słowa kluczowe: **antena, antena mikropaskowa, system 5G, mmWave, 6G, .**

Keywords: **antenna, microstrip antenna, 5G wireless system, mmWave, 6G.**

1. WSTĘP

Wraz z wdrażaniem systemów komunikacji mobilnej piątej generacji (5G), obserwuje się gwałtowny wzrost zapotrzebowania na transmisję danych. Sieci 5G wprowadzają istotne zmiany w zakresie przeciążenia dotychczas doświadczanego w starszych generacjach systemów komunikacyjnych. Jednym z kluczowych osiągnięć tej technologii jest znacząca poprawa szybkości transmisji danych. Widmo 5G obejmuje szeroki zakres częstotliwości radiowych, niezbędnych do realizacji usług technologii piątej generacji. Zakres ten został precyzyjnie podzielony na różne pasma, z których każde posiada unikalne cechy determinujące parametry transmisji, takie jak wysokie prędkości przesyłania danych, niskie opóźnienia oraz duża pojemność sieci [1].

Sieci 5G operują w dwóch podstawowych zakresach częstotliwości: pasmach sub-6 GHz (0,7–6 GHz) oraz pasmach milimetrowych (5G mmWave), przekraczających 24 GHz. Zakres sub-6 GHz obejmuje częstotliwości od 410 MHz do 7,125 GHz i zawiera zarówno pasma wykorzystywane wcześniej w technologiach 3G i 4G, jak i nowe zakresy dedykowane systemom 5G. Charakteryzuje się ono szerokim zasięgiem i dobrą penetracją sygnału, co czyni je szczególnie użytecznym w środowiskach miejskich i podmiejskich, gdzie wymagane są stabilna łączność i pokrycie. Pasma mmWave (24,25–71,0 GHz) oferują znacznie wyższe szybkości transmisji oraz wyjątkowo niskie opóźnienia w porównaniu do pasm sub-6 GHz. Są one szczególnie przydatne w gęsto zaludnionych obszarach miejskich oraz miejscach o dużym zapotrzebowaniu na przepustowość, takich jak stadiony, centra kongresowe i sale koncertowe. Dzięki możliwościom technologicznym pasm mmWave możliwe jest wdrażanie zaawansowanych aplikacji, takich jak rozszerzona rzeczywistość w czasie rzeczywistym, strumieniowanie wideo w jakości ultra-HD oraz rozbudowane sieci IoT [2][3].

W ostatnim czasie standaryzowano zakres pasma mmWave (24–71 GHz) dla technologii 5G New Radio, odpowiadając na potrzeby aplikacji multimedialnych wymagających wysokiej przepustowości [1]. W nadchodzących systemach komórkowych szóstej generacji (6G) planuje się rozszerzenie dostępnych technologii dostępu radiowego (RAT) o pasma subterahercowe (sub-THz), działające w zakresie 100–300 GHz, a szczególnie w przedziale 80–140 GHz [4][5].

Obecnie zarówno technologie mmWave, jak i przyszłe rozwiązania 6G są wdrażane w kontekście komunikacji 5G. Choć pasma mmWave oferują szerokie zasoby widma oraz bardzo wysoką przepustowość, cechują się także istotnymi ograniczeniami, takimi jak krótki zasięg, duża tłumienność sygnału oraz wrażliwość na czynniki środowiskowe. Z kolei pasma 6G charakteryzują się lepszą propagacją i szerszym zasięgiem przestrzennym. W związku z tym niektórzy badacze sugerują, że integracja

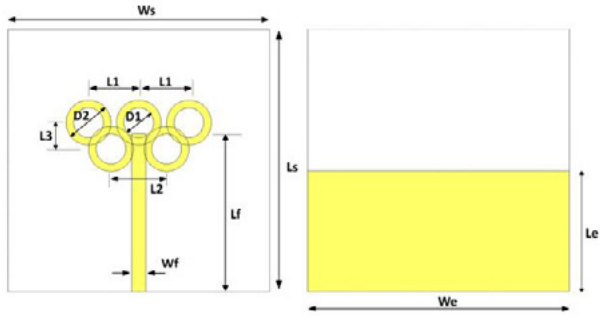
danych z sieci 6G może wspierać efektywną komunikację w systemach mmWave [5].

W niniejszym artykule przedstawiono propozycję szerokopasmowej anteny mikropaskowej przeznaczonej do pracy w systemie 5G mmWave i 6G oraz wyniki przeprowadzonych symulacji wybranych parametrów elektrycznych tej anteny. Oprócz pasma częstotliwości innym ważnym wymaganiem stawianym dla projektowanej anteny są wymiary, które powinny być nie większe niż 5 mm x 5 mm oraz dookólna charakterystyka promieniowania.

2. MIKROPASKOWA ANTENA DO ZASTOSOWAŃ W 5G mmWAVE I 6G

Pasma fal milimetrowych (mmWave) 5G, znane również jako FR2, mieszczą się w zakresie od 24,25 GHz do 71,0 GHz. Sygnały radiowe z pasma o wysokiej częstotliwości oferują znacznie wyższe szybkości transmisji i niższe opóźnienia w porównaniu z sygnałami z pasm sub-6 GHz, co czyni je idealnymi do zastosowań takich jak rozszerzona rzeczywistość w czasie rzeczywistym, strumieniowe przesyłanie wideo w rozdzielczości ultra HD i wdrożenia o dużej przepustowości w gęsto zaludnionych obszarach miejskich. Planowane pasma w przyszłych systemach sieci komórkowych 6G powinny działać w dolnej części pasma terahercowego. Aktualnie można już znaleźć w literaturze dotyczącej nowego systemu 6G, że będzie on wykorzystywany między innymi w zakresie częstotliwości od 80 GHz do 140 GHz a nawet do 300 GHz.

Ze względu na planowane zakresy częstotliwości pracy systemu 5G mmWave i 6G, głównym założeniem dla projektowanych anten przeznaczonych do pracy w tym systemie, jest częstotliwość pracy w zakresie pokrywającym to pasmo. Do projektów mikropaskowych anten wybrano profesjonalny laminat ROGERS RO4450F. Jest to materiał o grubości h = 0,102 mm, przenikalności elektrycznej równej ε_r = 3,7 oraz niskich stratach tan(δ) = 0,004. Składa się on z kilku warstw, gdzie środkowa część wykonana jest z prepregu a warstwy zewnętrzne wypełnione są niskoprzepływowym FR-4. Taki układ warstw zapewnia sporą trwałość, ochronę przed wilgocią, oraz łatwość w obróbce, co czyni go popularnym wyborem do produkcji anten mikropaskowych. Dla założonych wymagań na antenę opracowano jej wstępny model w oprogramowaniu CST Microwave Studio. W oprogramowaniu tym można wyznaczyć parametry elektryczne i charakterystyki promieniowania anteny. W oprogramowaniu CST Microwave Studio dokonano również ostatecznego procesu optymalizacji uzyskanego modelu anteny pod kątem stawianych jej wymagań modyfikując jej wymiary oraz sposób zasilania. W procesie optymalizacji zmieniane były głównie długość i szerokość całkowita anteny, długość i szerokość linii zasilającej, odległość pomiędzy promiennikami oraz wymiary promienników. Pozostałe elementy struktury anteny jednoelementowej, takie jak kształt promiennika, grubość i przenikalność elektryczna podłoża zostały bez zmian. W wyniku przeprowadzonego procesu optymalizacji ostatecznie otrzymano model anteny przedstawiony na rys. 1, którego wymiary przedstawiono w Tab. 1.



Rys. 1. Projekt szerokopasmowej anteny mikropaskowej na pasmo 5G mmWave i 6G

Tab. 1. Wymiary projektu anteny mikropaskowej

Nazwa komponentu anteny	Symbol	Wymiar (mm)
Szerokość podłoża	W_s	4,90
Długość podłoża	L_s	4,90
Szerokość ekranu	W_e	4,90
Długość ekranu	L_e	2,25
Średnica wewnętrzna promiennika	D_1	0,56
Średnica zewnętrzna promiennika	D_2	0,84
Grubość podłoża	h	0,102
Przenikalność podłoża	ε_r	3,70
Szerokość linii zasilającej	W_f	0,26
Długość linii zasilającej	L_f	2,95
Odległość między promiennikami	L_1	0,94
Odległość między promiennikami	L_2	1,04
Odległość między promiennikami	L_3	0,47

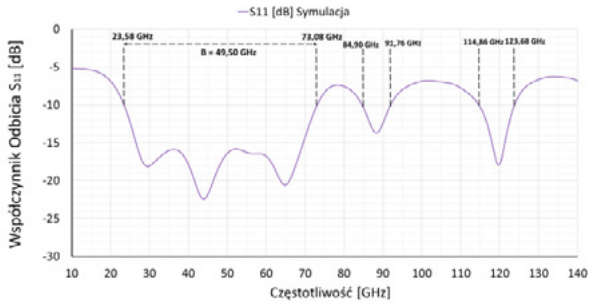
3. WYNIKI SYMULACJI DLA MIKROPASKOWEJ ANTENY DO ZASTOSOWAŃ W 5G mmWAVE I 6G

Zaprojektowana mikropaskowa antena na pasmo 5G mmWave i 6G składa się z trzech komponentów: ekranu, promiennika i podłoża. Dla tak zaprojektowanej anteny przeprowadzono proces symulacji z wykorzystaniem oprogramowania CST Microwave Studio. W rezultacie czego otrzymano wyniki wartości parametrów elektrycznych takich jak: współczynnik odbicia, współczynnik fali stojącej, impedancja wejściowa, zysk energetyczny oraz charakterystyki promieniowania. Wszystkie wyżej wymienione parametry przedstawiono poniżej w celu pokazania właściwości proponowanej anteny.

3.1. Współczynnik odbicia S_{11}

Pierwszym współczynnikiem, który pozwala stwierdzić jaki jest zakres pracy proponowanej anteny jest współczynnik odbicia S_{11}. Na rys. 2 przedstawiono jego wyniki symulacji oraz wyniki pomiarów dla proponowanej mikropaskowej anteny do zastosowań w 5G mmWave i 6G. Proponowana antena charakteryzuje się głównym pasmem pracy od częstotliwośći 23,58 GHz do 73,08 GHz oraz dwoma rezonansami na częstotliwościach 88 GHz i 120 GHz. Wartość strat odbiciowych w głównym paśmie pracy anteny jest na poziomie -18,00 dB, a przy dodatkowych rezonansach odpowiednio -14,00 dB i -17,00 dB. Proponowana antena ma szerokie pasmo pracy wynoszące 49,05 GHz co daje szerokość względną pasma pracy na poziomie 97,63%. Dla dodatkowych rezonansów pasma pracy wokół 88 GHz i 120 GHz wynoszą odpowiednio 6,86 GHz (7,76 %) i

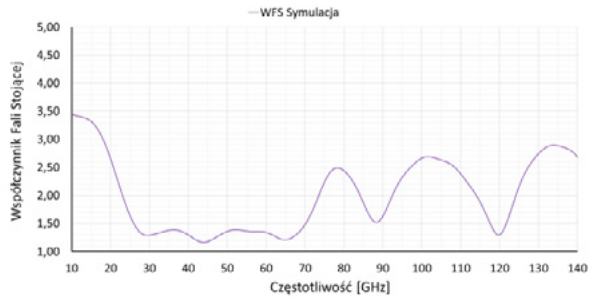
8,82 GHz (7,39 %). Na podstawie tego możemy powiedzieć, że jest to antena szerokopasmowa, którą można zastosować w całym paśmie 5G mmWave oraz w wybranych pasmach systemu 6G tj, 80 GHz i 120 GHz.



Rys. 2. Wartość współczynnika odbicia S_{11} w funkcji częstotliwości dla anteny mikropaskowej na pasmo 5G mmWave i 6G

3.2. Współczynnik fali stojącej WFS

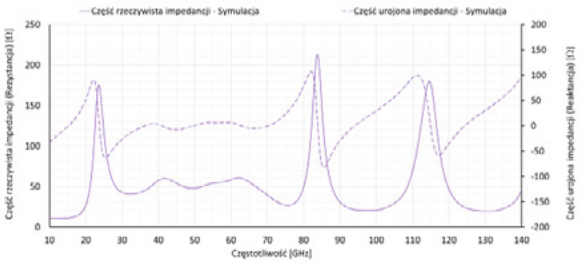
Kolejnym parametrem, który określa właściwości impedancyjne anten jest współczynnik fali stojącej (WFS). Na rys. 3 przedstawiono napięciowy współczynnik fali stojącej w funkcji częstotliwości dla proponowanej anteny. Jak można zaobserwować wartość WFS uzyskana przy częstotliwościach z głównego pasma pracy anteny jest na poziomie 1,38 Z powyższych wartości wynika, że proponowana antena pracuje w zakładanym wysokim paśmie częstotliwości 5G i 6G, w którym wymagane szerokości kanału wynoszą 50 MHz, 100 MHz, 200 MHz, 400 MHz i 800 MHz. Pasma pracy proponowanej anten pokrywają wszystkie szerokości kanałów stosowanych w paśmie 5G mmWave i 6G, co potwierdza możliwość ich zastosowania w tych aplikacjach.



Rys. 3. Wartość WFS dla anteny mikropaskowej na pasmo 5G mmWave i 6G

3.3. Impedancja wejściowa

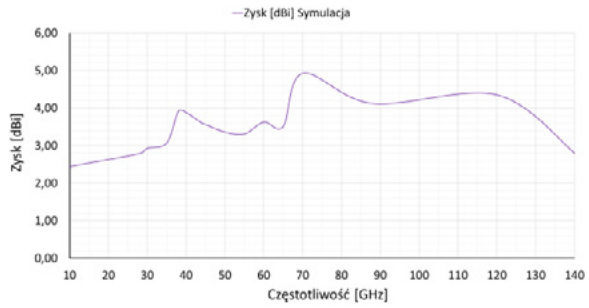
Impedancja wejściowa anteny składa się z części rzeczywistej i urojonej, których wartości zmieniają się wraz z częstotliwością. W projekcie anteny przyjęto, że impedancja linii zasilającej powinna wynosić 50 Ω. Zaprojektowane szerokości i długości linii zasilającej w antenie pozwoliło na uzyskanie impedancji wejściowej o wartości zbliżonej do wartości zakładanej tj. Z = 50,00 – j0,00 Ω w pasmach pracy proponowanej anteny. Szczegółowe wartości impedancji wejściowej dla proponowanej anteny w funkcji częstotliwości przedstawiono na rys. 4.



Rys. 4. Wartość impedancji wejściowej dla anteny mikropaskowej na pasmo 5G mmWave i 6G

3.4. Zysk energetyczny

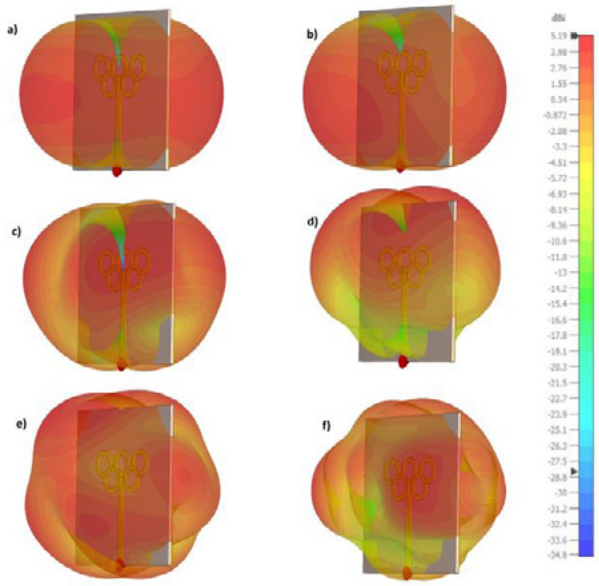
Jednym z istotnych parametrów branych pod uwagę podczas projektowania proponowanej mikropaskowej anteny jest zysk energetyczny. Zysk energetyczny anteny podawany jest w odniesieniu do anteny izotropowej i wyrażany jest w jednostkach dBi. Proponowana antena z założenia powinna mieć charakterystykę dookólną więc wartości zysku energetycznego w paśmie pracy takiej anteny powinny być na poziomie 2,4 dBi – 3,0 dBi. Uzyskane wartości tego parametru dla proponowanej anteny są uważane za dobre. Przebieg wartości zysku energetycznego w funkcji częstotliwości dla proponowanej anteny przedstawiono na rys. 5.



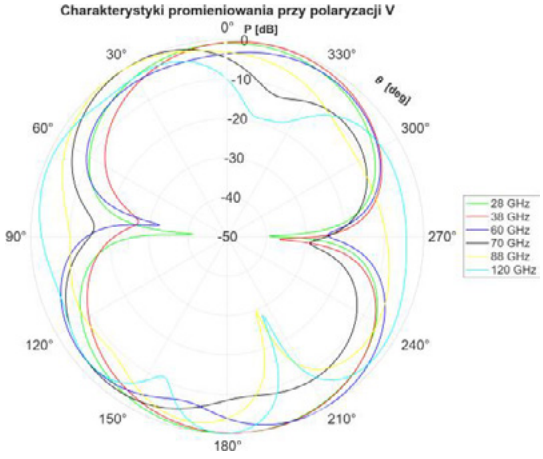
Rys. 5. Wartość zysku energetycznego dla anteny mikropaskowej na pasmo 5G mmWave i 6G

3.5. Charakterystyki promieniowania

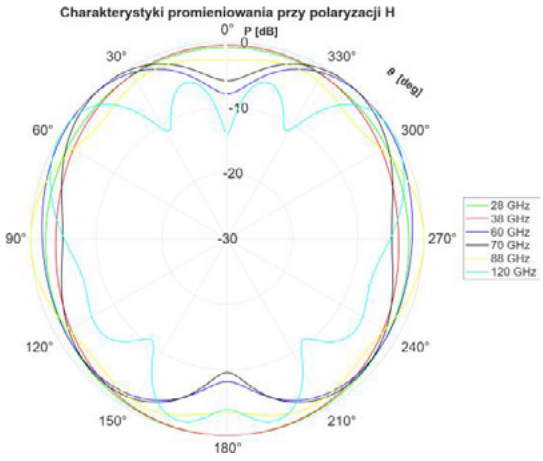
Charakterystyka promieniowania obrazuje, w jaki sposób antena promieniuje energię w zależności od kierunku. Przedstawia ona unormowany rozkład pola elektrycznego lub też względny rozkład powierzchniowej gęstości mocy. Charakterystyki są wyznaczane w dwóch płaszczyznach, poziomej i pionowej i mogą być także przedstawiane w postaci trójwymiarowej. Na rys. 6 przedstawiono charakterystyki promieniowania 3D proponowanej anteny dla częstotliwości z pasma pracy anteny: 28 GHz, 38 GHz, 60 GHz, 70 GHz, 88 GHz, 120 GHz uzyskane podczas przeprowadzonych symulacji w CST Microwave Studio. Na rys. 7 oraz rys. 8 przedstawiono wygląd unormowanych charakterystyk promieniowania szerokopasmowej anteny mikropaskowej na pasmo 5G mmWave i 6G odpowiednio w płaszczyźnie pionowej V i płaszczyźnie poziomej H. Z przedstawionych charakterystyk promieniowania widać, że kształt charakterystyki promieniowania proponowanej anteny jest bardzo powtarzalny co świadczy o dobrych właściwościach anteny w jej paśmie pracy.



Rys. 6. Charakterystyki promieniowania 3D proponowanej anteny dla częstotliwości: a) 28 GHz, b) 38 GHz, c) 60 GHz, d) 70 GHz, e) 88 GHz, f) 120 GHz



Rys. 7. Unormowane charakterystyki promieniowania szerokopasmowej anteny mikropaskowej na pasmo 5G mmWave i 6G w płaszczyźnie pionowej V



Rys. 8. Unormowane charakterystyki promieniowania szerokopasmowej anteny mikropaskowej na pasmo 5G mmWave i 6G w płaszczyźnie poziomej H

4. WNIOSKI

W ramach artykułu przedstawiono projekt szerokopasmowej anteny mikropaskowej na pasmo 5G mmWave i 6G. Antena zostały zaprojektowane specjalnie do odbioru sygnałów w paśmie 5G mmWave i wybranych pasmach 6G, które dają możliwość transmisji bardzo dużej ilości danych w krótkim czasie. Opracowana antena to bardzo kompaktowe urządzenie, które umożliwia odbiór sygnałów w sposób bardzo efektywny i niezawodny. Dzięki swoim konstrukcjom antena może być z łatwością zintegrowana z różnymi urządzeniami mobilnymi, takimi jak telefony, tablety czy laptopy. Wyniki przeprowadzonych symulacji i pomiarów dla proponowanej anteny są bardzo obiecujące i dają nadzieję, że zaprojektowana antena może służyć do komunikacji w gęsto zaludnionych miastach, gdzie znajduje się wiele przeszkód terenowych.

Opisana w artykule szerokopasmowa antena mikropaskowa na pasmo 5G mmWave i 6G wykonana została na podłożu dielektrycznym i przeznaczona jest do pracy w systemie 5G mmWave i wybranych pasmach systemu 6G. Opracowana antena posiada zwartą konstrukcję o wymiarach nie przekraczających (5 x 5 x 0,102) mm. W artykule przedstawiono analizę wyników symulacji parametrów elektrycznych i charakterystyk promieniowania zaprojektowanej anteny. Opisana w artykule antena mikropaskowa charakteryzuje się niskim współczynnikiem odbicia, dużym zyskiem energetycznym (jak dla anten dookólnych) od 2,7 dBi do 4,98 dBi i szerokim pasmem pracy na poziomie 97,63 % (49,05 GHz).

Praca została sfinansowana przez Wojskową Akademię Techniczną w ramach projektu nr UGB/22-055/2025/WAT pt. "Badanie odporności wybranych smartfonów na promieniowane pole elektromagnetyczne w aspekcie wymagań normy PN-EN 17025"

LITERATURA

[1] Ahmed et al., "A Survey on hybrid beamforming techniques in 5G:Architecture and system model perspectives," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3060–3097, 4th Quart., 2018.

[2] Y. Meng *et al.*, "Sub-6 GHz and mm-Wave Dual-Band Aperture-Shared Phone MIMO Antenna for 5G/6G Applications," in *IEEE Antennas and Wireless Propagation Letters*, vol. 23, no. 11, pp. 3639-3643, Nov. 2024.

[3] W. Hong *et al.*, "The Role of Millimeter-Wave Technologies in 5G/6G Wireless Communications," in *IEEE Journal of Microwaves*, vol. 1, no. 1, pp. 101-122, Jan. 2021.

[4] E. Sopin *et al.*, "User Persistence in 5G/6G mmWave/Sub-THz Systems With Blockage: Does It Pay Off?," in *IEEE Transactions on Wireless Communications*, vol. 23, no. 10, pp. 14864-14878, Oct. 2024.

[5] X. Xia *et al.*, "Millimeter-Wave and Sub-6-GHz Aperture-Shared Antenna and Array for Mobile Terminals Accessing 5G/6G-Enabled IoT Scenarios," in *IEEE Internet of Things Journal*, vol. 11, no. 10, pp. 18808-18823, 15 May15, 2024

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

SZESNASTOELEMENTOWY SZYK ANTENOWY Z MODULACJĄ CZASOWĄ I SKANOWANIEM WIĄZKI W PAŚMIE ISM

SIXTEEN-ELEMENT TIME-MODULATED ANTENNA ARRAY WITH BEAM STEERING IN ISM BAND

Stanisław Słowiński, Michał Bandych, Grzegorz Bogdan, Yevhen Yashchyshyn

Politechnika Warszawska, Instytut Radioelektroniki i Technik Multimedialnych
{stanislaw.slowinski.stud, michal.bandych.stud, grzegorz.bogdan, yevhen.yashchyshyn}@pw.edu.pl

DOI: 10.15199/59.2025.4.94

Streszczenie: **W artykule zaprezentowano szyk antenowy z modulacją czasową (szyk AMC) umożliwiający odchylanie listka głównego charakterystyki kierunkowej z krokiem 4,5 stopnia w płaszczyźnie azymutu. Szyk AMC składa się z dwóch ośmioelementowych liniowych szyków anten łatkowych połączonych z siecią zasilającą poprzez przełączniki jednobiegunowe-dwupołożeniowe tworzące pary elementów o przeciwnej polaryzacji. W danej chwili podłącza się okresowo tylko jedną antenę z pary. W rezultacie, sygnał na wyjściu przełącznika jest złożeniem sygnałów z obu anten i ma prążkowe widmo. Odpowiednia zmiana czasów przełączania powoduje zmianę charakterystyk przestrzennych wygenerowanych prążków. W artykule przedstawiono charakterystyki kierunkowe szyku AMC uzyskane za pomocą symulacji elektromagnetycznych oraz pomiarów przeprowadzonych w komorze bezodbiciowej.**

Słowa kluczowe: **anteny, anteny adaptacyjne, anteny z modulacją czasową, szyki antenowe**

Abstract: **The paper presents a time-modulated antenna array (TMAA) capable of beam steering with a step of 4,5 degrees in the azimuth plane. The TMAA consists of two eight-element linear patch antenna subarrays, which are connected to the feed network via single-pole double-throw switches. Each switch groups two inversely polarized elements into a pair. Switching between two elements within a pair is periodic. As a result, the signal at the switch output is a combination of signals received by two oppositely polarized antennas, with a discrete spectrum composed of sidebands. An appropriate change in the switching times causes a change in the spatial characteristics of the generated sidebands. The paper presents radiation patterns obtained from the electromagnetic simulation and measured inside an anechoic chamber.**

Keywords: **adaptive antennas, antennas, antenna arrays, time-modulated antenna arrays**

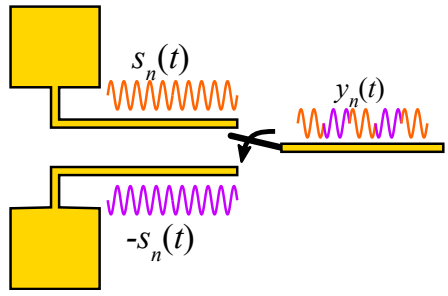
1. WSTĘP

Obecnie stosowane układy adaptacyjnego kształtowania charakterystyki kierunkowej, tj. szyki fazowane oraz układy cyfrowego kształtowania wiązki, charakteryzują się niską sprawnością energetyczną i dużym zużyciem energii. Z tego względu, w ciągu ostatnich dwóch dekad obserwuje się zwiększone zainteresowanie naukowe alternatywnymi sposobami kształtowania charakterystyki kierunkowej. Jednym z nich jest okresowa modulacja sygnałów w sieci zasilającej za pomocą przełączników mikrofalowych [1][4][5]. Szyk antenowy, w którym zaimplementowano tą metodę, nazywany jest szykiem antenowym z modulacją czasową (AMC). Cechą charakterystyczną szyków AMC jest wykorzystanie czasu jako dodatkowego parametru w syntezie pożądanego kształtu charakterystyki kierunkowej. Elementy szyku AMC są okresowo dołączane lub odłączane do/od sieci zasilającej. W rezultacie, sygnał na wyjściu sieci zasilającej (w przypadku odbiorczego szyku AMC) jest złożeniem sygnałów odebranych przez różne elementy w różnych odstępach czasu. Z uwagi na okresowość funkcji przełączających, sygnał po modulacji czasowej ma widmo prążkowe, a każde z prążków wykazuje inne charakterystyki przestrzenne. Regulacja czasów włączenia i wyłączenia modulacji sygnałów pobudzających ma kluczowy wpływ na kształt charakterystyk kierunkowych [1]-[3].

W artykule przedstawiono szesnastoelementowy szyk AMC pracujący na częstotliwości środkowej 2,45 GHz z ośmioma przełącznikami mikrofalowymi, które pozwalają na sterowanie listkiem głównym charakterystyki kierunkowej z krokiem 4,5 stopnia w płaszczyźnie azymutu bez konieczności stosowania przesuwników fazy.

2. OPIS MATEMATYCZNY

Podstawowy element szyku AMC składa się z dwóch przeciwlegle ustawionych anten łatkowych połączonych do sieci zasilającej przez mikrofalowy przełącznik jednobiegunowy-dwupołożeniowy, tak jak to zilustrowano na rys. 1. Załóżmy, że z kierunku prostopadłego do płaszczyzny na szyku AMC, składającego się z N podstawowych elementów, pada fala płaska o częstotliwości f_c. W takim scenariuszu jedna z anten n-tego podstawowego elementu szyku AMC ($n = 0…N\text{-}1$), odbiera sygnał $s_n(t)$, a druga ten sam sygnał, ale w przeciwnej polaryzacji.



Rys. 1. Podstawowy elementu szyku AMC

Przełącznik pracuje okresowo i równomiernie, a więc sygnał na jego wyjściu przez pierwszą połowę okresu ma wartość $s_n(t)$, a przez drugą $-s_n(t)$. Uzyskany przebieg $y_n(t)$ można opisać równaniem (1) jako sygnał $s_n(t)$ zmodulowany bipolarną funkcją prostokątną $m_n(t)$ przyjmującą wartość +1 oraz −1.

$$y_n(t) = s_n(t)m_n(t) \tag{1}$$

Przykładowe przebiegi na wyjściu pierwszych czterech przełączników zilustrowano na rys. 2. Zakładając, że funkcje modulujące $m_n(t)$ są okresowe, można je zapisać w postaci zespolonego szeregu Fouriera:

$$m_n(t) = \sum_{k=-\infty}^{\infty} M_n[k]e^{j2\pi f_0 kt} \tag{2}$$

$$M_n[k] = \frac{1}{T_0}\int_{-T_0/2}^{T_0/2} m_n(t)e^{-j2\pi f_0 kt}\,dt \tag{3}$$

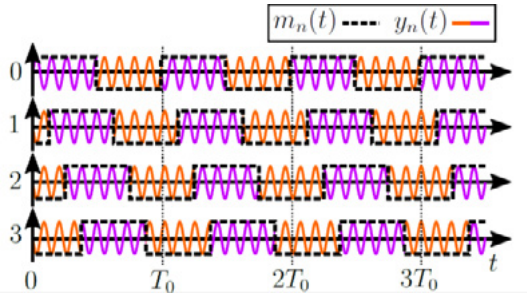
gdzie $M_n[k]$ to k-ty współczynniki n-tego sygnału, T_0 to okres prostokątnego sygnału modulującego, a $f_0 = 1/T_0$ to jego częstotliwość powtarzania. W dalszej części artykułu wartość k określana jest mianem wstęgi bocznej lub prążka. Rys. 3 przedstawia przykładowe widmo sygnału po modulacji czasowej, zawierające wstęgi boczne na częstotliwościach f_c+kf_0, o amplitudach odpowiadających wartościom współczynników Fouriera $M_n[k]$. Dzięki własności szeregu Fouriera o przesunięciu w dziedzinie czasu można dokładnie kontrolować fazę wstęg bocznych opóźniając sygnały modulujące w dziedzinie czasu:

$$m_n(t) = m_0(t - \Delta t_n) \tag{4}$$

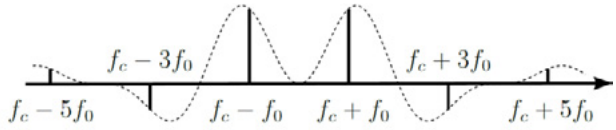
$$M_n[k] = M_0[k]e^{-j2\pi k\Delta t_n/T_0} \tag{5}$$

gdzie Δt_n jest opóźnieniem n-tej funkcji modulującej. W rezultacie, opóźniając funkcje modulujące w dziedzinie czasu, uzyskuje się przesunięcia fazowe na wstęgach bocznych w dziedzinie częstotliwości. Ta zależność pozwala uzyskać skanowanie wiązki analogiczne jak w typowym szyku fazowanym, ale bez konieczności stosowania mikrofalowych przesuwników fazy, które zastępowane są znacznie prostszymi przełącznikami [2].

Zastosowanie szyków szyków AMC w systemach radiokomunikacyjnych wymaga spełnienia dwóch warunków: częstotliwości modulacji powinna być znacznie mniejszej od częstotliwości nośnej ($f_0 \ll f_c$) oraz szerokości pasma odbieranego/nadawanego sygnału powinna być mniejsza od wartość częstotliwości modulacji czasowej [4]. Niespełnienie tych warunków skutkuje odpowiednio niesymetrycznością widma względem f_c oraz nakładaniem się replik widmowych.



Rys. 2. Przykładowe przebiegi na wyjściach pierwszych czterech elementów szyku AMC



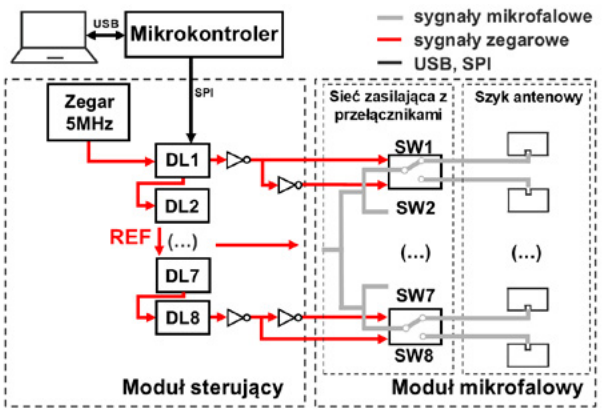
Rys. 3. Widmo amplitudowe sygnału na wyjściu przełącznika

3. PROJEKT SZYKU AMC

W oparciu o przyjętą koncepcję modulacji czasowej opracowano szyk AMC, którego architektura obejmuje dwa podstawowe moduły funkcjonalne zilustrowane na rys. 4: moduł sterujący oraz moduł mikrofalowy. Każdy z modułów został wykonany jako osobny obwód drukowany PCB (ang. *printed circuit board*) na laminacie FR4 o grubości 1 mm. Na obu obwodach drukowanych umieszczono przelotki, które po złożeniu ich ze sobą umożliwiły połączenie linii sterujących pomiędzy przełącznikami (moduł mikrofalowy), a liniami opóźniającymi (moduł sterujący). Zaletą takiego podejścia jest możliwość łatwej wymiany jednego z modułów oraz obniżenie kosztu produkcji w porównaniu do analogicznego obwodu czterowarstwowego.

3.1. Moduł sterujący

Funkcją modułu sterującego jest generacja sygnałów sterujących przełącznikami mikrofalowymi. Zdecydowano, że częstotliwość powtarzania przebiegu prostokątnego będzie wynosiła $f_0 = 5$ MHz, aby częstotliwości wstęg bocznych odpowiadały częstotliwościom środkowym kolejnych kanałów w paśmie ISM (patrz rys. 3). Jako źródło sygnału prostokątnego o częstotliwości 5 MHz i wypełnieniu 50% wybrano miniaturowy układ elektromechaniczny (ang. *microelectromechanical system,* MEMS) Microchip DSC1001 z uwagi na krótki czas narastania ($t_r = 1$ ns) i opadania ($t_f = 0{,}9$ ns) sygnału oraz stabilnością częstotliwościową ±50 ppm. Sygnał z generatora MEMS jest przekazywany na wejście pierwszej z programowalnych cyfrowych linii opóźniających Maxim Integrated DS1023S-500+. W całym module wykorzystano osiem kaskadowo połączonych układów, które pozwalają na opóźnianie sygnału wejściowego wartością z przedziału od 0 do 1275 ns z krokiem 5 ns, możliwe jest więc zastosowanie 255 różnych wartości opóźnienia sygnału. Każda z linii opóźniających ma dwa wyjścia: opóźniane oraz referencyjne.

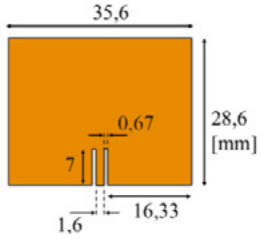


Rys. 4. Schemat blokowy szyku AMC

Połączenie kaskadowe (ang. *daisy-chain*) polega na wykorzystaniu wyjścia referencyjnego, jako źródła sygnału dla kolejnej linii opóźniającej i pozwala na uproszczenie układu dystrybucji sygnałów sterujących. Z uwagi na to, że przełączniki sterowane są parą komplementarnych sygnałów, do generacji odwróconych wersji sygnałów pochodzących z wyjść linii opóźniających zastosowano inwertery logiczne.

3.2. Moduł mikrofalowy

Na moduł mikrofalowy szyku AMC składają się szesnastoelementowy szyk anten łatkowych oraz sieć zasilającą z przełącznikami. Wszystkie anteny łatkowe w szyku mają wymiary 35,6 × 28,6 mm jak to pokazano na rys. 5. Szesnaście anten łatkowych rozmieszczono w odstępach 48 mm (niespełna 0,4 długości fali w wolnej przestrzeni) w konfiguracji dwóch liniowych podszyków, tak jak to przedstawiono na rys. 6. Do przełączania pomiędzy dwoma antenami w parze wybrano przełącznik mikrofalowych Skyworks AS214 z uwagi na zakres częstotliwości pracy od 0,1 do 3 GHz oraz czas opadania (od 90 % do 10 %) i narastania (od 10 % do 90 %) amplitudy sygnału oraz wynoszący, według producenta, 10 ns, czyli znacząco mniej niż zakładany czas pomiędzy przełączeniami $T_0/2 = 100$ ns. Straty wtrąceniowe na częstotliwości 2,45 GHz, również na podstawie danych producenta, wynoszą 0,4 dB, a izolacja pomiędzy wrotami przełączanymi jest na poziomie 25 dB. Przełącznik sterowany jest dwoma komplementarnymi liniami sterującymi (jedna linia w stanie logicznym wysokim, a druga w stanie niskim dla pierwszego stanu i odwrotnie dla drugiego stanu). Praca każdego z przełączników w szyku może być niezależnie opóźniania. W szczególnym przypadku wszystkie przełączniki mogą przełączać się w tym samym czasie lub być opóźniane kolejno (progresywnie) o wielokrotność zdefiniowanej wartości progresji opóźnienia Δt.



Rys. 5 Antena łatkowa z wymiarami (wymiary w mm)

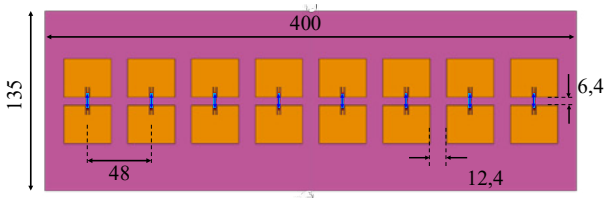
4. SYMULACJE

Symulacje elektromagnetyczne przeprowadzono w pierwszej kolejności w celu odpowiedniego zaprojektowania anteny łatkowej, którą zoptymalizowano pod kątem minimalizacji współczynnika odbicia jako elementu szyku. Optymalizację przeprowadzono na modelu bez sieci zasilającej, w którym dla każdej z szesnastu anten łatkowych zdefiniowano niezależnie źródło pobudzenia. Umożliwiło to zamodelowanie okresowej modulacji za pomocą zespolonych amplitud reprezentujących wstęgi boczne wynikające z modulacji okresowymi przebiegami prostokątnymi. Analizę przeprowadzono dla opóźnień przełączania wzrastających wraz z numerem porządkowym podstawowego elementu szyku AMC zgodnie ze wzorem:

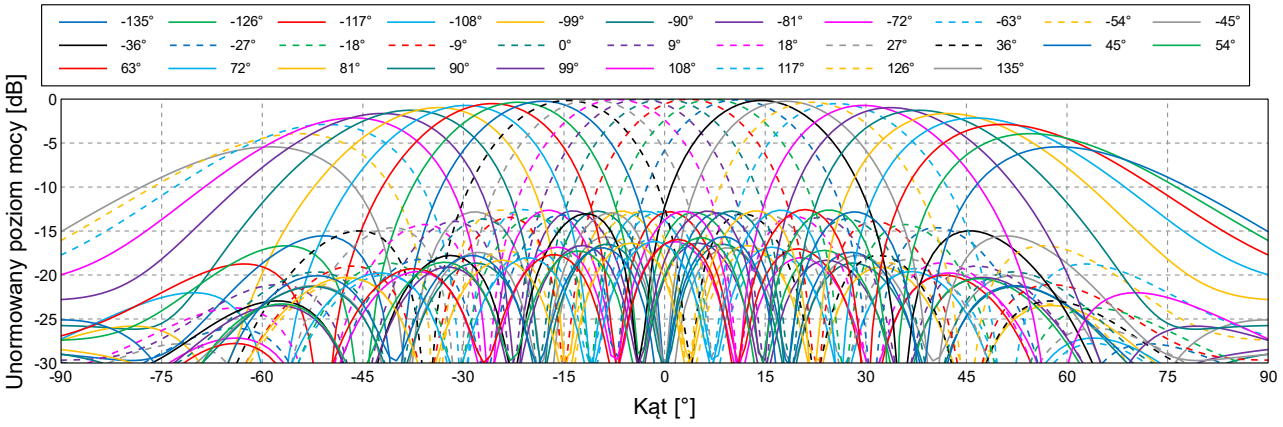
$$\Delta t_n = n\Delta t. \tag{6}$$

Symulacje charakterystyk promieniowania dla progresji fazy od -135° do 135° z krokiem 9° przeprowadzono za pomocą metody momentów w oprogramowaniu Altair FEKO. Wyniki pokazano na rys 7. Na ich podstawie można stwierdzić, że skanowanie jest możliwe z krokiem 4,5° w zakresie od -50° do 50° (dla tych odchyleń kierunkowość spada o 3 dB względem maksimum). Dla odchyleń wiązki o 0°, 30° i 59° trzydecybelowa szerokość listka głównego wynosi odpowiednio 16,2°, 18,73° oraz 25,58°. W praktyce progresja fazy jest uzyskiwana pośrednio, poprzez opóźnianie przebiegów modulujących. W przypadku gdy wszystkie przełączniki pracują równocześnie, generowana jest wiązka w kierunku 0°. Wprowadzenie opóźnienia przełączania sąsiednich elementów o Δt wprowadza progresję fazy na k-tej wstędze bocznej wyrażoną wzorem:

$$\beta = 2\pi k\Delta t/T_0. \tag{7}$$



Rys. 6. Model symulacyjny szyku AMC (wymiary w mm)



Rys. 7.Wyniki symulacji odchylenia listka głównego charakterystyki kierunkowej szyku antenowego dla wybranych wartości progresji fazy

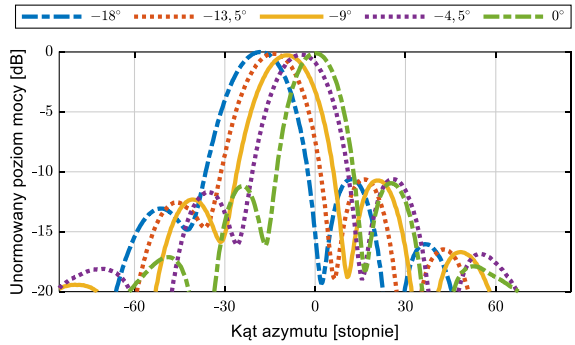
Na przykład, dla $\Delta t = 5$ ns progresja fazy na pierwszej wstędze bocznej wynosi 9°, co powoduje odchylenie wiązki o -4,5° względem kierunku głównego.

5. POMIARY

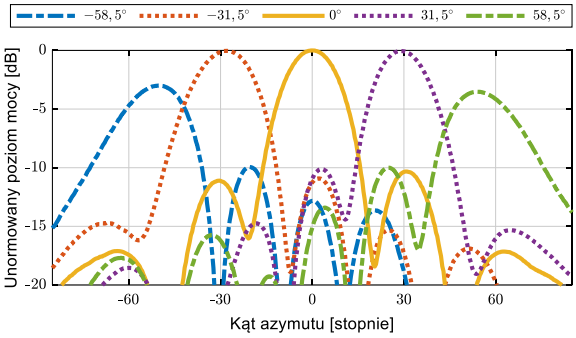
Pomiary charakterystyk kierunkowych szyku AMC przeprowadzono w komorze bezodbiciowej wykorzystując generator sygnałów SMBV100A, analizator widma Rohde&Schwartz FSP30 oraz antenę referencyjną RFspin RG18. Szyk AMC, połączony z wejściem analizatora widma, umieszczono na rotorze i zmierzono w trybie odbiorczym, natomiast generator sygnałów wraz z anteną referencyjną wykorzystano jako część nadawczą systemu pomiarowego, nastawioną do emisji sygnału sinusoidalnego o częstotliwości 2,45 GHz. Na ekranie analizatora widma zaobserwowano, że widmo sygnału na wyjściu szyku AMC ma widmo prążkowe, podobne do tego ilustrowanego na rys. 3, aczkolwiek z niewielkimi poziomami wstęg bocznych na częstotliwości środkowej oraz parzystych wstęgach bocznych. W celu zmierzenia charakterystyki kierunkowej pierwszego prążka (pierwszej wstęgi bocznej) ustawiono na analizatorze: częstotliwość środkową 2,455 GHz, parametr span na zero, a długość czasu przemiatania na 20 s, co odpowiada obrotowi rotora o 180 stopni. W celu dokonania pomiaru amplitudy tylko jednej ze wstęg bocznych należało ustawić szerokość filtra rozdzielczego analizatora (RBW) mniejszą od 5 MHz, a jednocześnie odpowiednio dużą, aby skrócić czas akwizycji, dlatego wybrano wartość 1 MHz. Szyk AMC w pozycji początkowej ustawiono pod kątem -90° względem anteny referencyjnej. Pomiar uruchamiano wyzwalając w tym samym momencie funkcję pomiaru na analizatorze widma oraz obrót rotora. Przebieg narysowany na ekranie analizatora w ciągu 20 sekund odpowiadał zmianom amplitudy sygnału nadawanego przez szyk AMC w trakcie obrotu od -90° do 90°. Proces ręcznego wyzwalania ruchu rotora wprowadza niepewność pomiaru kąta, którą oszacowano na ± 2° i uznano za dopuszczalną. Na rys. 8 przedstawiono wyniki pomiarów dla skanowania od -18° do 0° z najmniejszym możliwym krokiem 4,5°. Rys. 9 przestawia skrajne wartości odchylenia wiązki od -58,5° do 57,5°. Wyniki pomiarów są zgodne z wynikami symulacji pokazanymi na rys. 7.

6. WNIOSKI

W artykule przedstawiono szesnastoelementowy szyk antenowy z modulacją czasową w paśmie ISM na częstotliwości 2.45 GHz. Funkcję sterowania wiązką zrealizowano za pomocą techniki modulacji czasowej. Przełączanie współbieżne (bez opóźniania) generuje charakterystykę o listku głównym skierowanym w kierunku 0°. Odchylanie listka głównego charakterystyki kierunkowej zaobserwowano na częstotliwości pierwszej wstęgi bocznej. Wartość kąta odchylenia jest niemalże proporcjonalna do wprowadzonej progresji opóźnień funkcji modulujących. Uzyskana rozdzielczość skanowania wyniosła 4,5°. W ramach dalszych prac planowane jest zbadane możliwość zwiększenia rozdzielczości skanowania poprzez zmniejszenie kroku opóźniania funkcji modulujących, a także przeprowadzenie prób zastosowania wykonanego szyku AMC w urządzeniach bezprzewodowych.



Rys. 8. Wyniki pomiarów charakterystyk kierunkowych dla skanowania wiązką od -18° do 0°



Rys. 9. Wyniki pomiarów charakterystyk kierunkowych dla wybranych wartości odchyleń kątowych

PODZIĘKOWANIA

Praca została opracowana w ramach konkursu “2025 IEEE AP-S Student Design Contest” i została częściowo sfinansowana przez IEEE Antennas and Propagation Society.

LITERATURA

[1] Ball Edward A., Sumin David Joseph. 2025. „A mmWave Transmitting Time Modulated Array Using Bespoke GaAs Integrated Circuits—Prototype Design and Laboratory Trials at 73 GHz". *IEEE Open Journal of Antennas and Propagation* 6 (1): 144-162.

[2] Bogdan Grzegorz, Konrad Godziszewski i Yevhen Yashchyshyn. 2020. „Time-Modulated Antenna Array for Real-Time Adaptation in Wideband Wireless Systems—Part II: Adaptation Study”. *IEEE Transactions on Antennas and Propagation* 68 (10): 6973–6981.

[3] Bogdan Grzegorz, Yevhen Yashchyshyn, Roberto Maneiro-Catoira i Luis Castedo. 2023. „Time-Modulated Antenna Array with Beam Steering Based on Single-Pole Double-Throw Switches and Four-State Phase Shifters”. *XXXVth General Assembly and Scientific Symposium of the International Union of Radio Science (URSI GASS)*, 1–4. Sapporo, Japonia.

[4] Maneiro-Catoira Roberto, Julio Brégains, José A. García-Naya i Luis Castedo. 2017. „Time modulated arrays: From their origin to their utilization in wireless communication systems”. *Sensors* 17 (3): 590.

[5] Varma D. Suneel, Gopi Ram i G. Arun Kumar. 2022. „Time-Modulated Arrays: A Review”. *IETE Technical Review* 40 (1): 136–151.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

ZASADA WZAJEMNOŚCI W KONTEKŚCIE MODELOWANIA SPRZĘŻEŃ WZAJEMNYCH W UKŁADACH ANTENOWYCH

RECIPROCITY THEORY IN THE CONTEXT OF MUTUAL COUPLING MODELING IN ANTENNA ARRAYS

Mariusz Zamłyński[1,2]; Piotr Słobodzian[3]

[1] NOKIA, Mobile Networks, Wrocław, mariusz.zamlynski@nokia.com
[2] Politechnika Wrocławska, Wydział Informatyki i Telekomunikacji, Wrocław, mariusz.zamlynski@pwr.edu.pl
[3] Politechnika Wrocławska, Wydział Informatyki i Telekomunikacji, Wrocław, piotr.slobodzian@pwr.edu.pl

DOI: 10.15199/59.2025.4.95

Streszczenie: Artykuł dotyczy opisu zjawiska sprzężeń wzajemnych w układach antenowych złożonych z elementów quasi-jednomodowych. Został w nim zaproponowany nowy model zjawiska sprzężeń wzajemnych, który pozwala na opisanie rozkładu prądu powierzchniowego w elemencie aktywnym z błędem średniokwadratowym mniejszym o ok. $8 \div 17$ dB (w zależności od przeanalizowanego przypadku) w porównaniu z modelem klasycznym.

Abstract: The paper addresses the phenomenon of mutual coupling in antenna systems composed of quasi single-mode elements. It introduces a novel model for mutual coupling, enabling the characterization of surface current distributions on active elements with a mean-square error reduction of approximately $8-17$ dB, depending on the analyzed case, in comparison to the classical model.

Słowa kluczowe: rozkład prądów powierzchniowych, sprzężenia wzajemne, układ antenowy, zasada wzajemności.

Keywords: antenna array, current distribution, mutual coupling, reciprocity theorem.

1. WSTĘP

Prace nad opisem wpływu sprzężeń wzajemnych na parametry układów antenowych sięgają drugiej połowy lat 60 ubiegłego wieku [1]. Z pracy tej pochodzą do dziś przywoływane opisy zjawiska sprzężeń wzajemnych dla układu pracującego w trybie nadawczym oraz odbiorczym, które to zostały spopularyzowane przez [2]. Choć nie podważają one klasycznej zasady wzajemności Lorentza, to autorzy [3], analizując opisane mechanizmy, doszli do wniosku, że problem analizy (i później kompensacji) sprzężeń wzajemnych w nadawczych i odbiorczych układach antenowych jest inny ze względu na różnice w ścieżkach rozchodzenia się sygnału. Tym samym wytłumaczyli różnicę pomiędzy klasycznie zdefiniowanymi impedancjami wzajemnymi a impedancjami wyznaczanymi w trybie odbiorczym.

Z kolei technika MAIEM (z ang. Multiple Antenna Induced EMF Method) [4] bazuje na zastosowaniu sprzężeń wzajemnych zamodelowanych w trybie nadawczym do kompensacji tego zjawiska w odbiorczym układzie antenowym. Również metoda Steyskala i Herda, której pierwotny opis dotyczy odbiorczego układu antenowego [5], jest stosowana do kompensacji sprzężeń wzajemnych w układach pracujących w trybie nadawczym [6].

W dalszej części artykułu pokazujemy, jaki jest związek pomiędzy parametrami układu antenowego pracującego w trybie odbiorczym i nadawczym. Zaczynamy od wynikającego z klasycznej zasady wzajemności Lorentza zastosowania rozkładów prądów powierzchniowych do wyznaczenia wartości napięcia na zaciskach anteny odbiorczej. Następnie pokazujemy związek tych samych rozkładów prądów z charakterystyką promieniowania elementów układu antenowego pracujących w obecności pozostałych elementów. Na koniec adresujemy problem niemożliwości zastosowania zasady wzajemności do określania związku pomiędzy rozkładami prądów powierzchniowych dla anteny pracującej w trybie nadawczym i odbiorczym, poprzez zaproponowanie nowego modelu zjawiska sprzężeń wzajemnych.

2. TEORIA

2.1 Zastępczy model obwodowy anteny odbiorczej

Antenę odbiorczą można zamodelować korzystając z tw. Thevenina jako połączenie szeregowe impedancji obciążenia (Z_L), impedancji wejściowej anteny wyznaczonej w trybie nadawczym Z_{ant} oraz zastępczego napięcia na rozwartych zaciskach anteny odbiorczej, które wynosi:

$$v_{\mathrm{oc}} = -\frac{1}{I(0)} \iint\limits_{S} \vec{J}(\vec{r}) \cdot \vec{E}_{\mathrm{inc}}(\vec{r})\, dS \tag{1}$$

gdzie $\vec{J}(\vec{r})$ jest rozkładem prądu powierzchniowego wyznaczonego dla anteny pracującej w trybie nadawczym, $\vec{E}_{\mathrm{inc}}$ jest polem elektrycznym padającym na powierzchnię S anteny odbiorczej. Natomiast napięcie na obciążeniu podłączonym do zacisków anteny odbiorczej jest równe:

$$v_{\mathrm{t}} = \frac{Z_L}{Z_{ant} + Z_L} v_{\mathrm{oc}} \tag{2}$$

Jak wykazano w [7] oraz [8] opisana metodologia bazuje na klasycznej zasadzie Lorentza i sprowadza się do wyznaczenia obwodowego parametru anteny pracującej w trybie

odbiorczym (tj. napięcia odłożonego na obciążeniu podłączonym do zacisków tej anteny) na podstawie jej parametrów wyznaczonych w trybie nadawczym (tj. impedancji wejściowej oraz rozkładu prądu powierzchniowego).

Analogiczne podejście ma zastosowanie do układów antenowych. W przypadku układu antenowego pracującego w trybie odbiorczym mamy do czynienia z zagadnieniem wyznaczenia napięć odkładających się na każdym z obciążeń N-elementowego układu antenowego. Dla n-tego elementu układu równanie 2 przyjmuje postać:

$$v_{\mathrm{t}n} = \frac{Z_{Ln}}{Z_{wej_n} + Z_{Ln}} v_{\mathrm{oc}n} \tag{3}$$

gdzie Z_{Ln} jest impedancją obciążenia podłączoną do zacisków n-tego elementu, Z_{wej_n} impedancją wejściową n-tego elementu wyznaczoną w warunkach obciążenia pozostałych elementów układu. Natomiast napięcie, które wydzieliłoby się na zaciskach rozwartego n-tego elementu $v_{\mathrm{oc}n}$ wynosi [4]:

$$v_{\mathrm{oc}n} = -\frac{1}{I_n(0)} \sum_{m=1}^{N} \iint\limits_{S_m} \vec{J}_{nm}(\vec{r}) \cdot \vec{E}_m(\vec{r})\, dS \tag{4}$$

gdzie $\vec{J}_{nm}$ jest rozkładem prądu powierzchniowego na powierzchni m-tego elementu wyznaczonego przy pobudzeniu n-tego elementu i obciążeniu pozostałych elementów.

Z równań (3) oraz (4) wynika, że zamodelowanie zjawiska sprzężeń wzajemnych w trybie nadawczym w postaci impedancji wejściowej elementów aktywnych oraz rozkładów prądów powierzchniowych jest wystarczające do właściwego określenia parametrów sygnałów na wyjściach tego samego układu antenowego pracującego w trybie odbiorczym.

2.2 Modelowanie charakterystyki promieniowania układu antenowego w trybie nadawczym

Korzystając z zasady superpozycji pól, charakterystykę promieniowania układu antenowego $F(u)$ przy znanych napięciowych współczynnikach pobudzeń w_n można wyznaczyć, korzystając z zależności:

$$F(u) = \sum_{n=1}^{N} w_n g_n(u) \tag{5}$$

gdzie $g_n(u)$ jest charakterystyką promieniowania tzw. elementu aktywnego, tj. wyznaczoną w warunkach pobudzenia n-tego elementu układu przy obciążeniu pozostałych elementów układu antenowego. Przy czym równanie (5) zakłada, że w charakterystyce $g_n(u)$ uwzględniony jest człon fazowy związany z położeniem elementu w układzie. Analogicznie do równania (4), charakterystyki te można wyznaczyć na podstawie znajomości rozkładów prądów powierzchniowych:

$$g_n(u) = -\frac{j\eta}{4\pi k} \sum_{m=1}^{N} \iint\limits_{S_m} \vec{J}_{nm}(\vec{r}) \cdot G(s,u)\, dS \tag{6}$$

gdzie $G(s,u)$ jest funkcją Greena dla wolnej przestrzeni, η impedancją wolnej przestrzeni, a k liczbą falową.

2.3 Modelowanie rozkładów prądów powierzchniowych

Z równań (4) oraz (6) widać, że do poprawnego modelowania zjawiska sprzężeń wzajemnych na potrzeby układów antenowych pracujących zarówno w trybie nadawczym, jak i odbiorczym, potrzebne jest znalezienie związku pomiędzy rozkładami $\vec{J}_{nm}(\vec{r})$, a rozkładem prądu powierzchniowego elementu odizolowanego. Przy czym należy pamiętać, że rozkład $\vec{J}_{nm}$ jest rozkładem na powierzchni m-tego elementu przy pobudzeniu n-tego elementu układu. Oznacza to, że przy rozważaniu tego problemu antena $\#n$ pracuje w trybie nadawczym, a antena $\#m$ w trybie odbiorczym i powstałe na jej powierzchni prądy są indukowane w wyniku pobudzenia polem pochodzącym od anteny $\#n$. Nawet w przypadku, w którym odległość pomiędzy identycznymi antenami jest na tyle duża, że zjawisko sprzężenia wzajemnego można pominąć, to zasada wzajemności Lorentza nie pozwala na określenie rozkładu prądu na całej powierzchni anteny odbiorczej na podstawie znajomości rozkładu prądu na powierzchni anteny nadawczej przy pobudzeniu źródłem napięciowym tylko w jednym punkcie [7].

Dlatego też przy analizie zjawiska sprzężeń wzajemnych w układach antenowych przyjmuje się przeważnie, że układ taki złożony jest z identycznych elementów jednomodowych. Według definicji zawartej w [1] jest to taki element antenowy, dla którego warunki brzegowe wymuszają, że dowolny prąd płynący w strukturze anteny musi przyjąć jeden, ściśle określony rozkład przestrzenny. Z kolei w [9] przez element jednomodowy rozumiany jest element antenowy, którego pole wtórne w polu dalekim jest takie samo jak charakterystyka promieniowania wyznaczona w trybie nadawania. Wszystkie te definicje zakładają, że rozkład prądu powierzchniowego w antenie pracującej w trybie odbiorczym jest wprost proporcjonalny do rozkładu prądu powierzchniowego wyznaczonego dla anteny nadawczej, co można zapisać jako:

$$\vec{J}_{odb} \propto \vec{J}_{nad} \tag{7}$$

Zakładając, że obecność pozostałych elementów nie wzbudza innych modów, to można przyjąć, że:

$$\vec{J}_{nm} = c_{nm} \cdot \vec{J}_{izo} \tag{8}$$

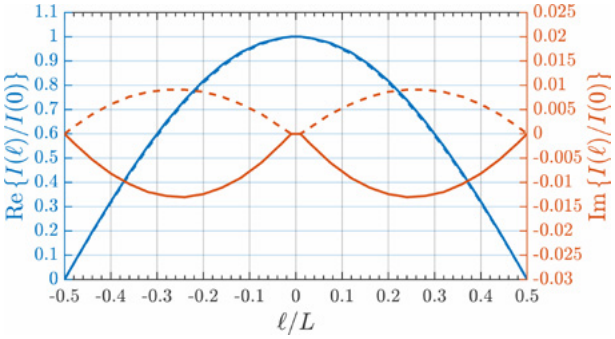
gdzie c_{nm} jest współczynnikiem sprzężenia, a $\vec{J}_{izo}$ rozkładem prądu powierzchniowego elementu odizolowanego. W poprzednich pracach wykazaliśmy już, że założenie to pozwala zarówno na skuteczną kompensację zjawiska sprzężeń wzajemnych [6], czy też modelowanie charakterystyk promieniowania układów antenowych w obecności zjawiska dyfrakcji [10].

Jednakże, jak pokazują wyniki symulacji elektromagnetycznych przedstawione w kolejnej sekcji, nawet w przypadku cienkiego dipola o długości zbliżonej do $\lambda/2$, z analizy rozkładu prądu powierzchniowego oprócz dominującego modu widoczny jest również dodatkowy mod o charakterze reaktywnym (tj. odpowiada on za część urojoną rozkładu prądu unormowanego względem środkowego segmentu). Choć w ujęciu Teorii Modów Charakterystycznych (z ang. TCM - Theory of Characteristic Modes) [11] element taki nie jest określany mianem elementu jednomodowego, to dalej możliwe jest wyznaczenie dla niego współczynnika sprzężenia wspólnego dla obu modów. Zakładając, że współczynniki dekomponujące rozkłady prądów dla anteny nadawczej i odbiorczej mają takie same amplitudy, a różnią się jedynie przesunięciem fazowym wynoszącym π dla modu odpowiedzialnego za część urojoną rozkładu, możemy wtedy przyjąć, że:

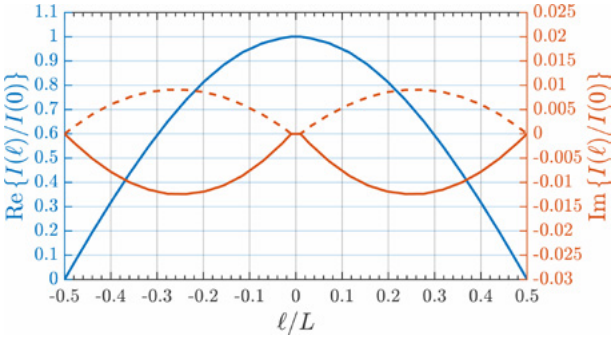
$$\vec{J}_{nm} \approx \begin{cases} c_{nm} \cdot \vec{J}_{\mathrm{izo}} & \text{dla } m = n, \\ c_{nm} \cdot \vec{J}^{*}_{\mathrm{izo}} & \text{dla } m \neq n. \end{cases} \tag{9}$$

Tabela 1: Parametry zamodelowanych dipoli półfalowych

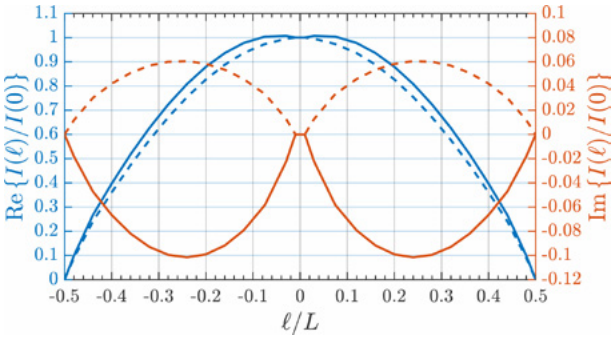
Parametr	*Dipol #1*	*Dipol #2*	*Dipol #3*	*Dipol #4*
Długość L	$\lambda/2$	$\approx 0{,}9888\lambda/2$	$\lambda/2$	$\approx 0{,}94\lambda/2$
Średnica d	$10^{-9}\lambda$	$\lambda/10^{9}$	$5\lambda/10^{3}$	$5\lambda/10^{3}$
Współczynnik L/d	$5 \cdot 10^{8}$	$\approx 4{,}944 \cdot 10^{8}$	100	≈ 94
Imp. wej. Z_{we}	$74{,}923 + j41{,}461\,\Omega$	$72{,}500\,\Omega$	$88{,}269 + j45{,}394\,\Omega$	$71{,}834\,\Omega$



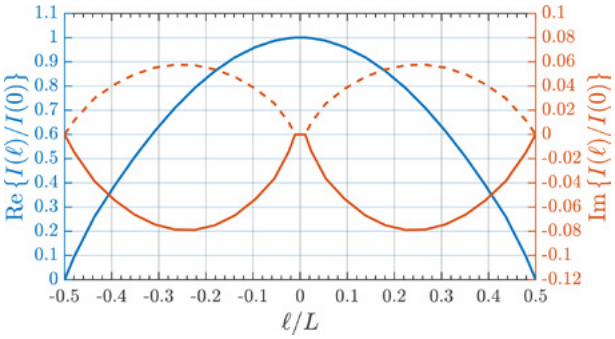
(a) Dipol #1



(b) Dipol #2



(c) Dipol #3



(d) Dipol #4

Rysunek 1: Rozkład prądu powierzchniowego dla poszczególnych dipoli pracujących w trybie nadawczym (linie przerywane) oraz odbiorczym (linie ciągłe).

3. SYMULACJE ELEKTROMAGNETYCZNE

3.1 Element odizolowany

W celu porównania rozkładów prądu powierzchniowego dla tej samej anteny pracującej w trybie nadawczym oraz odbiorczym zamodelowano w symulatorze elektromagnetycznym FEKO cztery warianty dipola półfalowego o parametrach zestawionych w Tabeli 1. Każdy z dipoli został podzielony na 25 segmentów, przy czym środkowy miał długość $\lambda/100$.

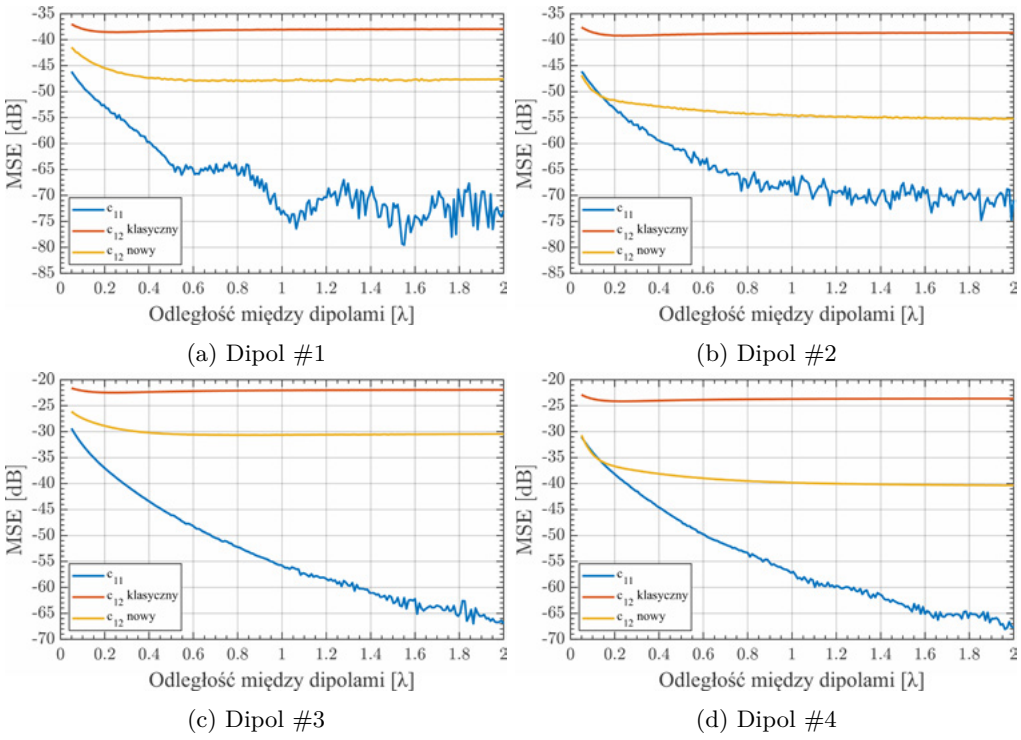
Rozkład prądu powierzchniowego dla anten nadawczych został wyznaczony w warunkach podłączenia idealnego źródła napięciowego (tj. o zerowej impedancji wewnętrznej) o amplitudzie 1 V. Natomiast rozkład prądu powierzchniowego dla anten odbiorczych został wyznaczony przy podłączeniu do zacisków każdej z anten impedancji 50 Ω i oświetleniu jednorodną falą płaską o polaryzacji liniowej zgodnej z orientacją anteny oraz padającej prostopadle do jej osi. Ich porównanie znajduje się na rys. 1. Każdy z rozkładów został unormowany do wartości prądu w środkowym segmencie. Dla wszystkich z elementów, z wyjątkiem przypadku #3 (który charakteryzuje się niskim współczynnikiem smukłości oraz brakiem rezonansu), część rzeczywista rozkładu prądu powierzchniowego zachowuje ten sam kształt zarówno w trybie nadawczym, jak i odbiorczym. Natomiast część urojona utrzymuje w przybliżeniu tą samą amplitudę i kształt, ale ma przeciwny znak. Oznacza to, że założenia poczynione pod koniec poprzedniej sekcji są w przybliżeniu spełnione.

3.2 Układ dwóch dipoli

w celu weryfikacji poprawności zależności (9) przeanalizowano sprzężenia wzajemne pomiędzy dwoma dipolami $\lambda/2$ o parametrach opisanych w poprzedniej podsekcji. Dipole zostały umieszczone w wolnej przestrzeni, równolegle do siebie i w tej samej orientacji. Za miarę oceny poprawności przejęto unormowany błąd średniokwadratowy wyrażony zależnością:

$$\mathrm{MSE} = 10\log_{10}\frac{\sum |J_{nm} - J^{*}_{izo}\cdot c_{nm}|^{2}}{\sum |J_{nm}|^{2}} \qquad (10)$$

Jej wartości dla modelowania prądów J_{11} oraz J_{12} przy użyciu zależności (8) i (9) przedstawiono na rysunku 2.



(a) Dipol #1 (b) Dipol #2

(c) Dipol #3 (d) Dipol #4

Rysunek 2: Średniokwadratowy błąd modelowania rozkładów prądów powierzchniowych.

4. PODSUMOWANIE

W artykule wykazano, że definicja jednomodowego elementu promieniującego nie jest spełniona nawet w przypadku półfalowych dipoli w sposób ścisły. Z zasady wzajemności nie wynika fakt, że rozkład prądu powierzchniowego w elemencie pracującym w trybie odbiorczym jest wprost proporcjonalny do rozkładu prądu powierzchniowego elementu pracującego w trybie nadawczym. Jednakże, jak pokazują wyniki symulacji elektromagnetycznych, stwierdzenie to jest z dość dużym przybliżeniem spełnione dla cienkich dipoli o długości $\lambda/2$, a w przypadku elementów o współczynniku L/d zbliżonym do 100, można stwierdzenie to jest spełnione z dokładnością rzędu $0,1\%$ (tj. przy unormowanym błędzie średniokwadratowym ≤ -30 dB) po uwzględnieniu sprzężenia zespolonego.

LITERATURA

[1] Allen J. L., B. L. Diamond. 1966. „Mutual Coupling in Array Antennas". Raport techniczny 424 (ESD-TR-66-443), MIT Lincoln Laboratory.

[2] Balanis C. A. 2016. *Antenna Theory: Analysis and Design*, wyd. 4. Wiley, Hoboken.

[3] Lui H. S., H. T. Hui, M. S. Leong. 2009. „A Note on the Mutual-Coupling Problems in Transmitting and Receiving Antenna Arrays". *IEEE Antennas and Propagation Magazine*, 51 (5) : 171–176.

[4] Henault S., Y. M. M. Antar, S. Rajan, R. Inkol, S. Wang. 2009. „The Multiple Antenna Induced EMF Method for the Precise Calculation of the Coupling Matrix in a Receiving Antenna Array". *Progress in Electromagnetics Research M*, 8 : 103–118.

[5] Steyskal H., J. Herd. 1990. „Mutual Coupling Compensation in Small Array Antennas". *IEEE Transactions on Antennas and Propagation*, 38 (12) : 1971–1975.

[6] Zamłyński M., P. Słobodzian. 2011. „Influence of the Aperture Edge Diffraction Effects on the Mutual Coupling Compensation Technique in Small Planar Antenna Arrays". *International Journal of Electronics and Telecommunications*, 57(1), s.115–120.

[7] Schelkunoff S. A., H. T. Friis, V. Twersky. 1953. *Antennas: Theory and Practice*. Wiley, New York.

[8] Kraus J. D. 1950. *Antennas*. McGraw-Hill, New York.

[9] Pfeiffer C., F. Dagefu, B. Tomasic, J. N. Twigg, B. I. Wu. 2021. „Virtual Impedance Method for Mutual Coupling Compensation". *IEEE Transactions on Antennas and Propagation*, 69 (8) : 4569–4579.

[10] Zamłyński M., P. Słobodzian. 2012. „Antenna Array Radiation Pattern Modeling Which Includes Mutual Coupling and Diffraction Effects". *Proceedings of the 6th European Conference on Antennas and Propagation*, s. 1627–1631, Praga.

[11] Harrington R., J. Mautz. 1971. „Theory of Characteristic Modes for Conducting Bodies". *IEEE Transactions on Antennas and Propagation*, 19(5), s.622–628.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

COMPARISON OF IMPROVEMENT ON IQA METRICS WITH DEEP LEARNING-BASED VISUAL QUALITY ENHANCEMENT MODELS

PORÓWNANIE POPRAWY METRYK IQA PRZY UŻYCIU MODELI POPRAWY JAKOŚCI WIZUALNEJ OPARTYCH NA GŁĘBOKIM UCZENIU

Doğukan Öztürk[1]; Mikołaj Leszczuk[2]

[1] AGH University of Krakow, al. Mickiewicza 30, PL-30059, Kraków, Poland, dozturk@agh.edu.pl
[2] AGH University of Krakow, al. Mickiewicza 30, PL-30059, Kraków, Poland, mikolaj.leszczuk@agh.edu.pl

DOI: 10.15199/59.2025.4.96

Abstract: Deep learning based architectures have been applied to improve image quality on underwater image photographs. After training with different architectures, the images obtained underwater are regenerated. The improvement in image quality was objectively measured by comparing the NIQE, BRISQUE, and NIQE IQA metrics of the obtained images. NIQE improved by 6.7%, BRISQUE by 12.3%, PIQE by 18.8%, and the overall score by 11.4%. Furthermore, it has been validated that cascade structures will increase the overall score.

Streszczenie: W pracy zastosowano modele głębokiego uczenia do poprawy jakości zdjęć podwodnych. Przeszkolono sześć różnych architektur sieci neuronowych, a uzyskane obrazy oceniono obiektywnie za pomocą bezreferencyjnych metryk NR-IQA: NIQE, BRISQUE i PIQE. Uzyskano średnie spadki: NIQE o 6,7%, BRISQUE o 12,3%, PIQE o 18,8%, co przełożyło się na poprawę ogólnego wskaźnika jakości o 11,4%. Dodatkowo wykazano, że struktury kaskadowe zwiększają skuteczność poprawy obrazu.

Keywords: Deep-models, Image restoration, NR-IQA, Underwater image enhancement, Quality metrics

Słowa kluczowe: Głębokie uczenie, Rekonstrukcja obrazu, NR-IQA, Ulepszanie obrazów podwodnych, Metryki jakości

1. INTRODUCTION

Factors such as fog and dust can introduce noise into images, resulting in a deterioration of quality attributed to inadequate lighting conditions while examining various unwanted signals categorized as noise. Similarly, underwater images experience noise and quality loss due to light refraction, scattering, and absorption effects caused by the water. As a result, achieving high-quality images underwater while preserving their natural appearance poses a considerable challenge. Enhancing the quality of these captured images is vital for both academic applications and industrial sectors, including coral mapping, pipeline inspection, submarine footage analysis, and deep-sea research. Moreover, such image enhancement methods hold importance in strategic domains such as imagery intelligence (IMINT) and electronic warfare (EW).

Over the last decade, underwater images have been reconstructed using a variety of learning architectures, relying on the significant increase in GPU processing power to significantly simplify image processing while simultaneously popularizing the fields of deep learning and generative artificial intelligence. The efficacy of all methods was assessed using no-reference metrics, specifically the Natural Image Quality Evaluator (NIQE), the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE), and the Perception-Based Image Quality Evaluator (PIQE). Additionally, an alternative approach was employed by examining the quality enhancement achieved through the cascading of certain architectures. Moreover, image quality assessment (IQA) metrics, visual attribute metrics, and low-level features.

2. MATERIAL AND METHODS

One of the strengths of this paper is its comparative analysis of seven distinct deep learning models in conjunction with three different cascade structures, which provides valuable insights for future research by elucidating the efficacy of these cascade structures. Moreover, the limitations inherent in various architectures may be alleviated through the utilization of alternative architectures, leading to the acquisition of higher quality images.

2.1 Data Preparation and patch-based training using center cropping to fixed resolution

The high-quality underwater images from the UIEB [1] and reina0112/PHISMID-PHISWID [2] datasets were integrated with 256 non-underwater images to reduce the risk of overfitting. The dataset comprises images of varying sizes. While there are notable advantages to utilizing the original dimensions of the images, it is essential

that the batch size during training remains fixed at 1. This restriction significantly accelerates the learning process owing to factors such as increased batch size, enhanced parallel processing capabilities, and minimized transitions between the CPU and GPU. Consequently, the dataset was standardized to a resolution of 384x384 pixels. To preserve the integrity of the data, the resizing process was avoided; instead, larger images were centrally cropped. With all images now uniform in size, various batch size configurations can be employed, facilitating different rates of gradient convergence, ultimately leading to a reduction in training duration. Furthermore, 309 of the 2853 images acquired were set aside for the validation set. In contrast, 247 images were allocated for the test set, distinct from those in the training set.

2.1.1 Augmentation

With the application of color jitters, brightness, contrast, saturation ±20%, and hue ±5% color tone shifts, photographs with various distortions were generated, thus expanding the dataset. The primary objective was to mitigate overfitting by training the system on diverse distortions. Data augmentation plays a vital role in enhancing data diversity; however, excessive augmentation can lead to overfitting, while insufficient augmentation may result in underfitting.

2.2 Deep learning-based visual quality enhancement models.

In this section, the models were trained using seven distinct deep learning architectures.

2.2.1 FusionNet-based residual enhancement network

This architecture was selected due to its ability to generate a natural image by processing features at multiple resolutions through a fusion structure that integrates various characteristics. The input block is comprised of convolutional layers utilizing the ReLU activation function, while the body section consists of stacked blocks, ultimately producing an RGB output at the final stage [3].

2.2.2 Residual convolutional network by WaterNet

WaterNet is a specialized network developed for the processing of underwater images. This advanced convolutional neural network (CNN) methodology is adaptable to various architectures and incorporates a structure that enhances performance by preserving image quality through the use of residual stacks, similar to FusionNet [4].

2.2.3 Residual CNN inspired by Retinexformer

It is an architecture designed to enhance performance in low-light conditions and to accommodate incompatible colors, based on the principles of ResNet. This compact neural network system, which consists of an input block, an encoder, and a decoder, extracts low-level features and subsequently generates an RGB image after learning the underlying details and distortions [5].

2.2.4 Attention-Gated U-Net (AttUNet)

This architecture is designed to emphasize significant features by bypassing irrelevant information and incorporating an attention mechanism within the encoder and decoder components of the U-Net framework. Such filtering can be effective in suppressing noise [6].

2.2.5 Lightweight convolutional model by DehazeNet

This system is typically applied to images characterized by high levels of noise and distortion, such as those encountered in environments with dense fog. In this paper, we propose that the noise created by water can be interpreted using fog logic, thereby improving image quality through this model with analogous reasoning [7]. Notably, this approach has yielded the most successful results when employed within single, non-cascaded architectures. Moreover, it has been demonstrated that this architecture is highly effective and possesses remarkable capabilities.

2.2.6 Multi-stage progressive network (MPR-Net)

It is an architecture characterized by multiple structures and is widely used today to improve image quality by addressing various parameters such as noise and contrast. It is a strong candidate that can increase the naturalness in images [8].

2.2.7 Transformer-based restoration network by Restormer

It is akin to the U-Net architecture and embodies an advanced transformer framework that integrates distinctive convolutional and attention blocks, as well as a feedforward layer. This architecture can be employed to enhance image quality and optimize lighting conditions in underwater images [9].

2.3 Cascade Models

Different models with cascade structures were trained, resulting in varying success percentages.

2.3.1 Cascade Retinexformer-to-Restormer

After generating the image with Retinexformer, the subsequent step involves employing various architectures utilizing Restormer. Additionally, incorporating two generators affords the capability to rectify the deficiencies observed in the earlier generation process, resulting in improved outcomes. These two architectures effectively complement each other's limitations to a significant degree.

2.3.2 Cascade AttUNet-to-Restormer

The transformer base of Restormer was integrated with AttuNet, and the performance metrics of the cascade structure were evaluated. It is essential to maintain a natural appearance by minimizing excessive processing corrections; however, generating a new image has the potential to enhance quality.

2.3.3 Cascade AttUNet-to-AttuNet

The image quality was assessed when the image was regenerated using the same architecture. Repetition was conducted to determine whether this would lead to overfitting or if the deficiencies observed in the initial image generation would be addressed. In instances where a cascade structure is applied to itself, only one of the three IQA scores exhibited an increase. Future research should explore the potential of improving shortcomings by integrating alternative architectures.

2.4 No-reference IQA metrics

Since reference-based metrics cannot be employed due to the lack of information regarding image quality and the absence of ground truth, we are compelled to utilize alternative approaches. In this context, it is essential to rely on no-reference quality metrics. The most reliable no-reference metrics are NIQE, BRISQUE, and PIQE [10].

2.4.1 Natural image quality evaluator (NIQE)

Because the unnaturalness score is employed, a low score, i.e. close to zero, indicates a natural image. Generally, any score above 10 is deemed artificial; as the score decreases, higher quality is achieved. This method detects blur and noise and incorporates perceptual blocks.

2.4.2 Blind/Referenceless image spatial quality evaluator (BRISQUE)

Natural correlations, differences between light and dark, and statistical deviations of parameters such as noise level are measured, and the image is assigned a Natural Scene Statistics (NSS) score on a scale from 0 to 100. A lower score indicates a higher measure of quality.

2.4.3 Perception-based image quality evaluator (PIQE)

Measures the quality of images by segmenting them into parts, utilizing a mechanized version of human perception to assign a score ranging from 0 to 100. Low scores, such as those produced by NIQE and BRISQUE, indicate lower image quality. The development of these metrics based on natural image statistics reinforces their alignment with human perceptual assessment.

3. RESULTS

The learning models described in the materials and methods section were trained sequentially using images from the training set under uniform conditions, without employing any image size reduction. Subsequently, the objective no-reference IQA metrics and image appearance metrics of the generated images were compared. To investigate the relationships among these metrics, a heatmap is presented in Figure 1.

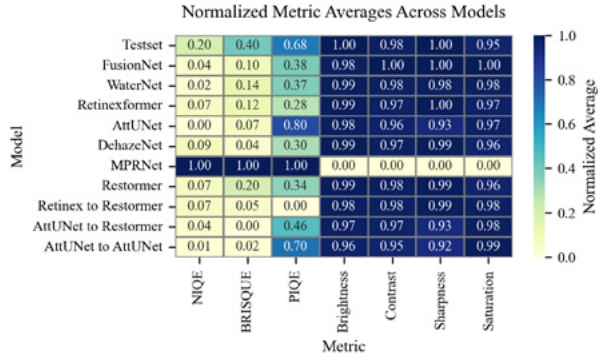


Figure 1: Heatmap of Normalized Visual Quality Indicators

As illustrated in Figure 1, AttUNet achieves a commendable NIQE score when evaluated independently and secures the highest BRISQUE score when combined with Restormer. Nonetheless, when all three IQA metrics are taken into account, AttUNet yields the lowest overall score. This finding suggests a relative success when used in conjunction with Restormer, as demonstrated by the comparatively poor performance observed under the PIQE.

Evaluating the success of image appearance metrics based solely on the absolute magnitude of the scores is inappropriate, particularly in the context of IQA metrics. Instead, these metrics should be assessed within a defined range that indicates naturalness. Although all the presented methods are within this acceptable range, the scores for MPR-Net are significantly worsened, leading to images that appear washed out. Consequently, this necessitates a re-evaluation of the research under different conditions.

The success of the enhancement is quantitatively assessed by evaluating the reduction in IQA metric scores relative to the test set, with the most substantial decrease indicating the most effective enhancement. Table 1 presents the average increase or decrease in the IQA metrics of images obtained through deep learning-based visual quality enhancement models.

Table 1: Visual quality improvements compared to test-set

Model	*NIQE*	*BRISQUE*	*PIQE*
FusionNet	↓ 5.3%	↓ 9.3%	↓ 8.2%
WaterNet	↓ 5.9%	↓ 8.0%	↓ 8.4%
Retinexformer	↓ 4.5%	↓ 8.7%	↓ 11.1%
AttUNet	↓ 6.7%	↓ 10.2%	↓ 3.4%
DehazeNet	↓ 3.7%	↓ 11.0%	↓ 10.5%
MPR-Net	↑ 27.6%	↑ 18.7%	↑ 8.9%
Restormer	↓ 4.3%	↓6.1%	↓ 9.2%
Retinexformer→Restormer	↓4.5%	↓ 10.8%	↓ 18.8%
AttUNet→Restormer	↓ 5.5%	↓ 12.3%	↓ 6.0%
AttUNet→AttUNet	↓ 6.4%	↓ 11.6%	↑ 0.6%

When examining Table 1, it is evident that AttUNet achieved the highest reduction in NIQE, demonstrating an improvement rate of 6.7%. In contrast, the most significant decrease in BRISQUE was observed from AttUNet to Restormer, with a rate of 12.3%. Furthermore, the highest reduction in PIQE occurred between Retinexformer and Restormer, resulting in an 18.8% improvement. In addition, when a cascade structure is established with AttUNet, the most favorable BRISQUE score is attained.

Figure 2 presents the average IQA metric scores for all evaluated models. The blue column indicates the most

effective technique, whereas the red column signifies the technique that exhibits the most significant degradation.

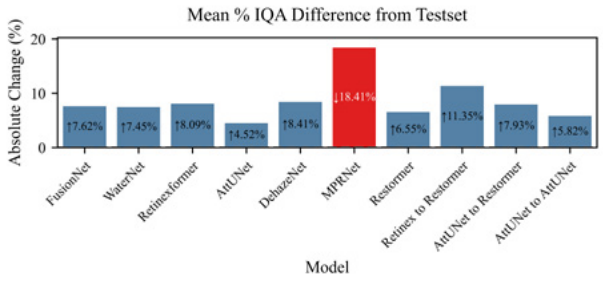


Figure 2: Overal score of quality indicators

When the importance of the three IQA metrics is evaluated equally, Retinex to Restormer receives the highest overall score, while Dehazenet differentiates itself as a cascade-free method. Consequently, future research can focus on smart cascade compatible with Dehazenet.

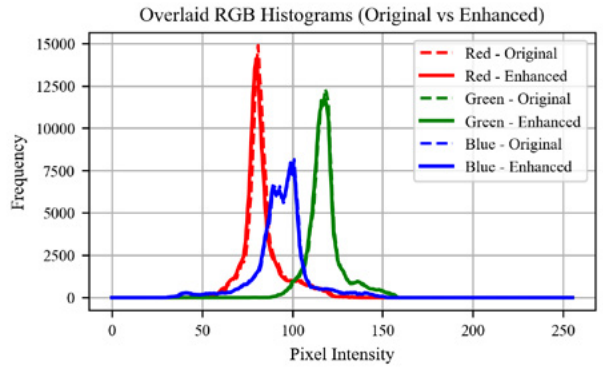


Figure 3: Histograms of best-performing image before and after training

In the histogram graph depicted in Figure 3, the dashed lines represent the generated image, whereas the continuous lines correspond to the original image. It is evident that the image exhibits increased vibrancy and reduced dullness, as indicated by the heightened frequency and intensity of red and green tones. Conversely, the increase in blue remains largely unchanged.

4. CONCLUSIONS

Deep learning-based visual quality enhancement models significantly improve IQA metrics, resulting in images that appear brighter and more natural. The most notable increases were observed as follows: a 6.7% improvement in NIQE when using AttUNet, 12.3% increase in BRISQUE from AttUNet to Restormer, and an 18. 8% enhancement in PIQE from Retinexformer to Restormer. The best average score across all improvements was achieved by Retinexformer to Restormer, with a mean increase of 11.4%, while DehazeNet demonstrated a significant improvement as a standalone method, achieving an 8.4% enhancement.

Since the MPR-Net method employs these techniques, there is a loss of liveliness and naturalness in the images. This powerful architecture necessitates testing with alternative techniques. Achieving higher quality images through DehazeNet-compatible intelligent cascaded structures could yield significant advancements in the field.

ACKNOWLEDGMENT

This paper was supported by the Polish Ministry of Science and Higher Education through subvention funds allocated to the Faculty of Computer Science, Electronics, and Telecommunications at AGH University of Krakow.

REFERENCES

[1] Li Chongyi, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, Dacheng Tao. 2019. „An underwater image enhancement benchmark dataset and beyond". IEEE Transactions on Image Processing, 29 : 4376–4389. Available at: https://li-chongyi.github.io/proj_benchmark.html, accessed: 21.05.2025.

[2] Reina0112. 2024. *PHISMID–PHISWID: Physics-Based Synthetic and Realistic Underwater Datasets*. GitHub repository. Available at: https://github.com/reina0112/PHISMID-PHISWID. Accessed: 21.05.2025.

[3] Bittner Ksenia, Fathalrahman Adam, Shiyong Cui, Marco Körner, Peter Reinartz. 2018. „Building footprint extraction from VHR remote sensing images combined with normalized DSMs using fused fully convolutional networks". IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11 (8) : 2615–2629.

[4] Lewandowski Roman, Ireneusz Kowalski. 2008. W poszukiwaniu obiektywnych metod pomiaru jakości usług medycznych. Współczesne wyzwania strukturalne i menedżerskie w ochronie zdrowia, 253–266. Uniwersytet Warmińsko-Mazurski w Olsztynie.

[5] Wang Zexin, Letu Qingge, Qingyi Pan, Pei Yang. 2024. „Retinex decomposition based low-light image enhancement by integrating Swin transformer and U-Net-like architecture". IET Image Processing, 18 (11) : 3028–3041.

[6] Lee Jongseok, Wahyu Wiratama, Wooju Lee, Ismail Marzuki, Donggyu Sim. 2023. „Bilateral Attention U-Net with Dissimilarity Attention Gate for Change Detection on Remote Sensing Imageries". Applied Sciences, 13 (4) : 2485.

[7] Chen Yutong, Zhang Wen, Chao Wang, Lei Gong, Zhongchao Yi. 2024. „PriorNet: a novel lightweight network with multidimensional interactive attention for efficient image dehazing". arXiv preprint arXiv:2404.15638.

[8] Mohamed Rana Reda, Omar M. Fahmy, Shereen Afifi. 2024. „MPR-Net-Based Enhancement Technique For Image Restoration". In: 2024 41st National Radio Science Conference (NRSC), vol. 1 : 136–142. IEEE.

[9] Zamir Syed Waqas, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang. 2022. „Restormer: Efficient transformer for high-resolution image restoration". In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition : 5728–5739.

[10] Jiménez-Gaona Yuliana, Diana Carrión-Figueroa, Vasudevan Lakshminarayanan, María José Rodríguez-Álvarez. 2024. „GAN-based data augmentation to improve breast ultrasound and mammography mass classification". Biomedical Signal Processing and Control, 94 : 106255.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

DATASET EVALUATING QoE IN SEGMENT BASED ADAPTIVE BITRATE ALGORITHMS

ZBIÓR DANYCH DO OCENY QoE W SEGMENTOWYCH ALGORYTMACH ADAPTACYJNEJ SZYBKOŚCI BITOWEJ

Syed Uddin[1]; Michal Grega[2]; Mikolaj Leszcuk[3]; Waqas ur Rahman[4];

[1] AGH University of Krakow, uddin@agh.edu.pl
[2] AGH University of Krakow, grega@agh.edu.pl
[3] AGH University of Krakow, leszczuk@agh.edu.pl
[4] Birmingham City University,waqas.rahman@bcu.ac.uk

. . .

DOI: 10.15199/59.2025.4.97

Streszczenie:

Wzrasta liczba aplikacji skoncentrowanych na treściach wideo, a dynamiczny rozwój technologii mobilnych stwarza znaczny potencjał wdrażania ich na urządzeniach przenośnych i smartfonach. Ponieważ użytkownicy odbierają strumienie wideo w różnych warunkach przepustowości sieci, kluczowe jest sprostanie ich oczekiwaniom oraz utrzymanie wysokiej jakości doświadczenia (Quality of Experience, QoE). Celem niniejszej pracy jest zrozumienie, w jaki sposób różne strategie adaptacyjnego bitrate'u (ABR) oraz rozmiar segmentów wpływają na postrzeganą przez użytkowników jakość strumieniowania. W ramach eksperymentów, przeprowadzonych z wykorzystaniem biblioteki dash.js, zastosowano algorytmy ABR oparte na pomiarze przepustowości, stanie bufora oraz minimalizacji opóźnień. Testowe sekwencje wideo udostępniono uczestnikom w środowisku crowdsourcingowym, którzy ocenili je za pomocą 5-punktowej skali Mean Opinion Score (MOS). Uzyskane wyniki dostarczają cennych wskazówek do optymalizacji QoE w adaptacyjnym strumieniowaniu wideo oraz wskazują kierunki dalszego rozwoju strategii ABR uwzględniających subiektywną percepcję użytkownika.

Abstract: There is a growth in the video centric applications. As mobile technology is trending, there is much potential to deploy the video applications on the mobiles and smart phones. As users access the video content under varying bandwidth conditions, it is crucial to align with their expectation and ensure the highest quality perceived by users. This study aims to understand how different adaptive bitrate (ABR) strategies and segment sizes affect the perceived quality by the user in the streaming domain. A series of experiments were conducted using dash.js with multiple ABR algorithms. These algorithms are based on throughput, buffer and low-latency based approaches. The video sequences were deployed in a crowdsource environment. The participants rated the video sequences using a mean opinion score (MOS) scale. The findings provide valuable insights for optimizing QoE in the adaptive streaming domain and suggest directions for future adaptive streaming strategies based on user perception.

Słowa kluczowe: Adaptacyjne przesyłanie strumieniowe, Percepcja, QoE, Algorytmy ABR

Keywords: Adaptive streaming, Perception, QoE, ABR Algorithms

1. INTRODUCTION

Online video streaming is growing exponentially [1] . This trend is expected to continue in the future. It is estimated that by 2030, there will be an increase in the use of video applications [2].

The HTTP standard is used to deliver the video stream from the server to the client. The streaming server hosts manifest files that contain information about video segments and quality levels [3]. The user sends a request to the server to retrieve information on video segments, their quality levels, and timing [4].

The ABR algorithm configured on the client side to select the appropriate video segments based on the network conditions. The ABR algorithms are classified into traditional HAS (HTTP Adaptive Streaming) algorithms and low-latency algorithms.

This research answers the following questions:

1. What is the impact of segments on user-perceived video quality in adaptive streaming?
2. What is the impact of ABR algorithms on video quality in adaptive streaming?

The motivation behind conducting this research is to explore how adaptive streaming algorithms impact the quality of experience (QoE) perceived by end users. In adaptive video streaming, ensuring high QoE is a central goal. Especially when users are accessing video content over unstable bandwidth. To achieve this, both objective evaluations and subjective opinions should be combined to create user-centric evaluation models. These models can potentially guide the design of adaptive algorithms that optimize for actual human perception rather merely focusing on system performance. The paper is structured as follows. Section 2 presents the background work, Section 3 outlines the methodology, Section 4 provides the discussion, and Section 5 concludes the article.

2. STATE OF THE ART

The ABR algorithms and QoE has been studied in various works. The author conducted research focused solely on the rebuffering metric [5]. A research [6] is conducted by comparing stalling and quality switching. The subjective evaluation is performed. The results revealed that stalling does not significantly impact the QoE. Additionally, a subjective evaluation [7] assessed the impact of variations in video quality on QoE. These bandwidth variations were derived from the ABR behavior observed during the evaluation, utilizing short segments and specific test scenarios. In another study [8] a model was proposed that leverages reinforcement learning combined with human feedback technologies to better align with users' opinion scores. Furthermore, the author of study [9] investigated the impact of playback rate on the video quality. A subjective assessment was conducted to measure user perception, and the results demonstrated that playback rate does indeed affect QoE.

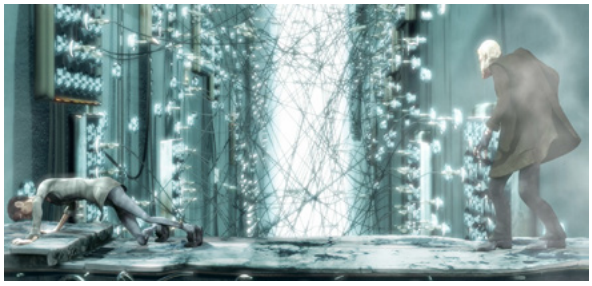
3. EVALUATION PROCEDURE

In this study, a subjective evaluation was conducted. A crowd-source platform is deployed. The Absolute Category Rating (ACR) method was employed in accordance with per ITU-T (P.910) guidelines [11]. The video sequences were presented to the test participants. The test participants were asked to rate each sequence on a 5-point Mean Opinion Score (MOS) scale.

3.1 Content Preparation

The source videos selected for this study include Big Buck Bunny, Tears of steel, Elephant Dream [12] and Sparks [13]. We chose video sequences from the Movies and Animation genre. Each clip has a duration of 60 seconds and operates at 24 Frames per second (FPS). The video frames screenshots shown in Figures 1a, 1b, 1c and 1d respectively.

(a) Big Buck Bunny



(b) Elephant Dream

(c) Tears of Steel

(d) Sparks

Figure 1: Sample video sequences used for evaluation

3.2 Procedure

The video sequences were prepared and deployed on a crowd-source platform called prolific. Each participant will watch the full sequence in a randomly generated order, and the same sequence will not be displayed a second time. The application on the crowd-source platform is capable of detecting whether that participants watch the entire video sequence within the stipulated time. The detailed steps are shown in Figure 2.

3.3 Test Participants

The test participants were recruited through the crowd-source platform prolific. A total of 70 participants were recruited. All participants watched videos online. The instructions were provided to the test participants [14]. Participants were required to complete the assessment using a desktop or laptop computer. The screen resolution was 1920x1080 which corresponds to the video

resolution used in this study [15]. The participants country information is provided in Figure 3.

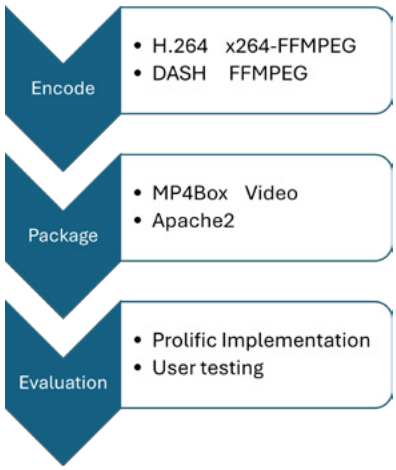


Figure 2: Study Evaluation Process

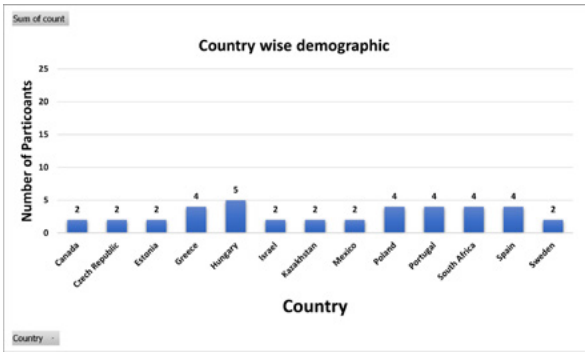


Figure 3: Test Participants Countries

4. OBSERVATION AND INSIGHTS

This section presents the study analysis and results. It will highlight the key observations and further insights. The regression analysis is conducted to identify the significant relationship between ABR algorithms and video sequences. The regression analysis of the Learn2Adapt-LowLatency (L2A-LL) and Low-on-Latency (LOL+) algorithms is presented. The equation 1 and equation 2, depicts the regression equations.

$$\text{Opinion Score}_{(\text{LOL+})} = 1.56 + 0.59 \cdot \text{Opinion Score}_{(\text{L2A-LL})} \quad (1)$$

$$S = 0.64, \quad R^2 = 47\%, \quad R^2_{\text{adj}} = 44.7\% \quad (2)$$

As F = 20.39, P < 0.001, this is statistically significant. The positive slope (0.59) demonstrate a moderate to strong positive correlation. The correlation fitted line plot is presented in the Figure 4.

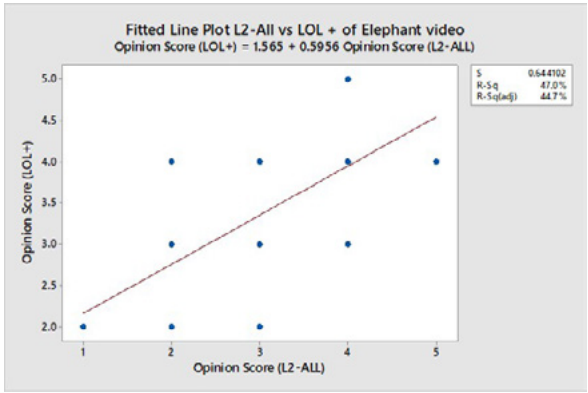


Figure 4: Fitted Line Plot L2-All vs LOL

Another intersting observation is recorded between L2A-LL and LOL+ algorithms. The equation 3 and equation 4, shows the regression equations.

$$\text{Opinion Score}_{(\text{LOL+})} = 1.307 + 0.67 \cdot \text{Opinion Score}_{(\text{L2A-LL})} \quad (3)$$

$$S = 0.81, \quad R^2 = 27\%, \quad R^2_{\text{adj}} = 23.8\% \quad (4)$$

As F = 20.39, P < 0.001, this is statistically significant. The positive slope (0.59) demonstrate a moderate to strong positive correlation. The correlation fitted line plot is presented in the Figure 5.

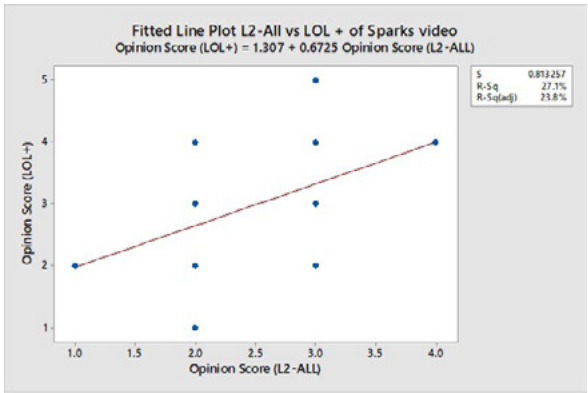


Figure 5: Fitted Line Plot L2-All vs LOL

As the above examples showing a positive linear relationship. The values scatter around the line. This indicates the variability in how the tester rated the video quality.

5. CONCLUSION AND FUTURE WORK

In this study, subjective evaluations are conducted to assess the performance of ABR algorithms under various network profiles, measuring their impact on the video quality. The research is conducted in a crowd-source environment. The results indicate that bandwidth profiles significantly influence user perception. The video dataset is created based on specific streaming policies. The video content is selected from various sources, including animations and movies. The prolific crowd-source platform is utilized for the evaluation. The low-latency algorithms evaluated and the regression analysis performed. The

findings suggest that the LOL+ outperforms the L2A-LL algorithm. The performance of LOL+ is significantly better compared to L2A-LL. Furthermore, the study findings highlight that user perception is not always aligned with the objective metrics. So it is crucial to build the user-centered video quality evaluation strategies.
In the future this research will be expanded to include additional content and genres. The number of test participants should be increased to gain a better understanding of the user experience. Subjective studies can be designed to measure video quality in the ultra-low latency streaming domain.
Acknowledgment This work was supported by the Polish Ministry of Science and Higher Education with the subvention funds of the Faculty of Computer Science, Electronics, and Telecommunications of AGH University.

REFERENCES

[1] Dao, Nhu-Ngoc & Tran, Anh-Tien & Hoang Tu, Ngo & Tran-Thien, Thanh & Bao, Vo Nguyen & Cho, Sungrae. (2022). A Contemporary Survey on Live Video Streaming from a Computation-Driven Perspective. ACM Computing Surveys. 54. 10.1145/3519552.

[2] Li, Xiangbo & Darwich, Mahmoud & Bayoumi, Magdy. (2020). A Survey on Cloud-Based Video Streaming Services. 10.48550/arXiv.2011.14976.

[3] Abdelhak Bentaleb, Zhengdao Zhan, Farzad Tashtarian, May Lim, Saad Harous, Christian Timmerer, Hermann Hellwagner, and Roger Zimmermann. 2022. Low Latency Live Streaming Implementation in DASH and HLS. In Proceedings of the 30th ACM International Conference on Multimedia (MM '22). Association for Computing Machinery, New York, NY, USA, 7343–7346. https://doi.org/10.1145/3503161.3548544

[4] Klink, Janusz. (2023). QoE-Driven Coding Bitrate Determination for Adaptive Video Streaming to Home Viewers. 10.23919/SoftCOM58365.2023.10271634.

[5] D. Ghadiyaram, J. Pan and A. C. Bovik, "A Subjective and Objective Study of Stalling Events in Mobile Streaming Videos," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 1, pp. 183-197, Jan. 2019, doi: 10.1109/TCSVT.2017.2768542.

[6] Taraghi, Babak & Haack, Selina & Timmerer, Christian. (2023). Towards Better Quality of Experience in HTTP Adaptive Streaming. 10.48550/arXiv.2301.13523.

[7] Lyko, Tomasz & Broadbent, Matthew & Race, Nicholas & Nilsson, Mike & Farrow, Paul & Appleby, Steve. (2023). Improving quality of experience in adaptive low latency live streaming. Multimedia Tools and Applications. 10.1007/s11042-023-15895-9.

[8] Tianchi Huang, Rui-Xiao Zhang, Chenglei Wu, and Lifeng Sun. 2023. Optimizing Adaptive Video Streaming with Human Feedback. In Proceedings of the 31st ACM International Conference on Multimedia (MM '23). Association for Computing Machinery, New York, NY, USA, 1707–1718. https://doi.org/10.1145/3581783.3611771

[9] T. Lyko, Y. Elkhatib, R. Ramdhany and N. Race, "Drop or Stop: Investigating the Impact of Playback Rate on QoE in Adaptive Video Streaming," 2024 16th International Conference on Quality of Multimedia Experience (QoMEX), Karlshamn, Sweden, 2024, pp. 111-117, doi: 10.1109/QoMEX61742.2024.10598252.

[10] Duanmu, Zhengfang & Liu, Wentao & Li, Zhuoran & Chen, Diqi & Wang, Zhou & Wang, Yizhou & Gao, Wen. (2020). Assessing the Quality-of-Experience of Adaptive Bitrate Video Streaming. 10.48550/arXiv.2008.08804.

[11] "ITU-T Rec. P.910 (04/2008) Subjective video quality assessment methods for multimedia applications." (2009).

[12] Stefan Lederer, Christopher Müller, and Christian Timmerer. 2012. Dynamic adaptive streaming over HTTP dataset. In Proceedings of the 3rd Multimedia Systems Conference (MMSys '12). Association for Computing Machinery, New York, NY, USA, 89–94. https://doi.org/10.1145/2155555.2155570

[13] Netflix, Sparks, special interest short film for HDR/HFR testing, 2016. [Online]. Available: https://www.netflix.com/title/80156943

[14] T. Lyko, Y. Elkhatib, R. Ramdhany and N. Race, "Differential QoE in Picture-in-Picture Gaming Videos: A Subjective Study," 2023 15th International Conference on Quality of Multimedia Experience (QoMEX), Ghent, Belgium, 2023, pp. 221-223, doi: 10.1109/QoMEX58391.2023.10178572.

[15] D. Ashimov, M. G. Martini and N. Barman, "Quality Assessment of Gaming Videos Compressed via AV1," 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX), Athlone, Ireland, 2020, pp. 1-4, doi: 10.1109/QoMEX48832.2020.9123112.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

Analiza i optymalizacja modelu sztucznej inteligencji do odwzorowania sylwetki człowieka na podstawie obrazu z kamery internetowej

Analysis and optimization of an artificial intelligence model for a human pose estimation based on a webcam imagery

Łukasz Seratowicz[1]; Jan Derkacz[2]

[1] AGH, Kraków, seratowicz@student.agh.edu.pl
[2] AGH, Kraków, jan.derkacz@agh.edu.pl

DOI: 10.15199/59.2025.4.98

Streszczenie: W artykule przeanalizowano i zoptymalizowano model YOLOv8-Pose do estymacji pozy ludzkiej w czasie rzeczywistym. Oceniono wydajność i dokładność różnych wariantów modelu, a także zbadano ich odporność na przypadki brzegowe, takie jak rotacja obrazu. Przedstawiono również rozszerzenie systemu o klasyfikator pozy jogi (Random Forest), który dzięki analizie kątów między stawami jest odporny na położenie osoby w kadrze i osiągnął skuteczność 95% (F1-Score).

Abstract: The paper analyzes and optimizes the YOLOv8-Pose model for real-time human pose estimation. Performance and accuracy of different model variants were evaluated, and their robustness to edge cases, such as image rotation, was also investigated. Furthermore, a system extension featuring a yoga pose classifier (Random Forest) is presented. This classifier, leveraging inter-joint angle analysis for invariance to the person's position within the frame, achieved an F1-Score of 95%.

Słowa kluczowe: Detekcja w czasie rzeczywistym, Estymacja pozy ludzkiej, Random Forest, Sztuczna inteligencja, YOLOv8-Pose

Keywords: Artificial Intelligence, Human Pose Estimation, Random Forest, Real-Time Detection, YOLOv8-Pose

1. WSTĘP

Od czasu, gdy w 1991 roku na Uniwersytecie Cambridge uruchomiono pierwszą „kamerę internetową" – tzw. Trojan Room Coffee Pot Camera – służącą wyłącznie do zdalnego podglądu poziomu kawy w zbiorniku, wizja komputerowa doświadczyła gwałtownego rozwoju. Początkowo opierała się na prostych metodach przetwarzania obrazu, takich jak wykrywanie krawędzi czy progowanie, a w latach 2000 zyskała ogromny impuls dzięki wprowadzeniu SIFT i SURF do detekcji cech. Rewolucję zapoczątkował jednak gwałtowny wzrost zainteresowania głębokimi sieciami neuronowymi po 2012 roku, gdy AlexNet znacząco przewyższył konkurencję na konkursie ImageNet, otwierając drzwi do powszechnego stosowania Convolutional Neural Networks. Współczesne podejścia do analizy obrazu bazują na zaawansowanych architekturach głębokiego uczenia, takich jak ResNet, YOLO, modele oparte na Transformerach czy sieci generatywne GAN. Dzięki ich rosnącej wydajności i elastyczności, systemy widzenia komputerowego znalazły zastosowanie w szerokim spektrum dziedzin – od pojazdów autonomicznych i diagnostyki medycznej, przez robotykę przemysłową i systemy nadzoru, po interaktywne aplikacje wykorzystujące rzeczywistość rozszerzoną i wirtualną. Przekształcenie komputerowego „widzenia" z eksperymentalnego laboratorium w kluczową technologię wielu sektorów pokazuje, jak dalece zmieniło się nasze postrzeganie i wykorzystanie danych wizualnych w ciągu niespełna trzech dekad.

2. CEL I ZAKRES BADAŃ

2.1 Scenariusze i kryteria ewaluacji modeli

Jednym z obszarów wizji komputerowej jest estymacja pozycji człowieka – zdolność do wykrywania kluczowych punktów anatomicznych (takich jak łokcie, kolana czy nadgarstki) i rekonstrukcji cyfrowego „szkieletu". Technologia ta ma szerokie zastosowania: od inteligentnych systemów monitoringu wideo, systemów wspomagających rehabilitację medyczną, po zastosowania militarne. Warto podkreślić, że rozwiązania oparte na różnych modelach można łączyć szeregowo, uzyskując hybrydowe architektury o rozszerzonych możliwościach.

Jako kluczowe założenia badawcze przyjęto:

Ewaluację wydajności, dokładności i pewności modelu YOLOv8-Pose na autorskim zbiorze danych, odzwierciedlającym różnorodne warunki oświetleniowe jak i jakości pozyskane z różnych kamer.

Analizę przypadków brzegowych („edge cases"), obejmującą skalowane, obrócone lub przechylone ujęcia, obrazy czarno-białe, analiza odporności modelu na prześwietlenia i niedoświetlenia obrazu, oraz wpływ kompresji wideo na precyzję i pewność detekcji.

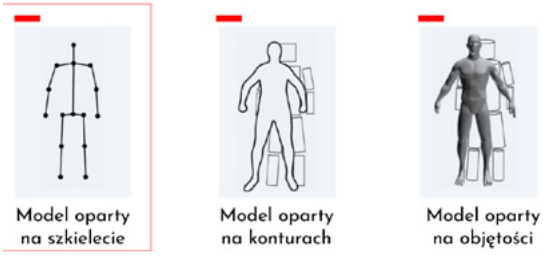
Praktyczną rozbudowę modelu, w tym opracowanie mechanizmu odpornego na zmienne położenie i orientację osób w kadrze, służącej do precyzyjnej i szybkiej detekcji rodzaju pozy na podstawie modelu Random Forest.

2.2 Metody estymacji pozy człowieka

W przeciwieństwie do tradycyjnych detektorów obiektów, które jedynie wyznaczają prostokątne ramki

(bounding boxes) wokół wykrytych elementów, systemy estymacji pozy człowieka dostarczają znacznie bogatsze informacje. W literaturze wyróżnia się trzy dominujące podejścia do analizy postawy (Rysunek 1):

- **szkieletowe** (skeleton-based) – zwracające uporządkowane współrzędne punktów stawów i relacji między nimi;
- **konturowe** (contour-based) – skupiające się na wykryciu zarysu sylwetki i segmentacji obszarów ciała;
- **wolumetryczne** (volumetric) – budujące pełne reprezentacje 3D na podstawie wielu strumieni wideo lub map głębokości.



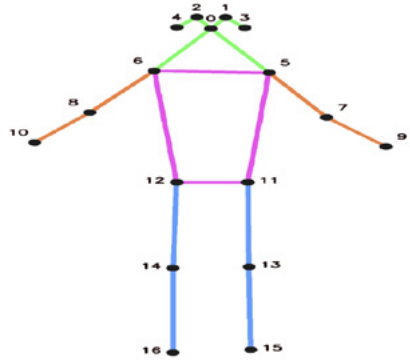
Rysunek 1: Modele Ciała Człowieka

Model YOLOv8-Pose, opracowany przez zespół Ultralytics, wykorzystuje podejście szkieletowe w architekturze, łącząc w jednym etapie detekcję osób z precyzyjną lokalizacją punktów anatomicznych –współrzędne kluczowych stawów (keypoints) oraz ich wzajemne powiązania, co umożliwia odtworzenie cyfrowego „szkieletu" osoby. Rodzina modeli obejmuje warianty o różnej skali (Tabela 1) – od najmniejszego „n" (3,3 M parametrów, 6,5 MB) po największy „x" (69,4 M parametrów, 133 MB) – co pozwala dobrać model do ograniczeń sprzętowych i wymagań wydajnościowych.

Tabela 1: Warianty modeli YOLOv8-Pose

Model	*Param. [M]*	*Plik [MB]*
YOLOv8**n**	3.3	6.51
YOLOv8**s**	11.6	22.42
YOLOv8**m**	26.4	50.8
YOLOv8**l**	44.4	85.26
YOLOv8**x**	69.4	132.96

Wszystkie warianty modelu YOLOv8-Pose przyjmują obraz wejściowy o stałej rozdzielczości 640×640 pikseli, niezależnie od oryginalnych wymiarów zdjęcia — biblioteka Ultralytics automatycznie skaluje dowolne wejście do tego formatu. Model zwraca dokładnie 17 punktów kluczowych w kolejności określonej przez twórców. Dla przejrzystości i ułatwienia analizy wyników opracowaliśmy graficzne przedstawienie zwracanych punktów (Rysunek 3). Każdy punkt opatrzono numerem, a segmenty odpowiadające nodze, ręce, tułowi i głowie wyróżniono odrębnymi kolorami.



Rysunek 2: Graficzne przedstawienie zwracanych przez model punktów

3. EWALUACJA WARIANTÓW MODELI YOLOV8-POSE

3.1 ZBIÓR DANYCH

Zgromadziliśmy zestaw ponad 1 000 zdjęć wykonanych ręcznie w zróżnicowanych warunkach oświetleniowych, przy użyciu różnych modeli kamer (smartfony, kamery internetowe, kamery sportowe, aparaty cyfrowe). Wszystkie obrazy miały rozdzielczość większą niż domyślny rozmiar wejściowy (640×640), co pozwoliło na przetestowanie wewnętrznego mechanizmu skalowania biblioteki YOLO oraz ocenę wpływu interpolacji na jakość detekcji. Każde zdjęcie zostało ręcznie zadnotowane 17 punktami kluczowymi w wykonanym przez nas do tego programie — w sumie oznaczyliśmy ponad 17 000 punktów stawów, co stanowiło istotnie czasochłonny etap pracy.

3.2 METODOLOGIA EWALUACJI

Ewaluację modeli podzieliliśmy na trzy główne kryteria:

Pomiar prędkości detekcji: Aby zminimalizować wpływ czynników zewnętrznych (temperatura CPU/GPU, obciążenie systemu), każdy test wykonywaliśmy kilkukrotnie, a za wynik ostateczny przyjmowaliśmy wartość średnią czasów inferencji. Szczególną uwagę zwróciliśmy na najmniejsze warianty (n, s), których bardzo krótki czas interferencji jest najbardziej podatny na błędy.

Odsetek niewykrytych punktów: Dla każdej klatki porównaliśmy liczbę wykrytych punktów kluczowych N_{det} z oczekiwanym $N_{\text{exp}} = 17$. W przypadku całkowitego braku detekcji $N_{\text{det}} = 0$, zaś przy braku tylko wybranych stawów była to różnica $17 - N_{\text{det}}$. Dla zbioru wszystkich danych testowych (1), była to suma różnic wszystkich różnic detekcji w stosunku do wszystkich ilości cech.

$$\text{Pewność} = \frac{\sum_{n=1}^{N}(N_{\text{exp}} - N_{\text{det},n})}{N \times N_{\text{exp}}} \times 100\% \qquad (1)$$

gdzie: N - liczba zdjęć badanych, $N_{\text{det},n}$ – liczba punktów rozpoznanych przez model dla zdjęcia n

Precyzja lokalizacji: Precyzję lokalizacji definiujemy jako średnią odległość euklidesową obliczoną dla

wszystkich punktów wykrytych przez model (punkty niewykryte są pomijane). Dla każdego wykrytego punktu $(x_{p,i}, y_{p,i})$ i odpowiadającego mu wzorcowego $(x_{t,i}, y_{t,i})$ normalizowanego względem szerokości W i wysokości H obrazu w pikselach obliczamy

$$d_i = \sqrt{\left(\frac{x_{p,i}-x_{t,i}}{W}\right)^2 + \left(\frac{y_{p,i}-y_{t,i}}{H}\right)^2}. \quad (2)$$

Ostatecznie:

$$\text{Precyzja} = \frac{1}{N_{\text{det}}} \sum_{i=1}^{N_{\text{det}}} d_i, \quad (3)$$

gdzie N_{det} to liczba rzeczywiście wykrytych punktów.

3.3 WYNIKI

W tabelach 2–4 zestawiono odpowiednio czas inferencji, pewność detekcji oraz precyzję lokalizacji, zdefiniowanych w sekcji 3.2.

Wzrastająca skala wielkości modeli przekłada się na istotne, choć malejące przyrosty dokładności (Tabela 2): największy skok odnotowaliśmy między najmniejszymi wariantami: `n` do `s`, podczas gdy różnica między największymi: `m` do `x` okazała się minimalna. Dokładność modeli utrzymywała się w granicach 1–1,7 %, co oznacza, że wszystkie modele poradziły sobie bardzo dobrze z identyfikacją położeń wykrytych punktów. W praktyce wartość poniżej 1 % najprawdopodobniej jest porównywalna z błędem powstałym podczas ręcznych adnotacji.

Tabela 2: Wyniki Podstawowych Badań Modeli

Model	*Czas [s]*	*Precyzja [%]*
YOLOv8**n**	49.68	1.67
YOLOv8**s**	64.22	1.28
YOLOv8**m**	76.42	1.22
YOLOv8**l**	77.36	1.11
YOLOv8**x**	82.3	1.06

Warianty `n` i `l` charakteryzowały się największą pewnością, czyli najniższym odsetkiem nierozpoznanych punktów (Tabela 3) - liczba niewykrytych punktów w relacji do łącznej możliwej puli. Wszystkie modele w tej kategorii mieściły się w granicy 4 do 5 %.

Tabela 3: Wyniki Podstawowych Badań Modeli

Model	*Pewność [%]*
YOLOv8**n**	4.08
YOLOv8**s**	4.44
YOLOv8**m**	4.57
YOLOv8**l**	4.15
YOLOv8**x**	4.94

Wariant `l` zasługuje na szczególną uwagę: oferuje lepszą pewność detekcji niż największy model `x`, działa szybciej i tylko nieznacznie ustępuje mu pod względem dokładności. Model `n` stanowi optymalny wybór przy bardzo rygorystycznych wymaganiach sprzętowych, chociaż zdarza mu się generować fałszywe pozy (tzw. halucynacje), co obniża jego precyzję. Wariant `s` łączy z kolei szybkość `n` z lepszą, lecz przeciętną precyzją i pewnością. Jeśli celem jest maksymalna precyzja niezależnie od czasu rzeczywistego, najlepiej radzi sobie model `x`, a gdy poszukujemy kompromisu między pewnością detekcji a wydajnością, najlepszym wyborem będzie model `l` lub `s` w zależności od mocy obliczeniowej urządzenia.

3.4 PRZYPADKI BRZEGOWE

Aby zweryfikować odporność modeli na nietypowe warunki, przygotowaliśmy rozszerzony zbiór ponad 22 000 obrazów syntetycznie zmodyfikowanych w następujący sposób: ręczne skalowanie do wejściowej rozdzielczości 640×640 (z zachowaniem proporcji), konwersja do odcieni szarości (luminancja, dithering i binaryzacja), zmiany jasności (underexposure 0,25×–0,75×, overexposure 1,25×–2,0×), kompresja JPEG (jakość od 10 % do 100 %) oraz obrót obrazu o losy kąt. Każdy z obrazów przetworzyliśmy wszystkimi pięcioma wariantami YOLOv8-Pose.

Jednym z ciekawszych rezultatów było obracanie obrazów o losowy kąt (Tabela 4). Każdy obraz był najpierw poddawany losowej rotacji, a następnie zapisywany w oddzielnym folderze, co zapewniało spójność i umożliwiało bezpośrednie porównanie wyników między poszczególnymi wariantami modeli. Takie przekształcenie znacząco obniżyło pewność modeli: wariant `n` odnotował największy spadek pewności (ponad 20% punktów nierozpoznanych), `s` pogorszył się niemal dwukrotnie, a `m` i `x` straciły około 25% względem wyników na obrazach oryginalnych. Jedynie model `l` wykazał się względną odpornością na rotację, zachowując wysoką pewność detekcji przy minimalnym pogorszeniu wyników. Wahania precyzji mieściły się w granicach 2–4% — w porównaniu z 1–1,5% dla obrazów bez rotacji.

Tabela 4: Wyniki Dla Rotacji

Model	*Precyzja [%]*	*Pewność [%]*
YOLOv8**n**	3.59	20.73
YOLOv8**s**	3.08	9.93
YOLOv8**m**	2.98	7.43
YOLOv8**l**	2.42	5.97
YOLOv8**x**	2.32	7.02

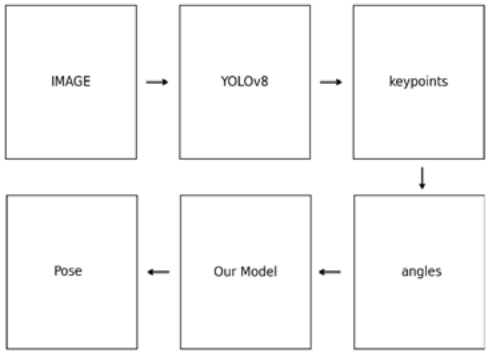
Obracanie obrazu pod dowolnym kątem znacząco zwiększa liczbę niewykrytych punktów, zwłaszcza w modelach o mniejszej liczbie parametrów. Jedną z możliwych przyczyn jest brak obróconych próbek w zbiorze treningowym, co spowodowało dezorientacje modelu. Aby zminimalizować ten efekt, wskazane jest zastosowanie wstępnej korekcji przy użyciu na przykład transformacji macierzowych. Dzięki temu model YOLOv8-Pose operuje na obrazie bez zakłóceń wynikających z niezamierzonych rotacji, co znacząco poprawia stabilność i precyzję estymacji pozycji. Jest to szczególnie istotne w przypadku mniejszych wariantów (np. `n`), dla których spadek dokładności detekcji jest najbardziej odczuwalny. Wariant `l` kolejny raz wyróżnia się jako optymalny kompromis — zapewnia niemal taką samą precyzję jak `x`, przy jednoczesnym skróceniu czasu inferencji i utrzymaniu najwyższej pewności wykryć.

4. ROZWINIĘCIE MODELI

4.1 Klasyfikacja rodzaju pozy

Detekcja kluczowych punktów anatomicznych za pomocą YOLOv8-Pose może służyć znacznie szerszym celom niż jedynie wizualizacja szkieletu. Aby pokazać potencjał takiego podejścia, zdecydowaliśmy się rozbudować standardowy model YOLOv8-Pose o moduł klasyfikacji typu pozy.

Naszym celem było stworzenie systemu rozróżniającego sześć wybranych pozycji jogi, wykorzystując otwarty zbiór danych, ok. 1200 obrazów. Wybraliśmy jogę ze względu na dostępność wspomnianego datasetu. Sama metodologia może zostać zaadaptowana do dowolnego scenariusza, np. analiz treningów, ćwiczeń rehabilitacyjnych czy detekcji pozy w systemach interakcji człowiek-komputer.



Rysunek 3: Schemat przetwarzania: detekcja punktów kluczowych → obliczanie kątów między segmentami → klasyfikacja Random Forest

W pierwszym kroku każdy obraz przechodzi przez model YOLOv8-Pose, który dostarcza współrzędne 17 punktów kluczowych. Następnie obliczamy kąty między wybranymi segmentami ciała (np. kąt kostka-kolano-udo), co pozwala na zamianę współrzędnych kartezjańskich na wektor kątowy niezależny od pozycji osoby w kadrze. Taki zestaw ośmiu wybranych kątów stanowi wejście do prostego, lecz efektywnego klasyfikatora Random Forest.

Do trenowania klasyfikatora wykorzystaliśmy zbiór 1200 próbek, dzieląc go w proporcji 85/15 na zbiór treningowy i testowy. Uzyskaliśmy dokładność (4) klasyfikacji (f1-score) na poziomie 95 %, co pokazuje skuteczność modelu Random Forest.

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

gdzie: Precision – precyzja, czyli odsetek poprawnych pozytywnych predykcji, Recall – czułość, czyli odsetek poprawnie wykrytych przypadków pozytywnych.

4.2 Wyniki dla rozwiniętych modeli

Zaproponowane rozwiązanie cechuje się wysoką skutecznością i wydajnością. Dzięki kompaktowej architekturze modelu YOLOv8-Pose oraz niewielkiej złożoności klasyfikatora Random Forest, cały system jest zdolny do działania w czasie rzeczywistym, przy zachowaniu wysokiej dokładności klasyfikacji.

Ze względu na charakter pracy klasyfikacja pozy nie była analizowana pod kątem częściowych braków danych. Random Forest wymaga ich wcześniejszego uzupełnienia lub zastosowania odpowiednich metod imputacji, co stanowi kolejny obszar do dalszych badań.

Rezultaty sugerują, że zastosowana metoda przetwarzania punktów kluczowych oraz klasyfikacji typu pozy ma potencjał adaptacyjny do innych obszarów. Przykładowo, może być użyteczna w systemach dla osób niewidomych do analizy otoczenia, systemach wspomagania rehabilitacji, nadzorze nad poprawnością wykonywania ćwiczeń fizycznych, czy w kontekście interakcji człowiek–komputer w aplikacjach opartych na rozpoznawaniu ruchu.

5. PODSUMOWANIE

Przeprowadzona analiza i badania modelu YOLOv8-Pose potwierdziły jego wysoką skuteczność w zadaniu estymacji pozy ludzkiej w czasie rzeczywistym. Ewaluacja różnych wariantów modelu na autorskim zbiorze danych wykazała kompromisy między szybkością, precyzją i pewnością detekcji. Model YOLOv8l-Pose wyróżnił się jako optymalny wybór, oferując najlepszy balans między tymi metrykami, zwłaszcza w kontekście odporności na trudne warunki, takie jak rotacja obrazu, która okazała się istotnym wyzwaniem dla mniejszych wariantów.

Zbadano również praktyczne zastosowanie wyników estymacji, opracowując system klasyfikacji sześciu pozycji jogi. Wykorzystanie kątów między kluczowymi punktami ciała jako cech dla klasyfikatora Random Forest pozwoliło na osiągnięcie wysokiej skuteczności (95% F1-Score) przy zachowaniu wydajności obliczeniowej i odporności pozycji człowieka na obrazie. Pokazuje to potencjał łączenia YOLOv8-Pose z innymi modelami do tworzenia bardziej złożonych systemów analizy ruchu.

Wyniki pracy wskazują, że YOLOv8-Pose jest wszechstronnym narzędziem, które może być efektywnie stosowane w różnorodnych aplikacjach – od systemów bezpieczeństwa, przez analizę sportową, aż po interaktywne systemy rozrywkowe czy wspomagające rehabilitację, nawet na urządzeniach o ograniczonych zasobach obliczeniowych.

LITERATURA

[1] Cordova-Esparza Diana, Terven Juan. 2024. „A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS". Machine Learning and Knowledge Extraction (MAKE)

[2] Redmon Joseph, Divvala Santosh, Girshick Ross, Farhadi Ali. 2016. „You Only Look Once: Unified, Real-Time Object Detection". ArXiv preprint 1506.02640v1.

[3] Albawi Saad, Tareq Abed Mohammed, and Saad Al-Zawi. 2017. „Understanding of a convolutional neural network". 2017 International Conference on Engineering and Technology (ICET), 1–6. DOI: 10.1109/ICEngTechnol.2017.8308186.

[4] Shannon Claude E. 1950. „Programming a Computer for Playing Chess". Philosophical Magazine, Ser. 7, Vol. 41, No. 314. (March 1950).

[5] Li Danyu, Yunci Ma, Jiajun Zou, and Zhiming Xu. 2024. „An Improved Lightweight Human Normative Pose Estimation Algorithm". 7th International Conference (CISAT), Hangzhou, China. IEEE. DOI: 10.1109/CISAT62382.2024.10695434.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

ANALIZA MOŻLIWOŚCI ZRÓWNOLEGLENIA OBLICZEŃ W KODERZE I DEKODERZE JPEG

ANALYSIS OF THE POSSIBILITY OF PARALLELIZING COMPUTATIONS IN JPEG ENCODER AND DECODER

Jakub Stankowski, Adrian Śliwiński, Adam Grzelka, Tomasz Grajek

Instytut Telekomunikacji Multimedialnej, Politechnika Poznańska, Poznań
{jakub.stankowski, adam.grzelka, tomasz.grajek}@put.poznan.pl

DOI: 10.15199/59.2025.4.99

Streszczenie: **Artykuł przedstawia metodę przyspieszenia kodera i dekodera JPEG poprzez zrównoleglenie przetwarzania oraz opcjonalne zastosowanie znaczników restartu. Przeanalizowano potencjalne problemy i możliwe podejścia, zaprezentowano szczegóły implementacji z wykorzystaniem podziału na niezależnie kodowane fragmenty oraz ocenę wydajności, która wykazała znaczące skrócenie czasu kodowania i dekodowania. Przyspieszenie nie miało wpływu na sprawność kompresji. Proponowane podejście pozwala na przyspieszenie kodowania z użyciem współczesnych wielordzeniowych procesorów.**

Abstract: **The paper presents a method for accelerating the JPEG encoder and decoder through parallel processing and the optional use of restart markers. Potential issues and possible approaches were analyzed, implementation details utilizing division into independently encoded segments were presented, and performance evaluation demonstrated significant reductions in encoding and decoding times. The acceleration did not affect compression efficiency. The proposed approach enables faster encoding using modern multicore processors.**

Słowa kluczowe: **kompresja obrazu, JPEG, kodowanie równoległe, znaczniki restartu.**

Keywords: image compression, JPEG, parallel encoding, restart markers.

1. WSTĘP

Technika JPEG [6, 14], pomimo że istnieje na rynku od ponad 30 lat, jest nadal bardzo popularnym kodekiem do kompresji obrazów statycznych. Największą jej zaletą jest ogromna popularność i kompatybilność – praktycznie każde urządzenie i oprogramowanie radzi sobie z tym formatem kompresji [15].

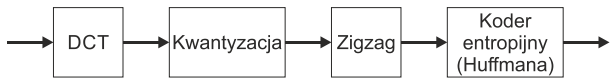
Technika JPEG cechuje się niską złożonością obliczeniową w porównaniu do nowszych technik takich jak HEVC Intra (HEIF) [9].Warto jednak zauważyć, że istnieją urządzenia, które operują na obrazach o rozmiarze >50Mpix, takie jak cyfrowe lustrzanki (DLSR) i nadal stosują technikę JPEG. Kodowanie i dekodowanie tak dużego obrazu z użyciem jednego rdzenia nadal jest czasochłonne.

Technika JPEG powstawała w czasach kiedy nikt nie brał pod uwagę zrównoleglenia obliczeń z podziałem na wiele rdzeni. W literaturze można znaleźć prace związane z problemem zrównoleglenia obliczeń w technice JPEG [2, 11, 12], jednak są to analizy teoretyczne lub próby implementacji nie całkowicie zgodne z normą.

Współczesnym trendem jest wyposażanie komputerów (i urządzeń mobilnych) w wielordzeniowe procesowy posiadające często kilkanaście (lub więcej) jednostek. W związku z tym istotna jest analiza czy możliwe jest wykorzystanie owych wielordzeniowych procesorów do przyspieszenia kodowania i dekodowania.

2. ZRÓWNOLEGLENIE OBLICZEŃ W KODEKU JPEG

Typowy koder JPEG składa się z czterech najistotniejszych etapów przetwarzania: transformacji DCT, kwantyzacji, uporządkowania zygzakowatego i kodera entropijnego (Rys. 1.). Twórcy techniki JPEG przewidzieli możliwość zastosowania dwóch różnych koderów entropijnych – kodera Huffmana i arytmetycznego. W praktyce jednak można przyjąć, że wariant z koderem arytmetycznym nie jest często stosowany – głównie ze względu na istniejące przez wiele lat utrudnienia prawne i licencyjne. Lista patentów dotyczących kodowania arytmetycznego w JPEG jest zawarta w normie [6] (aneks L).

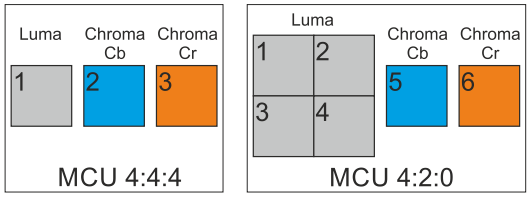


Rys. 1. Schemat blokowy kodera JPEG

W przypadku trzech pierwszych etapów przetwarzania zrównoleglenie obliczeń jest relatywnie łatwe. Zarówno transformacja DCT, kwantowanie, jak i zmiana kolejności współczynników odbywa się w blokach o rozmiarze 8x8. Te bloki mogą być przetwarzane niezależnie od siebie, co stanowi ogromne ułatwienie w przypadku implementacji wielowątkowej. Przykładowym (i stosowanym przez autorów) podejściem jest podzielenie obrazu na kolejne wiersze bloków 8x8 i potraktowanie każdego wiersza jako osobnego zadania obliczeniowego.

Zdecydowanie gorzej wygląda sytuacja związana z próbą zrównoleglenia kodowania Huffmana.

JPEG definiuje strukturę nazywaną „minimalną jednostką kodowania” (Minimum Coded Unit – MCU) [6]. Rozmiar i budowa jednostki MTU zależy od wybranego próbkowania chrominancji. Przykładowo, dla formatu 4:4:4 jednostka MCU ma rozmiar 8x8 i zawiera po jednym bloku luminancji i obu chrominancji. Z kolei dla popularnego formatu 4:2:0 jednostka MCU ma rozmiar 16x16 i zawiera cztery bloki luminancji i po jednym bloku dla obu chrominancji (Rys. 2.).

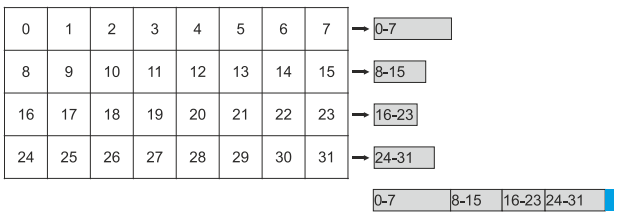


Rys. 2. Typowe jednostki MCU (Minimum Coded Unit) w JPEG dla obrazu kodowanego z próbkowaniem chromy 4:4:4 i 4:2:0

Kodowanie entropijne realizowane jest przez przetwarzanie kolejnych jednostek MCU zaczynając od składowej luminancji. Kolejność przetwarzania bloków została zaznaczona na Rys. 2. Co więcej, współczynniki DC (składowa stała dla każdego bloku) są kodowane predykcyjnie, w ramach danego komponentu. Wysyłana jest różnica pomiędzy obecnie kodowanym współczynnikiem DC a współczynnikiem poprzednio zakodowanego bloku tego samego komponentu. Powoduje to istnienie pewnej zależności pomiędzy kolejnymi blokami.

2.1. Zrównoleglenie kodowania Huffmana w koderze JPEG

Pomimo opisanych wcześniej niedogodności, możliwe jest zrównoleglenie kodowanie entropijnego przez zakodowanie wielu fragmentów obrazu, a następnie scalenie ich w jednym buforze. Ilustracja takiego rozwiązania została przedstawiona na Rys. 3. Podział obrazu na fragmenty powinien odbywać się na granicach MCU. Stosowane przez autorów podejście zakłada podział obrazu na wiersze jednostek MCU, co w przypadku próbkowania 4:2:0 odpowiada wycinkom obrazu o wysokości 16 pikseli.



Rys. 3. Schemat działania równoległego kodera JPEG i łączenia strumieni powstałych w wyniku kodowania poszczególnych fragmentów. Niebieskim kolorem zaznaczono wypełnienie do pełnych bajtów.

Wartości współczynników DC używane do predykcji można pobrać z poprzedzającego bloku (po wcześniejszym wyliczeniu jego położenia).

Niestety sama operacja łączenia zakodowanych fragmentów strumieni jest dość uciążliwa i czasochłonna. Główną przyczyną jest to, że kolejne fragmenty strumieni mają długość, która nie jest liczona w pełnych bajtach, co skutkuje koniecznością przeprowadzenia czasochłonnej manipulacji na pojedynczych bitach przy operacji łączenia. Po operacji łączenia konieczne jest zrealizowanie kolejnego kroku – analizy i unikania emulacji znaczników (markerów). Oba wspomniane kroki muszą być zrealizowane szeregowo przez jeden wątek, a to, zgodnie z prawem Amdahla [4], zmniejsza potencjalne zyski ze zrównoleglenia.

3. ZASTOSOWANIE ZNACZNIKÓW RESTARTU

Wiele technik kompresji zawierały w sobie rozwiązania poprawiające odporność na błędy transmisji. Przykładem jest tu H.263 Annex H („Forward Error Correction for coded video signal”) [7]. Popularnym rozwiązaniem jest również podział obrazu na niezależne fragmenty – plastry (slice) stosowane między innymi w MPEG-4 AVC/H.264 [8], HEVC/H.265 [9] czy VVC H.266 [10].

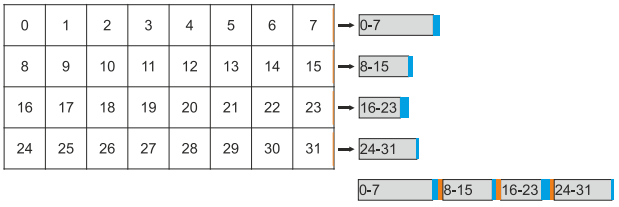
Technika JPEG również przewiduje takie rozwiązanie, a nazywane jest ono znacznikami restartu (Restart Markers – RST). RST zostały przewidziane jako sposób poprawy odporności na błędy transmisji. Pozwalają wznowić dekodowanie od kolejnego znacznika i uniknąć sytuacji gdzie jeden błąd powoduje utratę całego obrazu.

Zgodnie z normą [6] możliwe jest zdefiniowane wartości Restart Interval (RI). W takim przypadku RST jest wstawiany do strumienia co RI jednostek MTU. Co więcej, znaczniki restartu są wyrównane do pełnych bajtów, więc w razie potrzeby są dodatkowo poprzedzane odpowiednią liczbą bitów (jedynek).

Ze względu na analogię do funkcjonalności zdefiniowanej w MPEG-4 AVC/H.264 [8], poszczególne fragmenty obrazu oddzielone RST będą nazywane plastrami. W technice JPEG plastry są kodowane całkowicie niezależnie – RST zrywa predykcję DC i wymusza dodanie wypełnienia do pełnych bajtów do poprzedzającego plastra.

Obie powyższe cechy powodują, że możliwe jest zastosowanie RST niezgodnie z intencją twórców – do zrównoleglenia obliczeń.

Zachowując ten sam sposób podziału jak opisany w podrozdziale 2.1. autorzy przyjęli wartość RI równą liczbie MCU w jednym wierszu (Rys. 4.).

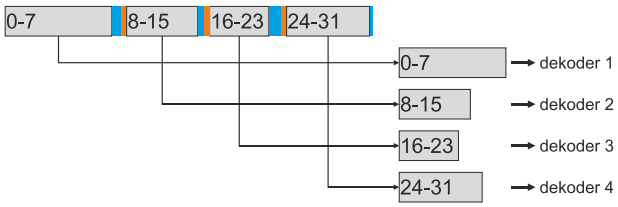


Rys. 4. – Schemat działania równoległego kodera JPEG wykorzystującego znaczniki restartu. Niebieskim kolorem zaznaczono wypełnienie do pełnych bajtów a pomarańczowym – RST.

Zastosowanie podejścia ze znacznikami restartu umożliwia zrównoleglenie większej liczby operacji. Możliwe jest niezależne kodowanie poszczególnych plastrów oraz analiza i unikanie emulacji znaczników. Natomiast samo łączenie strumieni jest prostą i szybką operacją kopiowania bloków pamięci.

Oczywiście opisane powyżej podejście ma pewne wady. Największą z nich jest zmniejszenie sprawności kompresji. Owa utrata sprawności ma dwie przyczyny. Pierwszą jest zerwanie predykcji DC na granicy plastrów, skutkujące większą liczbą bitów zużytych na zakodowanie tego elementu składni. Drugim jest konieczność wysłania samych znaczników (i ewentualnie wyrównania do pełnych bajtów) co przekłada się na dodatkowy wzrost rozmiaru strumienia (bądź pliku).

Z drugiej strony nie można wspomnieć o jeszcze jednej zalecie użycie RST. Podział obrazu na plastry umożliwia jego równoległe dekodowanie. W takiej sytuacji dekoder może w pierwszej kolejności sparsować strumień/plik i wyszukać w nim wszystkie znaczniki. Następnie można wydzielić poszczególne plastry i każdy z nich dekodować niezależnie (Rys. 5.).

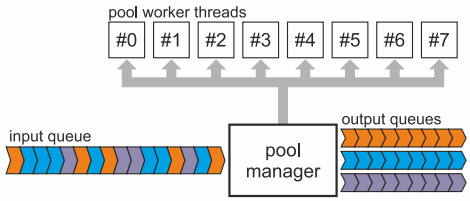


Rys. 5. Schemat działania zrównoleglonego dekodera przetwarzającego strumień ze znacznikami restartu

4. IMPLEMENTACJA

W ramach weryfikacji przedstawionych wcześniej koncepcji przygotowano implementację wielowątkowego kodeka JPEG. Kodek powstał jako rozwinięcie opracowanego przez autorów oprogramowania JOptEnc (https://github.com/jstankowski/joptenc).

Opracowane oprogramowanie pozwala na równoległe kodowanie zarówno w wariancie z użyciem znaczników restartu, jak i bez ich stosowania. Dodatkowo przygotowano dekoder, który pozwala na równoległe dekodowanie obrazów zawierających znaczniki restartu.



Rys. 6. Idea wzorca projektowego „pula wątków” (thread pool)

Oba moduły (koder i dekoder) wykorzystują własną implementację puli wątków (thread pool) w celu minimalizacji narzutu związanego z przekazywaniem zadań do dodatkowych wątków (Rys. 6.). Dokładny opis tej techniki można znaleźć w [13].

5. WERYFIKACJA EKSPERYMENTALNA

5.1. Metodologia

Eksperyment został przeprowadzony na bazie testowych sekwencji Netflix [1]. Zawiera ona 9 zróżnicowanych sekwencji w formacie FullHD. Przygotowane w ramach tej pracy oprogramowanie kodera zostało uruchomione dla trzech punktów pracy, związanych z siłą stosowanej kwantyzacji, odpowiednio dla $Q_L \in \{20, 25, 30, 35\}$, $Q_M \in \{50, 55, 60, 65\}$, $Q_H \in \{80, 85, 90, 95\}$. Warto zauważyć, że w technice JPEG parametr Q oznacza „jakość”. Wyższa wartość Q wiąże się ze słabszą kwantyzacją i większym strumieniem.

Test był przeprowadzany na komputerach z procesorem AMD Ryzen R9-7950X, który posiada 16 rdzeni. Każde kodowanie było realizowane niezależnie, czyli jeden eksperyment był realizowany w danym momencie. Pomiar czasu był wykonany z wykorzystaniem odczytu licznika TSC procesora, czyli instrukcji RDTSCP [5]. Eksperyment zrealizowano w trzech trybach pracy:

- Tryb A – tryb szeregowy, traktowany jako odniesienie do wyliczenia przyśpieszenia uzyskanego w wyniku zrównoleglenia.
- Tryb B – wykorzystujący zrównoleglenia, ale bez użycia znaczników restartu.
- Tryb C – wariant wielowątkowy, dodatkowo wykorzystujący znaczniki restartu.

W obu zrównoleglonych wariantach zastosowano podział obrazu na wiersze jednostek MCU, co w przypadku obrazów FullHD z próbkowaniem chromy 4:2:0 oznacza podział na 68 jednostek pokrywających obszar 1920x16 pikseli.

Warto zaznaczyć, że wersja wielowątkowa kodera i dekodera JPEG działa identycznie jak wariant jednowątkowy – w tym sensie, że zakodowany strumień i zdekodowany obraz są binarnie identyczne. W związku z tym samo zrównoleglenie nie wpływa na jakość zakodowanego obrazu, rozmiar strumienia czy sprawność kompresji.

5.2. Wpływ stosowania znaczników restartu

Pierwszym krokiem weryfikacji przedstawionej koncepcji jest określenie, jak samo użycie znaczników restartu wpływa na sprawność kompresji. Jak wynika z Tabeli 1, wpływ ten jest zauważalny, ale niezbyt istotny. W najgorszym przypadku – dla bardzo niskiej jakości obrazu – wiąże się z 2% wzrostem rozmiaru skompresowanego pliku. Jednak w bardziej realistycznym przypadku – dla wysokiej jakości – wzrost rozmiaru obrazu to tylko 0,6%. Wpływ użycia RST na czas kodowania jest minimalny i wynosi ~31µs przy czasie kodowania jednego obrazu wahającym się pomiędzy 5 a 10 ms na ramkę.

Tab. 1. Wpływ użycia RST na rozmiar zakodowanego obrazu dla RST wstawianego co jeden wiersz MCU

Punkt pracy	Średni wzrost rozmiaru jednego obrazu [B]	Średni względny wzrost rozmiaru jednego obrazu [%]	Średni wzrost czasu kodowania jednego obrazu [µs]
Q_L	179,13	1,950	
Q_M	183,34	1,252	31,24
Q_H	192,31	0,594	

5.3. Przyśpieszenie uzyskane przez zrównoleglenie kodeka

Wyniki zawierające ocenę wydajności implementacji wielowątkowej przedstawiono w Tabeli 2. Zauważalne jest, że przyspieszenie jest większe dla wyższego zakresu jakości. Jest to spodziewane zachowanie, gdyż złożoność obliczeniowa kodera i dekodera entropijnego jest bezpośrednio skorelowana z liczbą bitów na obraz. Można też zauważyć, że zastosowanie znaczników restartu pozwala na znaczne zwiększenie zysków wynikających z przetwarzania wielowątkowego.

Co istotne, zastosowanie znaczników restartu pozwala również na przyspieszenie procesu dekodowania, co nie jest możliwe w przypadku obrazów zakodowanych bez użycia RST.

Tab. 2. Przyśpieszenie kodowania i dekodowania uzyskane dla implementacji wielowątkowej

Punkt pracy	Przyśpieszenie kodera (tryb B)	Przyśpieszenie kodera + RST (tryb C)	Przyśpieszenie kodera RST (tryb C)
Q_L	3,90×	4,87×	2,94×
Q_M	3,78×	4,93×	3,03×
Q_H	3,42×	5,24×	3,42×

W Tabeli 3. zebrano wyniki dla zrównoleglonego kodera JPEG wykorzystującego technikę RDOQ, która jest zaimplementowana w JOptEnc. Warto zauważyć, że użycie RDOQ większa sprawność kompresji o około 10%, ale kosztem 10, do 20-krotnego wydłużenia czasu kodowania [3]. W tym przypadku użycie wielowątkowej implementacji pozwala na częściowe skompensowanie wzrostu złożoności obliczeniowej spowodowanego przez użycie RDOQ.

Tab. 3. Przyśpieszenie kodera używającego RDOQ uzyskane dla implementacji wielowątkowej

Punkt pracy	Przyśpieszenie kodera (tryb B)	Przyśpieszenie kodera + RST (tryb C)
Q_L	6,72×	6,80×
Q_M	7,56×	7,74×
Q_H	9,02×	9,31×

Jak pokazano w powyższych tabelach, możliwe jest znaczne przyspieszenie działania kodeka JPEG poprzez podział obliczeń na wiele wątków i wykorzystanie możliwości współczesnych procesorów wielordzeniowych. W skrajnym przypadku uzyskana prawie dziesięciokrotne przyśpieszenie obliczeń.

6. Podsumowanie

W artykule przedstawiono koncepcję zrównoleglenia obliczeń w kodeku JPEG. Przeanalizowano dwa podejścia do zrównoleglenia – z wykorzystaniem znaczników restartu oraz bez ich użycia. W celu weryfikacji opracowanej koncepcji przygotowano wielowątkową implementację kodeka JPEG i przeprowadzono eksperymenty mające na celu ocenę potencjalnego przyśpieszenia kodowania i dekodowania.

Wyniki eksperymentalne wskazują na to, że nawet technika kompresji, która nie powstawała z myślą o ułatwieniu zrównoleglenia obliczeń pozostawia twórcom kodeka pewne możliwości w kwestii zrównoleglenia. W opisywanym przypadku uzyskano ponad pięciokrotne przyśpieszenie obliczeń dla prostego kodera JPEG oraz prawie dziesięciokrotne dla kodera wykorzystującego technikę RDOQ. Pokazano również potencjał zrównoleglenia dekodera JPEG.

LITERATURA

[1] Bampis C. G., Li Z., Moorthy A. K., Katsavounidis I., Aaron A. Bovik A.C. 2016. „LIVE Netflix Video Quality of Experience Database„. Online: http://live.ece.utexas.edu/research/LIVE_NFLXStudy/index.html.

[2] Castells-Rufas D., Joven J., Carrabina J. 2010. „Scalability of a Parallel JPEG Encoder on Shared Memory Architectures”. *39th International Conference on Parallel Processing*. 502-507.

[3] Grajek T., Stankowski J. 2022. „Rate-Distortion Optimized Quantization in Motion JPEG”. *30th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG)*.

[4] Gustafson, J.L. 2011. Amdahl’s Law. In: Padua, D. (eds) *Encyclopedia of Parallel Computing*. Springer.

[5] Intel. 2015. „Intel 64 and IA-32 Achitectures Software Deceloper's Manual„. 2B : 4-302.

[6] ISO/ITU. 1992. „Information Technology — Digital Compression and Coding of Continuous-Tone Still Images — Requirements and Guidelines”, ISO 10918-1, także jako ITU T.81.

[7] ITU-T Rec. H.263:2005. „Video coding for low bit rate communication”. International Telecommunication Union.

[8] ITU-T Rec. H.264:2021. „Advanced video coding for generic audiovisual services„. International Telecommunication Union.

[9] ITU-T Rec. H.265:2023. „High efficiency video coding„. International Telecommunication Union.

[10] ITU-T Rec. H.266:2023. „Versatile video coding„. International Telecommunication Union.

[11] Monnes P., Furht B. 1994. „Parallel JPEG algorithms for still image compression” *SOUTHEASTCON '94*. 375-379.

[12] Shin H., Lee J. 2024. „Hardware Multi-Threaded System for High-Performance JPEG Decoding”. Journal of Signal Processing Systems. 96 : 67–79.

[13] Stankowski J., Grzelka A., Mieloch D., Wegner K. 2018. „Processing Pipeline for Real-Time Remote Delivery of Virtual View in FTV Systems”. *2018 International Conference on Signals and Electronic Systems (ICSES)*. 118-123

[14] Wallace G.K. 1992. „The JPEG still picture compression standard”, *IEEE Transactions on Consumer Electronics*, 38 (1) : xviii – xxxiv.

[15] W3Techs. 2025. „Usage statistics of JPEG for websites”, https://w3techs.com/technologies/details/im-jpeg.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

DETEKCJA MANIPULACJI W OBRAZACH WYSOKIEJ ROZDZIELCZOŚCI Z UŻYCIEM TRANSPARENTNEGO ZNAKU WODNEGO

High-Resolution Image Manipulation Detection via Transparent Watermarking

Paweł Duszejko[1]; Zbigniew Piotrowski[2]

[1] Wojskowa Akademia Techniczna, Warszawa, pawel.duszejko@wat.edu.pl
[2] Wojskowa Akademia Techniczna, Warszawa, zbigniew.piotrowski@wat.edu.pl

DOI: 10.15199/59.2025.4.100

Streszczenie: **W artykule przedstawiamy system detekcji semantycznie istotnych manipulacji w obrazach wysokiej rozdzielczości z wykorzystaniem transparentnego znaku wodnego. Metoda opiera się na głębokiej sieci neuronowej do osadzania niewidocznego znacznika oraz binarnym klasyfikatorze, który ocenia, czy wykryte zmiany mają znaczenie semantyczne. W architekturę systemu wkomponowano reprezentację obrazu w rozdzielczości pośredniej, co wpływa korzystnie na jego efektywność i odporność. Skuteczność proponowanego podejścia potwierdzamy wynikami eksperymentalnymi oraz analizą ilościową.**

Abstract: **In this paper, we present a system for detecting semantically meaningful manipulations in high-resolution images using a transparent watermark. The method employs a deep neural network to embed an imperceptible watermark and a second binary classifier to assess whether any alterations are semantically significant. The design incorporates an intermediate-resolution representation, which contributes to both the efficiency and robustness of the system. We validate the effectiveness of the approach through experimental results and quantitative evaluation.**

Słowa kluczowe: **znak wodny, detekcja manipulacji, obrazy wysokiej rozdzielczości, sieci neuronowe, analiza semantyczna, klasyfikacja binarna, przetwarzanie obrazu.**

Keywords: **watermarking, manipulation detection, high-resolution images, neural networks, semantic analysis, binary classification, image processing.**

1. WSTĘP

Współczesna przestrzeń cyfrowa obfituje w obrazy wysokiej rozdzielczości, wykorzystywane zarówno w komunikacji codziennej, jak i w zastosowaniach profesjonalnych, takich jak archiwizacja, monitoring czy diagnostyka medyczna. Wraz z łatwością edytowania i rozpowszechniania treści wizualnych rośnie potrzeba narzędzi umożliwiających automatyczne wykrywanie przekształceń, które mogą wpływać na przekaz lub wiarygodność obrazu. Kluczowe staje się zatem opracowanie rozwiązań zdolnych do odróżniania zmian semantycznie istotnych od tych, które nie mają znaczenia z punktu widzenia odbiorcy [1].

W niniejszym artykule prezentujemy system oparty na transparentnym, półkruchym znaku wodnym, który pełni rolę sygnatury umożliwiającej detekcję ingerencji w treść obrazu. Zaprojektowany w taki sposób, aby pozostać odpornym na nieszkodliwe operacje przetwarzania (takie jak rozmycie, szum Gaussa czy zmiany jasności i kontrastu), system jednocześnie wykazuje wysoką czułość na istotne manipulacje takie jak wycinanie fragmentów, ich podmiana czy zmiany skali. Rozwiązanie zostało opracowane przy wykorzystaniu współpracy dwóch głębokich sieci neuronowych: jednej odpowiedzialnej za osadzanie znacznika, a drugiej za ocenę charakteru wykrytych zmian. Proponowana architektura została zaprojektowana z myślą o efektywności i skalowalności, a jej skuteczność potwierdzamy eksperymentalnie.

2. ARCHITEKTURA SYSTEMU

2.1. Przepływ danych

Aby opracować rozwiązanie odpowiadające na problem przedstawiony we wstępie, konieczne jest zobrazowanie typowego przebiegu przetwarzania obrazu w kontekście jego publikacji i późniejszej weryfikacji. Wysokopoziomowy schemat takiego toru zawiera trzy zasadnicze etapy:

1. utworzenie lub pozyskanie obrazu,
2. umieszczenie go w publicznie dostępnym zasobie, takim jak media społecznościowe lub inne otwarte repozytorium,
3. pobranie obrazu oraz wynikająca z tego potrzeba potwierdzenia jego autentyczności i integralności.

Proponowane w niniejszym artykule rozwiązanie zostało zaprojektowane tak, aby w sposób syntetyczny odzwierciedlać tę sytuację, umożliwiając tym samym testowanie i rozwój systemu zdolnego do działania w rzeczywistych warunkach.

W związku z tym cały algorytm podzielono na trzy zasadnicze komponenty, zgodnie z podejściem powszechnie stosowanym w dziedzinie znakowania wodnego [2,3]:

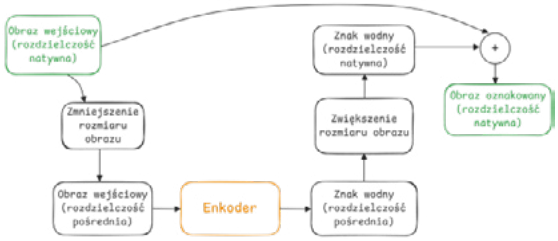
- **enkoder** który osadza transparentny znak wodny w obrazie wejściowym,
- **blok zakłóceń** modelujący typowe przekształcenia zachodzące w trakcie rozpowszechniania obrazu w otwartym środowisku (takie jak kompresja, filtrowanie czy zmiany jasności),
- **detektor**, którego zadaniem jest analiza pobranego obrazu oraz ocena czy zawiera on semantycznie istotne modyfikacje.

Każdy z tych komponentów pełni odrębną, istotną funkcję w całym procesie przetwarzania i zostanie szczegółowo opisany w kolejnych podrozdziałach.

W analizowanym scenariuszu mamy do czynienia z obrazami o wysokiej rozdzielczości, co stawia istotne wymagania zarówno wobec etapu osadzania znaku wodnego, jak i jego późniejszego odczytu. Wydajna i skuteczna obsługa takich danych wymaga specjalnych mechanizmów pozwalających na uproszczenie i standaryzację procesu przetwarzania. W tym celu w proponowanym rozwiązaniu wprowadzono koncepcję rozdzielczości pośredniej, która pełni rolę wewnętrznego standardu reprezentacji danych w systemie. Mechanizm ten polega na skalowaniu każdego obrazu wejściowego zarówno podczas etapu znakowania, jak i detekcji do ustalonej, stałej rozdzielczości. W prezentowanej implementacji jest to format 256×256 pikseli, niezależnie od oryginalnych proporcji obrazu. Zabieg ten umożliwia pracę sieci neuronowych na spójnym formacie danych, co upraszcza proces uczenia oraz zwiększa stabilność i skuteczność działania systemu. W praktyce rozwiązanie to pozwala zachować zgodność pomiędzy komponentami algorytmu oraz znacząco ograniczyć zużywane zasoby niezależnie od źródła czy formatu obrazu wejściowego.

2.2. Enkoder

Enkoder, oznaczony na rysunku 1 kolorem pomarańczowym, odpowiedzialny jest za generowanie znaku wodnego na podstawie obrazu wejściowego. W proponowanym rozwiązaniu znak wodny przyjmuje postać maski szumowej, która po dodaniu do oryginalnego obrazu nie wprowadza widocznych dla człowieka artefaktów, zachowując tym samym pełną transparentność.



Rys. 1. Przepływ danych enkodera systemu

Rysunek ten przedstawia również kluczowy element architektury, jakim jest zastosowanie rozdzielczości pośredniej. W tym podejściu obraz wejściowy jest w pierwszej kolejności skalowany do ustalonego, mniejszego rozmiaru (256×256 px), co umożliwia przetwarzanie go przez sieć neuronową enkodera w sposób efektywny pod względem zasobów obliczeniowych. Na podstawie przeskalowanego obrazu sieć generuje maskę szumową, która następnie jest interpolowana z powrotem do natywnego rozmiaru obrazu wejściowego. Pozwala to na osadzenie znaku wodnego również w obrazach o bardzo wysokiej rozdzielczości, bez konieczności bezpośredniego przetwarzania ich przez sieć, co znacząco ogranicza zużycie pamięci GPU. W praktycznej implementacji enkoder realizowany jest przez standardową architekturę typu U-Net [4], szeroko stosowaną w zadaniach przetwarzania obrazu ze względu na swoją efektywną strukturę i zdolność do precyzyjnej lokalizacji informacji.

2.3. Blok zakłóceń

Istotnym komponentem systemu jest blok zakłóceń, którego zadaniem jest symulacja realistycznych scenariuszy degradacji i manipulacji obrazu. Element ten, odzwierciedla proces, jaki może zachodzić w momencie, gdy obraz trafia do środowiska publicznego, np. mediów społecznościowych.

W odróżnieniu od typowych podejść, blok zakłóceń zaprojektowano z myślą o rozróżnieniu dwóch fundamentalnie różnych klas modyfikacji:

Ataki typu fałszywego – obejmujące operacje przetwarzania sygnału, które nie zmieniają semantycznej treści obrazu, lecz mogą pogarszać jego jakość lub strukturę niskopoziomową. Typowe przykłady to:

- kompresja JPEG (z różnymi poziomami jakości),
- dodanie szumu Gaussa,
- zmiany jasności i kontrastu,
- konwersja do odcieni szarości.

Ataki typu rzeczywistego – polegające na lokalnych manipulacjach prowadzących do celowego zaburzenia znaczenia wizualnego obrazu. Przykłady obejmują:

- wymazywanie lub zamiana fragmentów obrazu,
- wstawianie obcych elementów graficznych (np. clipartów),
- mieszanie treści z innych obrazów,
- lokalne rozmycie lub szum,
- przybliżenie i wykadrowanie fragmentu sceny.

Rozróżnienie tych dwóch typów ataków pozwala nie tylko wykrywać manipulacje, ale również ocenić ich charakter i potencjalne znaczenie semantyczne. Blok zakłóceń działa w sposób pseudolosowy, intensywność, lokalizacja i kombinacja perturbacji są zmienne, co umożliwia modelowi uczenie się w warunkach zbliżonych do rzeczywistych. Co istotne, każda próbka może być poddana kombinacji modyfikacji z obu kategorii, co lepiej odwzorowuje rzeczywisty obieg obrazu w sieci.

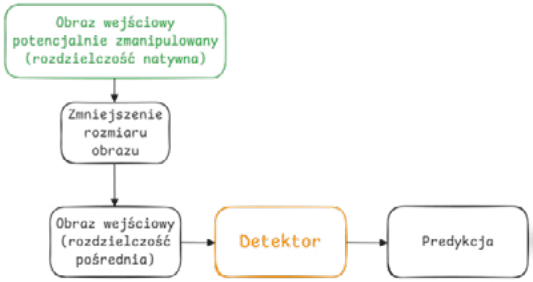
Dla poprawy generalizacji oraz stabilności modelu, w części przypadków pomija się proces znakowania - obrazy bez znaku wodnego również trafiają do detektora po modyfikacjach. Dzięki temu model uczy się rozróżniać rzeczywiste manipulacje od naturalnych cech niezabezpieczonych obrazów.

Większość operacji w module zakłóceń zaimplementowano tak, aby umożliwić pełen przepływ gradientu (pełna różniczkowalność) [5], co zapewnia integrację z procesem trenowania i przyspiesza zbieżność modelu.

2.4. Detektor

Detektor stanowi ostatni komponent systemu i odpowiada za identyfikację ataków typu rzeczywistego, czyli manipulacji, które wpływają na semantyczną treść obrazu. Celem tego modułu nie jest ogólna klasyfikacja jakości obrazu, lecz wyłącznie detekcja zmian, które mogą

świadczyć o celowym naruszeniu integralności treści wizualnej.



Rys. 2. Przepływ danych detektora systemu

Zasada działania detektora została zilustrowana na rysunku 2. Obraz wejściowy, potencjalnie zawierający znak wodny oraz ślady manipulacji, zostaje najpierw przeskalowany do wcześniej zdefiniowanej rozdzielczości pośredniej (256×256 px). Tak przygotowany obraz trafia następnie na wejście sieci detekcyjnej, a wynikiem jest prawdopodobieństwo świadczące o ocenie manipulacji. W prezentowanym rozwiązaniu rolę detektora pełni architektura MobileNetV3 [6], wybrana ze względu na korzystny kompromis pomiędzy wydajnością a efektywnością obliczeniową. Sieć ta, dzięki swojej kompaktowej strukturze, umożliwia wdrożenie systemu również w środowiskach o ograniczonych zasobach, takich jak urządzenia brzegowe czy systemy wbudowane.

3. PROCES TRENINGU

3.1. Zbiór danych

Do trenowania oraz oceny działania systemu wykorzystano zestaw danych obejmujący ponad 324 000 obrazów, wybranych z publicznie dostępnej kolekcji Public Domain 12M [7]. Jest to obszerne źródło danych obraz–tekst, opracowane pierwotnie na potrzeby trenowania modeli generatywnych. Na potrzeby eksperymentów wyodrębniono podzbiór składający się z wysokiej jakości obrazów o szerokim zakresie rozdzielczości — od standardu HD aż po 8K. Uwzględnione zostały m.in. następujące warianty:

- **Proporcje 16:9**: (1280×720), (1600×900), (1920×1080), (2560×1440), (3200×1800), (3840×2160), (5120×2880), (7680×4320)
- **Proporcje 1:1**: (256×256), (400×400), (600×600), (720×720), (1080×1080), (1440×1440), (2160×2160), (2880×2880), (4320×4320)

Zestaw danych został podzielony na trzy części:

- **Zbiór treningowy**: 250 000 obrazów
- **Zbiór walidacyjny**: 25 000 obrazów
- **Zbiór testowy**: 49 000 obrazów

3.2. Funkcje kosztu

Ze względu na odmienną rolę poszczególnych komponentów systemu, każdy z nich trenowany był z użyciem dedykowanej funkcji kosztu. Celem działania enkodera jest generowanie transparentnego znaku wodnego — takiego, który po dodaniu do obrazu nie wprowadza widocznych zakłóceń. Ponieważ brak jest bezpośredniej referencji dla samego znaku wodnego, proces ten traktowany jest jako uczenie częściowo nadzorowane.

Zastosowana funkcja kosztu składa się z dwóch części, odpowiadających ocenie jakości znaku wodnego w dwóch rozdzielczościach:

$$L_{natywny} = MSE + \frac{PSNR}{60.0} \quad (1)$$

$$L_{pośredni} = SSIM + LPIPS \quad (2)$$

$$L_{enkodera} = L_{natywny} + L_{pośredni} \quad (3)$$

gdzie:

MSE – błąd średniokwadratowy
$SSIM$ – strukturalny indeks podobieństwa
$LPIPS$ – percepcyjna miara podobieństwa)

Moduł detekcyjny działa jako binarny klasyfikator, którego zadaniem jest wskazanie, czy dany obraz zawiera semantycznie istotną manipulację (atak typu true). Jego funkcja kosztu opiera się na binarnym błędzie krzyżowym (Binary Cross-Entropy, BCE):

$$L_{detektora} = BCE(y_{pred}, y_{gt}) \quad (4)$$

gdzie:

BCE – binarna entropia krzyżowa

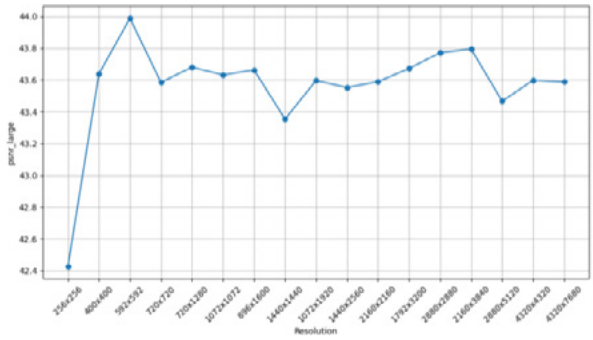
3.3. Parametry i środowisko treningowe

Proces treningu został przeprowadzony w ciągu 16 epok, co umożliwiło osiągnięcie satysfakcjonującej zbieżności zarówno lokalnych funkcji kosztu przypisanych poszczególnym modułom, jak i globalnej funkcji strat. Do optymalizacji parametrów wszystkich komponentów zastosowano optymalizator Adam z ustalonym współczynnikiem uczenia lr=0,0001. Trening prowadzono z rozmiarem mini partii równym 12, na pojedynczej karcie graficznej NVIDIA A40. System treningowy był oparty na serwerze wyposażonym w procesor Intel Xeon Gold 6342 (48 rdzeni fizycznych, 96 wątków logicznych) oraz 504 GB pamięci RAM. W celu zwiększenia stabilności i wydajności procesu treningowego, użyto mechanizmów automatycznego skalowania gradientów oraz automatycznego rzutowania typów.

4. OTRZYMANE REZULTATY

Ocena skuteczności rozwiązania została przeprowadzona w oparciu o dwie główne grupy metryk: metryki transparentności znaku wodnego oraz metryki skuteczności detekcji manipulacji.

Transparentność (niewidoczność osadzenia w obrazie) została oceniona zarówno na poziomie rozdzielczości pośredniej, jak i natywnej. Średnie wartości globalne (ujęte jako średnia ważona względem liczności próbek dla każdej rozdzielczości) wyniosły odpowiednio: PSNR 43.5616 dB, SSIM 0.9983.



Rys. 3. Wartości PSNR w zależności od rozdzielczości natywnej obrazu

Rysunek 3 przedstawia wartości metryki PSNR w funkcji różnych wariantów rozdzielczości występujących w zbiorze testowym. Jak pokazano, wahania wartości są niewielkie, co świadczy o stabilnej i wysokiej transparentności generowanego znaku wodnego niezależnie od rozdzielczości wejściowej obrazu.

W drugim zestawie eksperymentów oceniono skuteczność binarnego detektora manipulacji. Średnie wyniki na całym zbiorze testowym, zawierającym obrazy zmanipulowane w różnorodny sposób, zgodnie z opisem w rozdziale 2.3 wyniosły: F1-score: 0.8907, Accuracy: 0.9241. Wyniki te potwierdzają zdolność modelu do skutecznego wykrywania semantycznie istotnych ingerencji w obraz, przy jednoczesnej odporności na nieszkodliwe operacje przetwarzania.

Tab. 1. Porównanie transparentności opisywanej metody z uznanymi metodami watermarkingu

Metoda	*Rozdzielczość [px]*	*SSIM*	*PSNR [dB]*
HiDDeN [2]	256×256	0.9234	28.87
StegaStamp [8]	400×400	0.930	29.88
SteganoCNN [9]	256×256	0.981	35.852
HiNet [10]	256×256	0.9920	46.88
Nasza	Od 256×256 do 4320×7680	0.9983	43.5616

Tabela 1 ilustruje, że proponowana metoda oferuje porównywalną lub lepszą transparentność względem istniejących rozwiązań, mimo że operuje na znacznie wyższych rozdzielczościach. Osiągnięte wyniki potwierdzają skuteczność zastosowanego znaku wodnego pod względem jakości wizualnej i skalowalności.

Tab. 2. Porównanie dokładności detekcji opisywanej metody z typowymi metodami detekcji manipulacji

Metoda	*Accuracy [%]*
Zang et al. [11]	91.09
Bunk et al. [12]	94.86
Bappy et al. [13]	94.8
Mazumdar et al. [14]	98.0
Yan et al. [15]	80.0

Tabela 2 zawiera porównanie proponowanego rozwiązania z innymi podejściami klasyfikującymi tego typu, w zakresie dokładności detekcji i obsługiwanej rozdzielczości. Jak widać, nasza metoda oferuje konkurencyjne wyniki, szczególnie pod względem skalowalności do obrazów wysokiej rozdzielczości, co stanowi jej istotną przewagę.

5. PODSUMOWANIE

W pracy przedstawiono system detekcji semantycznie istotnych manipulacji w obrazach wysokiej rozdzielczości, oparty na transparentnym, półkruchym znaku wodnym. Kluczowym elementem rozwiązania jest zastosowanie rozdzielczości pośredniej, która umożliwia przetwarzanie dużych obrazów przy niewielkim obciążeniu obliczeniowym. Zaproponowana architektura skutecznie dyskryminuje różnicę pomiędzy nieszkodliwymi przekształceniami a rzeczywistą ingerencją w treść obrazu, zachowując jednocześnie wysoką jakość wizualną.

Uzyskane wyniki potwierdzają, że metoda łączy wysoką transparentność z precyzją detekcji, co czyni ją przydatną w praktycznych zastosowaniach, takich jak archiwizacja, monitoring czy analiza treści cyfrowych.

LITERATURA

1. Duszejko, P.; Walczyna, T.; Piotrowski, Z. Detection of Manipulations in Digital Images: A Review of Passive and Active Methods Utilizing Deep Learning. *Applied Sciences* **2025**, *15*, 881, doi:10.3390/app15020881.
2. Zhu, J.; Kaplan, R.; Johnson, J.; Fei-Fei, L. HiDDeN: Hiding Data With Deep Networks 2018.
3. Duszejko, P.; Piotrowski, Z. 4KSecure: A Universal Method for Active Manipulation Detection in Images of Any Resolution. *Applied Sciences* **2025**, *15*, 4469, doi:10.3390/app15084469.
4. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation 2015.
5. Riba, E.; Mishkin, D.; Ponsa, D.; Rublee, E.; Bradski, G. Kornia: An Open Source Differentiable Computer Vision Library for PyTorch 2019.
6. Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3 2019.
7. Meyer, J.; Padgett, N.; Miller, C.; Exline, L. Public Domain 12M: A Highly Aesthetic Image-Text Dataset with Novel Governance Mechanisms 2024.
8. Tancik, M.; Mildenhall, B.; Ng, R. StegaStamp: Invisible Hyperlinks in Physical Photographs 2020.
9. Duan, X.; Liu, N.; Gou, M.; Wang, W.; Qin, C. SteganoCNN: Image Steganography with Generalization Ability Based on Convolutional Neural Network. *Entropy* **2020**, *22*, 1140, doi:10.3390/e22101140.
10. Jing, J.; Deng, X.; Xu, M.; Wang, J.; Guan, Z. HiNet: Deep Image Hiding by Invertible Network.

Conference on Computer Vision (ICCV); October 2021; pp. 4713–4722.
11. Zhang Y, Goh J, Win LL, Thing VL (2016) Image region forgery detection: a Deep learning ap-proach. SG-CRC 2016:1–11
12. Bunk, J.; Bappy, J.H.; Mohammed, T.M.; Nataraj, L.; Flenner, A.; Manjunath, B.S.; Chandrasekaran, S.; Roy-Chowdhury, A.K.; Peterson, L. Detection and Localization of Image Forgeries Using Resampling Features and Deep Learning 2017.
13. Bappy, J.H.; Simons, C.; Nataraj, L.; Manjunath, B.S.; Roy-Chowdhury, A.K. Hybrid LSTM and Encoder-Decoder Architecture for Detection of Image Forgeries. *IEEE Trans. on Image Process.* **2019**, *28*, 3286–3300, doi:10.1109/TIP.2019.2895466.
14. Mazumdar, A.; Jacob, J.; Bora, P.K. Forgery Detection in Digital Images through Lighting Environment Inconsistencies. In Proceedings of the 2018 Twenty Fourth National Conference on Communications (NCC); February 2018; pp. 1–6.
15. Yan, Y.; Ren, W.; Cao, X. Recolored Image Detection via a Deep Discriminative Model. *IEEE Transactions on Information Forensics and Security* **2019**, *14*, 5–17, doi:10.1109/TIFS.2018.2834155.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

DZIAŁANIA MOTYWUJĄCE TESTERA JAKO ELEMENT PRZYGOTOWANIA EKSPERYMENTU QoE

SUBJECT'S MOTIVATING ACTIVITIES AS AN ELEMENT OF QoE EXPERIMENT PREPARATION

Tomasz Konaszyński[1]; Avrajyoti Dutta[2]; Dawid Juszka[3]; Mikołaj Leszczuk[4]

[1] Akademia Górniczo-Hutnicza, Tomasz.Konaszynski@agh.edu.pl
[2] Akademia Górniczo-Hutnicza, Avrajyoti.Dutta@agh.edu.pl
[3] Akademia Górniczo-Hutnicza, Dawid.Juszka@agh.edu.pl
[4] Akademia Górniczo-Hutnicza, Mikolaj.Leszczuk@agh.edu.pl

DOI: 10.15199/59.2025.4.101

Streszczenie: **Na podstawie wyników przeprowadzonego eksperymentu QoE wykazano związek między podejmowanymi działaniami motywującymi testera, a kwalifikowalnością rezultatów oraz wartościami uzyskanych wyników badań subiektywnych. Implikuje to konieczność rozważenia zasadności modyfikacji procedury organizacji badań w przyszłych eksperymentach QoE, w szczególności prowadzonych w warunkach ograniczonego budżetu i dostępności testerów, jak też odpowiedniego przygotowania testerów do planowanych eksperymentów.**

Abstract: **Based on the results of the conducted QoE experiment, a relationship was shown between the actions taken to motivate the subject and the eligibility of obtained results as well as values of the obtained subjective research results. This implies the need to consider the validity of modifying the procedure for organizing research in future QoE experiments, especially those conducted under conditions of limited budget and tester availability, as well as the appropriate preparation of testers for the planned experiments.**

Słowa kluczowe: **QoE, badania subiektywne, poziom motywacji testera, działania motywujące, język natywny**

Keywords: **QoE, subjective research, subject's motivation level, motivational activities, native language**

1. WPROWADZENIE

Zakres badań będących podstawą niniejszej publikacji wpisuje się w dziedzinę Quality of Experience [1]. W pracach badawczych QoE warto szczególnie pochylić się nad aspektem, który w badaniach subiektywnych jest obecny zawsze, tj. czynnikiem stanu mentalnego testera w momencie podchodzenia do oceny. Wartościowym poznawczo zagadnieniem jest identyfikacja wpływu stanu mentalnego, w jakim tester przystępuje do udziału w eksperymencie na kwalifikowalność uzyskanych wyników oraz na jakość i wartość poznawczą uzyskanych wyników badań subiektywnych. Jednym z czynników wpływu [10] wpływających na stan mentalny testera są **podejmowane wobec niego działania motywujące**, odwołujące się bezpośrednio do istoty badania, wzmocnienia rzetelności oraz sumiennego podejścia do wykonywanej oceny.

Istnieją prace wskazujące na role motywacji. Np. wyniki badania [4] dowodzą, że aby dokonać dokładnej oceny jakości usług wideo na urządzeniach mobilnych, oprócz jakości obrazu należy brać pod uwagę także motywację użytkownika. Wyniki badania [9] dowodzą, że poziom zainteresowania osoby badanej treścią silnie wpływa na jej subiektywną ocenę jakości obrazu w badaniu jakości obrazu. Prace te jednak nie dotyczą np. aspektu kwalifikowalności wyników testerów, co ma krytyczne znaczenie dla efektywności prowadzonych badań, w szczególności w kontekście ograniczonych zasobów badawczych [11].

2. ZAGADNIENIE BADAWCZE

Aby odpowiedzieć na zidentyfikowaną niepewność badawczą, postawiono pytanie badawcze:

- **RQ1: Czy na ocenę subiektywną wideo 2D ma wpływ motywowanie testera przed rozpoczęciem eksperymentu?**

Określono również powiązane hipotezy badawcze:

- H1: Podjęte działania motywujące testera mają wpływ na kwalifikowalność uzyskanych wyników
- H2: Podjęte działania motywujące testera mają wpływ na ocenę subiektywną

Zidentyfikowany wpływ może skutkować nie tylko odrzuceniem istotnej części wyników i „marnotrawstwem" ograniczonych zasobów, jak też może wpływać na uzyskiwane wyniki, zaburzając ich wartość poznawczą, stąd odpowiedź na postawione pytanie badawcze jest istotna dla jakości badań QoE.

3. OPIS EKSPERYMENTU

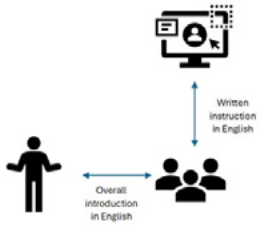
Wstępne założenia. Eksperyment jest kontynuacją cyklu badań, opisanych w [5]. W identycznej konfiguracji technicznej (środowisko, aplikacja badawcza, dataset) przeprowadzono badanie z udziałem dodatkowej grupy testerów, która została poddana czynnikowi różnicującemu – zewnętrznym działaniom motywującym.

Testerzy. W eksperymencie udział wzięło 61 testerów, w tym 6 kobiet oraz 55 mężczyzn. Nierównoliczność reprezentacji płci wynika z faktu, że eksperyment został przeprowadzony z udziałem grupy studentów kierunku związanego z telekomunikacją i

multimediami i jest determinowana składem osobowym grupy studenckiej.

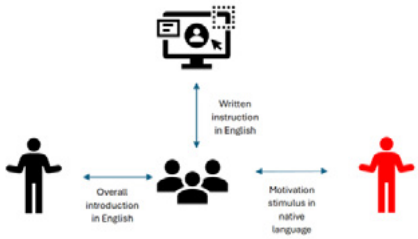
Projekt eksperymentu. Badania przeprowadzono zgodnie z zaleceniami Rek. ITU-T P.913 [3] dla badań QoE, według metody ACR, w której bodźce PVS (*Processed Video Sequences)* są prezentowane pojedynczo i oceniane w 5-stopniowej skali Likerta (5 – excellent, 4 – good, 3 – fair, 2 – poor, 1 – bad).

Czynnikiem różnicującym jest mechanizm dostarczenia części uczestników przed rozpoczęciem eksperymentu zewnętrznego bodźca motywującego. Testerów podzielono na 2 podgrupy, przy czym w obydwu podgrupach przewidziano ogólne wprowadzenie do celu i przebiegu eksperymentu w j, angielskim, niebędącym językiem natywnym dla uczestników, uzupełnione o udostępnienie instrukcji użytkownika w j. angielskim, zgodnie z rys. 1.



Rys. 1. Źródła informacji o przebiegu i celu eksperymentu – podgrupa 1 (bez dodatkowej motywacji)

Drugiej podgrupie testerów dodatkowo dostarczono zewnętrzny bodziec motywacyjny - silnie akcentową motywacyjnie wypowiedź autorów eksperymentu w języku natywnym testerów, tj. szczegółowe wyjaśnienie przebiegu eksperymentu, oraz mocno zaznaczoną prośbę o koncentrację i uczciwą ocenę jakości subiektywnej zgodnie z własnymi odczuciami testera (rys. 2).

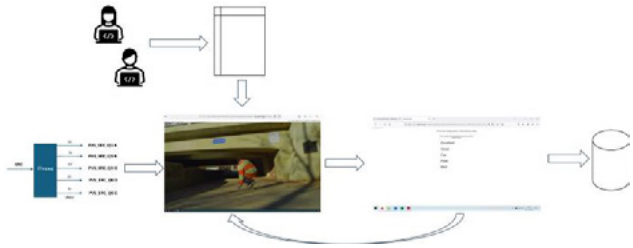


Rys. 2. Źródła informacji o przebiegu i celu eksperymentu – podgrupa 2 (z dodatkową motywacją)

W podgrupie 1 (bez bodźca motywacyjnego) przewidziano udział 50 testerów. W podgrupie 2 (z dodatkowym bodźcem motywacyjnym) zaplanowano udział 11 testerów. Nierównoliczność obu grup wynika z niskiej liczebności grupy studenckiej która została zaangażowana do tej części eksperymentu, na którą badający nie mieli wpływu.

Testerzy oceniali sekwencje wideo według jednej z dwóch naprzemiennych reguł: (1) każdy tester oceniał 170 sekwencji PVS jeden raz, (2) każdy tester oceniał 60 sekwencji PVS po trzy razy, tj. łącznie 180 filmów. Prezentowane filmy pochodziły ze wspólnej bazy źródłowej, a testerzy oceniali te same filmy.

Na rys. 3. przedstawiono schemat logiczny eksperymentu.



Rys. 3. Schemat logiczny przebiegu eksperymentu

Oceniane pliki wideo. Jako pliki źródłowe użyto zestawu 34 unikalnych sekwencji wideo SRC (*Source Reference Circuit*), pozyskanych z baz Netflix, CableLabs, SJTU Media Lab oraz Xiph.org Video Test Media. Pod względem technicznym źródłowe dane badawcze składały się z sekwencji wideo Full HD w standardzie MPEG-4 o rozdzielczości 1080p: 1920 × 1080 pikseli z prędkością odtwarzania 60 klatek/s i zmienną szybkością transmisji poszczególnych filmów.

Zbiór sekwencji PVS podlegających ocenie powstał poprzez przetworzenie plików SRC z wykorzystaniem narzędzia FFmpeg [6] i zdegradowanie jakości filmów do 5 poziomów obiektywnej jakości technicznej, wyznaczonych z wykorzystaniem pełno-referencyjnej metryki jakości wideo VMAF (ang. *Video Multimethod Assessment Fusion*) [13]], Jakość ocenianych filmów odpowiadała wartości najbliższej: (1) Gr. A - VMAF=90, (2) Gr. B - VMAF=70, (3) Gr. C - VMAF=50, (4) Gr. D - VMAF=30, (5) Gr. E - VMAF=10.

Warunki realizacji eksperymentu. Badanie przeprowadzono w kontrolowanych (wyciszenie, oświetlenie, klimatyzacja) warunkach laboratoryjnych, w laboratorium Instytutu Telekomunikacji AGH, z użyciem autorskiej aplikacji badawczej.

4. ANALIZA REZULTATÓW

Wyniki eksperymentu zostały przeanalizowane z uwagi na ich kwalifikowalność oraz ocenę subiektywną wyników zakwalifikowanych do dalszej analizy.

4.1. Wpływ motywacji na kwalifikowalność uzyskanych wyników do dalszej analizy

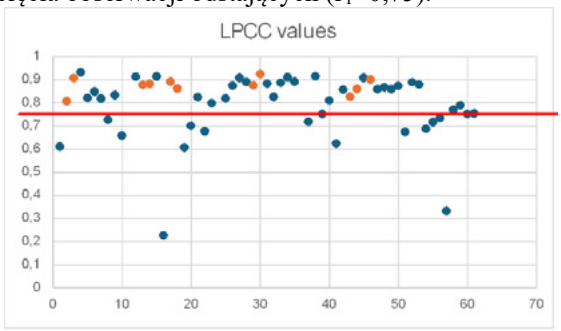
Weryfikacja kwalifikowalności wyników. Do weryfikacji kwalifikowalności wyników wykorzystano metodę screeningu post-eksperymentalnego za pomocą współczynnika liniowej korelacji Pearsona [3], względem pojedynczego bodźca (PVS).

Otrzymane wyniki. W efekcie uzyskano następujące rezultaty kwalifikowalności wyników:

Tab.1. Kwalifikowalność wyników w zależności od obecności czynnika motywującego

	Bez dodatkowej motywacji	Z dodatkową motywacją
Zakwalifikowani	33	11
Odrzuceni	17	0
% zakwalifikowanych	66,00%	100,00%

Rys. 4 przedstawia wartości współczynnika korelacji liniowej Pearsona (LPCC) w ostatnim cyklu procedury dla poszczególnych testerów, tj. w cyklu eliminującym wyniki danego testera (dla testerów niezakwalifikowanych) lub w ostatnim cyklu całej procedury (dla testerów zakwalifikowanych). Na niebiesko zaznaczono wyniki testerów wobec których nie podjęto działań motywujących, zaś na pomarańczowo wyniki testerów wobec których takie działania podjęto. Czerwona linia oznacza uznany w badaniach QoE próg odcięcia obserwacji odstających (r_1=0,75).



Rys.4. Uzyskane wartości LPCC dla poszczególnych testerów – w zależności od czynnika motywującego

Interpretacja wyników. Uzyskane dane wskazują na istotną różnicę kwalifikowalności testerów w obu grupach. Uzyskano 100% kwalifikowalności wyników testerów, wobec których podjęto działania motywujące oraz tylko częściową kwalifikowalność wyników testerów niepodlegających dodatkowej motywacji. Należy też podkreślić, że dla kwalifikowalności wyników różnoliczność grup nie ma znaczenia, gdyż weryfikacja obserwacji odstających odnosi wyniki pojedynczego testera do wyników całej grupy 61 testerów.

Potwierdzono zatem stawianą hipotezę, że podjęte działania motywujące testera mają pozytywny wpływ na kwalifikowalność uzyskanych wyników do dalszej analizy, skutkując istotnie wyższym odsetkiem zakwalifikowanych wyników w porównaniu z grupą niepodlegającą działaniom motywacyjnym.

4.2. Wpływ motywacji na ocenę subiektywną otrzymanych wyników

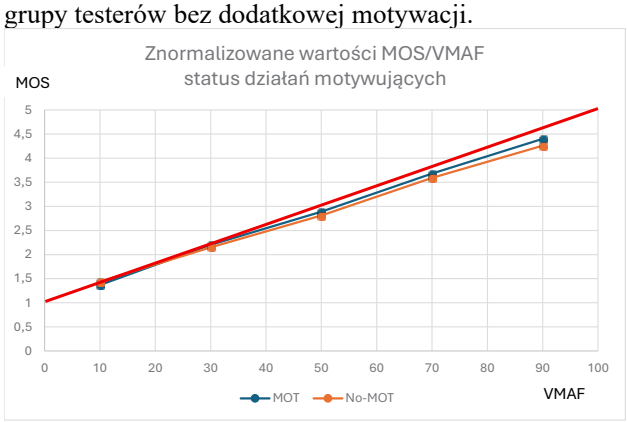
Weryfikacja wpływu na ocenę subiektywną. W celu weryfikacji statystycznej wyników przeanalizowano rozkład wyników próby badawczej pod kątem jego normalności. Przy spełnionych założeniach do weryfikacji rozkładu próby wybrano test Shapiro-Wilka, który wykazał, że rozkład badanej cechy w populacji nie ma charakteru normalnego. Do dalszych analiz dwóch niezależnych grup został wybrany test nieparametryczny U Manna-Whitheya porównujący mediany obu grup.

Otrzymane wyniki. Dla poszczególnych grup jakościowych otrzymano następujące wartości MOS:

Tab. 2. Wartości MOS w zależności od obecności działań motywacyjnych lub ich braku

	MOS_{mot}	$MOS_{no\text{-}mot}$	P
Grupa A	4,40	4,26	0,0580
Grupa B	3,68	3,59	0,1211
Grupa C	2,89	2,81	0,0951
Grupa D	2,20	2,15	0,3674
Grupa E	1,37	1,44	0,1350

Na rys. 5 przedstawiono wykres zależności uzyskanych wyników MOS w odniesieniu do znormalizowanej wartości VMAF, wyznaczanej przez zaznaczoną kolorem czerwonym linię trendu, odpowiadającą liniowej normalizacji skali VMAF w odniesieniu do skali ocen MOS. Wartość VMAF będąca wynikiem matematycznej estymacji wielu metryk pomiarowych nie ma charakteru liniowego w rozumieniu matematycznym. Dla czytelności analizy przyjęto jednak założenie, że skala VMAF jest pojęciowo liniowa w przedziale [0,100], podobnie jak wartości MOS w przedziale [1,5]. W związku z tym możliwe jest wytyczenie referencyjnej linii trendu VMAF/MOS, do której można odnosić uzyskane wyniki MOS. Na niebiesko zaznaczono wyniki grupy testerów objętej dodatkową motywacją. Kolor pomarańczowy to wyniki grupy testerów bez dodatkowej motywacji.



Rys. 5. Zależność uzyskanych wyników w odniesieniu do znormalizowanej wartości VMAF/MOS w zależności od czynnika motywującego

Pomimo, że w przeprowadzonej analizie nie potwierdzono istotności statystycznej uzyskanych wyników w poszczególnych grupach jakościowych, wartości MOS filmów zachowują trend lepszej korelacji wyników z wartościami obiektywnymi VMAF dla grupy testerów podlegających działaniom motywacyjnym. Uzyskane wartości p w grupach jakościowych A oraz C są silnie zbliżone do progu istotności.

Interpretacja wyników. Otrzymane wyniki wskazują, że oceny przyznawane przez testerów podlegających dodatkowym działaniom motywacyjnym są bliższe znormalizowanej linii trendu VMAF/MOS, niż wyniki osiągane przez testerów niepodlegających dodatkowej motywacji, tj. bardziej skoncentrowane wokół wartości wyznaczonych przez obiektywną jakość techniczną VMAF. W konsekwencji lepiej odwzorowują jakość zobiektywizowaną, co sugeruje wyższy poziom koncentracji testerów podlegających dodatkowej motywacji i oceny bardziej adekwatne w odniesieniu do jakości obiektywnej ocenianych filmów.

Pomimo braku formalnej istotności statystycznej, uzyskane prawidłowości uprawdopodobniają hipotezę, że istnieje korelacja między działaniami motywującymi a oceną subiektywną – wyniki grupy motywowanej są

bardziej zbliżone do jakości obiektywnej. Siłę uzyskanych wniosków osłabia jednak silna nierównoliczność grup, czego autorzy są świadomi.

5. DYSKUSJA I DALSZE KIERUNKI BADAWCZE

Uzyskane wyniki wskazują na niepomijalny wpływ aspektów związanych z poziomem motywacji na uzyskiwane wyniki badań subiektywnych, ich kwalifikowalność, jakość i stabilność. W kontekście efektywności badań QoE kluczowe jest zagadnienie aspektów mentalnych, w tym motywacja, które wpływają na prawdopodobieństwo uzyskania wysokiej jakości wyników. Aspekt ten należy usytuować w kontekście efektywności wykorzystania ograniczonych zasobów badawczych tj. dostępnego budżetu na opłacenie testerów płatnych oraz dostępnych testerów, np. wolontariuszy. Istniejące badania [12] wskazują na dostrzegane przez badaczy znaczenie tego aspektu, kluczowym wnioskiem jest więc **uzupełnienie obecnych procedur badawczych** [3] o wprowadzenie świadomych, proaktywnych działań motywujących testerów przed rozpoczęciem badań, w sposób maksymalizujący szanse uzyskania kwalifikowalnych danych o wysokiej jakości poznawczej.

Jednym z elementów działań motywujących jest również używanie **języka natywnego dla testerów**. Aspekt ten poruszono w [8] w kontekście wyników badań QoE w zakresie subiektywnej oceny algorytmów tłumienia szumów w eksperymentach QoE.

W eksperymencie nie był analizowany wpływ samego użytego języka ale istotne różnice w wynikach kwalifikowalności testerów podlegających motywowaniu w języku dla nich natywnym wskazują, że aspekt języka komunikacji przed badaniem również może być istotny, biorąc pod uwagę zróżnicowane umiejętności językowe testerów. W kontekście uzyskanych wyników warta dalszej analizy jest również kwestia **bezpośredniej komunikacji testera z badaczem**. W badaniach niebezpośrednich, jak np. badania crowdsourcingowe, brak bezpośredniej komunikacji może skutkować takimi odchyleniami jak zidentyfikowane w niniejszych badaniach dla grupy niepodlegającej motywowaniu.

Obszar analizy wpływu stanu mentalnego i odpowiedniego przygotowania eksperymentu ma zasadniczy i wielowątkowy wpływ na uzyskiwane wyniki, zasadnym jest zatem pogłębienie badań w tym zakresie i **udoskonalenie procedur badawczych pozwalających na minimalizację odchyleń wy**ników związanych z czynnikiem ludzkim w badaniach QoE.

Do rozważenia w przyszłych badaniach pozostaje również kwestia ewentualnej korekcji odchyleń powodowanych czynnikiem ludzkim, związanych ze stanem mentalnym. Ustalenie takiego **współczynnika korekcji** również może być potencjalnym obszarem przyszłych prac badawczych [2][7].

Finansowanie: Praca została sfinansowana przez polskie Ministerstwo Nauki i Szkolnictwa Wyższego w ramach subwencji dla Wydziału Informatyki, Elektroniki i Telekomunikacji AGH.

LITERATURA

[1] Brunnström, et al. 2013. Qualinet white paper on definitions of quality of experience.

[2] Husić, J. B., & Baraković, S. (2022). Multidimensional modelling of quality of experience for video streaming. Computers in human behavior, 129, 107155.

[3] ITU-T. Recommendation ITU-T P.913: Methods for the subjective assessment of video quality, audio quality and audio-visual quality of Internet video and distribution quality television in any environment, International Telecommunication Union, 2016.

[4] Kobayashi, F., et al. (2012, January). QoE assessment method for mobile video services based on user motivation. In Image Quality and System Performance IX (Vol. 8293, pp. 301-309). SPIE.

[5] Konaszyński, T., Dutta, A., Juszka, D., & Leszczuk, M. (2024). Czy w badaniach QoE powinno się ograniczać dostęp testerom w złym stanie psychofizycznym?. Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne.

[6] Lei, X., et al. (2013, December). Design and implementation of a real-time video stream analysis system based on FFMPEG. In 2013 Fourth World Congress on Software Engineering (pp. 212-216). IEEE.

[7] Menkovski, V., et al. (2010, August). Estimations and remedies for quality of experience in multimedia streaming. In 2010 Third International Conference on Advances in Human-Oriented and Personalized Mechanisms, Technologies and Services (pp. 11-15). IEEE.

[8] Naderi, B., et al. (2021, June). Effect of Language Proficiency on Subjective Evaluation of Noise Suppression Algorithms. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4140-4144). IEEE.

[9] Palhais, J., et al. (2011, September). Quality of experience assessment in internet tv. In International Conference on Mobile Networks and Management (pp. 261-274). Berlin, Heidelberg: Springer Berlin Heidelberg.

[10] Reiter, U. et al. 2014. Factors Influencing Quality of Experience. In: Möller, S., Raake, A. (eds) Quality of Experience. T-Labs Series in Telecommunication Services. Springer, Cham.

[11] Sackl, A., & Schatz, R. (2014, September). Evaluating the influence of expectations, price and content selection on video quality perception. In 2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX) (pp. 93-98). IEEE.

[12] Seufert, M., et al. (2016, June). Impact of test condition selection in adaptive crowdsourcing studies on subjective quality. In 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX) (pp. 1-6). IEEE.

[13] Zhang, F., Katsenou, A., Bampis, C., Krasula, L., Li, Z., & Bull, D. (2021, June). Enhancing VMAF through new feature integration and model combination. In 2021 Picture Coding Symposium (PCS) (pp. 1-5). IEEE.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

HD-dSEQUA: Baza zróżnicowanych sekwencji HD do obiektywnej oceny jakości wraz z indykatorami

HD-dSEQUA: High Definition Diverse Video Sequence Dataset for Quality Assessment with Indicators

Filip Korus[†1]; Dawid Juszka[†2]; Michał Grega[†3]; Mikołaj Leszczuk[†4]; Michał Suchoń[‡5]; Avrajyoti Dutta[†6]; Dominika Wanat[†7]

[†] AGH Akademia Górniczo-Hutnicza, Kraków [‡] tvhive sp. z o.o., Kraków

[1] fkorus@student.agh.edu.pl [5] msuchon@quickstream.tech
[2] dawid.juszka@agh.edu.pl [6] avrajyoti.dutta@agh.edu.pl
[3] michal.grega@agh.edu.pl [7] dominika.wanat@agh.edu.pl
[4] mikolaj.leszczuk@agh.edu.pl

DOI: 10.15199/59.2025.4.102

Streszczenie: Wzrost popularności treści wideo w Internecie oraz postęp w dziedzinie automatycznego przetwarzania obrazu powodują rosnące zapotrzebowanie na narzędzia służące do porównywania algorytmów kompresji. W odpowiedzi opracowano autorską bazę sekwencji wideo, zawierającą 131 materiałów o różnorodnych charakterystykach przestrzennych, czasowych i semantycznych. W artykule przedstawiono strukturę bazy, kryteria doboru sekwencji oraz metryki jakości wideo pozwalające na szybkie porównanie parametrów materiału wideo.

Abstract: The growing popularity of video content on the Internet and the development of automatic image processing systems have created a need for appropriate testing tools to enable comparison of compression algorithms. In response, a video author dataset has been developed that contains 131 sequences with various spatial, temporal, and semantic characteristics. The paper presents the structure of the data set, the sequence selection criteria, and video quality metrics that allow for quick comparison of video parameters.

Słowa kluczowe: baza sekwencji, kompresja wideo, metryki jakości wideo, ocena jakości, przetwarzanie obrazu

Keywords: video processing, quality assessment, video dataset, video compression, video quality metrics

1. INTRODUCTION

According to [3], global IP traffic reached 120 zettabytes in 2023 and is projected to increase to 181 zettabytes by 2025, with video content accounting for more than half of the total data volume. This substantial share highlights the exponential growth of video usage across diverse domains – such as entertainment, surveillance, education, healthcare, artificial intelligence, and more – which in turn underscores the need for diverse video datasets to support the evaluation and comparison of video processing systems.

Such systems consist of video encoding techniques, quality metrics, and machine learning (ML) models. Selecting a proper training database for an ML model is crucial because it directly influences the model's performance, generalizability, and robustness. A well-designed data set ensures that the model learns relevant patterns and features, avoids overfitting to narrow content, and performs reliably in diverse real-world scenarios.

A wide variety of video quality databases are available for research and evaluation purposes. Notable examples include

- **YouTube UGC** [10] – a data set comprising 1,500 user-generated videos, each 20 seconds long, with resolutions ranging from 360p to 1080p. The videos span diverse categories such as gaming, sports, and more, and have been evaluated using no-reference quality metrics.
- **BVI-HD** [8] – a collection consisting of 32 reference videos and 384 corresponding distorted versions, along with associated Differential Mean Opinion Scores (DMOS). This data set has been used to assess subjective video quality under HEVC compression and to benchmark the performance of various reference quality metrics.
- **KoNViD-150k** [4] — a dataset comprising 153,841 publicly available video clips, carefully selected from various sources to ensure a diverse range of content and visual characteristics. Each video includes five quality ratings, with a subset of 1,596 videos receiving up to 89 ratings each.

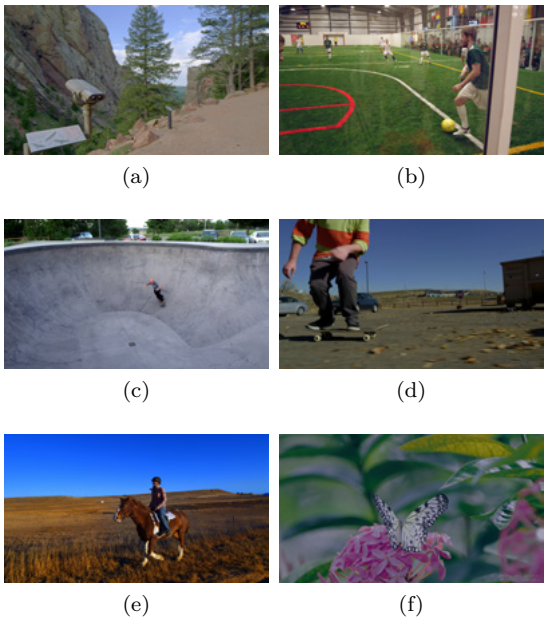
The main limitation of these databases, as well as many others, is the absence or very limited availability of videos in raw format. Raw video data are particularly valuable as they enable the introduction of various distortions – such as compression – into the source material, allowing for comprehensive testing of video quality under various conditions. One of the most widely used [1, 2, 9] full reference metrics for this purpose is the Video Multimethod Assessment Fusion (VMAF), developed by Netflix [5].

2. VIDEOS

The primary goal when selecting video sequences for HD-dSEQUA was to include as diversified video content as possible, not limited to animated or professionally generated content but also encompassing a wide variety of real-world footage. This approach was intended to ensure broad coverage of visual characteristics such as motion and spatial complexity, texture variation, and lighting conditions, making the data set robust and representative of different visual scenarios.

2.1 Sources

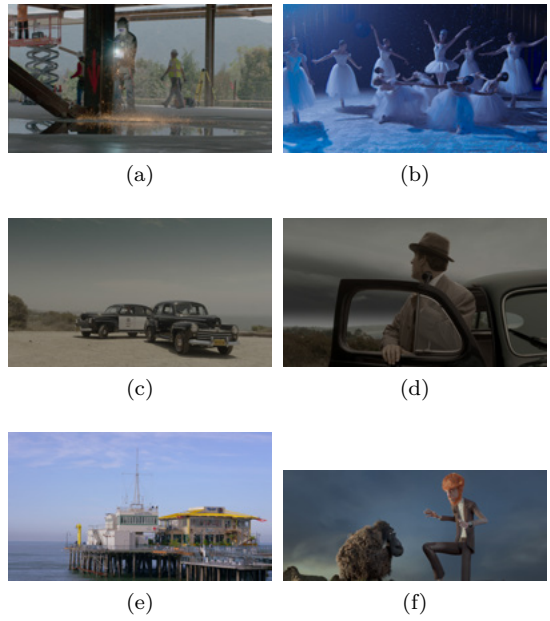
The video sequences were collected from several publicly available datasets and video content sources. A significant portion of the material originates from CableLabs, which offers a diverse range of 1080p videos, including both user-generated content (UGC) and professional productions. Representative sequences include Ancient Thought, Eldorado (Fig. 1a), Indoor Soccer (Fig. 1b), Moment of Intensity (Fig. 1c), Skateboarding (Fig. 1d), Unspoken Friend (Fig. 1e), and Wipe HDR (Fig. 1f). Among these, Moment of Intensity and Wipe HDR are high frame rate sequences recorded at 60 FPS, while the remaining videos are captured at 24 FPS.



(a) (b) (c) (d) (e) (f)

Figure 1: Frame examples of the CableLabs sequences.

Additional sequences were sourced from Netflix, featuring a wide variety of professionally produced content at different resolutions and frame rates. Notable 1080p sequences recorded at 60 FPS include Sparks (Fig. 2a), ballet (Fig. 2b), beach, cave, cliff, driving, inside, leaving, police (Fig. 2c), stairs, talkinghead, and telephone (Fig. 2d). The Chimera sequences, mostly at 1080p and 30 FPS, include several detailed scenes such as Aerial, Driving POV, Pier Seaside (Fig. 2e), Roller Coaster Passengers, and Toddler Fountain. The Cosmos Laundromat collection (Fig. 2f) consists of multiple 1920×804 sequences at 24 FPS, with natural and synthetic content. Additionally, 4K 60 FPS sequences include El Fuente scenes like Boat and Food Market, as well as Chimera Dancers.



(a) (b) (c) (d) (e) (f)

Figure 2: Frame examples of the Netflix sequences.

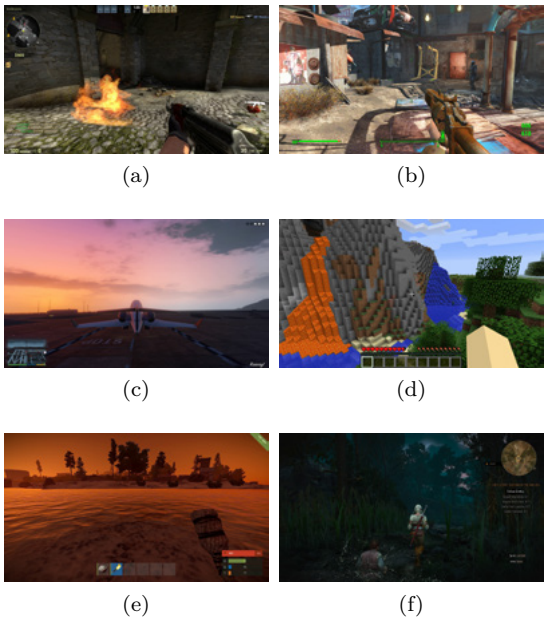
The SJTU Media Lab contributed a comprehensive set of sequences depicting both urban and natural environments, all captured in 1080p resolution at 25 FPS. Representative videos include Rush Hour (Fig. 3a), Campfire Party, and Tall Buildings (Fig. 3c), along with additional scenes such as Bund Nightscape (Fig. 3b), Library, Tree Shade, Construction Field, Fountains, Marathon, Residential Building, Runners (Fig. 3d), Scarf, Traffic Flow, Traffic and Building (Fig. 3e), and Wood (Fig. 3f).

(a) (b) (c) (d) (e) (f)

Figure 3: Frame examples of the SJTU Media Lab sequences.

In addition, dynamic gameplay and streaming content were incorporated from the Xiph.org Video Test Me-

dia and Ultra Video Group datasets, as well as a wide range of Twitch-captured streams, facilitating the analysis of real-time rendering and interactive media. The Twitch collection features 1080p sequences at 60 FPS across numerous popular game titles, including CS:GO (Fig. 4a), DOTA2, Euro Truck Simulator 2, Fallout 4 (Fig. 4b), GTA V (Fig. 4c), Hearthstone, Minecraft (Fig. 4d), Rust (Fig. 4e), Starcraft and The Witcher 3 (Fig. 4f). This collection covers a range of motion complexity, visual patterns, camera pans, and gameplay perspectives.



(a) (b) (c) (d) (e) (f)

Figure 4: Frame examples of the gameplay sequences.

2.2 Metadata and Statistics

The videos in the data set are stored in the Y4M format, which preserves the quality of uncompressed raw video. This format does not introduce compression artifacts, ensuring an accurate representation of the source content for processing. The use of Y4M guarantees consistency across the data set and is particularly well suited for research and benchmarking tasks in video quality assessment, encoding, and streaming performance evaluation.

Of the 131 videos in the dataset, 117 are in Full HD resolution (1920×1080), 11 have a resolution of 1920×804, and 3 are available in Ultra HD (3840×2160). In terms of frame rates, 39 videos are recorded at 24 FPS, 16 at 25 FPS, 5 at 30 FPS, and the remaining 71 at 60 FPS.

The durations of the video sequences are presented in a histogram in Figure 5. Most of the videos have durations between 10 and 20 seconds, with the highest concentration around 13 to 16 seconds – where the number of sequences peaks at 48. The interquartile range, from the 25th percentile at 11.93 seconds to the 75th percentile at 15 seconds, confirms that half of the videos are tightly clustered within this short duration window. The median duration is 14 seconds, while the mean is slightly higher at 17.12 seconds. There are also a noticeable number of shorter videos between 7 and 13 seconds, and very few sequences shorter than 7 seconds. During 20 seconds, the number of sequences drops dramatically. A distinct outlier exists at the 60-second mark, which aligns with the maximum value of 60.05 seconds and contributes to the elevated mean. The standard deviation of 12.80 seconds reflects the overall spread in video durations. In summary, the majority of the videos are short and fairly consistent in length, with a small number of much longer videos that skew the distribution.

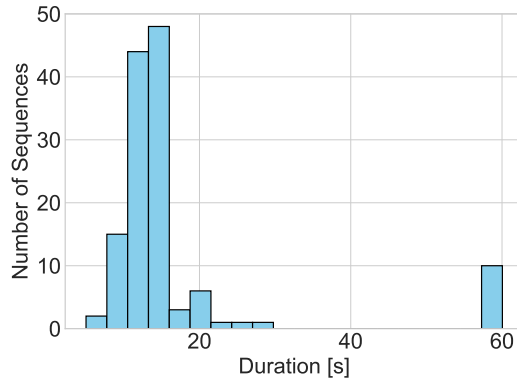


Figure 5: Distribution of video sequence durations.

3. VIDEO QUALITY INDICATORS

The video sequences were assessed using 10 no-reference Video Quality Indicators (VQIs) developed at AGH [7, 6]. **Spatial Activity (SA)** captures the amount of spatial detail within each frame, while **Temporal Activity (TA)** reflects the intensity of movement [7]. **Blockiness (BLN)** detects compression artifacts, and **Blockloss (BLL)** measures the impact of lost data blocks on visual quality [7]. **Blur (BLU)** assesses the loss of sharpness, often due to compression, removing high-frequency details [7]. **Exposure (EXP)** and **Contrast (CON)** evaluate the lighting balance and the variation of brightness, respectively [7]. **Interlace (INT)** identifies interlacing artifacts, and **Noise (NOI)** captures random brightness or color fluctuations [7]. Finally, **UGC** distinguishes user-generated content from professional video [6].

All of these indicators were computed using the aghvqis Python package [6], which can be easily installed via pip. The metrics are lightweight and efficient, making them well-suited for real-time video quality assessment.

Figure 6 presents the average values of SA and TA for various video sequences. Higher values indicate greater visual detail or more motion within a video. Several sequences exhibit notably high activity levels, such as sequence 107, which features dense tree coverage combined with rapid camera panning, and sequence 41, showcasing Fallout gameplay with intense detail and fast camera movements, including elements like trees, sand, fire, and rain. In contrast, sequences 101 and 104 show low SA and TA values, corresponding to a blurry recording of a caterpillar moving slowly along a branch.

Table 1 summarizes the descriptive statistics of the VQIs computed on the entire video sequence database.

The BLN and BLL metrics show relatively low mean values (0.99 and 1.50, respectively), although BLL displays high variability, as indicated by its large standard deviation (2.47) and maximum value (14.97). This suggests that while most sequences exhibit minor block loss, a subset suffers from severe degradation.

The BLU metric has a mean of 4.575 and a wide range, from 1.33 to 12.59, reflecting a significant variation in sharpness in the data set.

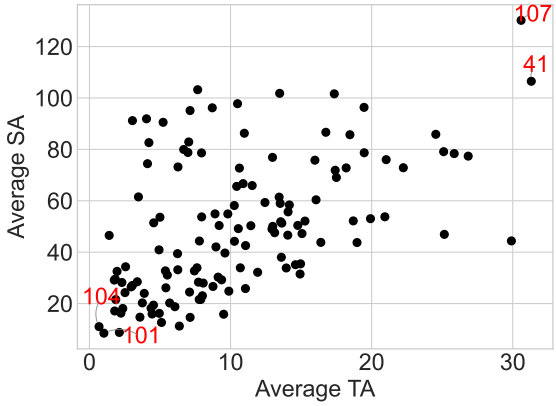


Figure 6: Average SA vs Average TA for each sequence.

EXP and CON both show moderate variability. EXP ranges from 64.77 to 143.34, with a mean of 111.32 and a standard deviation of 16.96, suggesting a well-lit data set, but with some under- and overexposed content. CON follows a similar pattern, ranging from 7.93 to 73.66.

The INT artifact is largely absent in the dataset, as reflected by its mean, median, and lower quartiles, all being zero. The maximum value of 0.03 indicates only rare occurrences of interlacing. The INT column was excluded from Table 1, as its values were almost entirely zero.

NOI shows a moderate mean of 1.44 and a substantial maximum of 6.52, indicating that some sequences are affected by significant noise, likely due to sensor limitations in UGC.

Lastly, the UGC classifier shows a bimodal distribution, with a median and 75th percentile of 1.00 and a mean of 0.67. This suggests that a large portion of the data set consists of UGC, but not all.

Table 1: Statistics of VQIs computed for the database.

	BLN	*BLL*	*BLU*	*EXP*	*CON*	*NOI*	*UGC*
mean	0.99	1.50	4.58	111.32	35.48	1.44	0.67
std	0.06	2.47	1.85	16.96	11.65	1.25	0.46
min	0.50	0.00	1.33	64.77	7.93	0.00	0.00
25%	0.99	0.36	3.40	100.51	26.87	0.65	0.00
50%	1.00	0.74	4.23	115.29	34.91	0.90	1.00
75%	1.00	1.68	5.25	125.58	44.21	1.70	1.00
max	1.09	14.97	12.59	143.34	73.66	6.52	1.00

A comprehensive table containing sequence IDs, names, and associated Video Quality Indicator values is provided in the CSV file available at:
https://qoe.agh.edu.pl/hd-dsequa.
All video sequences can also be accessed from the same location.

4. CONCLUSIONS AND FUTURE WORK

We introduced a diverse dataset of 131 raw video sequences designed for objective quality assessment and video compression research. Each sequence includes key no-reference quality metrics, enabling quick and consistent evaluation. The data set fills a gap in the availability of high-quality, diverse raw content suitable for benchmarking modern video systems. Future work will focus on expanding the dataset, adding distortion variants, and integrating subjective assessments to support broader research applications.

5. ACKNOWLEDGMENTS

The work was funded by the Ministry of Science and Higher Education as part of the subsidy for the Faculty of Computer Science, Electronics, and Telecommunications at AGH University of Krakow.

The presented research results come from a research and development project entitled „Technology for Secure and Reliable Delivery of Professional Audio/Video Contribution Live Transmissions with the Lowest Possible Latency." The project was carried out under Submeasure 1.1.1 of the Intelligent Development Operational Program 2014–2020 and co-financed from the Polish National Center for Research and Development. It was implemented as part of the National Center for Research and Development's Fast Path competition.

REFERENCES

[1] Amirpour Hadi et al. 2022. Between Two and Six? Towards Correct Estimation of JND Step Sizes for VMAF-based Bitrate Laddering. 2022 14th International Conference on Quality of Multimedia Experience (QoMEX), 1–4.

[2] Mercat Alexandre et al. 2021. Comparative Rate-Distortion-Complexity Analysis of VVC and HEVC Video Codecs. IEEE Access, 9: 67813–67828.

[3] Shirvani Moghaddam Shahriar. 2024. The Past, Present, and Future of the Internet: A Statistical, Technical, and Functional Comparison of Wired/Wireless Fixed/Mobile Internet. Electronics, 13(10): 1986.

[4] Götz-Hahn Franz et al. 2021. KonVid-150k: A Dataset for No-Reference Video Quality Assessment of Videos in-the-Wild. IEEE Access, 9: 72139–72160.

[5] Li Zhi et al. 2016. Toward a Practical Perceptual Video Quality Metric. The Netflix Tech Blog.

[6] Leszczuk Mikołaj et al. 2023. "In the Wild" Video Content as a Special Case of User Generated Content and a System for Its Recognition. Sensors, 23(4): 1769.

[7] Leszczuk Mikołaj et al. 2023. Objective Video Quality Assessment and Ground Truth Coordinates for Automatic License Plate Recognition. Electronics, 12(23): 4721.

[8] Zhang Fan et al. 2018. BVI-HD: A Video Quality Database for HEVC Compressed and Texture Synthesized Content. IEEE Transactions on Multimedia, 20(10): 2620–2630.

[9] Uddin Syed et al. 2022. Preliminary Study on Video Codec Optimization Using VMAF. Intelligent Information and Database Systems, 469–480.

[10] Wang Yilin et al. 2019. YouTube UGC Dataset for Video Compression Research. 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP), 1–5.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

SYSTEMY WIZYJNE SWOBODNEGO PUNKTU WIDZENIA O WYSOKIEJ POSTRZEGANEJ JAKOŚCI USŁUG

FREE-VIEWPOINT VIDEO SYSTEMS WITH HIGH QUALITY OF EXPERIENCE

Adam Grzelka

Politechnika Poznańska, Poznań,
adam.grzelka@put.poznan.pl

DOI: 10.15199/59.2025.4.103

Streszczenie: Artykuł dotyczy systemów swobodnego punku widzenia, które są systemami wizyjnymi umożliwiającymi użytkownikowi wybór kierunku i miejsca obserwowania prezentowanej sceny. Autor podejmuje za cel zapewnienie wysokiej postrzeganej jakości usług w takich systemach, który realizuje poprzez szereg propozycji oryginalnych rozwiązań. Rozwiązania te dotyczą między innymi optymalizacji ustawienia kamer systemu wielokamerowego, zbadania wpływu stratnej kompresji, opóźnień transmisji do terminala użytkownika oraz realizacji nowego systemu wielokamerowego i serwera swobodnej nawigacji.

Abstract: The work deals with free-viewpoint video systems, which allow a user to choose the location and viewing direction for observing the presented scene. The author proposes a set of original solutions to obtain a high-quality experience in such systems. These proposals relate to optimizing the positioning of the cameras in a multi-camera system, investigating the impact of lossy compression, the transmission delay to the user terminal, and implementing a new multi-camera system and a free-viewpoint server.

Słowa kluczowe: estymacja map głębi, postrzegana jakość usług, swobodny punkt widzenia, synteza widoku wirtualnego

Keywords: depth map estimation, free-viewpoint, quality of experience, virtual view syntesis

1. WSTĘP

Systemy swobodnego punktu widzenia (ang. Free-viewpoint Video – FVV; Free-viewpoint Television - FTV)[20][5] umożliwiają użytkownikowi samodzielny wybór miejsca i kierunku, z którego obserwuje on zarejestrowaną scenę. Widok prezentowany użytkownikowi nie musi pochodzić bezpośrednio z rzeczywistej kamery, lecz być wirtualnie wygenerowany na podstawie obrazów z wielu kamer, rozmieszczonych wokół sceny. Dzięki temu użytkownicy mogą swobodnie „poruszać się" w przestrzeni wizualnej – niezależnie od siebie i w czasie rzeczywistym.

Systemy tego typu znajdują zastosowanie m.in. w transmisjach wydarzeń sportowych czy koncertów, a także w rozwiązaniach immersyjnych, takich jak wyświetlacze nagłowne (HMD)[17]. Kluczowym wyzwaniem w ich projektowaniu jest zapewnienie wysokiej jakości doświadczenia użytkownika (Quality of Experience – QoE), która zależy nie tylko od parametrów technicznych, ale również od subiektywnego komfortu odbiorcy.

Celem pracy, jest znalezienie rozwiązań, które umożliwiają uzyskanie wysokiej jakości postrzeganej (QoE) w systemach swobodnego punktu widzenia. Autor skoncentrował się na analizie, jak różne elementy systemu – w tym ustawienie kamer, kompresja danych i opóźnienia transmisji – wpływają na odbiór jakości przez użytkownika. Autor przeprowadził badania zarówno teoretyczne, jak i eksperymentalne w czterech obszarach: konfiguracja kamer, kompresja stratna, opóźnienia w wyświetlaczach nagłownych oraz budowa prototypowego systemu swobodnej nawigacji. Pełen opis przeprowadzonych badań można znaleźć w [10].

2. SYSTEM SWOBODNEJ NAWIGACJI

Projektowanie systemów swobodnej nawigacji wymaga uwzględnienia szeregu aspektów technicznych: od rozmieszczenia i synchronizacji kamer, przez przetwarzanie obrazu (w tym estymację głębi[13]), aż po syntezę widoków[8] i transmisję danych do terminala użytkownika[5]. W literaturze opisano wiele różnych metod reprezentacji sceny trójwymiarowej. Należą do nich m.in. chmury punktów[21], siatki trójkątów, reprezentacje objętościowe[24] czy opisy geometryczne z teksturowaniem[19]. Dostępne są również rozwiązania oparte na wielowidokowej wizji z dodatkowymi mapami głębi (MVD – Multiview Video plus Depth)[16][12][3], które pozwalają na rekonstrukcję widoku z dowolnego punktu przy pomocy tzw. DIBR (Depth-Image-Based Rendering)[9] czyli syntezie widoku wirtualnego na podstawie obrazów i odpowiadających im map głębi[13][8].

W niniejszej pracy skoncentrowano się na podejściu DIBR – polegającym na przesyłaniu rzeczywistych widoków (np. z dwóch wybranych kamer), kompresowanych dowolną techniką stratną (np. AV1[2], HEVC[11], VVC[22]). Następnie, po stronie odbiornika, generowane są mapy głębi dla przesłanych widoków oraz widok wirtualny z punktu wybranego przez użytkownika. Taki schemat (referencyjne widoki i głębia) zapewnia rozsądny kompromis między jakością a złożonością systemu[1].

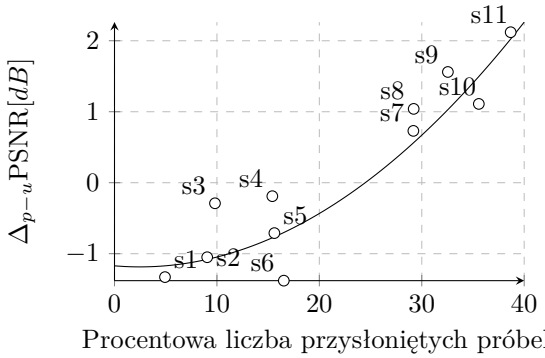
Do oceny jakości systemu wykorzystano obiektywne

miary jakości takie jak PSNR oraz IV-PSNR[6]. Wpływ opóźnień transmisji na postrzeganą jakość swobodnej nawigacji oceniano na podstawie testów subiektywnych.

3. OPTYMALIZACJA USTAWIENIA KAMER W SYSTEMACH SWOBODNEJ NAWIGACJI

W celu określenia optymalnego rozmieszczenia kamer w systemie FTV, w pracy przyjęto pewien model systemu wielokamerowego. Składa się on z czterech kamer. Kamery skrajne nie zmieniają swojej pozycji, natomiast kamery wewnętrzne mogą zostać ustawione tak, aby odległości między sąsiednimi kamerami w systemie były równe lub przesuwać się w kierunku kamer skrajnych, tworząc pary kamerowe. Model ten pozwala na teoretyczne wyprowadzenie wpływu zmiany położenia kamer na najważniejsze zjawiska wpływające na jakość widoku wirtualnego. Na jego podstawie wyznaczono liczbę przysłoniętych punktów obrazu oraz wyprowadzono dwa uproszczone modele teoretyczne opisujące wpływ wybranych parametrów systemu na jakość wirtualnego obrazu.

Pierwszy model uwzględnia zmianę odległości pomiędzy kamerami (odległości bazowej), analizując wpływ tej zmiany na teoretyczną dokładność mapy głębi i jakość wirtualnego widoku. Drugi skupiał się na analizie powierzchni obszarów przysłoniętych w scenie, których nie można poprawnie zsyntezować. Oba modele pozwoliły na uzyskanie wyników teoretycznych, które są sumą omówionych wyżej modeli. Wynik, w funkcji procentowej liczby przysłoniętych próbek dla kamer ustawionych równomiernie, jako Δ_{p-u}PSNR czyli zmiana wartości metryki PSNR pomiędzy systemem z kamerami ustawionymi w pary (ang. pair) oraz równomiernie (ang. uniform) przedstawiono w postaci linii na rysunku 1.



Rysunek 1: Porównanie wyników eksperymentalnych i modelu teoretycznego

Dla weryfikacji modeli teoretycznych przeprowadzono eksperymenty na wielokamerowych sekwencjach testowych, analizując wpływ ustawienia kamer na jakość syntezowanego widoku. Na wykresie 1 wyniki eksperymentalne są przedstawione w postaci kropek, każda oznacza inną sekwencje testową, do której odniesienie można znaleźć w [10]. Eksperyment wykazał zgodność pomiędzy modelami teoretycznymi a jakością ocenianą przez PSNR.

Wyniki teoretyczne, jak i eksperymentalne pokazują, że stosowanie par kamer o małej odległości bazowej daje zysk, kiedy w scenie występuje wiele przysłonięć. W przypadku małej liczby przysłonięć lepiej stosować układy kamerowe z równomiernym ich rozłożeniem. Liczba przysłoniętych punktów, dla kamer ustawionych równomiernie, dla których warto stosować pary kamer to około 25%. Jest to istotna wskazówka dla projektowania rzeczywistych systemów FTV[3][20][5].

4. WPŁYW KOMPRESJI NA JAKOŚĆ WIRTUALNEGO WIDOKU

W tym rozdziale przebadano, w jaki sposób jakość przesyłanych widoków rzeczywistych wpływa na jakość widoku syntetycznego generowanego techniką DIBR[9]. We wszystkich przypadkach przesyłano jedynie zakodowaną stratnie wizje, natomiast mapy głębi oraz widoki wirtualne były estymowane lokalnie po stronie odbiornika. W ramach tej części badań przeprowadzono trzy eksperymenty: wstępny, ogólny oraz wykorzystujący technikę MIV[14].

Eksperyment wstępny miał charakter eksploracyjny i został przeprowadzony z wykorzystaniem starszych technik kompresji, estymacji głębi oraz syntezy widoku wirtualnego[4]. Dla każdej sekwencji testowej zakodowano widoki przy użyciu trzech kodeków: MPEG-2[18], AVC (H.264)[23] i HEVC (H.265)[11]. Widok wirtualny generowano ze zdekodowanych widoków, i porównywano z widokami syntezowanymi z oryginalnych, niezakodowanych widoków. Wyniki pokazały, że kompresja w systemach swobodnego widzenia jest możliwa i nie powoduje dużej straty jakości widoku wirtualnego.

W eksperymencie ogólnym przeanalizowano wiele sekwencji testowych i cztery techniki kompresji: MPEG-2[18], AVC(H.264)[23], HEVC(H.265)[11] oraz VVC[22]. Dla różnych konfiguracji kodera mierzono PSNR widoków przesyłanych oraz IV-PSNR[6] widoku syntezowanego. Wyniki potwierdziły obserwację z eksperymentu wstępnego oraz pozwoliły na zebranie danych dla dalszych badań, z wykorzystaniem aktualnych metod estymacji map głębi i syntezy wirtualnego widoku.

W ostatnim eksperymencie wykorzystano technikę MPEG Immersive Video (MIV)[14][7]. W tym przypadku kodowano reprezentację sceny przy użyciu czterech kodeków: HEVC[11], VVC[22], AV1[2] i VP9[15]. Wykorzystano profil normy MIV GA (ang. Geometry Absent)[14], w którym mapy głębi nie są wysyłane a wyznaczane w dekoderze. Także w tym przypadku jakość widoku syntetycznego była silnie zależna od jakości zdekodowanych widoków.

Wyniki przedstawione w eksperymentach ogólnym oraz z użyciem MIV pozwoliły na dokładne przebadanie zależności pomiędzy jakością widoków przesyłanych i syntezowanych. W tym celu obliczono współczynnik korelacji Pearsona[25], który był bardzo wysoki dla prawie wszystkich przeprowadzonych eksperymentów. Następnie zaproponowano prosty model liniowy, który nie jest zależny od techniki kompresji. Pozwala on na oszacowanie jakości widoku syntezowanego na podstawie jakości przesyłanych widoków, dla danej sekwencji, zgodnie ze wzorem:

$$\text{IV-PSNR}_p \approx f_{AG}(\text{PSNR}_p) = a \cdot \text{PSNR}_p + b, \quad (1)$$

gdzie IV-PSNR$_p$ oraz PSNR$_p$ to zmierzone wartości metryk oceny jakości odpowiednio widoku wirtualnego oraz przesyłanego, a f_{AG} jest zaproponowanym liniowym modelem tej zależności[10].

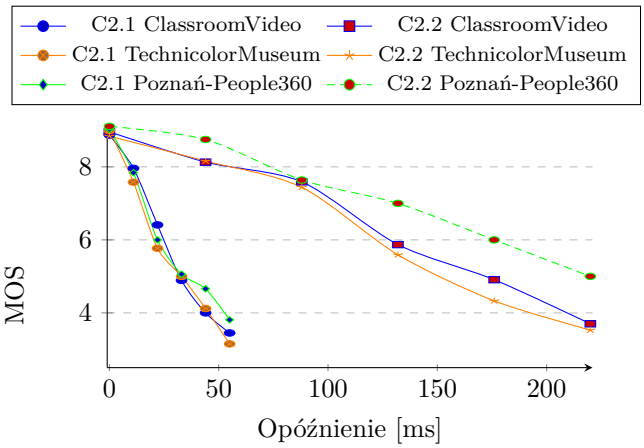
Podsumowując przeprowadzone badania, kluczowy wpływ na jakość widoku syntetycznego ma jakość przesyłanych widoków a metoda kompresji (np. VVC, AV1,

MPEG-2) ma drugorzędne znaczenie. Prosty model jakości może służyć jako skuteczne narzędzie predykcyjne, wspierające projektowanie i optymalizację systemów FTV.

5. ANALIZA WPŁYWU OPÓŹNIEŃ NA POSTRZEGANĄ JAKOŚĆ USŁUGI W WYŚWIETLACZACH NAGŁOWNYCH

W kolejnej części pracy przeanalizowano wpływ opóźnień na jakość swobodnej nawigacji użytkownika korzystającego z wyświetlacza nagłownego[17]. W ramach tych prac przeanalizowano różne scenariusze wymiany danych pomiędzy terminalem użytkownika i serwerem. Zidentyfikowano dwa rodzaje opóźnień, które mogą mieć różne wartości dla badanych scenariuszy. Są to opóźnienia wynikające ze zmiany pozycji użytkownika w scenie oraz zmiany kierunku patrzenia.

Następnie przeprowadzono eksperyment, który miał zbadać subiektywną jakość dla różnych scenariuszy transmisji, a co za tym idzie wartości opóźnień. Został on przeprowadzony za pomocą urządzenia Oculus Rift CV1. W trakcie eksperymentu uczestnicy oceniali w 11-punktowej skali MOS jakość swobodnej nawigacji w trzech sekwencjach testowych, dla dwóch scenariuszy transmisji oraz kilkunastu wariantów opóźnienia. Scenariusz C2.1 to taki, w którym opóźniona jest pozycja oraz kierunek patrzenia użytkownika o tę samą wartość. Wariant C2.2 to taki, w którym występuje jedynie opóźnienie pozycji użytkownika, a kierunek patrzenia syntezowany jest bez opóźnienia. Wyniki zostały przedstawione na rysunku 2.



Rysunek 2: Uśrednione wyniki testów subiektywnych dla wszystkich sekwencji i scenariusza C2.1 oraz C2.2

Pierwsza grupa krzywych to wyniki uzyskane dla przypadku z opóźnioną zarówno pozycją użytkownika w scenie, jak i kierunkiem patrzenia. Druga grupa krzywych obrazuje przypadek, gdzie jedynie opóźnienie wiąże się z pozycją użytkownika w scenie.

Wyniki wykazały, że opóźnienia rotacyjne (scenariusz C2.1) są znacznie bardziej odczuwalne i dokuczliwe dla użytkowników niż opóźnienie jedynie pozycji użytkownika w scenie (scenariusz 2.2). Wykorzystując scenariusz transmisji, który zapewnia niską wartość tego opóźnienia, wysoka jakość usługi swobodnej nawigacji może być zapewniona nawet przy czterokrotnie większym czasie komunikacji terminala użytkownika z serwerem.

6. EKSPERYMENTALNY SYSTEM SWOBODNEJ NAWIGACJI

Celem tej części badań było opracowanie i uruchomienie eksperymentalnego systemu FTV umożliwiającego swobodną nawigację widza w zarejestrowanej rzeczywistej scenie. System miał być możliwie prosty i tani w realizacji.

System składał się z dwóch głównych elementów, pierwszy to system wielokamerowy, drugi to serwer brzegowy, który dostarcza dane do terminala użytkownika, którym może być dowolne urządzenie z przeglądarką internetową.

System wielokamerowy został zbudowany z popularnych, konsumenckich kamer Go Pro Hero 4. Jest to kamera sportowa, która posiada złącze w tylnej części służące do podłączania akcesoriów, oraz umożliwiające złączenie dwóch kamer w stereopary[27]. Ta funkcjonalność została wykorzystana do budowy systemu wielokamerowego. Zaprojektowano dedykowany moduł „plecaków" kamer, czyli płytki elektronicznej dołączonej do kamery. Płytka ta odpowiadała za zarządzanie kamerą oraz odbiór informacji sterujących i synchronizacyjnych od zarządcy systemu. Z uwagi na duże odległości, zbudowano również moduły rozdzielające i regenerujące przesyłane sygnały. Zarządca systemu jest urządzeniem elektronicznym, które umożliwia sterowanie wszystkimi kamerami, a w szczególności zapewnia synchronizację rejestrowanych ramek.

Drugi element systemu to serwer swobodnej nawigacji. Korzysta on z przygotowanych widoków oraz map głębi. Użytkownicy systemu podłączają się do serwera przez stronę internetową. Zawiera ona kod skryptu JavaScript, który realizuje wyświetlanie nowych przesyłanych widoków wirtualnych, jak i umożliwia ich zmianę, czyli swobodną nawigację w scenie. Sam serwer jest wielowątkową implementacją zrealizowaną w języku C++. Serwer może być uruchomiony nawet na konsumenckim komputerze i obsługiwać jednocześnie kilku użytkowników.

Przeprowadzone badania pokazały, że budowa systemu swobodnego punktu widzenia jest możliwa, nawet z wykorzystaniem prostych konsumenckich komponentów. Serwer ten, mimo prostej wielowątkowej realizacji, może obsługiwać nawet kilkunastu użytkowników. Zarejestrowana scena musi jednak zostać odpowiednio przygotowana, w szczególności muszą zostać wyznaczone parametry systemu oraz mapy głębi.

7. PODSUMOWANIE

Przedstawione badania dotyczą systemów swobodnego punktu widzenia (FTV), ze szczególnym uwzględnieniem czynników wpływających na jakość odbioru usługi (QoE). Obejmują one zarówno prace teoretyczne, jak i praktyczne, prowadzone w czterech głównych obszarach: konfiguracji kamer, kompresji, wpływu opóźnień oraz implementacji systemu eksperymentalnego.

W pracy przeanalizowano wpływ rozmieszczenia kamer na jakość syntezowanego widoku. W szczególności wykazano, że układ kamer w postaci par o małej odległości bazowej jest korzystniejszy w scenach zawierających przysłonięcia. Poza tym zaproponowano model jakości widoku wirtualnego, który pozwala oszacować jakość widoku syntezowanego na podstawie jakości przesyłanych danych. Model ten, pomimo prostej liniowej definicji, wykazał dużą zgodność z wynikami eksperymentalnymi oraz co najistotniejsze, nie zależy od zastosowanej techniki kompresji. Również eksperymentalnie zbadano wpływ transmisji

w systemach z wyświetlaczem nagłownym z uwzględnieniem opóźnień kierunku patrzenia oraz pozycji użytkownika. Wyniki subiektywne pokazały około czterokrotnie większą wrażliwość użytkowników na opóźnienia kierunku patrzenia – co pozwala na dobranie takiego scenariusza transmisji, który poprzez minimalizację tego czasu może zapewnić wyższą postrzeganą jakość. W ramach pracy również zaprojektowano i uruchomiono prosty kompletny system FTV. System umożliwia synchroniczne nagrywanie rzeczywistych wydarzeń oraz ich dostarczenie do terminali użytkowników końcowych, z usługą swobodnej nawigacji w scenie. Stanowi on w pełni funkcjonalną platformę badawczo-testową.

Pełne wyniki przedstawionych badań można znaleźć w pracy doktorskiej [10].

LITERATURA

[1] Boyce J. M. i in. 2021. „MPEG Immersive Video Coding Standard". Proceedings of the IEEE 109.9. s. 1521–1536.

[2] Chen Y. i in. 2018. „An Overview of Core Coding Tools in the AV1 Video Codec". 2018 Picture Coding Symposium (PCS). s. 41–45.

[3] Domański M. i in. 2014. „Experiments on Acquisition and Processing of Video for Free-Viewpoint Television". 2014 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON). s. 1–4.

[4] Domański M. i in. 2016. „Optimization of Camera Positions for Free-Navigation Applications". 2016 International Conference on Signals and Electronic Systems (ICSES). s. 118–123.

[5] Domański M. 2015. „Widz wybiera punkt widzenia". Academia, Magazyn Polskiej Akademii Nauk, No. 2 (42)., s. 48-49.

[6] Dziembowski A. i in. 2022. „IVPSNR — The Objective Quality Metric for Immersive Video Applications". IEEE Transactions on Circuits and Systems for Video Technology 32.11. s. 7575–7591.

[7] Dziembowski A. i in. 2022. „Nowe techniki kompresji wizji dla rzeczywistości wirtualnej – MPEG Immersive Video". Przegląd telekomunikacyjny. s. 116-123.

[8] Dziembowski A. i in. 2018. „Szybka synteza widoków wirtualnych dla systemów telewizji swobodnego punktu widzenia". Przegląd Telekomunikacyjny, Vol. 91, No. 6. s. 330-333.

[9] Fehn C. 2003. „A 3D-TV Approach Using Depth-Image-Based Rendering DIBR". in Proceedings of 3rd IASTED Conference on Visualization, Imaging, and Image Processing, Benalmadena. s. 482-487.

[10] Grzelka A. 2024. „Systemy wizyjne swobodnego punktu widzenia o wysokiej postrzeganej jakości usług,". Praca doktorska. Politechnika Poznańska.

[11] Hu Q. i in. 2014. „Analysis and optimization of x265 encoder". 2014 IEEE Visual Communications and Image Processing Conference. s. 502–505.

[12] Merkle P. i in. 2007. „Multi-View Video Plus Depth Representation and Coding". 2007 IEEE International Conference on Image Processing, San Antonio, TX, USA, s. I - 201-I - 204.

[13] Mieloch D. i in. 2017. „Estymacja głębi dla systemów wielowidokowych". Przegląd Telekomunikacyjny, Vol. 86, No. 6. s. 479-482.

[14] Mieloch D. i in. 2022. „Overview and Efficiency of Decoder-Side Depth Estimation in MPEG Immersive Video". IEEE Transactions on Circuits and Systems for Video Technology 32.9. s. 6360–6374.

[15] Mukherjee D. i in. 2013. „The latest open-source video codec VP9 - An overview and preliminary results". 2013 Picture Coding Symposium (PCS). s. 390–393.

[16] Muller K. i in. 2011. „3-D Video Representation Using Depth Maps". Proceedings of the IEEE 99.4, s. 643–656.

[17] Shibata T. 2002. „Head Mounted Display". Displays 23.1, s. 57 –64. issn: 0141-9382. http://www.sciencedirect.com/science/article/pii/S0141938202000100

[18] Sikora T. 1997. „MPEG digital video-coding standards". IEEE Signal Processing Magazine 14.5. s. 82–100.

[19] Smolic A. i in. 2005. „3D Video Objects for Interactive Applications". 2005 13th European Signal Processing Conference. s. 1–4.

[20] Tanimoto M. 2010. „FTV (Free-viewpoint TV)". 2010 IEEE International Conference on Image Processing. s. 2393–2396.

[21] Wei K. i in. 2013. „Point-Based Model Construction for Free-Viewpoint TV". 2013 IEEE Third International Conference on Consumer Electronics Berlin (ICCE-Berlin). s. 220–221.

[22] Wieckowski A. i in. 2021. „Vvenc: An Open And Optimized Vvc Encoder Implementation". 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). s. 1–2.

[23] Wiegand T. i in. 2003. „Overview of the H.264/AVC video coding standard". IEEE Transactions on Circuits and Systems for Video Technology 13.7, s. 560–576.

[24] Wu G. i in. 2023. „4D Gaussian Splatting for Real-Time Dynamic Scene Rendering". arXiv preprint arXiv:2310.08528.

[25] N.A. Weiss. 2012. Introductory Statistics. Pearson Education, isbn:9780321691224. url: https://books.google.pl/books?id=_r5ucgAACAAJ

[26] C. Fehn i in. 2006. „Interactive 3-DTV-Concepts and Key Technologies". W: Proceedings of the IEEE 94.3, s. 524–538.

[27] Dual HERO System. Dostęp: 2020-05-06. https://gopro.com/content/dam/help/dual-hero-system/manuals/DualHeroSystem_UG_SPA_REVD_WEB.pdf.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

KLASYFIKACJA ZDARZEŃ AKUSTYCZNYCH W SYSTEMIE MONITORINGU BEZPIECZEŃSTWA OPARTA NA UCZENIU MASZYNOWYM
CLASSIFICATION OF ACOUSTIC EVENTS IN SAFETY-RELATED SCENARIOS USING MACHINE LEARNING TECHNIQUES

Józef Kotus, Maciej Szczodrak, Piotr Szczuko, Grzegorz Szwoch

Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki;
Katedra Systemów Multimedialnych, Gdańsk, Polska, jozkotus@pg.edu.pl

DOI: 10.15199/59.2025.4.104

Streszczenie: **W referacie przedstawiono rozwiązanie przeznaczone do systemów monitoringu bezpieczeństwa, które umożliwia klasyfikację zdarzeń akustycznych na podstawie analizy dźwięku z wykorzystaniem sieci neuronowych: model bazujący na EfficientNet, stosujący parametryzację zdarzeń akustycznych oraz na YAMNet, analizujący strumień próbek audio. Opracowano zbiór zawierający 5059 zdarzeń akustycznych w ośmiu klasach. Wyniki rozpoznawania zdarzeń (niebezpieczne lub bezpieczne) wynosiły 78% dla klasyfikatora EfficientNet i 82% dla klasyfikatora YAMNet.**

Abstract: **The paper presents neural network-based classifiers for recognition of security threats based on audio analysis. Two classifier models were tested: a model based on EfficientNet that analyzed parameters extracted from the detected audio events, and based on YAMNet that performed event recognition on raw audio data. A custom dataset containing 5059 acoustic events of eight classes was created. The classification results (dangerous or normal events) were 78% for the EfficientNet and 82% for the YAMNet.**

Słowa kluczowe: **detekcja zdarzeń akustycznych, klasyfikacja zdarzeń, sieci splotowe, YAMNet.**

Keywords: **acoustic events detection, CNN, event classification, YAMNet.**

1. WSTĘP

Systemy monitoringu bezpieczeństwa są zazwyczaj oparte na modalności wizyjnej. Modalność dźwiękowa jest wykorzystywana stosunkowo rzadko. Rozwój technik uczenia maszynowego daje możliwość skutecznego rozpoznawania niebezpiecznych zdarzeń na podstawie analizy strumienia dźwiękowego [10, 20]. Zaproponowano liczne modele do różnorodnych zastosowań, np. rozpoznawania ptaków [12] i dźwięków związanych z aktywnością człowieka [8, 10, 14] oraz modeli do dialogu człowieka z urządzeniem [1]. Powstają również modele do rozpoznawania potencjalnie niebezpiecznych zdarzeń [2, 3, 5, 9, 11, 19], wspierane przez systemy lokalizacji źródła akustycznego [4]. Problemem jest nie tylko poprawne rozpoznanie zdarzenia akustycznego, ale także odróżnienie zdarzeń niebezpiecznych od normalnych. Referat opisuje prace dotyczące opracowania takiego klasyfikatora oraz zbioru danych do uczenia i testowania klasyfikatora.

Większość istniejących klasyfikatorów opartych na sieciach neuronowych analizuje surowy strumień dźwiękowy. Podejście takie skutkuje wzrostem ryzyka powstawania fałszywie dodatnich wyników oraz dużym obciążeniem systemu dokonującego klasyfikacji zdarzeń. W niniejszym referacie zaproponowano hybrydowe rozwiązanie, oparte o: urządzenia brzegowego - instalowane typowo w miejscach występowania zdarzeń akustycznych, urządzenie to dokonuje wstępnej analizy sygnału, wydziela i parametryzuje potencjalnie niebezpieczne zdarzenia ze strumienia akustycznego, a następnie, na urządzeniu węzłowym, po odebraniu komunikatów z urządzeń brzegowych, klasyfikator oparty na sieci neuronowej dokonuje rozpoznania typu danego zdarzenia.

2. ALGORYTM KLASYFIKACJI ZDARZEŃ

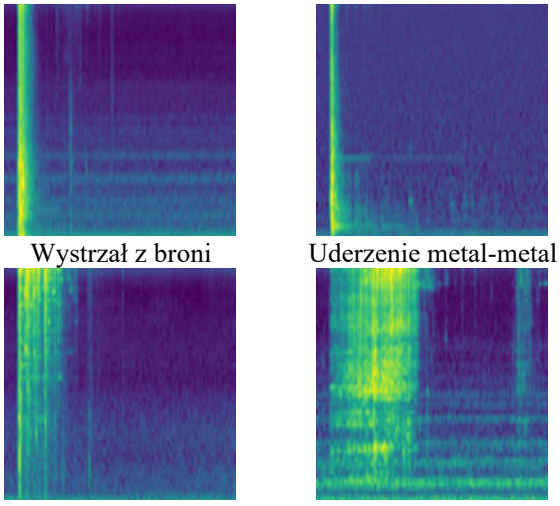
Algorytm klasyfikacji zdarzeń składa się z dwóch zasadniczych części. Pierwsza część, uruchamiana na urządzeniu brzegowym, dokonuje ekstrakcji i parametryzacji zdarzeń akustycznych. Druga część, przeznaczona do uruchamiania na urządzeniu węzłowym, dokonuje rozpoznawania zdarzeń i określania czy stanowią one potencjalne zagrożenie.

2.1. Detekcja i parametryzacja zdarzeń

W pierwszym etapie przetwarzane są strumienie dźwiękowe pozyskane z mikrofonów. Wykrywane są zdarzenia akustyczne wydzielone z tła akustycznego. Detektor zdarzeń jest adaptacyjnym układem analizującym energię sygnału w krótkich fragmentach (8 ms). Estymaty energii sygnału i energii szumu są na bieżąco uaktualniane metodą uśredniania wykładniczego. Detekcja zdarzeń polega na wydzieleniu fragmentów strumienia dźwiękowego, w których estymata energii sygnału jest większa od szumu. Przykłady wybranych zdarzeń akustycznych przedstawiono na rys. 1.

Wykryte zdarzenia akustyczne są poddawane parametryzacji. Unormowany czas zdarzenia jest równy 4 sekundom (co wynika z rozmiaru warstwy wejściowej klasyfikatora) – zdarzenia są przycinane lub uzupełniane zerami w razie potrzeby. Wydzielony fragment sygnału spróbkowanego z częstotliwością 16 kHz jest dzielony na 125 segmentów o długości 2048 próbek każdy, z zakładkowaniem 50%. Każdy segment jest mnożony przez okno Hamminga i poddawany przekształceniu Fouriera. Następnie kwadrat modułu widma sygnału jest przetwarzany przez bank 54 filtrów melowych. Całe czterosekundowe zdarzenie zostaje zatem opisane przez macierz 125×54 współczynników mel-cepstralnych (MFCC) [6]. Współ-

czynniki są następnie normalizowane do zakresu 0-1 i zamieniane na obraz w przestrzeni barwnej RGB, z wykorzystaniem mapy kolorów Viridis [15]. Obraz ten stanowi opis zdarzenia, przekazywany do klasyfikatora.



Rys. 1. Przykłady wybranych zdarzeń akustycznych

2.2. Klasyfikator zdarzeń

Klasyfikator zdarzeń akustycznych jest oparty na sieci neuronowej. Analizuje on zestaw parametrów zdarzenia w formie obrazu RGB. Klasyfikator składa się ze splotowej sieci neuronowej EfficientNet-B3 [7], działającej jako ekstraktor cech, oraz trzech warstw sieci zawierających odpowiednio 1280, 512 i 256 neuronów, dokonujących klasyfikacji zdarzeń. W procesie uczenia sieci zastosowano dodatkowo warstwę typu dropout pomiędzy warstwami 1280 i 512. Wynikiem parametryzacji czterosekundowego zdarzenia akustycznego jest tensor o rozmiarach 54×125×3 (125 segmentów po 54 współczynniki MFCC, kodowane za pomocą kanałów RGB). Tensory te są skalowane do rozmiaru 224×224×3, tak aby ich rozmiar odpowiadał rozmiarowi wejściowej warstwy klasyfikatora.

Aby umożliwić ocenę skuteczności klasyfikatora, wykorzystano drugi klasyfikator, oparty na modelu YAMNet [13]. W odróżnieniu od modelu EfficientNet, model YAMNet nie wykorzystuje sparametryzowanych zdarzeń, a operuje bezpośrednio na fragmentach strumienia dźwiękowego wyodrębnionych na etapie detekcji zdarzeń. Oryginalny model YAMNet, dokonujący klasyfikacji 521 typów zdarzeń, został wytrenowany z użyciem metryki AP@10, tzn. wynik klasyfikacji jest poprawny, jeżeli mieści się on wśród 10 najwyższych wyników [17]. Do celów eksperymentów opisanych w referacie dokonano treningu klasyfikatora z użyciem metryki AP@1 (dokładne dopasowanie wyniku) oraz dostosowano go do detekcji wybranych ośmiu klas zdarzeń. Model YAMNet dokonuje wewnętrznej parametryzacji dźwięku metodą MFCC, z wykorzystaniem 64 pasm melowych.

2.3. Zbiór nagrań i klasy zdarzeń

Do celów eksperymentów zdefiniowano osiem klas zdarzeń, przedstawionych w tabeli 1. Pierwsze cztery klasy reprezentują zdarzenia niebezpieczne, pozostałe cztery – zdarzenia normalne. Zdarzenie typu *sweep* jest sztucznie wygenerowanym sygnałem tonalnym o przestrajanej częstotliwości. W celu wytrenowania i testowania obu klasyfikatorów przygotowano własny zbiór przykładów dźwiękowych reprezentujących zdarzenia z tych klas. Nagrania wykonano w rzeczywistych warunkach, w pomieszczeniach i na zewnątrz, zapisano je w formacie monofonicznym, 32-bitowym, z częstotliwością próbkowania 32 kHz. Z nagrań wyodrębniono ręcznie czterosekundowe fragmenty i nadano im etykiety zdarzeń. Łącznie zebrano 5059 przykładów zdarzeń (tabela 1). Następnie utworzono zbiór zawierający po 500 przykładów dla każdej klasy, wybierając przykłady losowo lub uzupełniając klasę przykładami generowanymi metodą augmentacji poprzez dodanie szumu gaussowskiego. Z tego zbioru wydzielano zbiór uczący (po 400 przykładów z każdej klasy) oraz zbiór testowy (pozostałe 100 przykładów z klasy plus ew. przykłady nie wykorzystane do treningu).

Tab. 1. Klasy zdarzeń i ich liczebności

Klasa zdarzenia	*Wewn.*	*Zewn.*	*Razem*
Wystrzał	1229	1386	2615
Tłuczone szkło	442	319	761
Krzyk	727	0	727
Eksplozja	25	182	207
Metal-metal	128	0	128
Metal-beton	127	0	127
Sweep	244	0	244
Szum tła	0	250	250

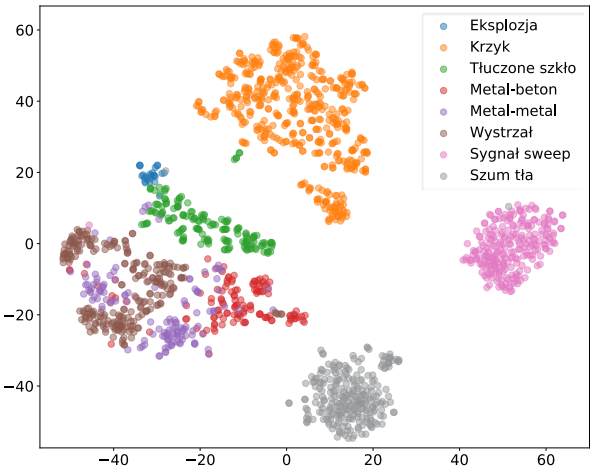
3. EKSPERYMENTY

W trakcie eksperymentów porównano dokładność rozpoznawania zdarzeń z użyciem obu klasyfikatorów, wykorzystując opracowany testowy zbiór przykładów dźwiękowych. Do oceny wykorzystano trzy standardowe metryki: precyzję (ang. *precision*), czułość (ang. *sensitivity*) oraz miarę F1 (ang. *F1-score*) [16]. Metryki te obliczano w trybie binarnym (zdarzenie niebezpieczne lub normalne) oraz w trybie mikro i makro, z uwzględnieniem poprawności oznaczenia klasy zdarzenia.

3.1. Klasyfikator EfficientNet

Na rys. 2 przedstawiono wyniki wstępnego eksperymentu z wykorzystaniem klasyfikatora EfficientNet. Wynikiem jego działania były wektory 1024 elementów, które zostały zredukowane do dwóch wymiarów za pomocą algorytmu t-SNE (ang. *t-distributed stochastic neighbor embedding*) [18], co pozwoliło na ich wizualizację oraz na ocenę separowalności klas zdarzeń. Klasyfikator był w stanie skutecznie wykryć klasy wybuchu, tłuczonego szkła oraz krzyku. Problemem jest natomiast rozróżnienie zdarzeń wystrzału oraz uderzenia metalu o metal, ze względu na podobny charakter akustyczny zdarzenia oraz z powodu użycia metalowego elementu do wyłapywania pocisków podczas nagrywania dźwięków wystrzałów na strzelnicy. W klasie krzyku obserwuje się największą wariancję, co jest spowodowane zróżnicowaniem charakteru głosu u różnych osób. Wyniki tego eksperymentu pozwoliły na zidentyfikowanie potencjalnych problemów w klasyfikacji zdarzeń. W tabeli 2 przedstawiono metryki dokładności klasyfikacji dla poszczególnych klas zdarzeń z wykorzystaniem klasyfikatora EfficientNet, opartego na

sparametryzowanych zdarzeniach akustycznych. Największą dokładność klasyfikacji uzyskano dla zdarzenia typu sweep. Jest to specyficzny sygnał, odmienny od pozostałych typów. Najczęściej mylone były zdarzenia: wystrzał i eksplozja oraz tłuczone szkło i eksplozja. Są to zdarzenia akustyczne o charakterze impulsowym, ich tensory parametrów są podobne do siebie.



Rys. 2. Rozkład cech klas wyznaczony za pomocą algorytmu t-SNE na podstawie modelu EfficientNet.

Tab. 2. Precyzja i czułość dla EfficientNet (EN) i YAMNet (YN)

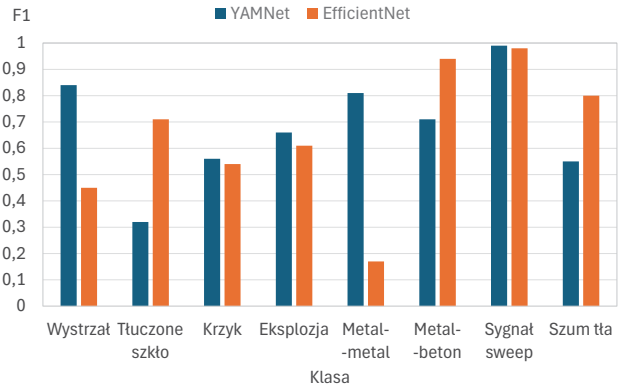
Klasa zdarzenia	*Precyzja*		*Czułość*	
	EN	*YN*	*EN*	*YN*
Wystrzał	0,30	0,95	0,92	0,75
Tłuczone szkło	0,76	1,00	0,66	0,19
Krzyk	0,93	0,92	0,38	0,40
Eksplozja	0,51	1,00	0,78	0,49
Metal-metal	0,09	0,84	0,69	0,78
Metal-beton	0,99	0,59	0,90	0,88
Sweep	0,96	0,97	0,99	1,00
Szum tła	0,75	0,38	0,85	1,00

3.2. Klasyfikator YAMNet

Wyniki rozpoznawania zdarzeń dla klasyfikatora YAMNet, działającego w oparciu o analizę niesparametryzowanych fragmentów strumienia dźwiękowego, przedstawiono w tabeli 2. Klasyfikator YAMNet okazał się skuteczny w wykrywaniu wystrzałów oraz uderzenia metalem o metal, z kolei w przypadku tłuczonego szkła i szumu tła, skuteczność była znacznie niższa niż dla klasyfikatora EfficientNet.

3.3. Porównanie klasyfikatorów

Na rys. 3 przedstawiono porównanie miar F1 w poszczególnych klasach, uzyskanych dla obu klasyfikatorów. W większości klas widoczne są wyraźne różnice między klasyfikatorami. YAMNet był bardziej skuteczny w rozpoznawaniu wystrzałów i uderzenia metalu o metal. EfficientNet wykazał lepszą dokładność klasyfikowania tłuczonego szkła i uderzania metalem o beton. Różnice te wskazują na to, że hybrydowy model klasyfikatora, wykorzystujący oba podejścia (parametryczne i bezpośrednie) mógłby poprawić skuteczność klasyfikacji. W tabeli 3 przedstawiono porównanie precyzji i miary F1 obliczonych zbiorczo dla wszystkich klas, w trybie binarnym (zdarzenie niebezpieczne lub normalne) oraz w trybie mikro i makro (poprawne rozpoznanie klasy). Klasyfikator oparty na YAMNet cechuje się wyższą precyzją w trybie mikro i makro od klasyfikatora opartego na EfficientNet. Test statystyczny Wilcoxona nie wykazał jednak istotnych różnic w metrykach dla sieci EfficientNet i YAMNet (precyzja: $p = 0{,}33$; czułość: $p = 0{,}67$).



Rys. 3. Porównanie miar F1 dla obu klasyfikatorów

Tab. 3. Porównanie skumulowanych metryk

Model	*EfficientNet*		*YAMNet*	
Metryka	*Precyzja*	*F1*	*Precyzja*	*F1*
Mikro	0,61	0,65	0,69	0,69
Makro	0,66	0,65	0,83	0,68
Binarna	0,96	0,78	0,97	0,82

4. WNIOSKI

Wyniki przeprowadzonych eksperymentów wskazują na to, że problem detekcji zdarzeń akustycznych z wykorzystaniem klasyfikatora opartego na sieci neuronowej jest złożony. Testowano dwa podejścia: jedno oparte na parametryzacji zdarzeń (EfficientNet) i drugie oparte na przetwarzaniu strumienia dźwiękowego (YAMNet). Żadne z tych podejść nie jest uniwersalne, nie daje konsekwentnie większej dokładności rozpoznawania dla wszystkich klas zdarzeń. Binarny klasyfikator zdarzeń „niebezpieczne lub normalne" daje wysokie wartości precyzji (0,96-0,97), jednak wyniki rozpoznawania konkretnych klas są znacznie mniej dokładne. Wartość miary F1 dla obu klasyfikatorów (0,65-0,69) jest zbyt niska dla praktycznych zastosowań omawianych klasyfikatorów. Oba typy klasyfikatorów miały problem z rozróżnianiem zdarzeń akustycznych o podobnym charakterze, np. wystrzał i eksplozja.

Jednym z celów przeprowadzonych eksperymentów było sprawdzenie czy parametryzacja zdarzeń akustycznych w klasyfikatorze EfficientNet poprawia skuteczność klasyfikacji w porównaniu z klasyfikatorem YamNet, w którym analizowany jest strumień dźwięku, a parametryzacja jest wykonywana wewnętrznie przez klasyfikator. Ponieważ miary skuteczności klasyfikacji w obu modelach są zbliżone, nie potwierdzono tej tezy. Jednak w rozproszonych systemach monitoringu możliwe jest wykonywanie parametryzacji na urządzeniach brzegowych, odciążając w ten sposób system klasyfikacji działający na

urządzeniu węzłowym, co umożliwia lepsze zbalansowanie obciążenia obliczeniowego w takim systemie, bez szkody dla dokładności klasyfikacji zdarzeń. Ponadto oszczędzana jest przepływność łącza w sieci.

Klasyfikator oparty na EfficientNet wykazał się dobrą skutecznością binarnego rozpoznawania zdarzeń (niebezpieczne lub normalne), co jednak nie przełożyło się na skuteczne rozróżnianie poszczególnych klas zdarzeń. Z kolei klasyfikator YAMNet ma bardzo zróżnicowaną skuteczność klasyfikacji – w niektórych klasach jest on bardziej dokładny od EfficientNet (wystrzały i eksplozje), w innych jest on mniej dokładny (krzyk, tłuczone szkło), co ogranicza jego praktyczne zastosowanie.

Istotnym problemem w trenowaniu klasyfikatorów jest obecność szumu tła. Problem ten można rozwiązać wprowadzając większą liczbę etykiet klas w modelu. W przypadku modelu YAMNet wykorzystującego AP@7 wymagałoby to zmodyfikowania struktury sieci. Można również rozważyć wyeliminowanie szumu tła z nagrań stosowanych do treningu sieci. Dalsze prace w tym temacie będą również dotyczyły oceny możliwości wykorzystania innej architektury sieci neuronowej do rozpoznawania sparametryzowanych zdarzeń akustycznych.

PODZIĘKOWANIA

Badania były współfinansowane przez Narodowe Centrum Badań i Rozwoju (NCBR) ze środków Europejskiego Funduszu Rozwoju Regionalnego, Programu Operacyjnego Inteligentny Rozwój, oś priorytetowa I: Wsparcie prowadzenia prac B+R przez przedsiębiorstwa, numer projektu POIR.01.01.01-00-0785/21.

Autorzy pragną podziękować pracownikom firmy Ambient System sp. z o.o. za wsparcie oraz udostępnienie sprzętu i pomieszczeń niezbędnych do przeprowadzenia eksperymentów.

LITERATURA

[1] Philippe Chabot et al., "Detection and Classification of Human-produced Nonverbal Audio Events," *Applied Acoustics* 171 (2020): 107643, https://doi.org/10.1016/j.apacoust.2020.107643.

[2] Antonio Greco et al., "DENet: A Deep Architecture for Audio Surveillance Applications," *Neural Computing and Applications* 33, no. 17 (2021): 11273–84, https://doi.org/10.1007/s00521-020-05572-5.

[3] John Irungu et al., "Gunshot Detection From Audio Excerpts of Urban Sounds Using Transfer Learning," *Proc. Meetings on Acoustics* 51 (2023): 045003, https://doi.org/10.1121/2.0001783.

[4] Jevon Mao, Marisabel Chang. "A Comparative Study of Deep Learning Models for Network Intrusion Detection." *Proc. Computer Science & Information Technology (CS & IT)* 13, (2023): 123–135. https://doi.org/10.5121/csit.2023.131310.

[5] Han Liang et al., "A Scene-Dependent Sound Event Detection Approach Using Multi-Task Learning," *IEEE Sensors Journal* 22, no. 18 (2021): 17483–89, https://doi.org/10.1109/jsen.2021.3098325.

[6] Paul Mermelstein, "Distance measures for speech recognition, psychological and instrumental", *Pattern Recognition and Artificial Intelligence* 116 (1976): 374–388.

[7] Mingxing Tan and Quoc V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *arXiv (Cornell University)*, January 1, 2019, https://doi.org/10.48550/arxiv.1905.11946.

[8] Sujoy Mondal and Abhirup Das Barman, "Deep Learning Technique Based Real-time Audio Event Detection Experiment in a Distributed System Architecture," *Computers & Electrical Engineering* 102 (2022): 108252, https://doi.org/10.1016/j.compeleceng.2022.108252.

[9] Alex Morehead et al., "Low Cost Gunshot Detection Using Deep Learning on the Raspberry Pi," *2021 IEEE International Conference on Big Data (Big Data)*, December 1, 2019, https://doi.org/10.1109/bigdata47090.2019.9006456.

[10] Wenjie Mu et al., "Environmental Sound Classification Using Temporal-frequency Attention Based Convolutional Neural Network," *Scientific Reports* 11, no. 1 (2021), https://doi.org/10.1038/s41598-021-01045-4.

[11] Suwan Park and Geonwoo Kim, "Pretrained Network-based Sound Event Recognition for Audio Surveillance Applications," *2021 International Conference on Information and Communication Technology Convergence (ICTC)*, October 20, 2021, 1306–9, https://doi.org/10.1109/ictc52510.2021.9621184.

[12] Karol J. Piczak, "Recognizing Bird Species in Audio Recordings Using Deep Convolutional Neural Networks," *CLEF (Working Notes)*, 2016, 534-43.

[13] Plakal M. and Ellis D., "YAMNet", https://github.com/tensorflow/models/tree/master/research/audioset/yamnet. Dostęp 21 maja 2025

[14] Justin Salamon and Juan Pablo Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," *IEEE Signal Processing Letters* 24, no. 3 (2017): 279–83, https://doi.org/10.1109/lsp.2017.2657381.

[15] Nathaniel Smith and Stéfan van der Walt, "A better default map for Matplotlib", SciPy 2015. https://youtu.be/xAoljeRJ3lU. Dostęp 21 maja 2025.

[16] Alaa Tharwat, "Classification Assessment Methods," *Applied Computing and Informatics* 17, no. 1 (2018): 168–92, doi.org/10.1016/j.aci.2018.08.003.

[17] "Testing the YAMNet model", https://github.com/tensorflow/models/blob/master/research/audioset/yamnet/yamnet_test.py. Dostęp 21 maja 2025.

[18] Laurens van der Maaten and Geoffrey Hinton, "Visualizing Data Using t-SNE," *Journal of Machine Learning Research* 9, no. 86 (2008): 2579–2605.

[19] Naing et al, "Real Time End-to-end Glass Break Detection System Using LSTM Deep Recurrent Neural Network," *International Journal of Advanced and Applied Sciences* 6, no. 3 (2019): 56–61, https://doi.org/10.21833/ijaas.2019.03.009.

[20] Jie You, Wenqin Wu, and Joonwhoan Lee, "Open Set Classification of Sound Event," *Scientific Reports* 14, no. 1 (2024), https://doi.org/10.1038/s41598-023-50639-7.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

ŁĄCZNE WYKORZYSTANIE TECHNIKI RDOQ I METODY OPTYMALIZACJI KODÓW HUFFMANA W KODERZE JPEG

COMBINED USE OF THE RDOQ TECHNIQUE AND THE HUFFMAN CODE OPTIMIZATION METHOD IN THE JPEG ENCODER

Jakub Stankowski, Artur Fojut, Damian Karwowski, Tomasz Grajek

Instytut Telekomunikacji Multimedialnej, Politechnika Poznańska, Poznań
{jakub.stankowski, damian.karwowski, tomasz.grajek}@put.poznan.pl

DOI: 10.15199/59.2025.4.105

Streszczenie: **Artykuł prezentuje autorskie badania dotyczące ulepszeń w kodowaniu JPEG. Zaproponowano połączenie techniki Rate-Distortion Optimized Quantization (RDOQ) z optymalizacją tablic kodów Huffmana w celu zwiększenia efektywności kompresji. Opracowano implementację tych metod i zintegrowano ją z autorskim kodekiem JPEG. Przeprowadzone eksperymenty na szerokim zestawie obrazów testowych wykazały poprawę efektywności kompresji od kilku do 20%, zależnie od ustawień kodera JPEG.**

Abstract: **The article presents original research on improvements to JPEG compression. The proposed approach combines the RDOQ technique with Huffman code table optimization to enhance coding efficiency. The methods were implemented and integrated into a custom JPEG codec. Experimental results on a diverse set of test images demonstrate compression gains ranging from several percent up to 20%, depending on the JPEG encoder settings.**

Słowa kluczowe: **Huffman (kodowanie), JPEG, kompresja obrazu, RDOQ, sterowanie koderem.**

Keywords: **Huffman (coding), image/video compression, image/video encoder control, JPEG, RDOQ.**

1. WSTĘP

Minęły ponad trzy dekady od opracowania normy kompresji JPEG cyfrowych obrazów [8, 17]. Pomimo pojawienia się znacznie bardziej wydajnych technik kodowania obrazu, takich jak np. JPEG 2000 [9], JPEG-XR [11], JPEG-XL [7], HEVC Intra (HEIF) [6], czy wreszcie AV1 Intra (AVIF) [4], kodowanie JPEG nadal cieszy się ogromną popularnością i jest powszechnie stosowane w różnych aplikacjach i produktach rynkowych [5, 15, 19]. Wynika to głównie z jego prostoty, niskich wymagań sprzętowych i obliczeniowych oraz zgodności ze starszym oprogramowaniem i sprzętem.

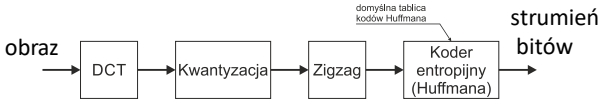
Aby w pełni wykorzystać potencjał danej techniki kompresji, niezbędny jest odpowiedni algorytm sterowania koderem [12, 14]. Ze względu na nienormatywny charakter tego algorytmu (nie jest on opisany w normie danej techniki kompresji), istnieje możliwość zwiększenia efektywności techniki kompresji obrazu poprzez udoskonalenie mechanizmu sterowania koderem — nawet długo po opracowaniu samej normy. Biorąc pod uwagę niesłabnącą popularność techniki JPEG, perspektywa poprawy współczynnika kompresji dzięki ulepszeniu algorytmu sterowania koderem wydaje się niezwykle atrakcyjna. Właśnie tego zagadnienia dotyczą badania, które przedstawiono w niniejszym artykule.

W literaturze można znaleźć różne podejścia do optymalizacji algorytmu sterowania w koderach obrazu statycznego i obrazu ruchomego, np. [2, 3, 13, 15, 18]. W przypadku techniki JPEG większość takich rozwiązań zakłada zdefiniowanie i przesłanie do dekodera obrazu nowych tablic kwantyzacji lub kodów entropijnych Huffmana albo zastosowanie wysokowydajnej zoptymalizowanej kwantyzacji (Rate-Distortion Optimized Quantization – RDOQ).

Celem niniejszej pracy jest eksperymentalna ocena wpływu jednoczesnego zastosowania wspomnianych mechanizmów, czyli techniki RDOQ oraz zmiany tablic kodów Huffmana na sprawność kompresji kodera JPEG. Przedstawione prace badawcze obejmują programową realizację zaproponowanych mechanizmów, ich włączenie do struktury oprogramowania kodera JPEG, oraz wykonanie serii eksperymentów kodowania obrazów testowych.

2. MOŻLIWOŚCI OPTYMALIZACJI KODOWANIA W TECHNICE JPEG

W najprostszym przypadku przedstawianym w wielu opracowaniach (np. [12]), koder JPEG składa się z czterech bloków funkcjonalnych: dyskretnej transformacji kosinusowej (DCT), kwantyzacji otrzymanych próbek widma, uporządkowania zygzakowatego współczynników transformaty po kwantyzacji oraz kodera entropijnego właściwych danych (Rys. 1.). Dodatkowo można jeszcze uwzględnić krok konwersji przestrzeni barw (np. z RGB do YCbCr), który dla czytelności pominięto w prezentowanych poniżej schematach. Kodowanie entropijne danych w JPEG zazwyczaj utożsamiane jest z kodowaniem Huffmana, ponieważ możliwy, alternatywny wariant wykorzystujący kodowanie arytmetyczne danych jest praktycznie niespotykany (z powodu obowiązujących jeszcze do niedawna ograniczeń patentowych).



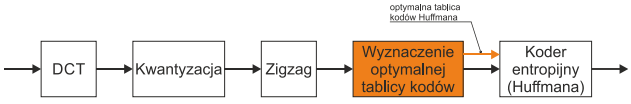
Rys. 1. – Schemat blokowy prostego kodera JPEG (bez optymalizacji)

Nawet tak prosta technika kompresji pozwala twórcom kodera obrazu na zastosowanie różnych algorytmów sterujących jego pracą. W kolejnych podpunktach przedstawiono dwie autorskie techniki pozwalające na znaczną poprawę sprawności kompresji z zachowaniem pełnej zgodności z normą [8] i gwarantujące pełną kompatybilność z dowolnym dekoderem JPEG.

2.1. Tablice kodów Huffmana

Technika JPEG definiuje zestaw domyślnych tablic kodów Huffmana. Jednak wiele formatów wymiany danych (w tym najpopularniejszy JFIF [10]) zakłada, że koder obrazu może wysłać własny zestaw tablic kodów Huffmana. Te indywidualnie wyznaczone tablice kodów powinny być lepiej dopasowane do charakterystyki kodowanego obrazu (charakteru jego treści). Tablice te przesyłane są w zwartej, pośredniej formie, a następnie są interpretowane przez dekoder obrazu i konwertowane na pełne tablice kodów.

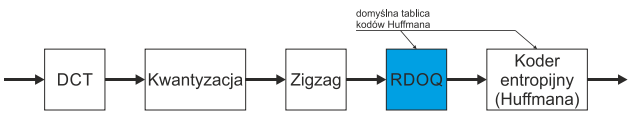
W celu zastosowania tej techniki konieczne jest uzupełnienie kodera obrazu o mechanizm analizy częstości występowania poszczególnych symboli danych (oznaczających poszczególne pary R-L (Run-Lenght) [12]) oraz dodatkowo o mechanizm wyznaczania tablicy kodów Huffmana, działający na podstawie wiedzy o prawdopodobieństwach wystąpienia poszczególnych symboli danych. Schemat blokowy takiego kodera został przedstawiony na Rys. 2.



Rys. 2. – Schemat blokowy kodera JPEG z optymalizacją tablic Huffmana

2.2. Rate-Distortion Optimized Quantization

Rate-Distortion Optimized Quantization (RDOQ) jest nienormatywną techniką pozwalającą na istotne zwiększenie wydajności kompresji danych poprzez zastosowanie dodatkowej analizy skwantowanych współczynników transformaty, wykonanej jeszcze przed kodowaniem entropijnym. RDOQ modyfikuje wartości skwantowanych współczynników DCT biorąc pod uwagę zarówno koszt bitowy ich zakodowania, jak i również ich wpływ na poziom zniekształceń zdekodowanego obrazu (Rys. 3.). Bardziej wyczerpujący opis techniki RDOQ można znaleźć w pracach [3, 16]. W szczególności zaś adaptacja techniki RDOQ do zastosowania w koderze JPEG została szczegółowo opisana w [3].



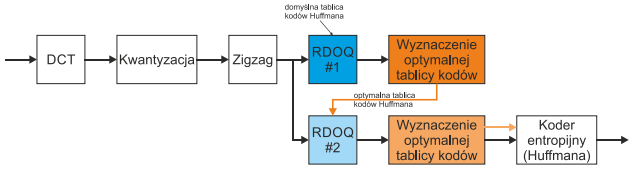
Rys. 3. – Schemat blokowy kodera JPEG z RDOQ

2.3. Łączne zastosowanie obydwu technik

Możliwe jest efektywne połączenie obu wyżej wymienionych technik, czyli jednoczesne zastosowanie RDOQ oraz optymalizacji tablic kodów Huffmana. W tym przypadku optymalne tablice kodów Huffmana są wyznaczane na podstawie skwantowanych współczynników DCT, wyznaczonych przez algorytm RDOQ (Rys. 4.). Warta zainteresowania jest też analiza zależności pomiędzy RDOQ a optymalizacją Huffmana i sprawdzenie, czy poprawa sprawności kompresji wnoszona przez każdą z technik jest niezależna (addytywna).

Rys. 4. – Schemat blokowy kodera JPEG z RDOQ i optymalizacją tablic kodów Huffmana

Kolejnym pytaniem jest, czy warto realizować wiele przebiegów wspomnianej optymalizacji? Możliwe jest, że RDOQ wykorzystujące zoptymalizowane tablice kodów Huffmana lepiej dopasuje się do krzywej prędkość bitowa-zniekształcenie (Rate-Distortion – RD). W tym celu zaproponowano dwuprzebiegową optymalizację kodowania, gdzie pierwszy przebieg wykorzystany jest do wyliczenia tablic kodów Huffmana, które to tablice są następnie wykorzystywane w optymalizacji RD przez drugi przebieg RDOQ (Rys. 5.). W ten sposób finalny krok RDOQ pracuje na tablicach słów kodowych Huffmana, które są lepiej dopasowane do charakterystyki kodowanego obrazu.



Rys. 5. – Schemat blokowy kodera JPEG z dwuprzebiegową optymalizacją RDOQ i tablic kodów Huffmana

3. WERYFIKACJA EKSPERYMENTALNA

Weryfikację przedstawionych w punkcie 2 koncepcji zrealizowano w dwóch krokach. W pierwszym przygotowana została implementacja kodera JPEG, a następnie przeprowadzone zostały eksperymenty mające na celu ocenę wpływu przyjętych rozwiązań na sprawność kompresji danych.

3.1. Oprogramowanie

Autorska implementacja kodera JPEG zawierająca zarówno technikę RDOQ, jak i mechanizm wyznaczania optymalnych tablic kodów Huffmana została przygotowana na bazie opracowanego przez autorów oprogramowania JOptEnc. Wspomniane oprogramowanie jest udostępnione na licencji BSD w repozytorium: https://github.com/jstankowski/joptenc.

Istniejąca wcześniej wersja JOptEnc zawierała już implementację techniki RDOQ, natomiast na potrzeby opisywanych w tej pracy badań zaimplementowano algorytm wyznaczania optymalnych tablic kodów Huffmana oraz rozwiązano wszystkie problemy, jakie pojawiły się w związku z interakcją pomiędzy dwiema technikami.

Między innymi takie, jak konieczność wyznaczenia kosztu bitowego zakodowania symbolu danych, który nie występuje w zoptymalizowanej tablicy kodów Huffmana. Zmodyfikowana wersja oprogramowania JOptEnc jest dostępna w wymienionym powyżej repozytorium.

3.2. Metodyka badań

W badaniach skupiono się na ocenie efektywności zaproponowanych rozwiązań, w kontekście kompresji sekwencji obrazów (obrazy ruchome), a nie kompresji obrazów statycznych. W przypadku takiego scenariusza koder JPEG koduje niezależnie od siebie kolejne obrazy sekwencji wizyjnej.

Jako materiał testowy do badań wykorzystano 17 sekwencji testowych używanych w ramach prac grupy MPEG. Wszystkie sekwencje miały rozdzielczość przestrzenną FullHD i zawierały co najmniej 100 obrazów. Sekwencje znacząco różniły się treścią, ilością detali, stopniem zaszumienia, itd. Pozwoliło to na przetestowanie wspomnianych technik na szerokim zakresie treści i zebranie dużego zestawu danych eksperymentalnych.

Jako metodę porównania sprawności kompresji przyjęto dobrze w kompresji znaną miarę Bjontegaarda [1], a dokładniej metrykę BD-rate. Wartość BD-rate wyznaczana była dla dwóch punktów pracy kodera odpowiadających: średniej jakości (wyższy stopień kompresji) – z widocznymi, acz nieuciążliwymi zniekształceniami oraz wysokiej jakości (niższy stopień kompresji) – bez widocznych zniekształceń zakodowanej treści. W technice JPEG jakością zakodowanych obrazów steruje się poprzez regulację parametru Q (Quality), który ma bezpośredni wpływ na zawartość tablic kwantyzacji [8]. W związku z tym przyjęto, że średnia jakość odpowiada zakresowi $Q \in \{50, 55, 60, 65\}$ a wysoka jakość zakresowi $Q \in \{80, 85, 90, 95\}$. W obu przypadkach użyto czterech wartości Q, gdyż tylu punktów pomiarowych wymaga obliczenie miary BD-rate.

Motywacją do przebadania dwóch różnych punktów pracy kodera był fakt występowania wpływu stopnia (siły) kwantyzacji danych na liczbę przesyłanych niezerowych współczynników próbek DCT. Zakres średniej jakości odpowiada małej liczbie przesyłanych współczynników DCT, zaś zakres wysokiej jakości – dużej liczbie przesyłanych współczynników DCT.

3.3. Scenariusze testowe

W ramach weryfikacji wpływu techniki RDOQ i metody zoptymalizowanych tablic kodów Huffmana na sprawność kompresji danych zdefiniowano pięć scenariuszy badawczych:

1. Odniesienia – w tym przypadku zastosowano prosty koder JPEG (przedstawiony na Rys. 1.)
2. OHT – koder JPEG z mechanizmem wyznaczania optymalnych tablic kodów Huffmana (przedstawiony na Rys. 2.)
3. RDOQ – koder JPEG, który wykorzystuje techniki adaptacyjnej (zoptymalizowanej) kwantyzacji RDOQ (przedstawiony na Rys. 3.)
4. RDOQ+OHT – koder JPEG używający obydwu technik optymalizacji (przedstawiony na Rys. 4.)
5. Dwuprzebiegowe RDOQ+OHT – rozbudowany wariant z pkt 4. (przedstawiony na Rys. 5.)

4. WYNIKI EKPERYMENTALNE

Wyniki eksperymentalne zostały zebrane w tabelach 1 i 2, odpowiednio dla wyższego i niższego stopnia kompresji danych.

Tab. 1. Wyniki dla wyższego stopnia kompresji (średniej jakości) $Q \in \{50, 55, 60, 65\}$. Ujemne wartości oznaczają poprawę efektywności kodowania obrazów.

Nazwa sekwencji	Poprawa sprawności kompresji (DB-rate) dla danego scenariusza [dB]			
	OHT	RDOQ	RDOQ + OHT	dwuprzebiegowe RDOQ + OHT
Poznan_Street	-5,41%	-10,16%	-14,43%	-14,93%
BasketballDrive	-6,69%	-11,03%	-17,10%	-17,71%
blue_sky	-1,79%	-6,56%	-8,33%	-8,76%
BQTerrace	-3,31%	-11,60%	-14,27%	-14,97%
Cactus	-2,81%	-10,49%	-12,96%	-13,34%
Kimono1	-6,47%	-8,89%	-14,95%	-15,12%
ParkScene	-3,38%	-9,55%	-12,59%	-12,89%
pedestrian_area	-7,94%	-8,39%	-15,91%	-15,91%
Riverbed	-4,66%	-8,19%	-12,57%	-12,71%
rush_hour	-10,46%	-9,01%	-19,15%	-19,25%
station2	-6,25%	-9,08%	-14,55%	-14,88%
Sunflower	-4,13%	-7,13%	-11,41%	-11,75%
Tenis	-9,61%	-9,22%	-18,30%	-18,72%
toys_calendar	-4,87%	-10,12%	-14,50%	-14,83%
Tractor	-4,07%	-7,81%	-11,59%	-11,95%
vintage_car	-6,68%	-9,40%	-15,33%	-15,84%
walking_couple	-4,84%	-10,10%	-14,37%	-14,66%
Średnia	**-5,49%**	**-9,22%**	**-14,25%**	**-14,60%**

Tab. 2. Wyniki dla niższego stopnia kompresji (wyższej jakości) $Q \in \{80, 85, 90, 95\}$. Ujemne wartości oznaczają poprawę efektywności kodowania obrazów.

Nazwa sekwencji	Poprawa sprawności kompresji (DB-rate) dla danego scenariusza [dB]			
	OHT	RDOQ	RDOQ + OHT	dwuprzebiegowe RDOQ + OHT
Poznan_Street	-0,72%	-7,69%	-8,23%	-8,76%
BasketballDrive	-2,73%	-10,10%	-11,58%	-12,61%
blue_sky	-2,45%	-9,35%	-11,57%	-12,44%
BQTerrace	-2,56%	-8,49%	-10,07%	-11,38%
Cactus	-1,26%	-9,64%	-10,26%	-11,19%
Kimono1	-1,41%	-11,87%	-12,67%	-13,47%
ParkScene	-0,89%	-8,95%	-9,61%	-10,37%
pedestrian_area	-1,60%	-11,05%	-11,93%	-12,70%
riverbed	-1,47%	-8,30%	-9,49%	-10,01%
rush_hour	-2,93%	-13,31%	-15,27%	-16,10%
station2	-1,04%	-9,84%	-10,56%	-11,19%
sunflower	-2,48%	-10,60%	-12,82%	-13,62%
tennis	-2,23%	-13,08%	-14,56%	-15,42%
toys_calendar	-1,49%	-10,95%	-11,71%	-12,55%
tractor	-1,55%	-8,96%	-10,21%	-10,92%
vintage_car	-1,61%	-9,03%	-10,05%	-10,75%
walking_couple	-1,51%	-9,50%	-10,48%	-11,20%
Średnia	**-1,76%**	**-10,04%**	**-11,24%**	**-12,04%**

Z przedstawionych danych jasno wynika, że zysk z zastosowania OHT jest znacznie wyższy dla sytuacji, gdy przesyłana jest niewielka liczba niezerowych współczynników DCT i wyraźnie maleje dla punktu pracy ze słabiej skwantowanymi współczynnikami DCT.

Jeśli chodzi o RDOQ to średnie zyski z zastosowania tej techniki są bardzo podobne w obu analizowanych zakresach parametru jakości *Q*.

Co istotne, można zauważyć, że łączne użycie obu technik (scenariusz RDOQ+OHT) wiąże się ze wzrostem sprawności kompresji zbliżonym do sumy zysków wynikających z niezależnego stosowania każdej z technik z osobna.

Zastosowanie dwuprzebiegowej optymalizacji pozwala na uzyskanie dodatkowego zwiększenia sprawności kompresji o około 1 punkt procentowy.

5. PODSUMOWANIE

Zastosowanie wybranych metod optymalizacji kodowania danych w technice JPEG umożliwia istotne zwiększenie efektywności kompresji przy zachowaniu pełnej zgodności ze standardem ISO/IEC 10918-1. Oznacza to, że zakodowany strumień danych może być poprawnie zdekodowany przez dowolny standardowy dekoder JPEG, co ma kluczowe znaczenie w kontekście praktycznego wykorzystania uzyskanych rezultatów.

W pracy zaproponowano połączenie dwóch rozwiązań: optymalizacji procesu kwantyzacji współczynników DCT z wykorzystaniem algorytmu sterowania RDOQ oraz adaptacyjnej modyfikacji tablic kodów Huffmana na podstawie rzeczywistego rozkładu statystycznego danych. Zaprojektowane warianty integracji tych technik w ramach kodera JPEG prowadzą do zauważalnej poprawy współczynnika kompresji.

Efektywność zaproponowanych rozwiązań została oceniona eksperymentalnie na szerokim zbiorze sekwencji testowych, przy różnych konfiguracjach kodera i poziomach jakości obrazu. Szczegółowe wyniki przedstawiono w Tabelach 1 i 2, a analiza danych wskazuje na wyraźne korzyści wynikające z zastosowania proponowanych metod – redukcja prędkości bitowej nawet o 20%.

LITERATURA

[1] Bjøntegaard G. 2001. „Calculation of average PSNR differences between RD curves", *ITU-T SG16/Q6*, dokument VCEG-M33.

[2] Canh T., Xu M., Jeon B. 2018. „Rate-Distortion Optimized Quantization: A Deep Learning Approach", *IEEE High Performance Extreme Computing Conference*.

[3] Grajek T., Stankowski J. 2022. „Rate-Distortion Optimized Quantization in Motion JPEG", *30th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*.

[4] Guyon Y., Barnes L., Chang W.-T. (edytorzy). 2021. „AV1 Image File Format (AVIF)", *Alliance for Open Media*, https://aomediacodec.github.io/av1-avif/.

[5] Hudson G., Léger A., Niss B., Sebestyén I., Vaaben J. 2018. „JPEG-1 standard 25 years: past, present, and future reasons for a success", *Journal of Electronic Imaging*, 27 (4) : 040901.

[6] ISO. 2017. „Information technology — High efficiency coding and media delivery in heterogeneous environments — Part 12: Image File Format", ISO/IEC 23008-12.

[7] ISO. 2022. „Information technology - JPEG XL image coding system", ISO/IEC IS 18181.

[8] ISO/ITU. 1992. „Information Technology — Digital Compression and Coding of Continuous-Tone Still Images — Requirements and Guidelines", ISO 10918-1, także jako ITU T.81.

[9] ISO/ITU. 2019. „Information technology - JPEG 2000 image coding system", ISO/IEC IS 15444 Edition 4, także jako ITU-T Rec. T.800, Edition 06/2019.

[10] ITU. 2011. „Information technology – Digital compression and coding of continuous-tone still images: JPEG File Interchange Format (JFIF)". ITU-T Rec. T.871.

[11] ITU. 2020. „Information technology - JPEG XR image coding system", ITU-T Rec. T.832.

[12] Karwowski D. 2019. „Zrozumieć kompresję obrazu". ISBN 978-83-953420-0-4.

[13] Luo X., Talebi H., Yang F., Elad M., Milanfar P. 2021 „The Rate-Distortion-Accuracy Tradeoff: JPEG Case Study",*Data Compression Conference*.

[14] Richardson I. 2002. „Video Codec Design: Developing Image and Video Compression Systems", *Wiley*.

[15] Richter T. 2016. „JPEG on STEROIDS: Common optimization techniques for JPEG image compression", *IEEE International Conference on Image Processing (ICIP)*, 61 – 65.

[16] Stankowski, J., Korzeniewski, C., Domański, M., Grajek, T. 2015. „Rate-distortion optimized quantization in HEVC: performance limitations", 31st Picture Coding Symposium, 85-89.

[17] Wallace G.K. 1992. „The JPEG still picture compression standard", *IEEE Transactions on Consumer Electronics*, 38 (1) : xviii – xxxiv.

[18] Wang Q., Liu P., Zhang L., Cheng F., Qiu J., Zhang X. 2022. „Rate–distortion optimal evolutionary algorithm for JPEG quantization with multiple rates", *Knowledge-Based Systems*, 244: 108500.

[19] W3Techs. 2025. „Usage statistics of JPEG for websites", https://w3techs.com/technologies/details/im-jpeg.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

OCENA JAKOŚĆI DOŚWIADCZENIA W WARUNKACH ZWIĘKSZONEGO REALIZMU ŻYCIOWEGO

Quality of Experience Assessment under Conditions of Increased Mundane Realism

Mateusz Olszewski[1,2]; Lucjan Janowski[1]; Rafał Figlus[1]; Mikołaj Leszczuk[1]

[1] AGH Akademia Górniczo-Hutnicza, Instytut Telekomunikacji, Kraków
[2] mateusz.olszewski@agh.edu.pl

DOI: 10.15199/59.2025.4.106

Streszczenie: Codziennie konsumujemy media w różnorodnych formach – oglądamy treści tuż po przebudzeniu, przy śniadaniu, w drodze do pracy, często wykonując jednocześnie wiele innych czynności. W samym tylko ubiegłym roku średni dzienny czas spędzany przez użytkowników w Internecie przekroczył sześć i pół godziny [1]. Rodzi się więc pytanie: czy mimo tak długiego czasu poświęcanego na oglądanie treści zwracamy uwagę na ich jakość i czy jesteśmy w stanie ją świadomie ocenić? W celu przybliżenia odpowiedzi na to pytanie przeprowadzono badanie postrzeganej jakości wideo (ang. Quality of Experience) w warunkach zwiększonego realizmu codziennego (ang. mundane realism) [2], [3]. Zastosowana metodologia rozszerza tradycyjne, laboratoryjne środowisko eksperymentalne o rzeczywiste warunki domowe, odtwarzając typowy kontekst konsumpcji treści multimedialnych przez użytkowników. W badaniu wykorzystano zasoby dostępne na platformie Netflix, oferującej bogaty katalog treści audiowizualnych. Dodatkowo, wprowadzono algorytm regulujący przepustowość sieci, co skutkowało dynamicznym obniżaniem rozdzielczości odtwarzanego materiału. Celem było wywołanie efektu pogorszenia jakości i skłonienie użytkowników do jej subiektywnej oceny. W artykule przedstawiono wyniki eksperymentu oraz przeanalizowano wpływ wielozadaniowości – tj. wykonywania innych czynności podczas oglądania – na percepcję treści wideo i ocenę ich jakości.

Abstract: We consume media in various formats on a daily basis—watching content right after waking up, during breakfast, or on our way to work—often while performing multiple tasks simultaneously. In the past year alone, the average daily time spent online exceeded six and a half hours per user [1]. This raises a crucial question: despite such extensive exposure to audiovisual content, do we actually pay attention to its quality, and are we capable of evaluating it? To address this question, a study was conducted to assess perceived video quality (QoE) under conditions of increased mundane realism [2], [3]. The adopted methodology extends traditional laboratory environments by incorporating real-life, in-home viewing conditions, thereby replicating the typical context in which users consume multimedia content. The study utilized resources from Netflix, a platform offering a rich library of video content, and applied a network throttling algorithm to dynamically reduce video resolution. This induced quality degradation was intended to provoke user reactions and facilitate subjective quality assessment. The article presents the results of the study and analyzes the impact of multitasking — i.e., engaging in concurrent activities while watching video—on user perception and quality evaluation.

Słowa kluczowe: Absolute Category Rating, Crowdsource, Mean Opinion Score, Netflix, Quality of Experience

Keywords: Absolute Category Rating, Crowdsource, Mean Opinion Score, Netflix, Quality of Experience

1. WSTĘP

Kwestia precyzyjnej oceny jakości doświadczenia zyskuje na znaczeniu w obliczu nieustannego rozwoju technologii. Jeszcze dekadę temu podstawową formą konsumpcji mediów, takich jak materiały wideo, były płyty DVD. Dziś jednak ten nośnik odchodzi w zapomnienie. W erze dystrybucji analogowej problem jakości dotyczył przede wszystkim producentów płyt czy kaset, ponieważ użytkownik końcowy otrzymywał produkt w gotowej formie. Obecnie, gdy dominującą formą dystrybucji treści jest streaming, kwestia jakości coraz bardziej dotyka samego użytkownika końcowego. Nawet przy największych staraniach ze strony dystrybutora, istnieje ryzyko, że odbiorca doświadczy spadku jakości wideo [4], [5]. Dlatego kluczowe staje się nie tylko mierzenie jakości materiałów udostępnianych przez dystrybutora, ale również zrozumienie, jak są one faktycznie postrzegane przez użytkowników.

Sam problem subiektywnej oceny jakości nie jest nowy [6], jednak obecnie nabiera nowego wymiaru — jak mierzyć jakość w warunkach możliwie najbardziej zbliżonych do rzeczywistego sposobu konsumpcji. Pierwsze próby nowego podejścia pojawiły się już w publikacjach z pierwszej dekady XXI wieku np. [7]. Wówczas stosowano metodologię opartą na dystrybucji płyt zawierających celowo zmanipulowany materiał wideo. Było to rozwiązanie innowacyjne jak na swoje czasy i — w świetle ówczesnych możliwości technologicznych — najlepiej odzwierciedlające naturalne warunki odbioru treści. Współcześnie jednak podejście to wydaje się mało praktyczne – zwłaszcza w kontekście możliwości, jakie oferują serwisy streamingowe. Dodatkowo rozwój technologii pozwala na prowadzenie badań na dużą większą skalę, z udziałem użytkowników w ich naturalnym środowisku domowym.

W artykule zaproponowano nową metodologię badawczą, przetestowaną pilotażowo na niewielkiej grupie

testerów o niskim zróżnicowaniu etniczno-kulturowym, z zaleceniem przeprowadzenia kolejnych badań na próbie reprezentatywnej.

2. POWIĄZANE PRACE

W klasycznych metodologiach badawczych dotyczących jakości wideo kwestia realizmu życiowego często zostaje pominięta. Jedną z prób opracowania nowego podejścia badawczego jest protokół przedstawiony przez G. Wielgus et al. [8]. Zaproponowana przez autorów metodologia stanowi adaptację rozwiązania opisanego przez W. Robitza'e et al. [9]. Zamiast modyfikować materiał źródłowy, jak miało to miejsce we wcześniejszych badaniach laboratoryjnych, zastosowano tu akwizycję danych z rzeczywistych sesji odtwarzania treści wideo przy pomocy rozszerzenia przeglądarkowego oraz manipulację przepustowością łącza internetowego za pomocą narzędzi dostępnych w przeglądarce Google Chrome. Narzędzie to umożliwia pomiar jakości transmisji w serwisie streamingowym YouTube, uwzględniając kluczowe wskaźniki jakości (takie jak opóźnienie wstępne, rozdzielczość i przepływność binarna) oraz zachowania użytkowników. Taka konfiguracja pozwala nie tylko odtworzyć naturalne warunki konsumpcji treści w środowisku laboratoryjnym, ale zapewnia także szeroką skalę danych przy jednoczesnym zachowaniu anonimowości uczestników badania. Jednym z ograniczeń tej metody pozostaje jednak brak pełnej kontroli nad środowiskiem użytkownika – np. w przypadku lokalnego buforowania wideo, użytkownik może nie doświadczyć pogorszenia jakości, mimo realnych zakłóceń w transmisji danych.

Alternatywne podejście do prowadzenia badań jakościowych w środowisku zdalnym zaproponowali B. Naderi et al. [10], opracowując aplikację w ramach programu Microsoft Open Source. Udostępnione rozwiązanie umożliwia prezentację i ocenę materiałów wideo, z opcją integracji badania z platformą Amazon Mechanical Turk (AMT), służącą do rekrutacji uczestników metodą crowdsourcingową. Aplikacja opiera się o odtwarzacz HTML5 i wczytywaniu plików wideo bezpośrednio do pamięci przeglądarki. W celu zbliżenia warunków badania do naturalnego środowiska korzystania z platform streamingowych, rozwiązanie zakłada gromadzenie danych o urządzeniu użytkownika oraz umożliwia mu samodzielny wybór ustawień. Niemniej jednak, rozwiązanie to posiada kilka istotnych wad, mogących wpływać na jakość odbioru. Pierwszym z nich jest sposób odtwarzania materiału – wideo pobierane jest w całości przed rozpoczęciem sesji. Proces ten może przeciążać infrastrukturę odbiorcy i wpływać negatywnie na komfort korzystania z platformy. Choć nie stanowi to przeszkody w przypadku krótkich filmów, przy materiałach średnio- lub pełnometrażowych metoda ta staje się niepraktyczna.

Dodatkowym wyzwaniem pozostaje konieczność częstej kalibracji urządzenia – autorzy wskazują, że ekran kalibracyjny powinien być wyświetlany nawet co kilka minut. Problematyczna jest również nadmierna liczba sesji treningowych. Choć aplikacja odwołuje się do wytycznych ITU-T P.910 [11], które przewidują przeprowadzenie jedynie jednej sesji treningowej na początku badania, w praktyce użytkownicy mogą być nimi nadmiernie obciążeni. Istotnym problemem jest także brak mechanizmu rejestrującego czas poświęcony przez uczestnika na ocenę poszczególnych materiałów. Tymczasem w badaniach realizowanych z wykorzystaniem crowdsourcingu informacja ta ma kluczowe znaczenie – umożliwia bowiem ocenę stopnia zaangażowania uczestnika w wykonywane zadanie.

W swoim opracowaniu R.R.R. Rao et al. [6] opisali eksperyment crowdsourcingowy z wykorzystaniem wysokiej jakości materiałów wideo. Słusznie wskazują oni, że przy odpowiednio dobranej platformie metoda ta może stanowić atrakcyjną alternatywę dla kosztownych i czasochłonnych badań laboratoryjnych. Choć wnioski te są zasadne, autorzy pozostają przy klasycznym modelu badawczym – prezentowaniu wcześniej przygotowanych nagrań wideo według z góry ustalonego scenariusza. Takie podejście, choć sprawdzone, nie oddaje w pełni rzeczywistego doświadczenia użytkownika w warunkach domowych. Korzystanie z gotowych materiałów może wpływać na percepcję odbiorcy w sposób zbliżony do badań laboratoryjnych, tym samym odwzorowując warunki akademickie w przestrzeni prywatnej uczestników. W efekcie, chociaż taka metodologia pozwala ograniczyć koszty związane z aparaturą badawczą, przenosi obciążenie czasowe z zespołu badawczego na osoby biorące udział w eksperymencie, co z kolei może prowadzić do ich rezygnacji z uczestnictwa.

3. METODOLOGIA BADAWCZA

Metodologia zastosowana w badaniu opiera się na rozwiązaniu zaproponowanym przez G. Wielgus et al. [8]. Przedstawione podejście, wykorzystujące platformę YouTube, zakłada pełną swobodę przeglądania treści przez uczestników. Metodologia ta została przekształcona w formę zdalną, co umożliwia całkowicie bezkontaktową dystrybucję oraz obsługę oprogramowania po stronie badanych.

3.1 Oprogramowanie

Oprogramowanie oparte jest na architekturze klient-serwer, gdzie serwer odpowiedzialny jest za gromadzenie i przechowywanie danych, a klient za modyfikacje prezentowanych treści. Platformą wybraną do prezentowania treści jest Netflix, która dostarcza dane takie jak przepustowość łącza, dane dotyczące odtwarzanego filmu wideo, a co najważniejsze udostępnia dane na temat obiektywnej metryki jakości, jaką jest Video Multimethod Assessment Fusion (VMAF) [12]. Dane te są kluczowe dla właściwej analizy subiektywnych ocen jakości ze względu na możliwość odwołania się do ustalonej wcześniej skali. Dodatkowo platforma Netflix umożliwia ręczne narzucenie przepływności binarnej, dla której został zoptymalizowany odtwarzacz, co skutkuje obniżeniem rozdzielczości, co dla tego rodzaju metodologii jest kluczowe ze względu na swobodę prezentowania treści.

Serwer danych został skonfigurowany z wykorzystaniem dystrybucji systemu operacyjnego Linux z rodziny Fedora. Oprogramowanie serwera zostało opracowane na podstawie środowiska wieloplatformowego Node.js oraz biblioteki Express.js. API serwera udostępniane było poprzez zabezpieczony port, przy użyciu szyfrowanego połączenia HTTPS. Do przechowywania danych wykorzystano bazę danych opartą na SQLite. Baza danych składała się z pięciu tabel połączonych ze sobą:

- **experiment**: która przechowywała dane na temat sesji poszczególnych użytkowników. Zawierała dane dotyczące samego użytkownika jak i prezentowanych treści,

- **video**: która przechowywała informacje na temat każdego z odtwarzanych wideo w ciągu całego badania,
- **assessment**: która przechowywała dane na temat ocen jakości w czasie badania, dodatkowo, zgodnie z rekomendacją Międzynarodowego Związku Telekomunikacyjnego, zawiera informacje dotyczące czasu w jakim tester oddał ocenę,
- **playback_data**: która przechowywała szczegółowe dane na temat odtwarzanych wideo, tj. dane VMAF, przepustowość, segment wideo, stan odtwarzacza. Dane zawarte w tabeli zostały podzielone, przy pomocy wyrażeń regularnych, na kolumny,
- **archive**: która przechowywała te same dane co tabela playback_data, jednak w nieuporządkowanej formie - były to surowe zrzuty danych dostarczonych przez platformę Netflix.

Aplikacja kliencka została oparta na narzędziu Webpack oraz bibliotece React. Środowisko Webpack zostało wykorzystane do efektywnego zarządzania zależnościami i budowy aplikacji, umożliwiając jednocześnie optymalizację oraz pakowanie zasobów. Do budowy interfejsu użytkownika wybrano bibliotekę React, która zapewnia deklaratywne i efektywne podejście do tworzenia dynamicznych komponentów. Aplikacja wykorzystuje dane dostarczone przez platformę Netflix dotyczące obsługiwanych przepustowości sieci, które następnie mapowane są na odpowiadające im wartości VMAF i dopasowywane do zadanych scenariuszy sesji badawczych. Całość odbywa się automatycznie po podaniu przez użytkownika linków do wybranych treści wideo.

Po rozpoczęciu sesji badawczej aplikacja, wykorzystując mechanizm adaptacyjnego strumieniowania (ang. adaptive streaming), dobiera wartości przepływności binarnej spośród opcji udostępnianych przez serwer dystrybutora. Równocześnie prezentuje uczestnikowi zapytania o ocenę jakości materiału wideo, zgodnie z międzynarodową skalą Absolute Category Rating (ACR) [11]. Całość jest zautomatyzowana i opisana w języku ojczystym osób badanych, dzięki czemu zredukowana zostaje możliwość niepowodzenia i niezrozumienia, a w rezultacie rezygnacji z sesji badawczej.

3.2 Dystrybucja

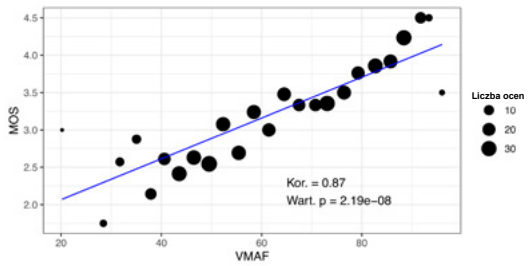
Dystrybucja oprogramowania została oparta na udostępnionej stronie internetowej stworzonej dla potrzeb badania. Oprócz plików niezbędnych do instalacji oprogramowania, udostępnione zostały szczegółowe instrukcje, które krok po kroku prowadzą przez proces instalacji oraz przygotowania do sesji badawczej.

4. ANALIZA DANYCH

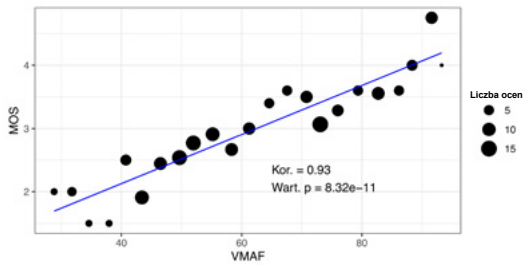
W badaniu wzięło udział siedmiu mężczyzn w wieku od 20 do 24 lat, zaznajomionych z platformą Netflix. W 89% przypadków użytkownicy wracali do wcześniej rozpoczętych filmów lub seriali. Z badania ilościowego wynika, że zaangażowanie użytkowników w prezentowaną treść było wysokie i wyniosło 8.18 w dziesięciostopniowej skali. Uczestnicy w znacznej mierze preferowali oglądanie seriali, które stanowiły 95% wszystkich materiałów odtwarzanych w ramach badania.

Analiza danych została przeprowadzona dwutorowo. W pierwszym wariancie uwzględniono cały zbiór danych. W drugim podejściu dokonano podziału sesji użytkowników na dwie grupy: (A) uczestników, którzy zadeklarowali pełne skupienie na materiale wideo, oraz (B) uczestników, którzy wykonywali inne czynności w trakcie oglądania (takie jak przeglądanie telefonu, sprzątanie czy granie w gry komputerowe).

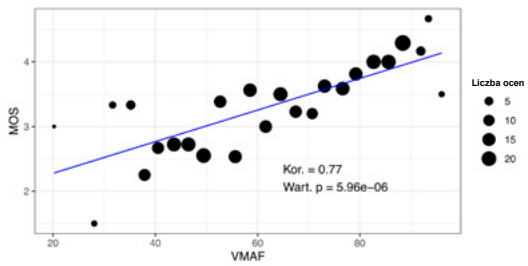
W analizie uwzględniono korelację między średnią oceną użytkowników (MOS) [13], uzyskaną na podstawie skali ACR, a średnią wartością VMAF w przedziałach czasowych pomiędzy ocenami. W pierwszym podejściu otrzymano korelacje na poziomie 0.87 przy istotności statystycznej wynoszącej 2.19e-08, natomiast odpowiednio dla grupy A uzyskano zgodność na poziomie 0.93, przy wartości p 8.32e-11, a dla grupy B: 0.77 z istotnością 5.96e-06. Wyniki przedstawiono na poniższych wykresach 1, 2, 3.



Rysunek 1: Współzależność wskaźników VMAF i MOS w analizowanym zbiorze danych



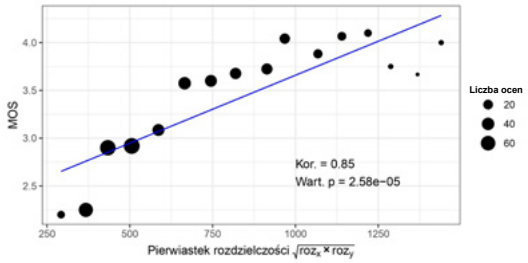
Rysunek 2: Współzależność wskaźników VMAF i MOS w grupie A



Rysunek 3: Współzależność wskaźników VMAF i MOS w grupie B

Analizując wyniki w porównaniu do badań G. Wielgus et al. [8], gdzie zanotowano korelację na poziomie 0.81, można zauważyć wyższą zgodność ocen pozyskanych w badaniu zdalnym. Jednakże należy pamiętać, że w cytowanym artykule badano korelację między rozdzielczością

strumienia wideo wysyłanego przez dostawcę a MOS. Relacja pomiędzy tymi zmiennymi w przypadku danych zebranych w eksperymencie croudsource'sowym została zobrazowana na wykresie 4. Współzależność między ocenami a rozdzielczością przesyłaną przez dostawcę w badaniu wyniosła 0.85.



Rysunek 4: Współzależność rozdzielczości i MOS w analizowanym zbiorze

5. PODSUMOWANIE

Wysoki poziom korelacji uzyskany w grupie uczestników wykonujących wyłącznie jedno zadanie – polegające na oglądaniu i ocenie treści – pozwala stwierdzić, że badania prowadzone w środowisku domowym dostarczają danych porównywalnych pod względem jakości z tymi uzyskiwanymi w warunkach laboratoryjnych.

Co istotne, również w grupie uczestników wykonujących dodatkowe czynności w trakcie oglądania odnotowano istotną statystycznie korelację pomiędzy subiektywnymi ocenami jakości a wskaźnikiem VMAF. Choć poziom tej korelacji był niższy niż w grupie skoncentrowanej wyłącznie na treści, uzyskane wyniki nie wykazały istotnych różnic pomiędzy grupami. W interpretacji danych należy jednak uwzględnić kontekst wykonywanego zadania oraz potencjalne czynniki rozpraszające uwagę użytkowników.

Uzyskane wyniki, mimo ograniczonego rozmiaru próby badawczej, są obiecujące i wskazują na zasadność przeprowadzenia dalszych badań na bardziej reprezentatywnej grupie, obejmującej przedstawicieli różnych grup wiekowych, zawodowych oraz etnicznych.

6. FINANSOWANIE

Badania prowadzące do otrzymanych wyników zostały dofinansowane ze środków Mechanizmu Finansowego Norweskiego na lata 2014-2021 w ramach projektu 2019/34/H/ST6/00599 oraz ze środków programu "Inicjatywa Doskonałości – Uczelnia Badawcza" dla Akademii Górniczo-Hutniczej w Krakowie.

LITERATURA

[1] Statista Search Department (2025) Average daily time spent using the internet by online users worldwide from 3rd quarter 2015 to 3rd quarter 2024 [Infographic]. Statista. https://www.statista.com/statistics/1380282/daily-time-spent-online-global/

[2] "APA dictionary of psychology," accessed: 2025-05-29. [Online]. Available: https://dictionary.apa.org/mundane-realism

[3] Brzezińska, A. I., & Brzeziński, J. M. (2011). Skale szacunkowe w badaniach diagnostycznych. W: JM Brzeziński (red.), Metodologia badań społecznych. Wybór tekstów, 299-399.

[4] Chang, S. F., & Vetro, A. (2005). Video adaptation: concepts, technologies, and open issues. Proceedings of the IEEE, 93(1), 148-158.

[5] Seeling, P., et al. (2004). Network performance evaluation using frame size and quality traces of single-layer and two-layer video: A tutorial. IEEE Communications Surveys & Tutorials, 6(3), 58-78.

[6] Rao, R.R.R., et al. (2021). Towards high resolution video quality assessment in the crowd. In 2021 13th International Conference on Quality of Multimedia Experience (QoMEX) (pp. 1-6). IEEE.

[7] Staelens, N., et al. (2010). Assessing quality of experience of IPTV and video on demand services in real-life environments. IEEE Transactions on broadcasting, 56(4), 458-466.

[8] Wielgus, G., et al. (2023). Proposing more ecologically-valid experiment protocol using YouTube platform. Electronic Imaging, 35, 1-6.

[9] Robitza, W., et al. "Are you still watching? Streaming video quality and engagement assessment in the crowd." 2020 twelfth international conference on quality of multimedia experience (QoMEX). IEEE, 2020.

[10] Naderi, B., & Cutler, R. (2024). A crowdsourcing approach to video quality assessment. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2810-2814). IEEE.

[11] ITU-T. (2023). P.910 (10/23): Subjective video quality assessment methods for multimedia applications. Geneva, Switzerland: ITU.

[12] Li, Z., Aaron, A., et al. (2016). Toward A Practical Perceptual Video Quality Metric. Netflix TechBlog.

[13] ITU-T. (2016). P.800.1 (07/16): Mean opinion score (MOS) terminology. Geneva, Switzerland: ITU.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

OD GENERACJI DO OCHRONY: MODELE DYFUZYJNE W ZNAKOWANIU WODNYM OBRAZÓW

FROM GENERATION TO PROTECTION: DIFFUSION MODELS IN IMAGE WATERMARKING

Marta Bistroń[1]; Zbigniew Piotrowski[2]

[1] Wojskowa Akademia Techniczna, Warszawa, marta.bistron@wat.edu.pl
[2] Wojskowa Akademia Techniczna, Warszawa, zbigniew.piotrowski@wat.edu.pl

DOI: 10.15199/59.2025.4.107

Streszczenie: **W artykule przedstawiono analizę potencjału modeli dyfuzyjnych w zadaniach znakowania wodnego obrazów. Omówiono najnowsze podejścia literaturowe — ZoDiac oraz SuperMark, wykorzystujące przestrzeń latentną oraz zamrożone modele generatywne. Zaprezentowano także autorską koncepcję metody opartą na strukturze mozaikowej, umożliwiającej osadzanie binarnych sygnatur o dużej pojemności. Propozycja łączy wysoką transparentność z odpornością na zakłócenia. Artykuł stanowi wprowadzenie koncepcyjne i punkt wyjścia do dalszych prac badawczych.**

Abstract: **This paper presents an analysis of the potential of diffusion models in the task of image watermarking. It discusses recent literature approaches — ZoDiac and SuperMark — which utilize the latent space and frozen generative models. An original method is also proposed, based on a mosaic structure that enables embedding high-capacity binary signatures. The approach combines high transparency with robustness to distortions. The article serves as a conceptual introduction and a starting point for further research.**

Słowa kluczowe: **DDIM inversion, modele dyfuzyjne, przestrzeń latentna, Stable Diffusion, znakowanie wodne.**

Keywords: **DDIM inversion, diffusion models, latent space, Stable Diffusion, watermarking.**

1. WPROWADZENIE

W dobie dynamicznego rozwoju modeli generatywnych, w szczególności obrazowych, rośnie potrzeba zabezpieczania cyfrowych treści przed nieautoryzowanym użyciem, kopiowaniem oraz fałszowaniem. Jednym z klasycznych podejść do ochrony treści wizualnych jest znakowanie wodne, polegające na ukryciu dodatkowej informacji w obrazie w sposób trudny do usunięcia, a jednocześnie niewidoczny dla odbiorcy.

Tradycyjne techniki znakowania wodnego, zarówno operujące w dziedzinie przestrzennej, jak i dziedzinie transformat, wykazują ograniczoną skuteczność w obliczu współczesnych wyzwań, takich jak kompresja stratna, przeskalowanie czy zakłócenia. Zastosowanie metod uczenia maszynowego, w szczególności modeli głębokich, takich jak generatywne sieci przeciwstawne (GAN) oraz autoenkodery wariancyjne (VAE), przyczyniło się do poprawy efektywności i adaptacyjności algorytmów znakujących. Niemniej jednak, nawet te zaawansowane podejścia okazują się niewystarczające w konfrontacji z nowymi typami ataków wynikającymi z zastosowania najnowszych modeli generatywnych.

Modele dyfuzyjne (diffusion models) stanowią jedną z najnowszych klas modeli generatywnych. Oferują wyjątkowe właściwości, takie jak iteracyjne generowanie obrazu od szumu oraz możliwość precyzyjnego sterowania procesem odtwarzania. Ich potencjał w kontekście znakowania wodnego jest dopiero rozpoznawany, lecz pierwsze badania wskazują, że mogą one stanowić silne narzędzie zarówno do ukrywania, jak i odzyskiwania informacji w sposób odporny i trudny do wykrycia.

Celem niniejszej pracy jest zarysowanie potencjału modeli dyfuzyjnych w zastosowaniach związanych ze znakowaniem wodnym obrazów. W artykule omówiono wybrane podejścia istniejące w literaturze, a także przedstawiono własną koncepcję, zarysowaną na tle dotychczasowych rozwiązań.

2. MODELE DYFUZYJNE W ZNAKOWANIU WODNYM

Modele dyfuzyjne zostały opracowane jako metoda generowania danych o wysokiej jakości, zdolna do pokonania ograniczeń wcześniejszych modeli, takich jak GAN i VAE. W odróżnieniu od nich, modele dyfuzyjne uczą się odwzorowywać rozkład danych wejściowych poprzez iteracyjne zakłócanie i odtwarzanie sygnału, co skutkuje wyraźnie lepszą stabilnością i możliwością kontroli procesu generacji.

Kluczową ideą modeli dyfuzyjnych jest dwufazowy proces: w fazie zaszumiania (ang. forward diffusion process) dane wejściowe — w tym przypadku obraz — zostają stopniowo przekształcone w biały szum przez dodawanie losowych zakłóceń w kolejnych krokach. Następnie, w fazie odwrotnej (ang. reverse diffusion process), model uczy się odtwarzać pierwotne dane z tej zaszumionej reprezentacji. Cały proces zachodzi iteracyjnie, co umożliwia bardzo precyzyjną kontrolę generowanego obrazu na każdym etapie.

Typowy model dyfuzyjny opiera się na architekturze U-Net, która umożliwia skuteczne przetwarzanie danych o wysokiej rozdzielczości dzięki mechanizmom przekazywania informacji pomiędzy warstwami (ang. *skip connections*). Dodatkowo wiele modeli korzysta z wbudowanego VAE, który służy jako warstwa kodująca dane do przestrzeni latentnej. Istotną rolę odgrywa również krok

dyfuzji, który przekazywany jest do modelu jako dodatkowe wejście — zakodowane za pomocą tzw. positional encoding, co umożliwia precyzyjne sterowanie etapami przekształceń.

Trenowanie modeli dyfuzyjnych polega na uczeniu sieci rekonstrukcji danych na podstawie losowo zaszumionych wersji wejścia, co zbliża ten proces do uczenia odwrotnej dynamiki systemu. Coraz większą rolę odgrywa tu warunkowanie — modele mogą być trenowane z dodatkowymi sygnałami wejściowymi, takimi jak klasy obrazów, maski semantyczne czy tekstowe opisy. Takie podejścia, jak Stable Diffusion czy Imagen, zyskały ogromną popularność dzięki wysokiej jakości wyników i elastyczności zastosowań.

Modele dyfuzyjne cechują się jednak istotną wadą - bardzo dużym kosztem obliczeniowym związanym z koniecznością przeprowadzania setek kroków odszumiania w fazie generacji. Długi czas próbkowania stanowi istotne ograniczenie praktyczne, zwłaszcza w zastosowaniach wymagających szybkiego kodowania lub dekodowania informacji, takich jak znakowanie wodne.

W odpowiedzi na ten problem rozwijane są techniki mające na celu przyspieszenie procesu generacji. Jednym z najczęściej stosowanych podejść jest tzw. strategia student–teacher, w której oddzielny, uproszczony model (student) uczy się na podstawie oryginalnego modelu (teacher), jak generować obrazy przy znacznie mniejszej liczbie kroków. Inne podejścia opierają się na redukcji liczby kroków przez inteligentne próbkowanie (np. FastDiff, Latent Diffusion Models), stosowanie predefiniowanych trajektorii dyfuzji lub przenoszenie operacji generacyjnych do przestrzeni latentnej, co istotnie zmniejsza wymiar problemu i czas przetwarzania.

Z perspektywy znakowania wodnego modele dyfuzyjne oferują szereg unikalnych właściwości, które odróżniają je od standardowych podejść opartych na głębokich sieciach neuronowych. Jedną z nich jest wysoka transparentność procesu osadzania informacji – w wielu wariantach kodowanie znaku wodnego nie odbywa się bezpośrednio w przestrzeni pikselowej obrazu, lecz w reprezentacji latentnej uzyskanej w wyniku przekształcenia obrazu przez warstwę VAE lub w trakcie modyfikacji rozkładu szumu. Takie podejście pozwala na subtelne, trudne do wykrycia osadzenie informacji bez wyraźnego wpływu na jakość wizualną końcowego obrazu.

Drugim istotnym aspektem jest naturalna odporność modeli dyfuzyjnych na typowe ataki stosowane w kontekście znakowania wodnego, takie jak przekształcenia geometryczne, zakłócenia czy kompresja. W przeciwieństwie do klasycznych metod opartych na głębokim uczeniu, w których sieci uczą się rozpoznawać lub przeciwdziałać zakłóceniom poprzez odpowiednio skonstruowaną warstwę „noise layer", w modelach dyfuzyjnych odporność ta wynika z samej struktury procesu generacyjnego. Iteracyjne, probabilistyczne odszumianie zakłada obecność szumu i rekonstruuje dane w sposób odporny na niewielkie zaburzenia, bez potrzeby uczenia się konkretnych wzorców ataku. Dzięki temu znak wodny może być trwale i niewidocznie zakodowany, a jednocześnie trudny do usunięcia lub zniekształcenia.

Początkowe zastosowania modeli dyfuzyjnych w znakowaniu wodnym koncentrowały się przede wszystkim na generowaniu syntetycznych treści opatrzonych unikalnym identyfikatorem. Celem takich rozwiązań była możliwość jednoznacznego określenia, że dana treść została wygenerowana przez konkretną architekturę oraz przypisania jej do źródła. Jednak w miarę rozwoju technologii i rosnących możliwości modeli dyfuzyjnych, zaczęto wykorzystywać je również do osadzania znaków wodnych w obrazach rzeczywistych. W literaturze pojawiło się kilka publikacji podejmujących to zagadnienie z różnych perspektyw.

Jednym z pierwszych podejść umożliwiających znakowanie wodne obrazów rzeczywistych przy użyciu modeli dyfuzyjnych jest metoda ZoDiac, zaproponowana przez Yu i in. [1]. Metoda ta wykorzystuje gotowy, zamrożony model Stable Diffusion bez potrzeby jego ponownego trenowania. Proces znakowania rozpoczyna się od zastosowania tzw. DDIM inversion, w wyniku którego z rzeczywistego obrazu uzyskiwana jest jego reprezentacja latentna, a następnie — poprzez replikację kroków dyfuzji — rekonstruowany jest odpowiadający mu początkowy rozkład szumowy.

To właśnie ten rozkład szumowy stanowi kluczowy obiekt, w którym osadzany jest znak wodny. ZoDiac realizuje znakowanie wodne poprzez dodanie do widma Fouriera latentnego szumu koncentrycznego, rotacyjnie symetrycznego wzorca. Rozkład ten jest traktowany jako obiekt trenowalny – optymalizowany przy użyciu funkcji straty obejmującej komponenty SSIM, LPIPS oraz różnicę w przestrzeni latentnej, tak aby zachować wysoką wierność między oryginalnym a zrekonstruowanym obrazem. Metoda nie wymaga żadnych danych treningowych ani dostrajania modelu generatywnego — znak wodny jest osadzany dynamicznie i indywidualnie dla każdego obrazu.

ZoDiac należy do klasy zero-bit watermarking, co oznacza, że nie koduje jawnej informacji binarnej, a jedynie pozwala na statystyczne stwierdzenie obecności znaku wodnego. Weryfikacja polega na ponownym przeprowadzeniu DDIM inversion na zakwestionowanym obrazie i analizie jego widma częstotliwościowego pod kątem obecności charakterystycznego wzorca. Autorzy przeprowadzili rozbudowane testy odporności na zakłócenia, wykazując wysoką skuteczność detekcji znaku po m.in. kompresji JPEG, przycinaniu, zmianie rozdzielczości, modyfikacjach koloru oraz atakach przeprowadzonych przez inne modele dyfuzyjne.

Drugą analizowaną publikacją jest praca Hu i in. [2], stanowiąca rozszerzenie ZoDiac i również oparta na wykorzystaniu gotowego, zamrożonego modelu dyfuzyjnego. W tym przypadku jest to Stable Diffusion Upscaler – model super-rozdzielczości (SR) oparty na architekturze latent diffusion, wykorzystywany do generowania obrazów wysokiej jakości z danych wejściowych o niższej rozdzielczości.

SuperMark należy już do klasy multi-bit watermarking, co oznacza, że zamiast wykrywać samą obecność znaku, umożliwia ekstrakcję binarnej sygnatury z obrazu

z osadzonym znakiem wodnym i porównanie jej z sygnaturą wejściową. Proces znakowania rozpoczyna się od przeskalowania oryginalnego obrazu do niższej rozdzielczości, który następnie trafia do modelu SR jako główne wejście. Równolegle generowany jest latentny rozkład szumu, w którym znak wodny zostaje subtelnie osadzony przy użyciu techniki Gaussian Shading. Oba składniki — obraz niskiej rozdzielczości oraz latent szumowy zawierający informację — są podawane do modelu Stable Diffusion Upscaler. Wyjście modelu stanowi obraz o zwiększonej rozdzielczości zawierający znak wodny.

Ważnym aspektem tej metody jest zastosowanie tzw. sygnału resztkowego, który uzyskiwany jest jako różnica między obrazem wygenerowanym przez SD Upscaler, a oryginalnym obrazem (po sprowadzeniu obu do tej samej rozdzielczości). Ten sygnał, odpowiednio przeskalowany, zostaje następnie dodany do oryginalnego obrazu wejściowego, co pozwala na zachowanie kompromisu między jakością obrazu a skutecznością zakodowania znaku wodnego. Takie podejście umożliwia uzyskanie znaku wodnego, który pozostaje niewidoczny dla obserwatora, a jednocześnie jest trudny do usunięcia przy typowych transformacjach i zakłóceniach.

Ekstrakcja sygnatury oparta jest na procedurze DDIM inversion, z tym że wykorzystywane są tutaj dwa wejścia: obraz wysokiej rozdzielczości zawierający znak wodny oraz jego przeskalowana wersja o niskiej rozdzielczości. Oba obrazy wprowadza się ponownie do modelu SR, co umożliwia odzyskanie przybliżonego rozkładu szumowego, w którym osadzona została sygnatura binarna. Następnie z tego rozkładu odczytywany jest ciąg bitów stanowiących zakodowany komunikat. Eksperymenty przeprowadzone przez autorów potwierdzają wysoką skuteczność ekstrakcji nawet po typowych zakłóceniach, takich jak kompresja JPEG, szum Gaussa, rozmycie czy zmiany kontrastu i jasności. Dodatkowo SuperMark utrzymuje wysoką jakość wizualną obrazów (PSNR powyżej 32), co czyni metodę praktyczną i trudną do wykrycia.

Omówione powyżej metody pokazują rosnące zainteresowanie wykorzystaniem modeli dyfuzyjnych w zadaniach znakowania wodnego. Zarówno podejścia typu zero-bit, jak i multi-bit korzystają z wysokiej transparentności, odporności na zakłócenia oraz elastyczności architektury dyfuzyjnej — przy jednoczesnym braku potrzeby ponownego trenowania modelu. Jednocześnie zauważalne są różnice w zakresie strategii osadzania, formatu sygnatury oraz mechanizmów ekstrakcji. W kolejnej części zaprezentowano wstępną koncepcję metody znakowania wodnego, która wykorzystuje i rozwija wybrane aspekty przedstawionych podejść.

3. ALGORYTM ZNAKOWANIA WODNEGO WYKORZYSTUJĄCY MODEL DYFUZYJNY

W oparciu o analizę dotychczasowych podejść wykorzystujących modele dyfuzyjne do znakowania wodnego, opracowano autorską koncepcję algorytmu, która łączy wybrane elementy metod ZoDiac i SuperMark, wprowadzając jednocześnie własne usprawnienia. Projektowana metoda opiera się na wykorzystaniu zamrożonego modelu Stable Diffusion Upscaler oraz mechanizmu latentnego zakodowania informacji w przestrzeni szumu. Kluczową innowacją w proponowanym podejściu jest zastosowanie struktury mozaikowej do reprezentacji znaku wodnego, w celu zwiększenia pojemności osadzanej informacji przy zachowaniu wysokiej transparentności oraz oczekiwanej odporności na zakłócenia.

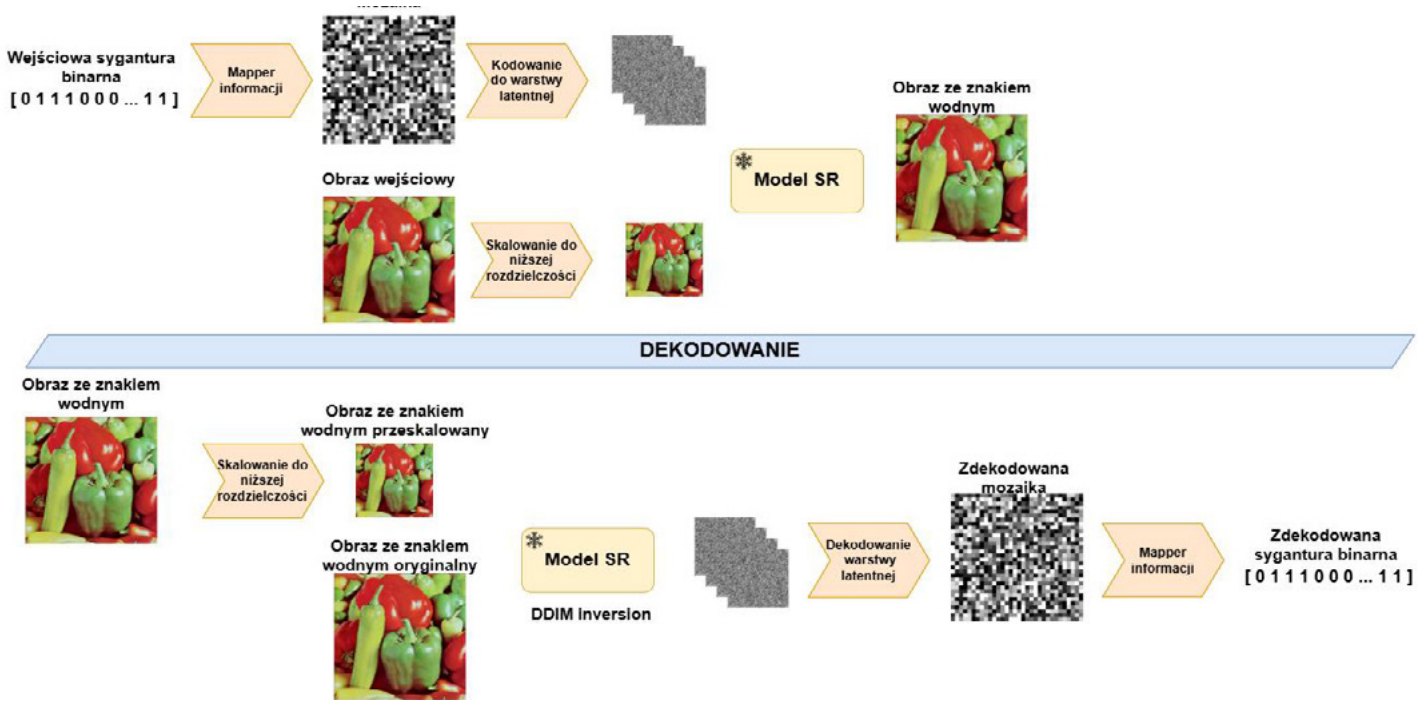
Struktura mozaiki reprezentującej binarną sygnaturę znaku wodnego została wcześniej zaproponowana jako element algorytmu znakującego bazującego na strukturze Koder-Dekoder z wykorzystaniem sieci konwolucyjnych oraz GAN, co opisano w publikacji [3].

Kluczowym elementem generowania znaku wodnego jest moduł mappera, który przekształca sygnaturę binarną o zadanej długości w statyczny obraz mozaiki. Algorytm dzieli sygnaturę na symbole, z których każdy reprezentuje ciąg bitów o zadanej długości. Na podstawie liczby symboli wyznaczana jest odpowiednia siatka bloków kwadratowych, w których umieszczane są wartości pikseli odpowiadające danym symbolom. Każda możliwa kombinacja symboli binarnych jest odwzorowana na konkretną wartość pikseli według ustalonego porządku, co pozwala na spójne zakodowanie informacji w postaci wzorca mozaikowego.

Dzięki temu pojedynczy znak wodny, dla rozbudowanej sygnatury binarnej, uzyskuje wysoką entropię lokalną, co sprawia, że jego struktura naturalnie przypomina losowy rozkład szumu. Ta właściwość jest szczególnie istotna w kontekście osadzania znaku wodnego w przestrzeni latentnej modeli dyfuzyjnych, które opierają się na iteracyjnym przekształcaniu rozkładu gaussowskiego w dane rzeczywiste. Mozaikowy układ o nieregularnych wartościach pikseli może być zatem osadzany w taki sposób, by pozostawał trudny do wykrycia, a jednocześnie zgodny z charakterem danych przetwarzanych przez model. Ponadto, rozproszona struktura pozwala ukrywać długie sygnatury binarne w sposób niejednorodny, co ogranicza ryzyko łatwego usunięcia znaku wodnego przez lokalne przekształcenia obrazu.

Dodatkową zaletą zastosowania mozaikowego znaku wodnego jest jego odporność na częściowe zniekształcenia. Nawet jeśli struktura mozaiki ulegnie lokalnym zakłóceniom lub zostanie tylko częściowo poprawnie zdekodowana, możliwe jest odtworzenie pełnej binarnej sygnatury. Wynika to ze sposobu działania dekodera (mappera), który operuje na zbiorze fragmentów, a nie na pojedynczych bitach w ustalonej lokalizacji [3].

W kolejnym etapie opracowywanej metody, wygenerowana mozaika zostaje przekształcona do postaci wektora latentnego o charakterze pseudolosowym. Tak uzyskany sygnał pełni funkcję nośnika znaku wodnego, który następnie zostaje osadzony w procesie zwiększania rozdzielczości obrazu z wykorzystaniem zamrożonego modelu Stable Diffusion Upscaler. Podobnie jak w metodzie SuperMark, wykorzystywane są równoległe wejścia: obraz o obniżonej rozdzielczości oraz komponent latentny



Rysunek 1. Schemat metody znakowania wodnego.

zawierający zakodowaną informację. Proces ekstrakcji informacji polega na ponownym przetworzeniu obrazu znakowanego oraz jego przeskalowanej wersji w modelu SR, w procedurze DDIM inversion, co umożliwia odzyskanie przybliżonej reprezentacji latentnej. Na jej podstawie rekonstruowana jest struktura mozaiki, z której odczytywana jest sygnatura binarna. Schemat metody przedstawiono na rysunku – Rys. 1.

Skuteczność proponowanej metody planuje się oceniać z wykorzystaniem zarówno metryk obiektywnych, jak i testów subiektywnych. Transparentność wizualna znakowanego obrazu będzie weryfikowana na podstawie wartości PSNR SSIM oraz LPIPS. Odporność algorytmu na zakłócenia oceniać się będzie za pomocą wskaźnika BER, mierzącego jakość odzyskiwanej sygnatury binarnej po różnorodnych transformacjach i atakach. Dodatkowo przewiduje się przeprowadzenie testów subiektywnych z udziałem uczestników (tzw. mean opinion score, MOS), w których badani będą oceniać wizualną jakość obrazów oryginalnych i znakowanych.

Proponowane podejście znajduje się obecnie na etapie weryfikacji koncepcyjnej. W niniejszym artykule przedstawiono ogólną ideę algorytmu oraz jego teoretyczne podstawy. Szczegóły implementacyjne, jak również kompletna analiza eksperymentalna obejmująca skuteczność ekstrakcji i odporność na ataki, zostaną zaprezentowane w odrębnej publikacji. Wstępne testy wskazują na wysoką transparentność generowanych obrazów, natomiast otwartym zagadnieniem pozostaje poprawa jakości rekonstrukcji struktury mozaiki w celu zapewnienia bezbłędnego odczytu informacji.

4. PODSUMOWANIE

Modele dyfuzyjne, dzięki swojej zdolności do precyzyjnego generowania i kontrolowania struktury obrazu, otwierają nowe możliwości w zakresie znakowania wodnego. Ich działanie w przestrzeni latentnej, iteracyjność oraz odporność na szum sprawiają, że mogą stanowić bardziej naturalne i skuteczne środowisko dla ukrywania informacji niż tradycyjne modele głębokiego uczenia. W artykule omówiono wybrane podejścia z literatury, które stanowią pierwsze kroki w kierunku praktycznego wykorzystania tej klasy modeli w zadaniach ochrony treści wizualnych. Analiza wskazuje, że dyfuzja może zasadniczo zmienić sposób myślenia o znakowaniu wodnym, umożliwiając zarówno zero-bit watermarking, jak i osadzanie sygnatur o dużej pojemności, przy zachowaniu wysokiej transparentności. W tym kontekście zaprezentowana w artykule koncepcja mozaikowego znakowania w przestrzeni latentnej stanowi jeden z możliwych kierunków dalszej eksploracji tych możliwości.

LITERATURA

[1] Zhang, Lijun, Xiao Liu, Antoni Viros Martin, Cindy Xiong Bearfield, Yuriy Brun, i Hui Guan. 2024. „Attack-Resilient Image Watermarking Using Stable Diffusion”. arXiv. https://doi.org/10.48550/arXiv.2401.04247.

[2] Hu, Runyi, Jie Zhang, Yiming Li, Jiwei Li, Qing Guo, Han Qiu, i Tianwei Zhang. 2024. „SuperMark: Robust and Training-free Image Watermarking via Diffusion-based Super-Resolution”. arXiv. https://doi.org/10.48550/arXiv.2412.10049.

[3] Bistroń, Marta, i Zbigniew Piotrowski. 2023. „Efficient Video Watermarking Algorithm Based on Convolutional Neural Networks with Entropy-Based Information Mapper”. *Entropy* 25 (2): 284. https://doi.org/10.3390/e25020284.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

PORÓWNANIE ALGORYTMÓW ZNAKOWANIA WODNEGO OBRAZÓW I ICH ODPORNOŚĆ NA ZNIEKSZTAŁCENIA

COMPARISON OF IMAGE WATERMARKING ALGORITHMS AND THEIR RESISTANCE TO DISTORTION

Weronika Soczek[1]; Jan Derkacz[2]

[1]Akademia Górniczo-Hutnicza, wsoczek@student.agh.edu.pl
[2]Akademia Górniczo-Hutnicza, jan.derkacz@agh.edu.pl

DOI: 10.15199/59.2025.4.108

Streszczenie: **Artykuł dotyczy porównania dwóch metod znakowania wodnego obrazów - Rozkładu Wartości Osobliwych (SVD) oraz Dyskretnej Transformaty Kosinusowej (DCT). Techniki zostały porównane pod względem jakości oznakowanego obrazu oraz jak radzą sobie w przypadku zniekształceń takich jak dodanie różnego szumu, rozmycie oraz obrót obrazu o różne kąty. Na podstawie badań nie udało się jednoznacznie określić, która metoda znakowania jest lepsza, natomiast można określić, która technika radzi sobie lepiej z wybranym atakiem.**

Abstract: **The article is devoted to the comparative analysis of two watermarking algorithms - Singular Value Decomposition (SVD) and Discrete Cosine Transform (DCT). The techniques were compared in terms of visual quality watermarked image and how they deal with distortions such as adding different types of noise, blurring and rotating the image by different angles. Based on the research, it was not possible to clearly determine which marking method is better, but it can be determined which technic copes better with the selected attack.**

Słowa kluczowe: **bezpieczeństwo, obraz cyfrowy, ochrona, znak wodny,**

Keywords: **digital image, protection, safety, watermark**

1. WSTĘP

W dobie dynamicznego rozwoju technologii cyfrowych, obrazy odgrywają kluczową rolę w komunikacji, marketingu oraz w sztuce. Wraz z ich powszechnym wykorzystaniem pojawiają się jednak wyzwania związane z ochroną praw autorskich, autentycznością treści i identyfikacją źródeł. Jednym z narzędzi, które może pomóc w przeciwdziałaniu fałszywym treściom, jest znakowanie obrazów. Techniki takie jak znakowanie wodne, tagowanie metadanych czy steganografia umożliwiają zarówno ochronę własności intelektualnej oraz identyfikowalność materiałów wizualnych. Poruszany problem badawczy polega na analizie algorytmów znakowania wodnego i ocenie, które zachowują najwyższą odporność na zniekształcenia obrazu, przy jednoczesnym minimalnym wpływie na jego jakość wizualną. W rozdziale drugim przedstawiona została literatura, w której porównano różne metody oraz ich odporność na wybrane ataki. Na podstawie zapoznanej literatury wybrano dwie metody do analizy, które były najskuteczniejsze według autorów oraz w części trzeciej artykułu opisany został sposób ich działania. Czwarty rozdział dotyczy opisu badania oraz głównych wniosków, a na końcu znajduje się podsumowanie.

2. ANALIZA LITERATURY

Problem bezpieczeństwa danych cyfrowych poruszyłojuż wielu badaczy, którzy w swoich pracach skupiali się na porównywaniu różnych metod znakowania wodnego oraz ich odporność na ataki.

W artykule [5] autorzy porównali dwie techniki cyfrowego znakowania wodnego: Dyskretną Transformatę Kosinusową - DCT (ang. Discrete Cosine Transform) i Dyskretną Transformatę Falkową - DWT (ang. Discrete Wavelet Transform) oraz ich odporność m.in. na różne rodzaje kompresji oraz szumu. Wydajność obu technik została oceniona przy użyciu metryk PSNR (ang. Peak Signal-to-Noise Ratio) i NCC (ang. Normalized Cross-Correlation). W przypadku wskaźnika PSNR, który określa poziom zakłóceń wprowadzony do obrazu, metoda DCT ma większe wartości co oznacza, że ma mniejszy spadek jakości zdjęcia niż DWT. Zmierzone wartości NCC dla różnych przeprowadzonych ataków dotyczą wyodrębnionych znaków wodnych i w tym przypadku również algorytm DCT cechuje się wyższą odpornością na zniekształcenia.

Autorzy pracy naukowej [3] skupili się na analizie wybranych trzech metod: DWT, Rozkładzie Wartości Osobliwej - SVD (ang. Singular Value Decomposition) oraz połączenia wspomnianych metod DWT-SVD w przypadku obrazów medycznych. Poziom zakłóceń wprowadzony do obrazu został zbadany za pomocą PSNR. Z badań wynika, że metoda SVD powoduje najmniejszy spadek jakości zdjęcia a kolejnym skutecznym algorytmem jest hybryda DWT-SVD. W przypadku analizy jak znaki wodne reagują na przekształcenia autorzy oceniali odporność na rozmycie i obrót. W obu przypadkach najwyższą odporność na wymienione ataki wykazała technika SVD, natomiast metody DWT oraz połączone metody DWT-SVD nie są odporne i wykazują się znaczną utratą danych.

Kolejne porównanie technik cyfrowego znakowania wodnego dotyczyło DCT oraz SVD i opisane zostało w artykule [7]. W badaniu sprawdzano jak algorytmy znakowania wpływają na jakość zdjęcia. Porównano zmiany

PSNR w algorytmach dla znakowania wodnego pojedynczych i wielu obrazów. Wyniki eksperymentu wykazały, że algorytm oparty na SVD ma lepszy znak wodny w pojedynczych i wielu obrazach.

3. ALGORYTMY ZNAKOWANIA

3.1 Rozkład Wartości Osobliwych (SVD)

Technika SVD [2][4] polega na dekompozycji macierzy bazowej na czynniki. Matematyczny zapis rozkładu macierzy A o wymiarach m×n można zapisać jako:

$$A=U\Sigma V^T \tag{1}$$

gdzie:
- U - ortogonalna macierz o wymiarach m×m, której kolumny nazywane są wektorami własnymi macierzy AA^T,
- Σ - macierz diagonalna o wymiarach m×n zawierająca wartości osobliwe macierzy A,
- V^T - transpozycja ortogonalnej macierzy V o wymiarach n×n, której kolumny są wektorami własnymi macierzy A^TA.

Metoda SVD jest jednym z wykorzystywanych narzędzi w przetwarzaniu obrazów dzięki swojej zdolności do dekompozycji macierzy na istotne składniki. Znakowanie obrazów polega na modyfikacji wartości osobliwych macierzy Σ w celu osadzenia znaku wodnego.

3.2 Dyskretna Transformata Kosinusowa (DCT)

Metoda DCT [1][6] to technika matematyczna stosowana do przekształcania sygnału z dziedziny czasowej lub przestrzennej na dziedzinę częstotliwościową. Jest szeroko wykorzystywana w przetwarzaniu obrazów, wideo oraz audio. Matematyczna reprezentacja metody wyrażona jest wzorem:

$$X(k) = \sum_{n=0}^{N-1} x(n) \cdot \cos\left[\frac{\pi}{N} \cdot \left(n+\frac{1}{2}\right) k\right] \tag{2}$$

gdzie:
- $X(k)$ - współczynniki DCT w dziedzinie częstotliwości,
- k - indeks częstotliwości od *0* do N-1,
- N - liczba próbek,
- $x(n)$ - wartości wejściowe np. jasność pikseli.

W przypadku znakowania obrazów umożliwia osadzanie znaku wodnego w domenie częstotliwościowej, co czyni go mniej widocznym i bardziej odpornym na próby usunięcia.

4. METODYKA BADAWCZA

Badaniu poddano zbiór plików graficznych w formacie JPEG o takiej samej rozdzielczości. Wybrany zbiór zdjęć różni się szczegółowością informacji zawartych w obrazie i przedstawia m.in. krajobrazy oraz budowle. Natomiast przygotowane znaki wodne to czarny okrąg na białym tle, a drugi jest bardziej złożony i zawiera logo AGH w skali szarości. Grafiki za pomocą których znakowano obrazy przedstawione zostały na Rysunku 1.

Rys. 1. Znak wodny prosty (po lewej) oraz znak wodny złożony (po prawej).

Wybrane metody znakowania wodnego obrazów SVD oraz DCT zaimplementowano z użyciem języka Python. Przygotowany zbiór zdjęć oznakowano osobno dla każdego znaku wodnego za pomocą obu technik. Proces znakowania związany jest z wprowadzeniem znaku wodnego do obrazu poprzez modyfikację współczynników związanych z częstotliwościami obrazu, a sam proces różni się w szczegółach w zależności od wybranej metody. Po osadzeniu znaku wodnego plik graficzny jest rekonstruowany do swojej pierwotnej postaci z zachowaniem jak najmniejszej ilości różnic wprowadzonych do obrazu. Oceny jakości obrazów dokonano z wykorzystaniem następujących metryk [8]:

- PSNR definiowany jest w decybelach (dB) i określony wzorem:

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX^2}{MSE}\right) \tag{3}$$

gdzie:
- MAX oznacza maksymalną wartość piksela, czyli 255 dla obrazów 8-bitowych,
- MSE to błąd średniokwadratowy między oryginalnym obrazem a przetworzonym, określony wzorem:

$$MSE = \frac{1}{N} \cdot \sum_{i=1}^{N} (x_i - y_i)^2 \tag{4}$$

gdzie:
- N - liczba pikseli w analizowanej klatce,
- x_i - i-ty piksel obrazu oryginalnego,
- y_i -i-ty piksel obrazu zakłóconego.

- SSIM (*ang. Structure Similarity Index Measure*) to miara oparta na podobieństwie strukturalnym, która odwzorowuje percepcję ludzkiego wzroku, określona jest wzorem:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{5}$$

gdzie:
- μ- uśrednione wartości intensywności wzdłuż osi x i y
- σ_x^2, σ_y^2 - wariancja x i y,
- σ_{xy} – cowariancja pomiędzy x i y,
- C_1, C_2 – zmienne stabilizujące, aby uniknąć dzielenia przez zero.

5. PORÓWNANIE SVD I DCT

5.1 Ocena jakości oznakowanych obrazów

W Tab. 1. oraz Tab. 2. przedstawione zostały średnie wartości wybranych metryk oceny jakości obrazu po dodaniu znaków wodnych za pomocą każdej z technik w porównaniu do oryginalnej treści.

Tab. 1. Średnie wartości metod oceny jakości obrazu ze znakiem wodnym dla SVD

Parametr	*Znak wodny prosty*	*Znak wodny złożony*
PSNR [dB]	28,42	28,63
SSIM	0,97	0,98

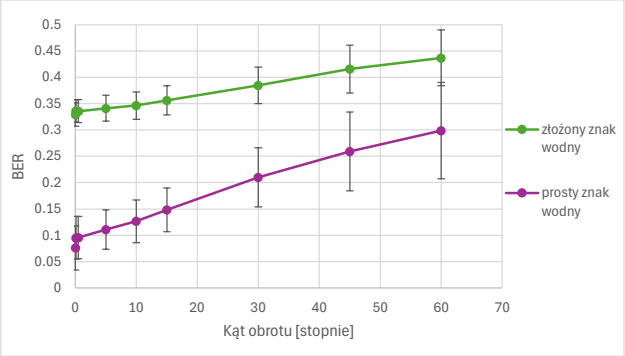
Tab. 2. Średnie wartości metod oceny jakości obrazu ze znakiem wodnym dla DCT

Parametr	*Znak wodny prosty*	*Znak wodny złożony*
PSNR [dB]	51,17	51,48
SSIM	1	1

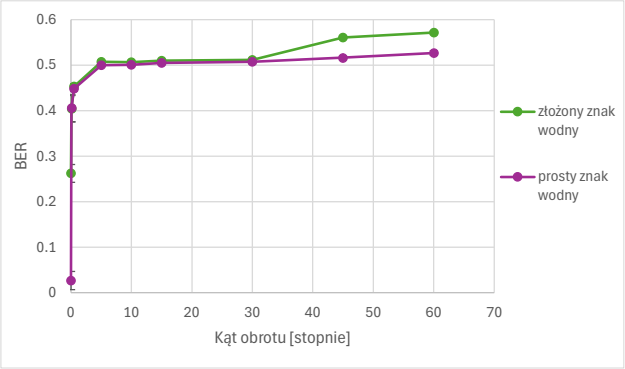
W przypadku analizy parametru PSNR lepsza jest ta metoda, która uzyskała wyższy wynik. Na podstawie danych liczbowych zawartych w powyższych tabelach można zauważyć, że w przypadku znaku wodnego prostego jak i złożonego metoda DCT osiągnęła znacznie wyższe wartości. Technika SVD uzyskała niższe wyniki co oznacza, że po wprowadzeniu znaku wodnego jakość pierwotnego obrazu się pogorszyła i widać niewielkie różnice. Natomiast parametr SSIM przyjmuje wartości od 0 do 1. Im wyższy jest uzyskany wynik to zapewnia większe podobieństwo obrazów. Obie techniki uzyskały wysokie wartości co świadczy o dużym podobieństwie obrazów, a w przypadku metody DCT można mówić o ich identyczności.

5.2 Ocena odporności na obrót

Następnie sprawdzono jak obie techniki radzą sobie z wyodrębnieniem znaku wodnego z obróconego oznakowanego obrazu. Badanie przeprowadzono w zależności od kąta obrotu obrazu o wartościach: 0,01; 0,1; 0,5 stopnia oraz 5, 10, 15, 30, 45, 60 stopni. Wykres 1. oraz Wykres 2. przedstawiają średnie wartości parametru BER wyizolowanych znaków wodnych z obrazów względem oryginalnego znaku wodnego od kąta obrotu obrazu dla obu analizowanych technik.



Wykres 1. Wykres zależności wartości BER prostego i złożonego znaku wodnego od kąta obrotu obrazu dla metody SVD.



Wykres 2. Wykres zależności wartości BER prostego i złożonego znaku wodnego od kąta obrotu obrazu dla metody DCT.

Parametr BER wskazuje na poziom naruszenia osadzonego znaku wodnego. Po analizie powyższych wykresów można zauważyć, że metoda SVD uzyskała niższe wyniki dla obu badanych znaków wodnych oraz dla każdego ze sprawdzanych kątów obrotu co oznacza, że jest bardziej odporna na tego typu manipulacje. Dodatkowo prosty znak wodny dla małych kątów obrotu obrazu wykazuje jego nieznaczne naruszenie. Technika DCT uzyskała wyższe wartości wskaźnika BER oraz można zauważyć, że różnice między prostym a złożonym znakiem wodnym są niewielkie dla analizowanych kątów, co oznacza słabą odporność metody na atak.

5.3 Ocena odporności na inne ataki

Ostatnia część badania polegała na przeprowadzeniu kilku ataków na obraz ze znakiem wodnym. Wśród nich znalazły się dodanie szumu Gaussa z poziomem 0,02, dodanie szumu typu sól&pieprz z poziomem 0,02 i 0,04 oraz rozmycie obrazu. W Tab. 3. oraz Tab. 4. przedstawione zostały średnie wartości parametru BER znaków wodnych wyodrębnionych z zaatakowanego obrazu ze znakiem wodnym.

Tab. 3. Średnia wartość parametru BER znaku wodnego wyodrębnionego z zaatakowanego obrazu ze znakiem wodnym dla SVD

Atak na obraz ze znakiem wodnym	*Wartość dla znaku wodnego prostego*	*Wartość dla znaku wodnego złożonego*
Szum Gaussa 0,02	0,0759	0,3289
Szum sól & pieprz 0,02	0,2296	0,3905
Szum sól & pieprz 0,04	0,2983	0,4135
Rozmycie	0,1738	0,3589

Szum Gaussa wyraźnie bardziej wpływa na złożone znaki wodne, co skutkuje znacząco wyższym BER w porównaniu do prostych znaków wodnych. Wraz ze wzrostem szumu sól&pieprz jakość odczytu obu znaków wodnych pogarsza się. Rozmycie ma umiarkowany wpływ na BER w przypadku prostych znaków wodnych, natomiast złożone znaki wodne są bardziej podatne na degradację, co skutkuje wyższym BER.

Tab. 4. Średnia wartość parametru BER znaku wodnego wyodrębnionego z zaatakowanego obrazu ze znakiem wodnym dla DCT

Atak na obraz ze znakiem wodnym	*Wartość dla znaku wodnego prostego*	*Wartość dla znaku wodnego złożonego*
Szum Gaussa 0,02	0,0	0,0001
Szum sól & pieprz 0,02	0,4065	0,4058
Szum sól & pieprz 0,04	0,4533	0,4534
Rozmycie	0,4589	0,4206

W przypadku analizy szumu Gaussa zarówno dla prostych, jak i złożonych znaków wodnych parametr BER jest bliski zeru, co wskazuje na wysoką skuteczność metody. Dodanie szumu typu sól&pieprz o intensywności 0,02 skutkuje podobnym poziomem zakłóceń dla obu typów znaków wodnych, co przekłada się na wysokie wartości BER. Przy wyższej intensywności szumu sól&pieprz wartości BER bardziej rosną, co potwierdza niską odporność na tego rodzaju zakłócenia. Odporność na rozmycie w tej metodzie jest niska, co można zauważyć po wysokich wynikach w tabeli.

Po analizie powyższych tabel można zauważyć, że w przypadku dodania szumu Gaussa większą odporność na zniekształcenia uzyskała metoda DCT, natomiast na dodanie różnych poziomów szumu typu sól&pieprz, rozmycie oraz obrót obrazu bardziej odporna jest metoda SVD. W metodzie SVD złożony znak wodny jest bardziej podatny na błędy niż prosty, a w przypadku drugiej omawianej techniki złożoność znaku wodnego nie wpływa znacząco na jego odporność na ataki.

6. WNIOSKI

Podsumowując, na podstawie uzyskanych wyników nie można jednoznacznie określić, która metoda zachowuje wyższą odporność na zniekształcenia obrazu, przy jednoczesnym minimalnym wpływie na jego jakość wizualną. Technika DCT zapewnia wyższą jakość odtworzonego obrazu, niezależnie od zastosowanego typu znaku wodnego, przy jednoczesnym zachowaniu niewidocznych różnic wizualnych w porównaniu do oryginalnego obrazu. Dodatkowo metoda DCT wykazuje praktycznie pełną odporność na szum Gaussowski, zarówno dla prostych, jak i złożonych znaków wodnych, natomiast technika znakowania SVD jest bardziej odporna na rozmycie, różne poziomy szumu typu sól&pieprz oraz kąty obrotu obrazu. Złożone znaki wodne są bardziej podatne na degradację w SVD, natomiast w DCT różnice między prostymi i złożonymi znakami wodnymi są nieznaczne co oznacza słabą odporność na ataki.

LITERATURA

[1] Alisha Parnami, Ankit Gupta, Gagan Parnami. 2014. *„A Robust DCT based Digital Image Watermarking using Random Mid-band Coefficient Exchange Scheme for Gray Scale Images”.* International Journal of Computer Applications. vol.100 – no.8 pp. 6-10.

[2] D.V.S. Chandra. 2002. “*Digital image watermarking using singular value decomposition*”. Proceedings of the 45th Midwest Symposium on Circuits and Systems. vol. 3, pp. 264–267.

[3] Kulbaka Weronika Zofia, Paluch Paulina, Kozieł Grzegorz. 2021.” *Analiza porównawcza metod znakowania wodnego obrazów RTG*”. JCSI - Journal of Computer Sciences Institute, vol. 20, pp. 189–196.

[4] Lange Kenneth. 2010. “*Numerical Analysis for Statisticians - Chapter 9. Singular Value Decomposition”.* Springer. pp.129-142.

[5] Rinki Kumari, Verma Puspneel. 2021. „ *A DWT and Coefficient Exchange Block Based DCT Image Watermarking and their Comparative Assessment*”. 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE – 51525.

[6] Syed Ali Khayam. 2003. “*The Discrete Cosine Transform (DCT): Theory and Application”.* Department of Electrical & Computer Engineering Michigan State University. ECE 802 – 602.

[7] Tarun Rathi, Rudra P. Maheshwari, Manoj Tripathy, Rahul Saraswat, X. Felix Joseph5. 2019. „*A Comparative Analysis of Watermarked and Watermark Images Using DCT and SVD Based Multiple Image Watermarking*”. 6th EAI International Conference (ICAST 2018). pp. 574-581.

[8] Umme Sara, Morium Akter, Mohammad Shorif Uddin. 2019. “*Image quality assessment through FSIM, SSIM, MSE and PSNR—a comparative study*”. Journal of Computer and Communications. pp. 8-18

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

REJECTION OUTLIERS DETECTION IN VIDEO QUALITY ASSESSMENT: A CROWDSOURCING STUDY

ODSIEWANIE WARTOŚCI ODSTAJĄCYCH W SUBIEKTYWNEJ OCENIE JAKOŚCI WIDEO: BADANIE CROWDSOURCINGOWE

Avrajyoti Dutta[1]; Burak Gizlice[2] Dawid Juszka[3]; Mikołaj Leszczuk[4]

[1] AGH University of Krakow, avrajyoti.dutta@agh.edu.pl
[2] Alanya Alaaddin Keykubat University, gizliceburak@gmail.com
[3] AGH University of Krakow, dawid.juszka@agh.edu.pl
[4] AGH University of Krakow, mikolaj.leszczuk@agh.edu.pl

DOI: 10.15199/59.2025.4.109

Streszczenie: W artykule przeanalizowano wartości odstające w eksperymentach oceny jakości wideo (ang. *Video Quality Assessment*, VQA) prowadzonych na platformie crowdsourcingowej. Odchylenia identyfikowano zgodnie z progami korelacji Pearsona ($< 0,75$) z Rekomendacji ITU-T P.910. W badaniu wzięło udział 296 uczestników z całego świata. Analiza ujawniła zależności czasów reakcji od płci. Wyniki wskazują, że wiarygodność VQA rośnie dzięki połączeniu filtrów technicznych i behawioralnych.

Abstract: This study examines rejection outliers in Video Quality Assessment (VQA) experiments, using a crowdsourcing platform. Outliers were detected using Pearson's correlation thresholds of the ITU-T P.910 guidelines. A total of 296 participants from all over the world participated in the study. Gender-dependent response time patterns were identified by outlier identification using Pearson's correlation (< 0.75). The results observed that the reliability of the VQA is increased by combining technical and behavioral filters.

Słowa kluczowe: bezwzględna skala kategorii (ACR), czas reakcji, eksperyment crowdsourcingowy, ocena jakości wideo (VQA), przetworzone sekwencje wideo (PVS)

Keywords: Absolute Category Rating (ACR), Response Time, Crowdsourcing Experiment, Video Quality Assessment (VQA), Processed Video Sequences (PVS)

1. INTRODUCTION

Subjective Video Quality Assessment (VQA) is an essential method for assessing user perceptions of visual defects in processed video sequences (PVS). Although conventional laboratory investigations provide controlled settings, crowdsourcing has emerged as a scalable alternative, but has issues with respect to the data dependability due to the conditions and behaviors of uncontrolled participants [1]. Recent research highlights the integration of technical pre-screening, such as hardware and bandwidth assessments, with response-based outlier detection, including response consistency [2].

In this study, rejection outliers are examined in four crowd-sourced VQA tests conducted on a bespoke platform that adheres to ITU-T P.910 criteria. Our platform incorporates automated verifications for full HD resolution ($\geq 1920 \times 1080$) and minimum internet speed (≥ 40 Mbps), resulting in the exclusion of 20% of participants during pre-screening. After screening, we implement the Pearson correlation coefficient to identify rating outliers, in accordance with ITU-T P.910 standards [3].

Our primary contribution is the identification of gender-specific response time trends among outliers. Previous research has investigated demographic biases in video quality assessment (VQA) [4], although none have correlated response times with outlier rates. Our results show:

Males show maximum dependability at 4000 to 5000 ms (~41.7% rejection rate), with significant decreases at both faster and slower speeds.

Females exhibit a U-shaped tendency, characterized by maximum dependability at 2000 – 4000 ms (~57% rejection) and elevated outliers at the extremities (<2000 ms: 73.7%; >5000 ms: 70.3%).

These insights enhance the VQA technique by illustrating how filters based on participant actions (for example, response time windows) can complement initial data quality screening procedures. In addition, we offer open source tools to duplicate the verifications of our platform.

This paper is organized as follows. Section 2 delineates the platform design and outlier identification methodologies, Section 3 outlines the methodology, and Section 4 findings and examines the implications for future crowd-sourced VQA research.

2. CROWDSOURCING PLATFORM DESIGN

The AGH university server was used to host the crowd-sourcing platform, which was developed using PHP, JavaScript, CSS, and HTML5. It enforced strict system-level and procedural controls to ensure data integrity [5] and was integrated with Prolific for participant recruitment.

2.1 Key Features: Pre-Screening Workflow

The crowdsourcing platform was designed with a structured pre-screening workflow to ensure eligibility and consistency in participant engagement. The process began with an introduction page that clearly outlined the ITU-T P.910 guidelines and specified the system requirements for participation. Participants were instructed to use Google Chrome in Incognito mode on desktop or laptop devices, explicitly excluding mobile phones to maintain the uniformity of the test.

Following this, participants were asked to complete a brief demographic survey. This step collected essential metadata including age group, gender, education level and field, subjective mood, fatigue level, and a list of interests. These parameters were later used for participant profile analysis and outlier correlation studies.

The monitor resolution check was implemented using JavaScript to automatically detect screen dimensions. Users whose monitors did not meet the Full HD requirement (1920×1080) were disqualified from proceeding. This restriction was necessary to ensure consistent video display quality and accurate assessment conditions.

Similarly, a speed test on the Internet was conducted to measure the participant's connection in real time. Users with download speeds below 40 Mbps were excluded from the study to avoid buffer artifacts that could affect quality ratings.

Participants who passed these checks were directed to the video evaluation interface, where they rated each clip using the Absolute Category Rating (ACR) method. This method asked users to score perceived video quality on a five-point scale: Excellent, Good, Fair, Poor, and Bad.

Upon completion of the experiment, a final confirmation message was displayed to express gratitude for the participant's time and contribution. This closing message marked the successful conclusion of the quality assessment and ensured that users received clear acknowledgment of their participation.

2.2 Data Acquisition at the Backend

(a) Response Times (ms), Timestamps, and Session ID.
(b) Technical data: Browser, operating system, and Internet speed (Mbps).
(c) The participant monitoring system uses a prolific ID.

3. METHODOLOGY

The following section describes the experimental methodology for video quality perception under controlled degradation conditions [6]. Section 2 describes the crowdsourcing platform and this section builds on that. There are three phases to it: the first is the generation of quality-degraded videos, the second is participant exposure through various groups for different experimental purposes, and the third is outlier detection to ensure high-quality validated data.

3.1 Content and Encoding PVS

The research used more than 400 processed video sequences (PVS) created from high-quality source clips with FFmpeg to achieve controlled quality degradation [7]. Video quality was methodically altered using three techniques: progressive quality degradation (e.g. 100% to 50% VMAF), sudden quality reductions (e.g. 80% to 30% VMAF) and constant quality levels (10 - 90% VMAF in 10-point increments). The encoder parameters comprised H.264 compression with constant rate factors (CRF 18 - 40), resolution scaling (720p to 4K) and variable bitrate modulation (0.5 - 15 Mbps). All PVS were validated via VMAF to confirm compliance with the specified quality levels (±3 VMAF points tolerance).

3.2 Experiment Design

The experiment used a design between groups with three conditions (groups A-C) to examine the effects of memory on quality judgment. Group A evaluated 80 PVS (60 at 20 seconds each + 20 at 60 seconds each) in a random sequence. Group B evaluated 120 PVS with dynamically rotating quality ratios (80/20% vs 20/80% high/low quality). Group C evaluated 40 PVS with triple repetitions for consistency analysis. Participants were automatically assigned to groups by sequential assignment [8]. The project implemented rigorous quality standards, including full HD resolution certification, a minimum bandwidth of 40 Mbps, and standardized ACR questionnaires. All sessions were limited to 60 minutes (40 minutes for video presentation and evaluation). The identification of outliers was carried out according to the ITU-T P.910 rules utilizing Pearson correlation coefficient levels ($r<0.75$).

4. RESULTS

This study involved 296 qualified participants from diverse geographical backgrounds, selected after automated eligibility checks. Subsequent to prescreening, only those possessing full HD resolution ($\geq 1920 \times 1080$) and an internet connection of ≥ 40 Mbps were allowed to advance. 114 individuals were accepted and 182 were rejected after outliers were identified using Pearson correlation coefficient levels.

Gender and response time were identified as key factors influencing behavioral trends in the analysis. Males had the highest accuracy in the 4000 – 5000 ms interval (only 41.7% rejection), with rejection rates increasing again for exceedingly sluggish answers (5000+ ms: 76.7%). The females demonstrated a U-shaped rejection curve. Their accuracy was highest in the 2000 – 4000 ms range (~57% rejection), but rejection rates were markedly elevated for hasty (<2000 ms: 73.7%) and exceedingly sluggish (>5000 ms: 70.3%) response.

- Evaluations made in under 2000 ms were most often classified as outliers for all gender groups, pointing to potential impulsiveness or inattention.
- Groups with less than five participants were omitted from the statistical summaries to maintain validity.
- Stacked bar graphs and radar graphs corroborated these trends in all tests, providing visual evidence of demographic and temporal influences on rating consistency.

Table 1: Response Time vs Gender

Response Time (ms)	Female	Male	Others
5000 ms+	37	30	4
4000 - 5000 ms	21	24	3
3000 - 4000 ms	35	31	2
2000 - 3000 ms	23	39	4
1000 - 2000 ms	19	19	2
0 - 1000 ms	2	1	0

4.1 Analysis of Male Outlier Rejection Rates by Response Time

Male participants demonstrated varying outlier rejection rates based on their response speed. A significant trend was a large increase in rejections (69.2%) within the 2000 - 3000 ms interval, rendering this the most unreliable response window. This suggests that male participants responding at moderate speed were more likely to produce ratings that deviated from the group consensus. Their lowest rejection rate (41.7%) was observed in the 4000 – 5000 ms range, suggesting that marginally slower and more measured replies from men corresponded more closely to the majority. However, when response durations exceeded 5000 ms, the rejection rates escalated to 76.7%, the highest recorded among all male response times. These findings imply that although mid-slow responses were optimal for males, extreme delays were linked to greater rating inconsistencies.

A possible explanation for this trend may be that men within the 2000 – 3000 ms interval exhibited confidence but inconsistency, while those in the 4000 - 5000 ms range allocated sufficient time to deliver considered and reliable evaluations. The elevated rejection rate for responses above 5000 ms implies that excessively prolonged response times may signify preoccupation, disengagement, or challenges in evaluating video quality.

4.2 Analysis of Female Outlier Rejection Rates by Response Time

The female participants exhibited distinct rejection rate patterns across response times, characterized by the highest outlier rejection rates at both ends of the response time spectrum. Upon swift reaction (1000 - 2000 ms), the females had a rejection rate of 73.7%, the highest in all response categories. Female participants' fast judgments tended to deviate from the collective assessment of the group. Rejection rates then decreased during the 2000 – 4000 ms interval, which means that women rendered the most consistent evaluations when given a moderate duration. However, once response times were above 4000 ms, the rejection rates escalated again, reaching 70.3% for responses greater than 5000 ms.

The data imply that optimal rating accuracy for female participants occurs within a mid-range response window (2000 – 4000 ms), avoiding both extremes. Increased outlier rates at the shortest and longest response times may reflect that:

Such rapid reaction times (1000 – 2000 ms) may point to a lack of careful consideration or a desire to complete the task quickly. Prolonged response times (exceeding 5000 ms) can suggest indecision, ambiguity, or external disturbances.

These findings show that women tend to provide more accurate quality ratings when responding at a moderate pace, rather than too quickly or too slowly.

4.3 Comparison Between Genders: Key Differences and Implications

When analyzing the rejection patterns between men and women, clear differences become apparent.

Hasty Responses (1000 – 2000 ms): Females (73.7%) exhibited a significantly higher propensity to be outliers compared to males (57.9%) within this range.

Implication: Females may require more stringent cautions against rushing while forming judgments, whereas males exhibited significantly greater consistency at this pace.

Moderate reactions (2000 - 4000 ms): Males had elevated rejection rates (69.2% at 2000–3000 ms, 61.3% at 3000 – 4000 ms) in contrast to females (56.5% and 57.1%, respectively).

Implication: Males within this range had greater variability in their ratings, whereas females demonstrated superior performance when given a moderate period of time.

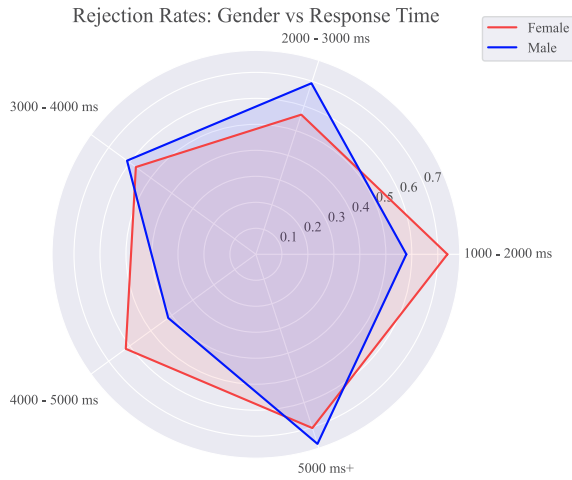


Figure 1: Radar chart depicting gender-specific response patterns across temporal intervals in ACR video quality assessments (Table: 1). Individual response times are represented by the median duration per participant. Rejection rates denote the proportion of participants excluded (Pearson $r < 0.75$) within each gender-time cohort.

Prolonged Responses (4000 – 5000 ms): The lowest rejection rate for men was 41.7%, indicating their most reliable response period. Conversely, women experienced a rise (61.9%), indicating that this range was suboptimal for them.

Implication: Moderately slow responses improved male consistency while marginally decreasing female reliability.

Extremely Delayed Responses (5000+ ms): Both genders had elevated rejection rates, and men (76.7%)

demonstrated slightly greater inconsistency than women (70.3%).
Implication: Increasingly sluggish responses pose problems for both genders, although males may be marginally more affected.

5. DISCUSSION

Prescreening Effectiveness: Prescreening with system-level versus participant-based filtering (resolution/speed) eliminated 20% of noise initially; however, response pattern outliers (62.2%) predominated after screening. Response time and gender were significant determinants of reliability, indicating that men may require time limitations (4000 to 5000 ms optimum) and females benefit from alerts regarding hasty actions (<2000 ms) or prolonged waits (>5000ms).
Comparison with Previous Research: Our gender-specific results diverge from research that presumes uniform participant behavior [5]. ITU-T P.910's Pearson's correlation coefficient suggestion demonstrated efficacy but could be improved with the incorporation of response time thresholds.
Constraints: The analysis of the small category of 'Other' gender is constrained. Fixed speed and resolution assessments may omit legitimate participants in low-bandwidth areas.
Practical ramifications of improvements to the platform: Several suggested platform improvements might have significant advantages. Giving participants instant feedback on how long it took them to respond could increase participation and improve the quality of the data. Users with restricted bandwidth may experience smoother performance if adaptive video quality is implemented. Furthermore, current research shows that hybrid outlier filters, which include demographic characteristics, temporal dynamics, and correlation metrics, could improve the dependability of upcoming crowdsourced video quality assessment (VQA) studies. Preliminary analysis showed no significant correlation between outlier rates and subjective mood, fatigue level, or education background; these factors will be explored further in follow-up work.

6. CONCLUSIONS

This study revealed that the integration of technological prescreening (resolution, speed) with participant-based outlier detection (correlation, reaction time) markedly improves the reliability of crowd-sourced video quality assessment (VQA).

This study contributes to ACR development by demonstrating that demographic-specific timing patterns (e.g., gender-based reliability windows) can enhance data quality through adaptive pacing and warning mechanisms. Our profiling strategy significantly improved result reliability, which had previously been inconsistent in open crowdsourced setups. Due to space limitations, more detailed analyses of fatigue, mood, and education impacts will be presented in a forthcoming journal paper.

Prospective Trajectories:

- Conduct adaptive studies (e.g., modifying video duration according to response velocity).
- Investigate machine learning techniques to identify outliers in pre-test surveys.
- Our open-source platform and data set offer a basis for reproducing and enhancing this research.

Table 2: Key Stats Summary

Metric	*Value*
Total Participants	296
Pre-Screening Rejections	20%
Post-Screening Outliers	62.2%
Male Optimal Time	4000 - 5000 ms (41.7% rejection)
Female Optimal Time	2000 - 4000 ms (57% rejection)

Note: Total participant number and the percentages for rejection rates are calculated from qualified participants after prescreening.

Funding: This work was funded by the Ministry of Science and Higher Education as part of a subsidy for the Faculty of Computer Science, Electronics and Telecommunications, AGH University of Krakow.

REFERENCES

[1] Dutta Avrajyoti, Dawid Juszka, and Mikołaj Leszczuk. 2024. "Crowdsourcing Evaluation of Video Summarization Algorithm." *International Journal of Electronics and Telecommunications* 70 (4): 1063–1069.

[2] Dutta Avrajyoti, Tomasz Konaszyński, Dawid Juszka, and Mikołaj Leszczuk. 2024. "A Crowdsourcing Study of Video Quality." *Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne* 97 (4): 325–328.

[3] ITU-T. 2024. *ITU-T Recommendation P.910: Methods for the Subjective Assessment of Video Quality, Audio Quality and Audio-Visual Quality*. Geneva: International Telecommunication Union.

[4] Dutta Avrajyoti, Leszczuk Mikołaj, Juszka Dawid, and Grega Michał. 2023. "Subjective Quality Assessment of Video Summarization Algorithms: A Crowdsourcing Approach." *Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne* 96 (4): 333–336.

[5] Naderi Babak, and Ross Cutler. 2024. "A Crowdsourcing Approach to Video Quality Assessment." In *Proceedings of ICASSP 2024*, 2810–2814.

[6] Grega Michał, Janowski Lucjan, Leszczuk Mikołaj, Romaniak Piotr, and Papir Zdzisław. 2008. "Quality of Experience Evaluation for Multimedia Services." *Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne* 81: 142–153.

[7] Baraković Husić Jasmina and Baraković Sabina. 2022. "Multidimensional Modelling of Quality of Experience for Video Streaming." *Computers in Human Behavior* 129: 107155.

[8] Wang Haiqiang and Liu Shan. 2021. "Subjective and Objective Study of Sharpness Enhanced UGC Video Quality." *Proceedings of SPIE 11842, Applications of Digital Image Processing XLIV*, August 1, 2021.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

REKOMENDACJA POLSKIEJ SKALI ACR NA PODSTAWIE BADANIA JAKOŚCI DOŚWIADCZEŃ Z WYKORZYSTANIEM PROTOKOŁU GŁOŚNEGO MYŚLENIA

RECOMMENDATION OF A POLISH ACR SCALE BASED ON A QUALITY OF EXPERIENCE STUDY USING THE THINKING ALOUD PROTOCOL

Natalia Jakubiec[1]; Lucjan Janowski[2]; Jakub Nawała[3]; Kamil Koniuch[4]

[1] Akademia Górniczo-Hutnicza, Kraków, njakubiec@agh.edu.pl
[2] Akademia Górniczo-Hutnicza, Kraków, ljanowsk@agh.edu.pl
[3] University of Bristol, Bristol, jakub.nawala@bristol.ac.uk
[4] Akademia Górniczo-Hutnicza, Kraków, koniuch@agh.edu.pl

DOI: 10.15199/59.2025.4.110

Streszczenie: **Standardową metodą w testach subiektywnych Quality of Experience (QoE) jest Absolute Category Rating (ACR). Pięciostopniowa skala opisana przymiotnikami nie posiada polskiej wersji językowej, co powoduje przyjmowanie własnych wersji skal przez polskich badaczy. Aby dowiedzieć się, jaka skala przymiotnikowa jest preferowana przez osoby badane, przeprowadzono badanie jakości wizji z wykorzystaniem Thinking Aloud Protocol (TAP). Uzyskane wyniki są przesłanką do rekomendacji skali symetrycznej w polskiej wersji językowej skali ACR.**

Abstract: **The standard method used in subjective Quality of Experience (QoE) testing is the Absolute Category Rating (ACR). The five-point scale, described using verbal labels, does not have an official Polish-language version, which leads Polish researchers to adopt their own versions of the scale. To identify which adjective-based scale is preferred by test participants, a video QoE study was conducted using the Thinking Aloud Protocol (TAP). The results provide grounds for recommending a symmetrical version of the ACR scale in Polish.**

Słowa kluczowe: **jakość doświadczeń użytkownika, jakość wideo, protokół głośnego myślenia, skala przymiotnikowa, testy subiektywne.**

Keywords: **Quality of Experience, video quality, Thinking Aloud Protocol, adjective scale, subjective tests.**

1. WSTĘP

Wraz z rosnącym zapotrzebowaniem na usługi multimedialne wysokiej jakości, coraz większego znaczenia nabierają zagadnienia związane z optymalizacją i kompresją treści wizualnych. W tym kontekście kluczową rolę odgrywa badanie jakości doświadczeń użytkownika końcowego, określane jako Quality of Experience (QoE). Qualinet definiuje QoE jako „stopień satysfakcji lub irytacji użytkownika końcowego podczas korzystania z danej aplikacji lub usługi. Jest wynikiem spełnienia (lub niespełnienia) jego oczekiwań względem użyteczności i/lub przyjemności płynącej z użytkowania, z uwzględnieniem cech osobowości i aktualnego stanu użytkownika oraz kontekstu użycia" [14].

W badaniach nad QoE wyróżnia się dwa główne podejścia: obiektywne i subiektywne [2]. Podejście obiektywne wykorzystuje modele matematyczne oraz metody statystyczne w celu przewidywania wyników jakości. Natomiast podejście subiektywne opiera się na opiniach użytkowników i służy do oszacowania jakości wizji z perspektywy percepcji użytkownika. Dzięki uwzględnieniu opinii użytkowników metody subiektywne pozwalają lepiej zrozumieć poziom satysfakcji oraz preferencje, wskazując różnice między wynikami technicznymi a rzeczywistymi doświadczeniami użytkowników [5].

Procedury przeprowadzania testów subiektywnych w warunkach laboratoryjnych są normalizowane przez Międzynarodowa Unia Telekomunikacyjna (International Telecomunication Union, ITU). Opisano je szczegółowo między innymi w rekomendacjach BT.500-15, P.800, P.910 oraz [9,10,11]. Najczęściej stosowanym podejściem eksperymentalnym są testy z pojedynczym bodźcem (Single Stimulus, SS). W tego rodzaju eksperymentach uczestnicy oglądają materiały wizualne prezentowane jeden po drugim i przypisują im ocenę jakości. W obrębie tej kategorii istnieje kilka metod, z których najbardziej rozpowszechnioną jest Absolute Category Rating (ACR). W tej metodzie uczestnicy indywidualnie oceniają jakość prezentowanych materiałów – każda prezentowana sekwencja jest oceniana osobno, a uczestnik musi zintegrować różne aspekty odbioru (np. ostrość, płynność, kolory, artefakty) w jednej ogólnej ocenie. Ocena jest dokonywana na skali od 1 do 5, opisanej dodatkowo przymiotnikami, tak by zwiększyć jednoznaczność, zrozumiałość i spójność ocen dokonywanych przez użytkowników. Wyniki ACR i podobnych metod są zazwyczaj reprezentowane przez średni wynik Mean Opinion Score (MOS). MOS to skalarna wartość QoE, obliczana jako średnia wszystkich ocen użytkowników, zwykle w bezwzględnej, porządkowej pięciopunktowej skali.

W rekomendacjach ITU dla skali ACR są zaprezentowane przymiotniki w języku angielskim, francuskim, a także hiszpańskim. W angielskiej wersji językowej są to kolejno: *Bad, Poor, Fair, Good* oraz *Excellent*. Niestety, nie powstała oficjalna polska wersja językowa skali.

Z tego powodu polskie zespoły badawcze przeprowadzające testy z polskimi użytkownikami tworzą własne wersje skal. Większość dostępnej literatury nie zawiera informacji o wykorzystanych w badaniu przymiotnikach [1, 3,7,18], co może wynikać z faktu, że artykuły naukowe są najczęściej publikowane w języku angielskim, a autorzy uznają przedstawienie pełnej wersji skali za zbędne. Tymczasem brak jawności co do użytej wersji skali utrudnia porównywalność wyników i powtarzalność badań.

Przegląd badań, w których zaprezentowano wykorzystane skale, ujawnia natomiast, że pomiędzy zespołami badawczymi brak jest konsensusu co do przymiotników opisujących kolejne punkty skali. I tak na przykład w badaniu Pinson i współpracowników [15], skalę opisywały kolejno przymiotniki: *zły, słaby, średni, dobry, bardzo dobry*. W badaniu jakości usług telefonii pakietowej Janowski [12] wykorzystał skalę: *zła*, *słaba*, *znośna*, *dobra*, *wspaniała*, co jest jednym z bezpośrednich tłumaczeń angielskiej wersji skali. Innym tłumaczeniem jest skala wykorzystana przez Zielińskiego [19]: *zła, słaba, dostateczna, dobra, doskonała*. Bułat i współpracownicy [6] kolejne punkty skali podpisali w następujący sposób: *zła, słaba, umiarkowana, dobra, wspaniała*. Konaszyński i współpracownicy [13] użyli jeszcze innej skali, zmieniając ponadto formę osobową przymiotnika, wiążąc ją z filmem, a nie jakością: *zły, słaby, przeciętny, dobry, doskonały* (wyświetlane pytanie brzmiało: „*Podaj Twoją ocenę obejrzanego filmu*"). Jak można zauważyć, szczególne trudności w przypisywaniu przymiotników sprawiają punkty środkowy i najwyższy (3 i 5 na skali ACR).

Jako że te różnice językowe mogą wpływać na częstość przypisywania danych ocen przez użytkowników, postanowiliśmy przeprowadzić badanie eksploracyjne, aby dowiedzieć się, jakich przymiotników używają uczestnicy badania, jeśli nie mają narzuconej skali. W tym celu wykorzystaliśmy protokół myślenia na głos (Thinking Aloud Protocol, TAP) [7]. Jest to jakościowa technika badawcza polegająca na werbalizowaniu przez uczestników swoich myśli podczas wykonywania zadania, co umożliwia wgląd w ich procesy poznawcze, strategie i decyzje.

2. METODY

2.1. Procedura eksperymentalna

W badaniu wzięło udział 15 uczestników (3 mężczyzn, 12 kobiet) w wieku od 19 do 65 lat (M = 27, SD = 12,01). Laboratorium, w którym przeprowadzono eksperyment zostało przygotowane zgodnie z rekomendacjami BT.500-15 oraz P.913 [10,11].

Procedura eksperymentalna została oparta na klasycznym schemacie eksperymentu ACR, z modyfikacją co do sposobu udzielania oceny jakości. Uczestnicy badania oglądali 160 krótkich sekwencji wizyjnych bez dźwięku. Każdy film był oceniany niezależnie od reszty sekwencji w eksperymencie. Ocen jakości nie dokonywano na tradycyjnej pięciostopniowej skali, lecz zgodnie z procedurą TAP, wypowiadając ją na głos. Uczestnicy mogli wyrazić swoją opinię na temat oglądanego filmu w trakcie i po jego prezentacji. Zadaliśmy to samo pytanie, co w eksperymencie ACR: „*Jaka jest Twoja opinia na temat jakości filmu?*". Poinformowano, że ocenie podlega wyłącznie jakość techniczna materiału wizyjnego, a treść filmu powinna zostać zignorowana. Ponadto podkreślono, że nie istnieje prawidłowy format odpowiedzi i wyrażona opinia może mieć dowolny charakter. Czas na udzielenie odpowiedzi nie był ograniczony — po dokonaniu oceny i gotowości do przejścia dalej, uczestnik samodzielnie uruchamiał kolejny film za pomocą przycisku na ekranie. Czas trwania badania był uzależniony od długości udzielanych wypowiedzi. Aby ograniczyć zmęczenie, w środku eksperymentu automatycznie uruchamiała się przerwa, którą uczestnicy mogli zakończyć w dowolnym momencie.

2.2. Materiały

Sekwencje wizyjne prezentowane uczestnikom badania pochodziły z dwóch baz: HDTV5 (80 wybranych filmów) [17] i ITS4S (80 wybranych filmów) [16]. Baza HDTV5 zawiera powtarzające się sceny, natomiast baza ITS4S zawiera filmy unikalne. Filmy w obrębie baz były prezentowane w losowej kolejności.

Aby zapewnić spójność, wszystkie filmy były wyświetlane w natywnej rozdzielczości laptopa, która odpowiada 1920x1080 pikseli. Filmy pochodzące z bazy HDTV5, na których widoczne były zniekształcenia specyficzne dla utraty pakietów, trwały 10 sekund. Filmy z bazy ITS4S trwały 4 sekundy i nie miały utraty pakietów — obserwowane zniekształcenia były wyłącznie efektem kompresji.

2.3. Przygotowanie danych

Uzyskane dane miały format nagrań wypowiedzi uczestników. W celu dalszej analizy przygotowano transkrypcje nagrań, przypisując do każdej wypowiedzi obejrzany film oraz jego MOS, a następnie przeprowadzono proces kodowania odpowiedzi stosując standardową metodę analizy tematycznej, szeroko stosowaną w badaniach jakościowych [4]. Do dalszej analizy wykorzystano jedynie te fragmenty, które wyrażały w jakiś sposób ocenę jakości.

3. WYNIKI

W trakcie badania zarejestrowano łącznie 2399 wypowiedzi uczestników, z których każda odnosiła się do oceny jednej z prezentowanych sekwencji wizyjnych. Warto zaznaczyć, że jeden z uczestników pominął ocenę jednego filmu, przez co liczba wypowiedzi jest o jeden mniejsza niż pełna liczba możliwych odpowiedzi (2400).

W 72,9% przypadków uczestnicy oceniali jakość obrazu za pomocą przymiotników lub zwrotów oceniających – łącznie użyto ich 1750 razy. W pozostałych przypadkach oceny jakości dokonano w sposób opisowy (opowiadając o widocznych zniekształceniach i innych artefaktach, np. „*widoczne piksele na ekranie"*) lub na skali numerycznej, wyrażając ocenę w formie wartości liczbowej – przy czym w tej kategorii przeważały oceny w skali od 1 do 10 (jeden uczestnik) lub nawiązujące do rozdzielczości w popularnych serwisach internetowych (np. *480p, 720p*).

W wypowiedziach uczestników pojawiło się łącznie 105 unikalnych przymiotników oraz zwrotów, które zostały przez nich spontanicznie użyte do oceny jakości. Aby zaprezentować dane w bardziej przejrzysty i reprezentatywny sposób, w tabeli 1 zaprezentowano jedynie te przymiotniki i zwroty, które pojawiły się co najmniej pięciokrotnie. Tego typu wyrażeń było łącznie 34.

Pod każdym przymiotnikiem podano liczbę jego wystąpień, czyli informację o tym, jak często pojawiał się on w wypowiedziach. Przymiotniki zostały uporządkowane w tabeli według częstości ich użycia – od najczęściej stosowanych do najrzadziej.

Tab. 1. Przymiotniki wykorzystane przez uczestników badania do oceny jakości oraz liczba ich wystąpień z podziałem na 5-stopniową skalę ACR według MOS (wartość MOS w nawiasie), które wystąpiły przynajmniej 5 razy. Podkreśleniem zaznaczono przymiotniki, które wystąpiły przynajmniej 50 razy.

1 (1,07 - 1,83)	2 (1,83 - 2,59)	3 (2,59 - 3,34)	4 (3,34 - 4,10)	5 (4,10 - 4,85)
bardzo zła (1,74) 59 tragiczna (1,65) 10 beznadziejna (1,57) 7 niewyraźna (1,73) 5	słaba (2,07) 209 zła (1,98) 97 bardzo słaba (1,86) 74 niska (2,15) 15 kiepska (2,36) 15 gorsza(2,44) 7 niedobra (2,49) 7 niefajna (1,92) 6 dość słaba (2,18) 5	średnia (2,83) 249 przeciętna (3,16) 44 nienajlepsza (2,69) 33 dosyć słaba (2,67) 7 bardzo średnia (2,67) 6 zadowalająca (2,94) 6 optymalna (2,82) 5	dobra (3,68) 314 w porządku (3,74) 70 okay (3,80) 58 dość dobra (3,44) 25 całkiem dobra (3,99) 22 niezła (3,65) 15 wysoka (4,08) 10 w miarę dobra (3,40) 8 bardzo wysoka (3,92) 8	bardzo dobra (4,19) 192 bardzo ładna (4,24) 21 ładna (4,24) 10 super (4,60) 9 świetna (4,64) 6

Wszystkim przymiotnikom przypisano średnią wartość MOS i podzielono je na pięć równych przedziałów odpowiadających stopniom na skali ACR. Zakres MOS dla poszczególnych punktów skali zaczyna się od najniższej wartości przypisanej do filmów wykorzystanych w badaniu (1,07), a kończy na wartości najwyższej (4,85) — wartości te stanowią zatem skrajne oceny uzyskane dla prezentowanych materiałów wizyjnych. Pragniemy zwrócić uwagę, że oceny MOS pochodziły z niezależnego badania baz filmów, a nie od uczestników tego eksperymentu.

Jak można zauważyć, skala zagregowała się naturalnie wokół kilku przymiotników – tylko cztery z nich wystąpiły powyżej 100 razy, piąty uzyskał wynik 97 – co może być przesłanką na korzyść kilkustopniowych skal, dobrze odzwierciedlających faktyczne strategie oceniania. Te najczęściej użyte oceny jakości to kolejno: *dobra, średnia, słaba, bardzo dobra i zła*. W grupie powyżej 50 wystąpień znalazły się jeszcze: *bardzo słaba, w porządku, bardzo zła, okay*.

Rozkład liczby unikalnych przymiotników przypisanych do poszczególnych przedziałów skali ACR ukazuje istotną asymetrię w zakresie ekspresji językowej uczestników. Największa liczba różnych określeń została odnotowana w przedziałach odpowiadających wyższym ocenom jakości: 26 unikalnych przymiotników wystąpiło zarówno w przedziale 4 (MOS 3,34–4,10), jak i 5 (MOS 4,10–4,85). Co ciekawe, mimo największej różnorodności słownictwa, najczęściej pojawiające się w literaturze określenia 5 punktu skali ACR, czyli *wspaniała* i *doskonała*, nie pojawiły się w uzyskanych wynikach ani razu. Dla przedziałów średnich i niższych odnotowano mniejszą różnorodność: 20 przymiotników w przedziałach 2 i 3 oraz tylko 13 w przedziale 1 (MOS 1,07–1,83).

4. WNIOSKI

Uzyskane wyniki potwierdzają, że procedura TAP stanowi efektywne narzędzie do pozyskiwania szczegółowych i spontanicznych danych na temat subiektywnej jakości wizji. Pomimo braku narzuconej instrukcji dotyczącej formatu odpowiedzi, zdecydowana większość uczestników spontanicznie oceniała jakość za pomocą przymiotników, które naturalnie grupowały się wokół pięciu głównych kategorii ocen. Wyniki te sugerują, że skala pięciostopniowa jest wystarczająca do oddania zróżnicowania ocen, a jednocześnie nie ogranicza swobody wyrażania opinii przez użytkowników.

Zróżnicowanie sposobów formułowania ocen jakości może świadczyć o spontanicznym stosowaniu indywidualnych strategii oceny i tym samym podkreśla znaczenie językowego aspektu w projektowaniu narzędzi do subiektywnej oceny QoE.

Zaobserwowana różnica w liczbie unikalnych przymiotników pomiędzy przedziałami ocen sugeruje, że uczestnicy częściej i bardziej różnorodnie opisywali wysoką jakość niż niską. Zauważona asymetria w liczbie unikalnych przymiotników może sugerować, że pozytywne doświadczenia jakościowe skłaniają uczestników do większej kreatywności językowej. Może to wynikać z tego, że pozytywne doświadczenia – np. dobra jakość obrazu – są bardziej złożone i subtelne, co sprzyja używaniu bardziej zróżnicowanych i kreatywnych określeń (takich jak: „*genialna*", „*idealna*", „*ładna*" czy „*super*"). Natomiast w przypadku ocen negatywnych uczestnicy częściej sięgali po proste, jednoznaczne słowa, jak „*beznadziejna*" czy „*okropna*", co może wskazywać na ograniczony zasób gotowych etykiet w języku potocznym do opisu złej jakości.

Niemniej jednak warto zwrócić uwagę, że najczęstsze określenia dla 5 stopnia skali ACR, jakie można spotkać w literaturze, nie wystąpiły w spontanicznych ocenach uczestników badania. Może to świadczyć o istotnym wpływie czynników kulturowo-językowych na dobór słownictwa w subiektywnych ocenach jakości, i wskazuje na potrzebę ostrożnego podejścia do przenoszenia gotowych skal semantycznych między językami i kontekstami kulturowymi.

Na podstawie uzyskanych wyników rekomendujemy więc ujednolicenie słownictwa stosowanego w polskiej skali ACR. Wydaje się uzasadnione, aby w polskiej wersji językowej używać skali symetrycznej – *bardzo zła, zła, średnia, dobra, bardzo dobra*. Taka wersja powinna zwiększyć spójność i jednoznaczność interpretacji ocen. Z uwagi na częstość występowania, dopuszczalne, choć nierekomendowane przez nas, wydaje się zastąpienie drugiego punktu skali, oceny *zła*, oceną *słaba*.

LITERATURA

[1] Andruloniw Paweł, Karol Kowalik, Piotr Zwierzykowski. 2022. "Unsupervised Learning Data-Driven Continuous QoE Assessment in Adaptive Streaming-Based Television System". *Applied Sciences,* 12 (16) : 8288.

[2] Barman Nabajeet, Maria G. Martini. 2019. "QoE modeling for HTTP adaptive video streaming–a survey and open challenges". *IEEE Access,* 7 : 30831-30859.

[3] Brachmański Stefan. 2010. "Subiektywna ocena wpływu wypadania pakietów na jakość mowy przesyłanej przez sieć Internet". Przegląd Telekomunikacyjny, Wiadomości Telekomunikacyjne, 83 (6).

[4] Braun Virginia, Victoria Clarke. 2014. „What Can 'Thematic Analysis' Offer Health and Wellbeing Researchers?". *International Journal of Qualitative Studies on Health and Well-Being*, 9 (1): 26152.

[5] Brooks Peter, Bjørn Hestnes. 2010. "User measures of quality of experience: why being objective and quantitative is important". *IEEE network,* 24 (2) : 8-13.

[6] Bułat Jarosław, Natalia Cieplińska, Rafał Figlus, Lucjan Janowski. 2024. „Daily Video: A tool for quality of experience (QoE) in long-term context research". *SoftwareX*, 25, 101637.

[7] Ericsson K. Anders, Herbert A. Simon. 1998. "How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking". *Mind, Culture, and Activity,* 5 (3) : 178-186.

[8] Falkowski-Gilski Przemysław, Grzegorz Debita. 2023. "Quality Evaluation of Speech Transmission via Two-way BPL-PLC Voice Communication System in an Underground Mine". *Archives of Acoustics*, 48 (4) *:* 585-592.

[9] International Telecommunication Union. 1996. *Methods for Subjective Determination of Transmission Quality*. ITU-T Recommendation P.800. Geneva: ITU.

[10] International Telecommunication Union. 2008. Subjective Video Quality Assessment Methods for Multimedia Applications. ITU-T Recommendation P.910. Geneva: ITU.

[11] International Telecommunication Union. 2023. *Methodology for the Subjective Assessment of the Quality of Television Images*. ITU-R Recommendation BT.500-15. Geneva: ITU.

[12] Janowski Robert. 2013. "Ocena jakości usług telefonii pakietowej." *Zeszyty Naukowe Warszawskiej Wyższej Szkoły Informatyki,* 7 (10) : 7-24.

[13] Konaszyński Tomasz, Dawid Juszka, Mikołaj Leszczuk. 2023. "Impact of the stimulus presentation structure on subjective video quality assessment." *Electronics,* 12 (22) : 4593.

[14] Le Callet Patrick, Sebastian Möller, Andrew Perkis, Kjell Brunnström, Sergio Beker, et al. 2013. "Qualinet White Paper on Definitions of Quality of Experience". Qualinet. https://hal.science/hal-04638470.

[15] Pinson Margaret, Lucjan Janowski, Romuald Pépion, Quan Huynh-Thu, Christian Schmidmer, Philip Corriveau, Audrey Younkin, Patrick Le Callet, William Ingram. 2012. "The influence of subjects and environment on audiovisual subjective tests: An international study". *IEEE Journal of Selected Topics in Signal Processing*, 6 (6) : 640-651.

[16] Pinson Margaret H. 2013. "The Consumer Digital Video Library [Best of the Web]". *IEEE Signal Processing Magazine*, 30 (4) : 172–174.

[17] Webster Arthur, Filippo Speranza. 2010. *Report on the Validation of Video Quality Models for High Definition Video Content.* Technical Report. Boulder, CO: Video Quality Experts Group.

[18] Wielgus Gabriela, Lucjan Janowski, Kamil Koniuch, Mikołaj Leszczuk, Rafał Figlus. 2023. „Proposing More Ecologically-Valid Experiment Protocol Using YouTube Platform". *Electronic Imaging*, 35: 1-6.

[19] Zieliński Sławomir. 2015. "Is a Multi-Slider Interface Layout Responsible for a Stimulus Spacing Bias in the MUSHRA Test?". *Archives of Acoustics*, 40 (4) : 585-594.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

REKONSTRUKCJA OBRAZU NA PODSTAWIE ZNAKOWANIA WODNEGO ZERO-BIT Z WYKORZYSTANIEM SIECI NEURONOWEJ

IMAGE RECONSTRUCTION BASED ON ZERO-BIT WATERMARKING USING A NEURAL NETWORK

Tomasz Walczyna[1], Zbigniew Piotrowski[2]

[1] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, tomasz.walczyna@wat.edu.pl
[2] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, zbigniew.piotrowski@wat.edu.pl

DOI: 10.15199/59.2025.4.111

Streszczenie: **W niniejszym artykule przedstawiono metodę znakowania wodnego typu zero-bit, umożliwiającą rekonstrukcję wyciętego obszaru obrazu na podstawie ukrytego znaku wodnego. Proponowane rozwiązanie wykorzystuje architekturę sieci neuronowej typu U-Net, uczoną do odtwarzania brakujących fragmentów obrazu. Przeprowadzono analizę wpływu rozmiaru wyciętego obszaru na jakość rekonstrukcji. Oceniono również odporność metody na przekształcenia oraz skuteczność rekonstrukcji w różnych warunkach degradacji danych. Uzyskane wyniki pokazują, że możliwe jest odzyskiwanie kluczowych informacji wizualnych nawet przy istotnym uszkodzeniu obrazu.**

Abstract: **This paper presents a zero-bit watermarking method designed to reconstruct cut-out image regions using an embedded watermark. The proposed approach employs a U-Net-based neural network trained to restore missing image areas. An analysis of the relationship between the size of the removed region and reconstruction accuracy is provided. The method's robustness to transformations and its effectiveness under various image degradation scenarios are also evaluated. Results show that the technique can recover key visual information even when significant portions of the image are removed.**

Słowa kluczowe: **rekonstrukcja obrazu, sieci neuronowe, U-Net, znak wodny, zero-bit**

Keywords: **image reconstruction, neural networks, U-Net, watermark, zero-bit**

1. WSTĘP

Znakowanie wodne (ang. watermarking) jest szeroko stosowaną techniką w dziedzinie przetwarzania obrazów, wykorzystywaną głównie do ochrony praw autorskich, wykrywania manipulacji oraz śledzenia pochodzenia danych [10]. Tradycyjne podejścia koncentrują się na osadzaniu danych binarnych w obrazie w sposób odporny na różnego rodzaju transformacje i kompresję [6]. W ostatnich latach coraz większą uwagę przyciągają techniki znakowania typu zero-bit, w których nie przenosi się jawnej informacji, a sama detekcja obecności znaku pełni funkcję sygnału sterującego lub weryfikacyjnego [4].

W prezentowanej pracy proponujemy nowe zastosowanie znakowania wodnego zero-bit – jako nośnika informacji umożliwiającego rekonstrukcję wyciętego fragmentu obrazu. Znakowanie wodne typu zero-bit (ang. zero-bit watermarking) polega wyłącznie na tym, że system potrafi stwierdzić obecność lub brak znaku—nie przenosi się żadnych jawnych bitów danych. Dzięki temu cała energia sygnału może zostać przeznaczona na zwiększenie odporności na degradację i ataki, co jest kluczowe, gdy znak ma później sterować rekonstruktorem obrazu.[3] Głównym założeniem metody jest to, że oryginalny obraz zostaje oznakowany w taki sposób, aby zakodowane w nim cechy mogły zostać wykorzystane do odtworzenia utraconego fragmentu, nawet jeśli nie zawiera on bezpośrednio danych binarnych. Takie podejście znajduje zastosowanie m.in. w scenariuszach zabezpieczania tożsamości wizualnej, gdzie pewne fragmenty mogą zostać celowo wycięte lub zamaskowane, a ich rekonstrukcja musi być możliwa jedynie dla systemu posiadającego wiedzę o znaku [1].

Do realizacji tej koncepcji wykorzystujemy architekturę sieci neuronowej typu U-Net, znaną z wysokiej skuteczności w zadaniach rekonstrukcji obrazu [11] dla czynnika degradującego typu usunięcie obrazu (ang. inpainting). W odróżnieniu od klasycznych podejść, które wymagają dużej ilości informacji kontekstowej wokół brakującego obszaru, nasza metoda bazuje na subtelnych cechach zakodowanych w całym obrazie, będących wynikiem działania znakującej sieci wodnej. Sieć ta działa jako funkcja zakłócająca, osadzająca sygnał umożliwiający dekodowanie brakującego fragmentu przez rekonstruujący dekoder.

W artykule analizujemy wpływ wielkości wyciętego obszaru na dokładność rekonstrukcji oraz identyfikujemy granicę informacyjną – czyli jak dużą część obrazu można usunąć, aby system nadal był w stanie poprawnie go odtworzyć. Eksperymenty pokazują, że możliwe jest odzyskanie istotnych semantycznie informacji wizualnych nawet przy utracie znacznej części obrazu.

Podejście to może być również postrzegane jako forma „regeneracyjnego" znakowania wodnego, w którym znak nie tylko służy do identyfikacji autentyczności, ale także umożliwia praktyczną rekonstrukcję treści wizualnej. Praca wpisuje się tym samym w nurt badań nad inteligentnym znakowaniem wodnym opartym na uczeniu głębokim, a także rozszerza zastosowania technik inpaintingu poza klasyczne przypadki brakujących danych [5, 8].

2. PROPONOWANA METODA

Celem proponowanej metody jest umożliwienie rekonstrukcji fragmentu obrazu, który został celowo usunięty, przy zachowaniu wysokiej przejrzystości i odporności znaku wodnego na degradację. W odróżnieniu od klasycznych metod znakowania wodnego, które skupiają się na detekcji manipulacji lub przenoszeniu danych binarnych, nasza metoda koncentruje się na tym, by umożliwić odtworzenie brakującego fragmentu na podstawie subtelnych cech zakodowanych w pozostałej części obrazu.

2.1. Architektura

System składa się z dwóch głównych komponentów: modułu osadzającego (ang. embedder) oraz modułu rekonstruktora (ang. reconstructor). Oba moduły mają identyczną strukturę, opartą na rozszerzonej architekturze U-Net, przystosowanej do zadania osadzania i dekodowania informacji ukrytej w obrazie.

Architektura składa się z:

- bloku kodującego (encoder block), odpowiedzialnego za ekstrakcję rózno wymiarowych reprezentacji obrazu,
- bloku dekodującego (decoder block), który stopniowo odzyskuje strukturę przestrzenną,
- bloków DenseNet [7], zapewniających silne propagowanie gradientów i wykorzystanie wcześniej uzyskanych reprezentacji,
- mechanizmu uwagi, w tym CBAM [12], umożliwiającego dynamiczne skupienie się na istotnych cechach kanałowych i przestrzennych,
- bloków zmiany rozdzielczości (UpSample i DownSample), opartych na konwolucji z normalizacją i aktywacją Mish [9], zapewniających płynne przejścia między poziomami.

Charakterystyczną cechą architektury jest rozszerzony system połączeń pomijających (ang. skip connections), który łączy nie tylko standardowe poziomy pomiędzy kodowaniem a dekodowaniem, ale także przekazuje cechy wysokiego poziomu bezpośrednio do końcowych etapów rekonstrukcji. Cały system przyjmuje obraz 256×256 jako wejście i zwraca zarówno wersję z osadzonym znakiem, jak i rekonstrukcję brakującego fragmentu. Wymiary wejścia są spowodowane naturą sieci neuronowych, które w domyślnych konfiguracjach przyjmują stałe wymiary.

2.2. Funkcje kosztu

Skuteczność metody opiera się na jednoczesnym balansowaniu między przejrzystością (ang. visibility) znaku a jakością rekonstrukcji. W tym celu zastosowano dwa zestawy funkcji kosztu:

Funkcje przejrzystości (ang. visibility loss): MSE (Mean Squared Error) – klasyczna miara różnicy pikselowej oraz LPIPS – miara percepcyjna, lepiej korelująca z ludzkim odbiorem jakości [13]. Liczona pomiędzy obrazem na który chcemy nałożyć znak wodny i tym już z nałożonym. LPIPS (Learned Perceptual Image Patch Similarity) to miara różnicy między obrazami, która wykorzystuje cechy wyekstrahowane przez sieci neuronowe i lepiej koreluje z ludzkim postrzeganiem jakości niż klasyczne metryki pikselowe.

Funkcje odporności (ang. robustness loss): MSE oraz LPIPS liczone pomiędzy oryginalnym fragmentem a jego rekonstrukcją z obrazu z wyciętym obszarem.

Równowaga między tymi dwoma typami strat jest kontrolowana przez hiperparametr, który w ramach eksperymentów był dobierany tak, aby maksymalizować dokładność rekonstrukcji przy minimalnej ingerencji wizualnej w oryginalny obraz.

2.3. Funkcja szumu

Kluczowym elementem procesu treningowego jest symulacja usunięcia fragmentu obrazu, którego rekonstrukcja stanowi cel zadania. W tym celu zaprojektowano algorytm generujący maskę, która wycina losowe prostokątne fragmenty o łącznej powierzchni odpowiadającej zadanej proporcji względem całego obrazu.

Wycinanie realizowane jest iteracyjnie poprzez losowy dobór prostokątnych fragmentów o zróżnicowanych proporcjach boków (wsp. kształtu od 0.3 do 3.0), aż do osiągnięcia docelowego procentu zakrytej powierzchni. Dzięki temu model uczy się rekonstrukcji różnych typów ubytków – zarówno lokalnych jak i rozległych.

3. EKSPERYMENTY

W niniejszym rozdziale przedstawione są parametry konfiguracyjne treningu, a także wyniki eksperymentów.

3.1. Ustawienia

Wszystkie eksperymenty przeprowadzono z wykorzystaniem zestawu danych Classified Images Dataset (ImageNet 256×256) dostępnego na platformie Kaggle [2], który zawiera 540 000 wysokiej jakości obrazów sklasyfikowanych w 1000 kategorii. Wszystkie obrazy są dostępne w rozdzielczości 256 × 256 pikseli, co umożliwia bezpośrednie ich wykorzystanie bez konieczności skalowania czy interpolacji. Zbiór ten zawiera szeroki zakres scen i obiektów, co pozwala przetestować ogólność proponowanej metody znakowania i rekonstrukcji na różnorodnych przypadkach wizualnych – od przedmiotów codziennego użytku po krajobrazy i obiekty biologiczne.

Zbiór danych został podzielony na trzy części: zbiór treningowy (80%), walidacyjny (10%) oraz testowy (10%), przy czym każdy podzbiór zawiera reprezentatywny przekrój klas. Dla walidacji i testów zapewniono brak powtórzeń obrazów względem danych treningowych, co pozwala dokładnie ocenić zdolność generalizacji modelu na wcześniej nieznanych danych.

Architektura sieci osadzającej i rekonstruującej (embedder i extractor) była trenowana na jednej karcie graficznej NVIDIA RTX 4090. Wykorzystano rozmiar batcha równy 16. Do optymalizacji wag sieci wykorzystano optymalizator Adam z parametrami: współczynnik uczenia 0.0001, $\beta_1 = 0.5$, $\beta_2 = 0.9$. Trening prowadzono do momentu ustabilizowania metryk walidacyjnych.

3.1. Ocena przejrzystości i odporności rekonstrukcji

W fazie eksperymentów oceniono skuteczność proponowanej metody pod kątem dwóch aspektów: przejrzystości znaku wodnego oraz jakości rekonstrukcji obszarów brakujących.

W celu oceny przejrzystości osadzonego znaku, obliczono wartość PSNR (Peak Signal-to-Noise Ratio) między obrazem oryginalnym a obrazem z osadzonym znakiem. Dla każdego poziomu zakrycia wynik pozostawał stały – 31.56 dB, co oznacza, że znak wodny nie wprowadza zauważalnych zniekształceń niezależnie od dalszych etapów przetwarzania.

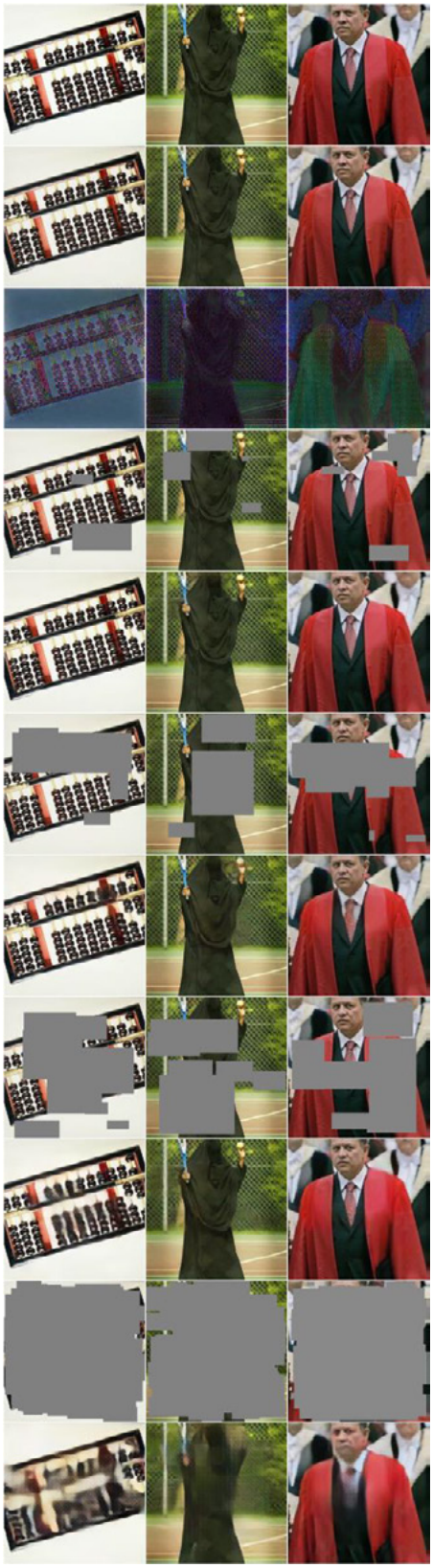
Następnie przeprowadzono szczegółową analizę wpływu procentowej wielkości wyciętego obszaru na jakość rekonstrukcji. W każdym przypadku wycięty obszar był losowy, zgodnie z procedurą opisaną w sekcji 2.3, a jego zawartość rekonstruowano na podstawie pozostałej części obrazu. Uzyskane wartości PSNR na obszarze brakującym zestawiono w tabeli 1.

Tab. 1. PSNR rekonstrukcji brakującego obszaru w zależności od stopnia zakrycia (średnia dla zbioru testowego)

Procent brakujących danych	*PSNR – rekonstrukcja [dB]*
5 %	19,29 ± 1.21
10 %	18,81± 1.10
20 %	18,26 ± 0.89
30 %	17,76 ± 0.88
40 %	17,56 ± 0.84
50 %	17,26 ± 0.89
60 %	16,89 ± 0.81
70 %	16,60 ± 0.81
80 %	16,21 ± 0.76
90 %	15,72 ± 0.77
95 %	15,29 ± 0.73
99 %	14,62 ± 0.72

Wyniki wskazują na stopniowy, ale relatywnie łagodny spadek jakości rekonstrukcji w miarę zwiększania się obszaru brakującego. Co istotne, nawet przy 50% brakujących danych model osiągał PSNR powyżej 17 dB, co odpowiada jakości wizualnej akceptowalnej w kontekście zastosowań inżynierskich, np. w systemach inpaintingu, analizy sceny czy automatycznego raportowania wizji.

Dla lepszej ilustracji działania metody, na rys. 1 przedstawiono przykłady rekonstrukcji dla wybranych wartości zakrycia: 10%, 30%, 50%, 95%. Wiersz 1 zawiera obrazy oryginalne, wiersz 2 – wersje z osadzonym znakiem wodnym, a wiersz 3 – wizualizację różnicy absolutnej |O - W| (wzmocnionej × 5). Kolejne pary wierszy przedstawiają obrazy z usuniętym fragmentem (masked) oraz odpowiadające im rekonstrukcje (reconstructed) dla poziomów degradacji: 10%, 30%, 50%, 95%. Szczególną uwagę zwraca odporność modelu na duże ubytki danych – nawet przy 95% zakrycia, rekonstruowany obraz zachowuje ogólny układ przestrzenny i estetyczną spójność.



Rys. 1. Przykładowe wyniki przetwarzania i rekonstrukcji dla trzech losowych obrazów.

4. WNIOSKI

W niniejszej pracy zaprezentowano metodę znakowania wodnego typu zero-bit, umożliwiającą rekonstrukcję brakujących fragmentów obrazu na podstawie cech ukrytych w całej strukturze wizualnej. W odróżnieniu od

klasycznych podejść opartych na detekcji binarnych znaków lub bezpośrednim przenoszeniu danych, proponowane rozwiązanie pełni funkcję „regeneracyjnego" znaku wodnego, którego obecność pozwala na odtworzenie usuniętej zawartości bez jawnej informacji o jej treści.

Przeprowadzone eksperymenty pokazały, że:

- osadzenie znaku wodnego nie prowadzi do zauważalnej degradacji obrazu (stały PSNR ≈ 31,56 dB),
- jakość rekonstrukcji stopniowo spada wraz ze wzrostem obszaru brakującego, jednak nawet przy 50% zakrycia możliwe jest odzyskanie semantycznie poprawnych i wizualnie spójnych rekonstrukcji (PSNR > 17 dB),
- model zachowuje zdolność odtwarzania ogólnej struktury sceny nawet przy bardzo dużych usunięciach.
- osadzenie informacji w sposób rozproszony i niejawny czyni metodę odporną na losowe degradacje i utratę danych w pojedynczych lokalizacjach.

Warto podkreślić, że podejście to nie wymaga znajomości lokalizacji brakującego obszaru ani obecności kontekstu w jego bezpośrednim sąsiedztwie. Umożliwia to wykorzystanie metody w zastosowaniach takich jak:

- ochrona tożsamości wizualnej,
- cyfrowa analiza śledcza (forensyka) i weryfikacja integralności zawartości wizualnej,
- adaptacyjne kodowanie treści w systemach o niskiej przepływności danych,
- znakowanie danych obrazowych wrażliwych na manipulacje.

Jednocześnie należy zaznaczyć, że skuteczność metody w kontekście bardziej złożonych zakłóceń (np. silnej kompresji stratnej, deformacji geometrycznych, czy przekształceń kolorystycznych) wymaga dalszych badań. Obecna wersja systemu opiera się na założeniu dostępności obrazu oryginalnego, co może ograniczać zastosowania w środowiskach otwartych.

W przyszłych pracach planowane jest rozszerzenie metody o dynamiczne warunkowanie procesu znakowania oraz integrację z systemami uwierzytelniania opartymi na embeddingach semantycznych. Celem jest stworzenie uniwersalnego frameworka pozwalającego nie tylko na rekonstrukcję obrazu, ale również na weryfikację jego autentyczności.

LITERATURA

[1] Asmaa Jawad, Rasha Thabit, Muntadher Al-Hadaad, and Khamis Zidan. 2024. 'A New Face Image Manipulation Localization and Recovery Algorithm Using Image Watermarking and Integer Wavelet Transform | IIETA'. 29 September 2024. https://doi.org/10.18280/mmep.110920

[2] 'Classified Images Dataset (ImageNet 256×256)'. n.d. Accessed 14 May 2025. https://www.kaggle.com/datasets/dimensi0n/imagenet-256

[3] 'Concept and Application of Zero-Watermark'. n.d. ResearchGate. Accessed 25 June 2025. https://www.researchgate.net/publication/292061969_Concept_and_application_of_zero-watermark

[4] Duszejko, Paweł, and Zbigniew Piotrowski. 2025. '4KSecure: A Universal Method for Active Manipulation Detection in Images of Any Resolution'. *Applied Sciences*, 15 (8). https://doi.org/10.3390/app15084469

[5] Duszejko, Paweł, Tomasz Walczyna, and Zbigniew Piotrowski. 2025. 'Detection of Manipulations in Digital Images: A Review of Passive and Active Methods Utilizing Deep Learning'. *Applied Sciences*, 15 (2). https://doi.org/10.3390/app15020881

[6] Faheem, Zaid Bin, Mubashir Ali, Muhammad Ahsan Raza, Farrukh Arslan, Jehad Ali, Mehedi Masud, and Mohammad Shorfuzzaman. 2022. 'Image Watermarking Scheme Using LSB and Image Gradient'. *Applied Sciences*, 12 (9). https://doi.org/10.3390/app12094202

[7] Huang, Gao, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2018. 'Densely Connected Convolutional Networks'. https://doi.org/10.48550/arXiv.1608.06993

[8] Kaczyński, Maciej, and Zbigniew Piotrowski. 2022. 'High-Quality Video Watermarking Based on Deep Neural Networks and Adjustable Subsquares Properties Algorithm'. *Sensors*, 22 (14). https://doi.org/10.3390/s22145376

[9] Misra, Diganta. 2020. 'Mish: A Self Regularized Non-Monotonic Activation Function'. https://doi.org/10.48550/arXiv.1908.08681

[10] Qureshi, Amna, David Megías, and Minoru Kuribayashi. 2021. 'Detecting Deepfake Videos Using Digital Watermarking'. In *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. https://ieeexplore.ieee.org/abstract/document/9689555

[11] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. 2015. 'U-Net: Convolutional Networks for Biomedical Image Segmentation'. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. https://doi.org/10.1007/978-3-319-24574-4_28

[12] Woo, Sanghyun, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. 'CBAM: Convolutional Block Attention Module'. https://doi.org/10.48550/arXiv.1807.06521

[13] Zhang, Richard, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. 'The Unreasonable Effectiveness of Deep Features as a Perceptual Metric'. https://doi.org/10.48550/arXiv.1801.03924

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

Statistical Evaluation of Subjective Video Quality on a Crowdsourcing Platform

Statystyczna ocena subiektywnej jakości wideo na platformie crowdsourcingowej

Avrajyoti Dutta[1]; Tomasz Konaszynski[2]; Dawid Juszka[3]; Mikołaj Leszczuk[4]

[1] AGH University of Krakow, avrajyoti.dutta@agh.edu.pl
[2] AGH University of Krakow, tomasz.konaszynski@agh.edu.pl
[3] AGH University of Krakow, dawid.juszka@agh.edu.pl
[4] AGH University of Krakow, mikolaj.leszczuk@agh.edu.pl

DOI: 10.15199/59.2025.4.112

Streszczenie: Zaprezentowano rozszerzoną analizę czasu reakcji wzgledem subiektywnych ocen jakości wideo na platformie crowdsourceingowej. Wyniki wykazały, że subiektywna ocena jakości klipów o „dostatecznej" jakości wymagała istotnie dłuższego czasu reakcji. Średni czas reakcji wyniósł 1867,19 ms, co sugeruje, że uczestnicy doświadczali większego obciążenia poznawczego i musieli włożyć więcej wysiłku w proces oceny. Badanie to pogłębia zrozumienie mechanizmów związanych z pozyskiwaniem subiektywnych ocen jakości wideo.

Abstract: An extended analysis of response time in relation to subjective video quality ratings on a crowdsourcing platform was presented. The results showed that subjective quality ratings of clips of "Fair" quality required significantly longer reaction times. The average reaction time was 1867.19 ms, suggesting that the participants experienced a higher cognitive load and had to put more effort into the evaluation process. This study deepens our understanding of the mechanisms involved in eliciting subjective assessments of video quality.

Słowa kluczowe: Jakość doświadczenia (QoE), przetworzone sekwencje wideo (PVS), eksperyment z crowdsourcingiem, subiektywna ocena jakości wideo (VQSA), poziom bezwzględnej oceny kategorii (ACR), czas reakcji.

Keywords: Quality of experience (QoE), Processed Video Sequences (PVS), Crowdsourcing Experiment, Video Quality Subjective Assessment (VQSA), Absolute Category Rating (ACR) Level, Response Time.

1. INTRODUCTION

The growing consumption of video content across platforms has made subjective video quality assessment increasingly important. One of the key concepts in this domain is Quality of Experience (QoE), which measures a user's perceived satisfaction with multimedia content. The ITU (International Telecommunication Union)-T P.910 recommendation provides standardized methods for conducting subjective assessments of video, audio, and audiovisual quality. This study aligns with these guidelines and investigates how response time during video evaluation can reflect underlying cognitive effort, particularly for videos of intermediate quality.

2. APPLICATION ARCHITECTURE

The application developed for this extended study is essential to allow subjective assessment of video quality via an efficient and intuitive interface [1]. The system architecture is illustrated in Figure 1.

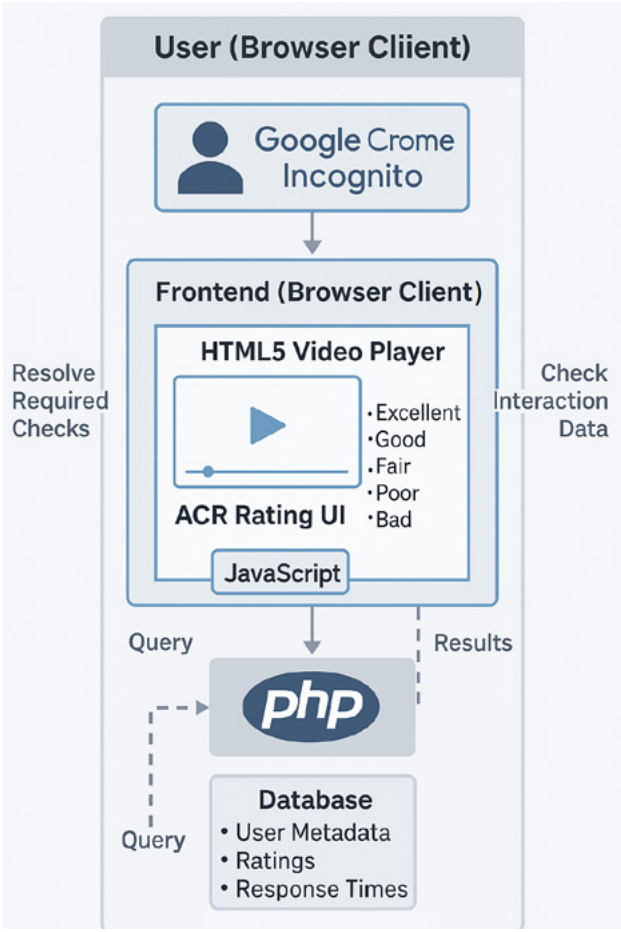


Figure 1: System architecture of the web platform for subjective video quality assessment (ACR-based) using Google Chrome Incognito and a PHP backend.

The design process emphasized usability, consistency, and data integrity, facilitating participant navigation of the experiment and enabling researchers to ob-

tain reliable high-quality data. This program was built according to the ITU-T P.910 Guidelines, which establish defined processes for performing subjective video quality assessments [2]. This compliance with industry standards guarantees the accuracy and comparability of the results.

Several previous studies, such as [3][4], have explored crowdsourced QoE assessments. Our approach differs by focusing on response time as a cognitive parameter, using a platform designed for full HD (High Definition) evaluation under ITU-T P.910 standards.

2.1 User Interface and Experience

The application interface was carefully designed to provide a smooth and entertaining user experience [5]. Upon initiating the application, users are presented with a friendly welcome message:

"Greetings! Welcome to the Video Quality Assessment Test. The results of your assessment will be utilized in our ongoing research on video quality algorithm systems."

This introductory message aims to guide participants by elucidating the purpose of the experiment and its importance in the furthering research on video quality algorithms. Participants are instructed to utilize the latest version of the Google Chrome browser in incognito mode to guarantee compatibility and optimal performance, with a link provided to download the browser if needed. Using incognito mode mitigates any distractions from cached data or browser extensions, thereby standardizing the testing environment for all users.

The interface prioritizes simplicity and clarity, guiding participants through the experiment with straightforward cues and few distractions [6]. The instructions are clear, specifying the anticipated duration of the test, approximately 55 minutes, and emphasizing the voluntary nature of participation. Participants are urged to complete the study, but are reminded of their right to withdraw at any time, to ensure an ethical methodology for data collection.

2.2 Experimental Guidelines

The application offers a succinct set of instructions to prepare participants for the assignment.

(a) Participants will consecutively observe 170 short video clips, each approximately 9 to 21 seconds long, encompassing diverse genres such as games, sports, reportage, and news. (b) The aim is to evaluate the quality of these video summaries, abridged copies of the original content. (c) Following each video, participants must select the video quality they deem preferable. (d) Participants determine the criteria for these quality assessments, allowing subjective evaluation based on individual perception. (e) Participants are cautioned to refrain from using mobile phones or the browser's back button throughout the experiment to avoid disruptions and maintain data integrity. (f) Participants are instructed to use F11 to activate full-screen mode for an improved viewing experience.

2.3 Technological Framework

The program was implemented using with PHP (Hypertext Preprocessor), JavaScript, CSS (Cascading Style Sheets), and HTML5 (Hypertext Markup Language). This technical stack was selected for its adaptability, cross-platform interoperability, and ability to provide a responsive and engaging user experience. The backend, developed in PHP, manages server-side functions, encompassing data storage, user session administration, and the aggregation of demographic and technical data. JavaScript facilitates client-side interactivity, allowing real-time data acquisition and video playback.

2.4 Assessment of Resolution and Environmental Conditions

An essential aspect of the application is the preliminary resolution verification, which guarantees that all participants meet the minimum display criteria for the experiment. At the commencement of the exam, the application automatically verifies the participant's monitor resolution to ensure it adheres to the 1080p (full HD) requirement. If the resolution fails to achieve this level, the participant is courteously barred from advancing in the experiment, so assuring that the video quality evaluations are performed under consistent settings.

2.5 Experimental Procedure

The experiment was performed solely on Google Chrome in incognito mode to eradicate potential biases like browser history, cookies or plugins. The participants received comprehensive instructions on how to perform the experiment, underscoring the need to present sincere and considered judgments.

This study used a crowdsourcing approach to ensure a rigorous and reliable evaluation of subjective video quality [3]. The experimental interface had an automatic sequential playback system that displayed video clips consecutively [6]. Following each clip, participants were instructed to evaluate the quality of the video on the ACR (Absolute Category Rating) scale, which comprised the categories Excellent, Good, Fair, Poor, and Bad. The interface was designed to be user-friendly, featuring explicit directions and straightforward controls to ensure seamless participation.

2.6 Data Collection

Two main data categories were collected: subjective quality assessments and reaction durations (in milliseconds). Subjective quality assessments were documented using the ACR scale, revealing categorical data on the participants' judgments about video quality [7]. Response times were recorded from the conclusion of each video clip until submission of the respective rating, thus quantifying the duration of the deliberation for each evaluation.

To guarantee the integrity and dependability of the data, multiple validation tests were integrated into the experiment. This encompassed attention checks with clearly distinguishable video quality levels and consistency assessments to detect any anomalous rating patterns. Participants who did not meet these criteria were omitted from the final analysis.

2.7 Data Analysis

The data collected underwent a thorough statistical analysis to reveal trends and patterns in subjective evaluations of video quality. Descriptive statistics offered a

summary of the distribution of ACR scores, while inferential statistics were used to investigate the correlations among video quality, response times, and other variables.

Response times were examined to comprehend the cognitive effort linked to varying quality levels. The investigation indicated that video clips classified as Fair Quality exhibited markedly longer response times, suggesting increased cognitive load and evaluative effort. This finding indicates that the participants exercised greater deliberation when evaluating videos of fair quality.

3. RESULTS

This section provides a comprehensive analysis of subjective video quality assessments gathered during the experiment. The results are structured to elucidate the distribution of quality ratings, reaction times, and the determinants that affect these assessments. The study provides a detailed understanding of subjective video quality perception by analyzing both the quantitative and qualitative dimensions of the responses of the participants.

3.1 Analysis of Response Time

The research focused on response times as an indicator of cognitive exertion in quality evaluations. Participants, on average, took 1,867.19 milliseconds to assess Fair quality videos, which was considerably longer than the durations for videos evaluated as Excellent (1,512.41 ms) or Bad (1,582.62 ms). The prolonged evaluation of Fair quality clips indicates that participants need extra time to analyze clips exhibiting intermediate quality, characterized by visual imperfections that were noticeable although not excessively distracting.

Table 1: Statistical Analysis Table [1]

ACR Video Quality Level	*Mean Response Time (ms)*	*Standard Deviation (ms)*
Excellent (5)	1512.41	870.142
Good (4)	1727.62	960.115
Fair (3)	1867.19	1042.44
Poor (2)	1767.75	1020.17
Bad (1)	1582.62	869.601

Table 1 denotes the average response times for each quality level, demonstrating a positive link between response times and the cognitive difficulty associated with evaluating fair quality videos [1]. The findings suggest that the participants were more inclined to participate in introspective decision making when confronted with moderate quality video content.

3.2 Statistical Analysis and Validation

In order to describe the extended results of the experiment in a more statistically significant way, measurement results were fitted to standard distributions. Such a statistical fit could be important to capture the trend of the distribution of the frequency of response time depending on the video quality. We fitted only the 'Fair' case which is the most significant one, the rest of the cases can be fitted similarly.

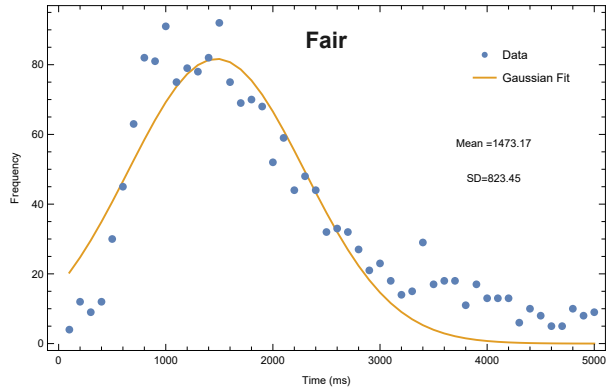


Figure 2: The distribution of participants' response times for votes corresponding to "Fair" video quality levels denoted by blue symbols. The solid curve represents a fit to the data by Gaussian distribution.

First, we assume that the sample distribution of the frequency of response time is a simple Gaussian. Fig. 2 shows the fit to the Gaussian distribution. The fit is reasonable, although not very good, as could be seen with an eye. As one moves towards larger response time, the fit gradually gets worse. The fit returns mean = 1473.17 ms and SD = 823.45 ms. Clearly, these values are farther apart from the sample mean presented in Table 1 [1]. One could clearly see a longer tail of the distribution at larger times. Therefore, a natural choice would be to fit with a gamma distribution. Fig. 3 represents a fit to the Gamma distribution. With the gamma distribution, the fit has improved significantly.

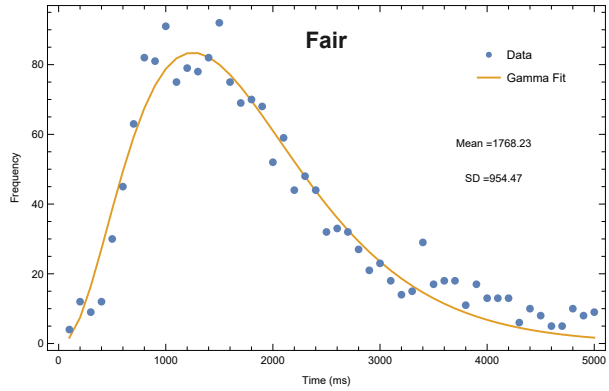


Figure 3: The distribution of participants' response times for votes corresponding to "Fair" video quality levels denoted by blue symbols. The solid curve represents a fit to the data by Gamma distribution.

The gamma fit returns Mean=1768.23 ms and SD=954.47 ms, which are close to the values corresponding to the sample mean and standard deviation. While the Gamma distribution visually offers a better fit than the Gaussian, especially for the longer tail of the response time data, it results in a slightly higher SD. This is due to the Gamma model capturing more of the extended range of response times, which increases the overall spread of the fitted data. The mean and standard deviation (SD) values shown in Table 1 are direct sample statistics derived from raw experimental data. In contrast, the mean and SD values in Figures 2 and 3 are calculated based on fitted Gaussian and Gamma distributions, respectively.

These fitted distributions aim to model the overall trend and may not exactly match the raw sample metrics.

Therefore, it would be pretty safe to infer that the results obtained from our measurement time on the response time for different cases (video quality) roughly follow a gamma distribution. While the Gamma distribution appears to fit the empirical data better than the Gaussian distribution based on visual inspection, no formal goodness-of-fit metrics (such as Kolmogorov–Smirnov or Chi-square tests) were computed in this study. We acknowledge this as a limitation and suggest incorporating such hypothesis testing in future work to quantitatively evaluate model suitability. Confidence intervals were not computed in this study but are planned for future work.

Although standard deviations are relatively high, ANOVA testing confirmed statistically significant differences between quality levels ($p < 0.05$), particularly for Fair-rated videos.

Subsequent statistical analyzes, corroborated that response times varied considerably between quality levels ($p < 0.05$), substantiating the study's conclusions that cognitive effort increased for videos classified as Fair quality. These statistical insights improve the results, emphasizing the importance of cognitive processing in subjective quality evaluation and the impact of content characteristics on evaluative effort.

4. CONCLUSIONS

This extended study conducts a comprehensive analysis of subjective video quality assessment using a crowdsourced evaluation of brief video clips. This research utilizes the Absolute Category Rating Scale (ACR) and measures response times to elucidate the cognitive processes associated with video quality assessment, providing practical insights for the advancement of video quality algorithms and user-focused multimedia applications.

We focused on the 'Fair' category as it exhibited the longest response times and the highest cognitive load. Future work may explore 'Good' and 'Poor' categories similarly. This tendency indicates that in subjective quality evaluations, viewers exert extra effort in assessing moderate-quality information, possibly due to nuanced flaws necessitating meticulous examination. This finding is crucial to comprehend the influence of cognitive load on quality assessment and underscores the need to recognize intermediate quality levels in subjective evaluations, which are often neglected in conventional assessment frameworks.

Understanding that 'Fair' quality causes increased cognitive load may help content creators improve visual elements such as sharpness, color consistency, or motion smoothness. These adjustments may elevate content perception from 'Fair' to 'Good' or 'Excellent'.

The implications of this research transcend technical aspects and encompass practical applications in the advancement of multimedia technology. Content creators, platform developers, and quality control professionals can utilize these insights to enhance their strategies for video content distribution, customizing experiences to align with user preferences and engagement trends.

Examining response times as a cognitive parameter offers a new perspective on subjective evaluations. Response times provide significant insights into the effort needed for evaluative decision making, acting as an indicator of the complexity of the viewer's cognitive processing. This measure enhances conventional quality ratings, augmenting the data and offering a more refined comprehension of user experiences. Subsequent research can build on this methodology by investigating the impact of additional variables, such as audio quality and relevancy of content, on response times and subjective evaluations.

Despite the substantial findings, it is crucial to acknowledge the limitations of this study. The sample size, while representative, is very limited and could be increased in a future study to encompass a wider demographic range. Moreover, dependence on a singular crowdsourcing platform may engender platform-specific biases. Extending this research to include varied demographics of participants and using additional data collection techniques could improve the generalizability of the findings.

This work advances our understanding of subjective video quality evaluation and its implications for multimedia technology. The findings underscore the importance of intermediate quality levels and the cognitive effort required to assess video quality, offering useful information for future study and practical applications. Crowdsourcing serves as a scalable and efficient approach to gathering subjective feedback, presenting a promising opportunity for video quality research by enabling swift data collection while maintaining rigorous standards of reliability and validity. As video content consumption expands, these insights will be essential in influencing future advancements in video quality evaluation and content distribution, thus improving the viewing experience for global audiences.

Funding: This work was funded by the Ministry of Science and Higher Education as part of a subsidy for the Faculty of Computer Science, Electronics and Telecommunications, AGH University of Krakow.

REFERENCES

[1] Dutta Avrajyoti, Tomasz Konaszyński, Dawid Juszka, and Mikołaj Leszczuk. 2024. "A Crowdsourcing Study of Video Quality." *Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne* 97 (4): 325–328.

[2] ITU-T. 2024. *ITU-T Recommendation P.910: Methods for the Subjective Assessment of Video Quality, Audio Quality and Audio-Visual Quality.*

[3] Naderi Babak, and Ross Cutler. 2024. "A Crowdsourcing Approach to Video Quality Assessment." In *Proceedings of ICASSP 2024*, 2810–2814.

[4] Hoßfeld Tobias, Keimel Christian, Hirth Matthias and Tran-Gia Phuoc. "Best Practices for QoE Crowdtesting: QoE Assessment with Crowdsourcing." *IEEE Transactions on Multimedia* 16, no. 2 (2014): 541–558.

[5] Dutta Avrajyoti, Dawid Juszka, and Mikołaj Leszczuk. 2024. "Crowdsourcing Evaluation of Video Summarization Algorithm." *International Journal of Electronics and Telecommunications* 70 (4): 1063–1069.

[6] Dutta Avrajyoti, Leszczuk Mikołaj, Juszka Dawid, and Grega Michał. 2023. "Subjective Quality Assessment of Video Summarization Algorithms: A Crowdsourcing Approach." *Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne* 96 (4): 333–336.

[7] Grega Michał, Janowski Lucjan, Leszczuk Mikołaj, Romaniak Piotr, and Papir Zdzisław. 2008. "Quality of Experience Evaluation for Multimedia Services." *Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne* 81: 142–153.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

SUBIEKTYWNA OCENA QOE WIDEO NA PODSTAWIE REAKCJI NA DEGRADACJĘ JAKOŚCI W RÓŻNYCH GRUPACH WIEKOWYCH

SUBJECTIVE VIDEO QOE ASSESSMENT BASED ON REACTION ON QUALITY DEGRADATION IN DIFFERENT AGE GROUPS

Dominika Wanat[1]; Dawid Juszka[2]

[1] AGH Akademia Górniczo-Hutnicza, Kraków, dominika.wanat@agh.edu.pl
[2] AGH Akademia Górniczo-Hutnicza, Kraków, dawid.juszka@agh.edu.pl

DOI: 10.15199/59.2025.4.113

Streszczenie: W Rekomendacjach ITU-T Absolute Category Rating (ACR) jest prezentowana jako wiodąca metoda przeprowadzania testów subiektywnych wideo. Niestety, standardowa procedura eksperymentu nie bierze pod uwagę czynników zaburzających ocenę. Z tego powodu, zaproponowaliśmy nową procedurę testów subiektywnych opartą o reakcję użytkownika na niesatysfakcjonującą jakość wideo. Zaadaptowana metoda dopasowania funkcji psychometrycznych pozwoliła porównać postrzeganie jakości wideo w trzech grupach wiekowych.

Abstract: ITU-T Recommendations suggest Absolute Category Rating (ACR) as a leading method of conducting subjective video quality tests. Unfortunately, the standard experiment procedure does not consider factors that could skew the ratings. Consequently, we proposed a novel subjective test procedure based on the user's reaction to dissatisfactory quality. The adapted method of fitting psychometric functions allows the perception of video quality in three age groups to be compared.

Słowa kluczowe: QoE (Quality of Experience), testy subiektywne, metoda największej wiarygodności, funkcje psychometryczne, postrzeganie jakości

Keywords: QoE (Quality of Experience), subjective tests, Maximum Likelihood Estimation (MLE), psychometric functions, quality perceptions

1. WSTĘP

Konwencjonalne testy subiektywne sprawdzające jak ludzie postrzegają jakość wideo, przeprowadzane są zazwyczaj według wytycznych zawartych w rekomendacjach ITU-T oraz ITU-R. Sekwencje wideo mogą być oceniane z użyciem metryk niereferencyjnych, takich jak: Naturalness Image Quality Evaluator (NIQE), Video BLIINDS (V-BLINDS) [7], Perception-based Image Quality Evaluator (PIQE) [9], Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [2] czy metryk wykorzystujących sieci neuronowe, oraz referencyjnych, takich jak Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Video Quality Metric (VQM) czy Video Multi-Method Assessment Fusion (VMAF) [5].

Jednak niektóre rekomendacje, takie jak ITU-T P.917 [3], zwracają także uwagę na szereg czynników, które mogą mieć dodatkowy wpływ zarówno na zachowanie uczestnika badań, jak i na udzieloną przez niego ocenę jakości wideo. Należą do nich:

- motywacje do obejrzenia danego wideo: zarówno wewnętrzna, jak i zewnętrzna,
- hipoteza testowa zaprezentowana użytkownikowi;
- realizm środowiska, w którym przeprowadzone są badania;
- przyjęte założenia oraz kontekst użycia.

Niniejsze czynniki mogą sprawiać, że użytkownicy będą mniej krytyczni (w przypadku braku motywacji, czy to zewnętrznej, czy wewnętrznej) lub będą one zaburzać oceny (w pozostałych przypadkach) [3].

Aby zniwelować ich wpływ, zaproponowaliśmy zupełnie nową procedurę eksperymentu subiektywnego. W tym celu odeszliśmy od klasycznego prezentowania krótkich, niemych sekwencji wideo oraz oceny ich jakości za pomocą pięciostopniowej skali ACR. W zamian pozwoliliśmy użytkownikom na wybór dowolnego filmu na platformie Netflix. Zupełnie nowym elementem w procedurze eksperymentu była możliwość reakcji na degradację jakości, kiedy stanie się ona dla użytkownika niesatysfakcjonująca. Pozwoliło to na obserwacje bardziej realistycznego zachowania użytkownika w obliczu degradacji jakości. Szczegółowa procedura eksperymentu jest opisana w Sekcji 2.1.

W rezultacie otrzymaliśmy dane dotyczące zachowania użytkowników, które wymagały nowego sposobu analizy danych. Zaadoptowaliśmy więc do naszych potrzeb metodę **dopasowania funkcji psychometrycznych**, znaną w psychologii, neurobiologii oraz naukach zajmujących się zachowaniem zwierząt i ludzi [4].

W wyniku dopasowania otrzymaliśmy funkcje obrazujące relację pomiędzy natężeniem bodźca oraz prawdopodobieństwem reakcji dla każdego użytkownika. Wykresy zostały dopasowane z użyciem pakietu R `Quickpsy` [6] za pomocą metody największej wiarygodności (ang. *Maximum Likelihood Estimation*, MLE). W przypadku naszego eksperymentu, za metrykę opisującą jakość wideo posłużyła nam odwrócona metryka VMAF (rVMAF) [5],

natomiast miarą reakcji na degradację jakości stało się prawdopodobieństwo naciśnięcia przycisku poprawiającego jakość. Szczegółowe informacje na temat modelu oraz dopasowania można znaleźć w Sekcji 2.2.

Zaproponowany eksperyment został przeprowadzony w laboratorium na grupie 34 osób. Otrzymane wykresy i ich parametry dostarczyły nam znacznie bardziej szczegółowych informacji o użytkowniku: jego konsekwencji (poprzez nachylenie funkcji – ang. *slope*) oraz punkcie, w którym użytkownik zaczyna zauważać różnicę w jakości (ang. *Point of Subjective Equality*, *PSE*) [1].

2. METODY

2.1 Procedura eksperymentu

W naszym nowym eksperymencie subiektywnym mogła wziąć udział każda osoba powyżej 18 roku życia nie będąca pracownikiem AGH (ograniczenie to wynikało z wymogów formalnych uczelni). Po zakończeniu eksperymentu otrzymywała ona wynagrodzenie w wysokości 200 PLN. Zgłoszenia odbywały się poprzez formularz online. Pierwszym etapem eksperymentu było sprawdzenie parametrów wzroku osoby chętnej do udziału w eksperymencie: wady wzroku oraz zaburzeń widzenia barw. W tym celu zostały użyte tablice Snellena oraz test Ishihary.

Następny etap obejmował dopełnienie formalności związanych z udziałem w eksperymencie. Jeśli uczestnik nie zadeklarował w formularzu jaki film chce obejrzeć podczas eksperymentu, eksperymentator odczytywał mu streszczenia trzech filmów: „The Higwayman" (thriller), „Faraway" (komedia romantyczna) oraz „The Old Guard" (film akcji/fantasy) i prosił o wybranie jednego z nich.

W końcu uczestnik był zapraszany do wygłuszonego pomieszczenia, gdzie znajdowały się dwie małe sofy, stolik RTV, na którym stał telewizor oraz przycisk pozwalający na reakcję i mały stolik boczny. Wygląd tego pomieszczenia został zaprezentowany na Rysunku 1.

Rysunek 1: Pomieszczenie eksperymentu. Po lewej stronie stołu z telewizorem widoczny jest przycisk pozwalający na reakcję na zmieniającą się jakość.

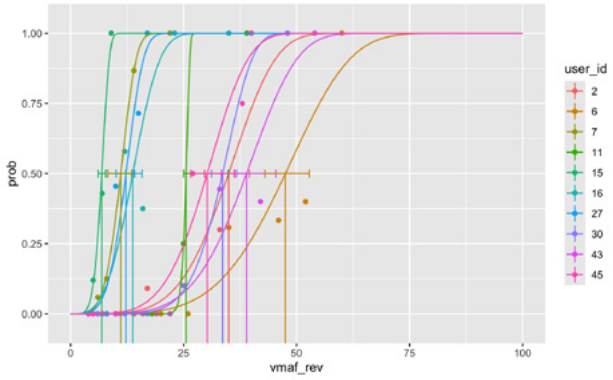
Uczestnicy byli zachęcani do swobodnego ustawienia stolika bocznego oraz drugiej sofy. Uczestnik oglądał instrukcję wideo wyjaśniającą przebieg eksperymentu, w szczególności, że jakość wideo może zmieniać się w czasie oglądania filmu. Kiedy jakość filmu stanie się nieakceptowalna dla uczestnika, może on przywrócić ją do najwyższej możliwej używając przycisku znajdującego się przy telewizorze. Przycisk mógł być używany w każdym momencie, zmiana jego lokalizacji i pozycji była jednak niedozwolona. Końcowy etap konfiguracji obejmował odpowiadanie na dodatkowe pytania uczestnika oraz preferencje dotyczące języka lektora oraz napisów. Po naciśnięciu przycisku „Start", następowała konfiguracja języka oraz napisów odtwarzacza wybranego filmu przez eksperymentatora i eksperymentator opuszczał pomieszczenie.

Tylko eksperymentator posiadał wiedzę, że zadaniem wtyczki do odtwarzacza serwisu Netflix w czasie trwania całego filmu była monotoniczna degradacja jakości do najgorszej dostępnej dla tego filmu (opisanej metryką VMAF). Każde naciśnięcie przycisku ponownie przywracało najlepszą dostępną jakość i rozpoczynało stopniową jej degradację.

Uczestnik był obserwowany w trakcie trwania całego filmu, aby w razie wystąpienia poważnych problemów technicznych lub zdrowotnych uczestnika miał możliwość szybkiej reakcji. Na zakończenie uczestnik wypełniał kwestionariusz dotyczący jego odczuć i wrażeń z eksperymentu oraz doświadczenia z platformami wideo.

2.2 Dopasowanie funkcji psychometrycznej

Wyniki uzyskane od każdego uczestnika eksperymentu zostały dopasowane funkcją psychometryczną. Opisuje ona prawdopodobieństwo reakcji uczestnika od jakości wyświetlanego wideo definiowaną poprzez odwróconą metrykę VMAF (rVMAF), wyrażoną wzorem 1. Przykładowe dopasowanie funkcji psychometrycznej do danych uzyskanych w eksperymencie widoczne jest na Rysunku 2.



Rysunek 2: Przykładowe funkcje psychometryczne uczestników eksperymentu.

$$rVMAF = 100 - VMAF \tag{1}$$

Z tego powodu używając metody największej wiarygodności możemy opisać tę relację:

$$f(\mathbf{k};\boldsymbol{\theta}) = \prod_{i=1}^{M} \binom{n_i}{k_i} \psi(x_i;\boldsymbol{\theta})^{k_i} (1-\psi(x_i;\boldsymbol{\theta}))^{n_i-k_i} \tag{2}$$

gdzie:

- f – funkcja gęstości prawdopodobieństwa;
- M – liczba prezentowanych użytkownikowi różnych poziomów rVMAF;
- x_i – *i-ta* sekwencja wideo o danej jakości opisanej przez rVMAF;
- n_i – liczba sekwencji wideo, w których prezentowany jest dany rVMAF (x_i);
- k – $(k_1, k_2, ..., k_M)$ wektor naciśnięć przycisku przy danej jakości;
- $\psi(x_i; \boldsymbol{\theta})$ – prawdopodobieństwo naciśnięcia przycisku kiedy *i-ty* rVMAF jest obserwowany – funkcja psychometryczna o następującej formie:

$$\psi(x; \boldsymbol{\theta}) = \psi(x; \alpha, \beta, \gamma, \lambda) = \gamma + (1 - \gamma - \lambda) F(x; \alpha, \beta) \tag{3}$$

gdzie:

  - θ – wektor definiujący rodzinę funkcji parametrycznych gęstości dla funkcji masy prawdopodobieństwa modelu;
  - α, β – parametry: pozycja i skala;
  - γ, λ – lewostronna i prawostronna asymptota ψ;
  - F – funkcja (rozkład) o sygmoidalnym kształcie, lewostronnej asymptocie w 0 oraz prawostronnej asymptocie w 1. Porównując grupy wiekowe do modelowania używamy rozkładu Weibulla, który zdefiniowany jest następująco:

$$F_W(x; \alpha, \beta) = 1 - e^{-\left(\frac{x}{\alpha}\right)^{\beta}} \tag{4}$$

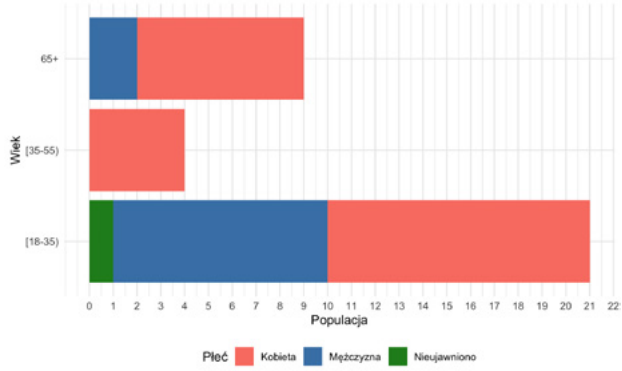
Rozkład ten ma dwa parametry: α oraz β. Parametr α zwiemy progiem psychofizycznym i określa on, jaki poziom jakości musi być zaprezentowany, aby otrzymać konkretne prawdopodobieństwo naciśnięcia przycisku. Parametr β nazywany jest nachyleniem funkcji i opisuje niepewność (niekonsekwencję) uczestnika eksperymentu. [4].

Estymacja parametru θ jest dokonywana poprzez metodę największej wiarygodności. Szukamy takiego parametru θ, który maksymalizuje prawdopodobieństwo L definiowane jako $L(\boldsymbol{\theta}) = f(\mathbf{k}; \boldsymbol{\theta})$, co jest jednoznaczne z maksymalizacją $\log L(\boldsymbol{\theta})$. Transformując równanie 2 otrzymujemy:

$$\log L(\theta) = \sum_{i=1}^{M} \left(\log \binom{n_i}{k_i} + k_i \log \psi(x_i; \theta) + (n_i - k_i) \log(1 - \psi(x_i; \theta)) \right) \tag{5}$$

3. WYNIKI

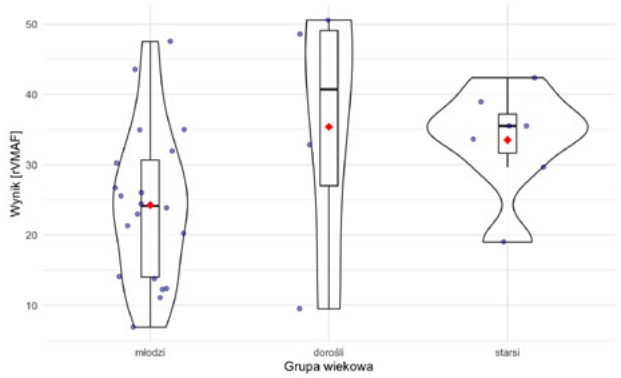
W eksperymencie opisanym w Sekcji 2.1 wzięło udział 34 uczestników: 11 mężczyzn, 22 kobiety oraz 1 osoba bez deklaracji płci. Uczestników eksperymentu podzielono na trzy grupy wiekowe: młodych dorosłych (ludzi pełnoletnich do ukończenia 35 roku życia), dorosłych (ludzi w wieku od 35 do ukończenia 55 roku życia) oraz ludzi starszych (65+). Finalny rozkład z podziałem na grupy wiekowe oraz płeć można zaobserwować na Rysunku 3.



Rysunek 3: Struktura płci i grup wiekowych uczestników eksperymentu.

Następnie dla każdej osoby, używając języka R oraz pakietu *Quickpsy*, stworzyliśmy funkcje psychometryczną opartą o metody opisane w sekcji 2.2 i przeprowadziliśmy analizę punktu równości subiektywnej (ang. PSE, *Point of Subjective Equality*). Jest to taki punkt, w którym uczestnik zaczyna zauważać różnicę w natężeniu bodźca (tutaj: zmianie jakości). Prawdopodobieństwo naciśnięcia przycisku w punkcie równości subiektywnej wynosi 50%, jest to zatem jeden z progów psychofizycznych.

Analiza dla średnich PSE spośród trzech grup została wykonana z użyciem ANOVA. Wcześniej sprawdziliśmy jednorodność wariancji za pomocą testu Bartlett'a, otrzymując $p = 0,1642$. Jako że $p > 0.05$, nie możemy odrzucić hipotezy, że wariancje w różnych grupach są równe. Przystąpiliśmy zatem do analizy poszczególnych grup, której efekty można obserwować na Rysunku 4.



Rysunek 4: Charakterystyka poszczególnych grup wiekowych: młodzi dorośli: $[18 - 35)$, dorośli: $< 35 - 55)$ oraz osoby starsze: 65+. Czerwony punkt reprezentuje wartość średnią, natomiast niebieskie punkty odpowiadają wartościom indywidualnym.

W grupie osób młodszych wartość PSE wynosi 24,2529, natomiast średnia szerokość przedziału ufności osiąga 6,5688. Zaobserwowaliśmy, że wariancja PSE w tej grupie jest najwyższa, przy stosunkowo równomiernym rozproszeniu wyników. W grupie tej zidentyfikowaliśmy również uczestnika o najwyższej wrażliwości (identyfikator 15), dla którego PSE wynosi 6,9039, a szerokość przedziału ufności 1,7274. Uczestnik ten charakteryzuje się wysoką spójnością odpowiedzi oraz zdolnością do wykrywania subtelnych zmian w degradacji jakości.

W grupie osób dorosłych średnia wartość PSE wyno-

si 35,3778, a średnia szerokość przedziału ufności 50,1072, co oznacza, że grupa ta charakteryzuje się najwyższą średnią w porównaniu z pozostałymi. W tej grupie wyróżnia się uczestnik o identyfikatorze 3, którego PSE wynosi 50,5685, a szerokość przedziału ufności osiąga 144,7674. Może to sugerować, że odpowiedzi tego uczestnika są niespójne, a jego reakcje na pogarszającą się jakość — chaotyczne.

W grupie osób starszych średnia wartość PSE wynosi 33,5172, a średnia szerokość przedziału ufności 8,1233. Rozkład wartości cechuje się prawoskośnością: większość uczestników w tej grupie wiekowej osiągnęła niższe wartości PSE niż średnia, a jedynie dwie osoby uzyskały wyższe wyniki. Ciekawym przypadkiem odstającym jest uczestnik o identyfikatorze 35, którego wyniki były zbliżone do średnich wartości uzyskanych przez osoby młodsze.

Wyniki testu ANOVA nie wykazały istotnych statystycznie różnic pomiędzy średnimi wartościami PSE w poszczególnych grupach wiekowych ($F(2;28) = 2,707; p = 0,0842$) przy założeniu $\alpha = 0,05$.

4. DYSKUSJA

W niniejszym artykule zaprezentowaliśmy nową procedurę eksperymentu opartego na reakcji na degradację jakości wideo oraz wykorzystanie zaadaptowanej metody dopasowania funkcji psychometrycznych. Na ich podstawie możemy porównać punkty, w których uczestnicy eksperymentu zaczęli dostrzegać degradację jakości: posługując się pierwotną metryką VMAF obserwujemy, że średnio grupa młodych ludzi zaczyna zauważać degradację jakości już przy wartości VMAF 75. Ciekawym wnioskiem jest fakt, że kolejną grupą są osoby starsze z wynikiem VMAF 66, a na końcu dorośli z wynikiem VMAF 50. Jest to zaskakujące, gdyż ze względu na podeszły wiek i związane z tym problemy ze wzrokiem oczekiwalibyśmy innych wyników.

Należy jednak zwrócić uwagę na fakt, że grupa dorosłych jest zdecydowanie mniej liczna od pozostałych. Związane jest to z faktem, że grupa ta jest aktywna zawodowo i bardzo trudno zachęcić ją do wzięcia udziału w tego typu eksperymencie (ang. *hard-to-reach population* [8] . Dodatkowa obecność chaotycznego i niespójnego uczestnika o identyfikatorze 3, ogranicza możliwość generalizowania wyników na populację.

Niewielka liczebność grupy dorosłych może być także powodem braku statystycznie istotnych różnic w teście ANOVA ($p = 0,0842$), który wykazuje jednak obiecujący wynik dla większych wartości parametru α.

5. PODZIĘKOWANIA

Referat został sfinansowany przez Ministerstwo Nauki i Szkolnictwa Wyższego w ramach subwencji dla Wydziału Informatyki, Elektroniki i Telekomunikacji AGH.

Badania prowadzące do tych wyników zostały sfinansowane z Norweskiego Mechanizmu Finansowego 2014-2021 w projekcie 2019/34/H/ST6/00599.

LITERATURA

[1] de'Sperati C. i Bovy G. B. 2017. „Low perceptual sensitivity to altered video speed in viewing a soccer match". Scientific reports 7, 1 (2017), 15379.

[2] Mittal A., Moorthy A. K., Bovik A. C. 2011 „Blind/Referenceless Image Spatial Quality Evaluator," 2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), Pacific Grove, CA, USA, 2011, pp. 723-727, doi: 10.1109/ACSSC.2011.6190099.

[3] ITU-T. 2019. P.917 „Subjective test methodology for assessing impact of initial loading delay on quality of experience". P.917.

[4] Klein S. A. 2001. „Measuring, estimating, and understanding the psycho- metric function: A commentary".Perception Psychophysics 63 pp. 1421–1455. issn: 0031-5117. doi: 10.3758/BF03194552.

[5] Li Z et al. 2016. „Toward A Practical Perceptual Video Quality Metric".url: https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652.

[6] Linares D. i López-Moliner J. 2016. „quickpsy: An R Package to Fit Psychometric Functions for Multiple Groups". The R Journal 8 (2016), 122. Issue 1. doi:10.32614/RJ-2016-008

[7] Saad M. A., Bovik A. C. , i Charrier C. , „Blind prediction of natural video quality," IEEE Trans. Image Process., vol. 23, no. 3, pp. 1352–1365, Mar. 2014.

[8] Shaghaghi A, Bhopal RS, Sheikh A. „Approaches to Recruiting 'Hard-To-Reach' Populations into Re-search: A Review of the Literature". Health Promot Perspect. 2011 Dec 20;1(2):86-94. doi: 10.5681/hpp.2011.009. PMID: 24688904; PMCID: PMC3963617.

[9] Venkatanath N, Praneeth D, Maruthi Chandrasekhar Bh, S. S. Channappayya and S. S. Medasani.2015. „Blind image quality evaluation using perception based features," Twenty First National Conference on Communications (NCC), Mumbai, India, 2015, pp. 1-6, doi: 10.1109/NCC.2015.7084843.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

SYNTION – TERMINOLOGIA I KLASYFIKACJA
SYNTHION – TERMINOLOGY AND CLASSIFICATION

Adrianna Sobczak[1]; Zbigniew Piotrowski[2]

[1]Wojskowa Akademia Techniczna, Wydział Elektroniki, Instytut Systemów Łączności, Warszawa
adrianna.sobczak@student.wat.edu.pl
[2]Wojskowa Akademia Techniczna, Wydział Elektroniki, Instytut Systemów Łączności, Warszawa
zbigniew.piotrowski@wat.edu.pl

DOI: 10.15199/59.2025.4.114

Streszczenie: **Referat przedstawia interdyscyplinarny system klasyfikacyjny Syntionów - treści generowanych cyfrowo-syntetycznie, z wykorzystaniem zarówno metod tradycyjnych, jak i systemów głębokiego uczenia. Zaproponowana została trójwymiarowa struktura porządkująca, obejmująca technologię tworzenia, modalność funkcjonalną oraz dziedzinę zastosowania. System ten umożliwia precyzyjne określenie rodzaju, pochodzenia i celu wszystkich jednostek i struktur cyfrowo-syntetycznych, a przede wszystkim wspiera rozwój technologiczny.**
Abstract: **The paper presents an interdisciplinary classification system for Synthions - digital-synthetic created content, using both traditional methods and deep learning systems. A three-dimensional ordering structure is proposed, encompassing the technology of creation, functional modality, and scope of application. This system enables precise identification of the type, origin, and purpose of all digitally synthetic units and structures and primarily supporting technological development.**
Słowa kluczowe: **cybersyntion, innowacje, nomenklatura, syntion.**
Keywords: **cybersynthion, innovations, nomenclature, synthion.**

1. WSTĘP

Współczesny dyskurs dotyczący cyfrowo-syntetycznych struktur i jednostek charakteryzuje się znacznym brakiem spójnej nomenklatury. Pomimo dynamicznego rozwoju modeli językowych, dźwiękowych, obrazowych i multimodalnych, nie został opracowany system pojęciowy, umożliwiający adekwatny opis, uporządkowanie i klasyfikację wytwarzanych jednostek.

Początki technologii generowania syntetycznych treści cyfrowych sięgają lat 70. XX wieku, natomiast przełomowe zmiany nastąpiły wraz z pojawieniem się sztucznej inteligencji. W 2014 roku została zaprezentowana przez Iana Goodfellowa architektura Generative Adversarial Networks (GAN), co zapoczątkowało nową erę technologii generatywnych oraz zostało uznane za jedno z najbardziej kontrowersyjnych osiągnięć współczesnej informatyki.

2. TERMINOLOGIA

W przestrzeni publicznej i dyskursie regulacyjnym stosowane są terminy nacechowane emocjonalnie, technicznie niedookreślone, bądź ograniczające znaczeniowo w odniesieniu do rzeczywistego spektrum zjawisk.

Określenie „deepfake" należy do najbardziej rozpoznawalnych i zarazem pejoratywnych pojęć związanych z technologią generatywną. Po raz pierwszy pojawiło się w 2017 roku jako pseudonim użytkownika, który opublikował kontrowersyjne treści pornograficzne - wygenerowane przy użyciu modeli głębokiego uczenia [1]. Pojęcie to szybko zostało spopularyzowane w mediach, a następnie wprowadzone do dokumentów prawnych. W amerykańskiej ustawie Protecting Americans from Deep Fakes Act (USA, 2019) zdefiniowano je jako syntetyczną treść audiowizualną tworzoną przez AI w celu imitowania autentyczności i wprowadzania odbiorcy w błąd [7]. Podobne ujęcie zastosowano w unijnej Ustawie o usługach cyfrowych (DSA, 2022), która nakłada obowiązek oznaczania treści syntetycznych z uwagi na ich potencjał dezinformacyjny [2]. Zgodnie z tymi regulacjami, określenie „deepfake" odnosi się wyłącznie do celowo szkodliwych i manipulacyjnych treści syntetycznych.

Wprowadzenie nowej, neutralnej i uporządkowanej taksonomii pojęciowej zostało zaproponowane w oparciu o dwa kluczowe terminy: „Syntion" oraz „Cybersyntion". System klasyfikacyjny tych jednostek opiera się na trójwymiarowej strukturze, w której wyróżnia się trzy niezależne osie porządkowe: metody tworzenia, modalności funkcjonalne oraz obszary zastosowania.

Definicja nadrzędna, obejmująca wszystkie treści, jednostki, struktury, bądź formy cyfrowo - syntetyczne:

Syntion - modalna jednostka lub struktura cyfrowo-syntetyczna, odwzorowująca rzeczywistość, symulację fikcji, bądź alternatywne reprezentacje percepcyjne. Może być wytwarzany z wykorzystaniem dowolnych technik cyfrowo-syntetycznych oraz występować zarówno w formie niematerialnej, jak i/lub materialnej.

Cybersyntion - syntion funkcjonujący w zintegrowanych środowiskach technologicznych, umożliwiających cyfrowe przechowywanie, przetwarzanie, transmisję, dystrybucję lub interakcję.

Syntion to nadrzędna jednostka obejmująca wszelkie formy syntetyczności - zarówno niematerialne (cyfrowe), jak i materialne. Termin ten odnosi się do każde-

go obiektu wytworzonego z użyciem technik cyfrowo-syntetycznych, niezależnie od jego aktualnej postaci lub nośnika - na przykład wydrukowany tekst wygenerowany przez model językowy.

CyberSyntion natomiast definiuje wyłącznie te jednostki lub struktury syntetyczne, które funkcjonują w przestrzeni cyfrowej w postaci niematerialnej. Rozróżnienie to umożliwia precyzyjną identyfikację, w zależności od ontologicznego zakotwiczenia, co stanowi podstawę dla projektowania stosownych ram prawnych, etycznych oraz systemów zarządzania.

Definicje uzupełniające - podział ze względu na metody tworzenia oraz udział czynnika ludzkiego:

DeepSyntion - syntion wygenerowany przez systemy głębokiego uczenia, bez bezpośredniej ingerencji człowieka w procesie realizacji. Powstaje autonomicznie, na podstawie danych wejściowych lub modeli percepcyjnych.

Tradycyjny Syntion - syntion powstały z wykorzystaniem tradycyjnych technik cyfrowych, przy bezpośrednim i wyłącznym udziale człowieka na każdym etapie tworzenia.

Hybrydowy Syntion - syntion powstały w wyniku zintegrowanego zastosowania tradycyjnych technik cyfrowych i metod sztucznej inteligencji, obejmujący jednocześnie działania człowieka i operacje generatywne.

„Tradycyjny Syntion" pozwala dokonać podziału ze względu na manualne metody tworzenia oraz bezpośredni wpływ człowieka na każdym etapie wytwarzania. „Deepsyntion" wyraźnie rozróżnia celowo zaprojektowane treści i w pełni generatywne materiały przez modele głębokiego uczenia (deep learning). „Hybrydowy Syntion" to synteza metod tradycyjnych z metodami algorytmicznymi. Terminologia posiada również znaczący potencjał skalowalności: może być stosowana w publikacjach naukowych, bazach danych, narzędziach detekcji i etykietowania, a także stanowić podstawę dla systemów normatywnych (ISO, IEEE, UNESCO).

Syntiony można także rozpatrywać w kontekście automatyzmu regulacyjnego czyli kluczowego mechanizmu opisanego przez Mariana Mazura (twórcę teorii systemów autonomicznych), który zakłada, że każdy system przetwarza napływające informacje zgodnie z wewnętrzną hierarchią wartości i celów (czyli *dominantą systemową*). W przypadku ludzi dominantą może być bezpieczeństwo, zysk, wiedza, emocje, w przypadku systemów sztucznej inteligencji to zadane cele algorytmiczne, architektura modelu czy parametryzacja funkcji straty. Oznacza to, że ten sam syntion (np. generatywny obraz lub instrukcja tekstowa) może mieć zupełnie odmienny efekt w zależności od systemu, do którego trafia i jego wewnętrznej dominanty.

W ujęciu M. Mazura istotny jest również podział informacji na energetyczną (realizującą działanie) oraz kierującą (inicjującą decyzję). Syntiony nie pełnią funkcji energetycznej – nie działają fizycznie – ale jako jednostki kierujące są nośnikami informacji determinujących reakcję, zachowanie lub reorganizację innego systemu np. użytkownika, modelu decyzyjnego, czy platformy społecznościowej [4].

3. KLASYFIKACJA

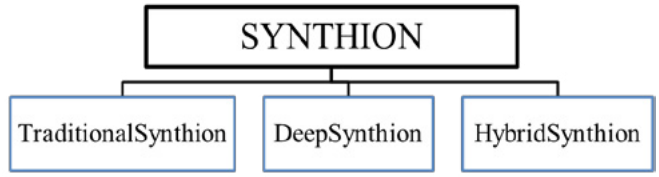
Proponowana klasyfikacja syntionów, oparta na trójwymiarowej strukturze porządkującej, uwzględnia trzy komplementarne wymiary opisu: sposób generowania (metody tworzenia), typ wykorzystywanych modalności (tekst, dźwięk, obraz, wideo, multimodalne) oraz dziedzinę zastosowania. Takie wielowymiarowe podejście umożliwia nie tylko systematyzację różnorodnych form syntionów, a także pozwala na analizę relacji między technologicznymi mechanizmami ich powstawania, a funkcjonalnym i społecznym kontekstem ich użycia.

Klasyfikacja ta wpisuje się w cybernetyczne ramy analityczne wypracowane przez Mariana Mazura. W jego ujęciu, każdy system autonomiczny charakteryzuje się zdolnością do samodzielnej regulacji na podstawie przetwarzanej informacji oraz relacji ze środowiskiem. Syntiony jako modalne jednostki cyfrowo-syntetyczne – mogą być interpretowane jako elementy informacyjne o charakterze kierującym, pełniące funkcje sygnałów sterujących w systemach odbiorczych (np. użytkownik, urządzenie, środowisko AI). Szczególnego znaczenia nabiera tutaj pojęcie informacji istotnej, czyli takiej, która wpływa na decyzje i zachowanie systemu. Modalność funkcjonalna syntionu (tekst, dźwięk, obraz, wideo, multimodalność) warunkuje kanał odbioru, natomiast dziedzina zastosowania determinuje możliwy efekt regulacyjny [4].

Dzięki swojej wielowymiarowej strukturze, klasyfikacja syntionów tworzy wspólny język opisu obiektów cyfrowo- syntetycznych.

3.1. Klasyfikacja Syntionu według metod tworzenia

Na podstawie poniższego schematu blokowego (Rys. 1.) zaprezentowana została klasyfikacja syntionów według metody ich tworzenia. W rezultacie wyodrębniono dwa odmienne paradygmaty tworzenia struktur syntetycznych: proceduralny - odpowiedzialność twórcza przypisywana jest człowiekowi (np. grafikowi, animatorowi), oraz algorytmiczny - opierający się na modelach głębokiego uczenia - proces tworzenia treści realizowany jest automatycznie przez system.



Rys. 1. Schemat blokowy przedstawiający podział ze względu na metodykę tworzenia Syntionu.
Źródło: opracowanie własne.

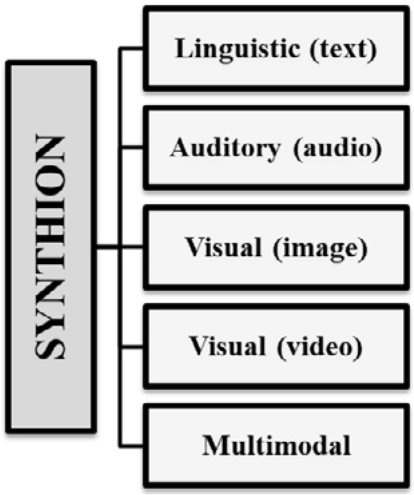
Strukturę pojęciową podzielono na trzy odrębne podzbiory. Syntiony Tradycyjne - wytwarzane w ramach manualnych procesów cyfrowych (np. CGI, ręczna animacja, cyfrowy montaż). DeepSyntiony - tworzone wyłącznie przy użyciu modeli głębokiego uczenia, bez bezpośredniej ingerencji człowieka w proces realizacji. Proces inicjowany jest przez użytkownika, natomiast

zasadniczy etap wytwarzania treści realizowany jest w sposób autonomiczny - struktura komponentów syntionu konstruowana jest bezpośrednio przez model głębokiego uczenia, działający niezależnie od człowieka na etapie projektowania i tworzenia. Syntiony Hybrydowe - konstruowane z wykorzystaniem metod mieszanych, integrujących komponenty manualne i algorytmiczne, pozwala na identyfikację struktur wytwarzanych z wykorzystaniem metod łączonych, co obecnie stanowi jeden z dominujących trendów technologicznych [5].

3.2. Klasyfikacja Syntionu według modalności funkcjonalnej

Modalność to nie tylko czynnik determinujący sposób percepcji treści przez człowieka, lecz również istotna kategoria technologiczna w strukturach systemów generatywnych. Z perspektywy detekcji syntetyczności, wpływ modalności na metody oceny autentyczności uwzględnia się m.in. poprzez analizę spektrum audio, identyfikację artefaktów obrazowych lub wykrywanie niespójności między kolejnymi klatkami materiału wideo.

Na schemacie blokowym (Rys. 2.) przedstawiono klasyfikację syntionów według modalności funkcjonalnej. Wyróżniono pięć podstawowych kategorii: językowe (tekst), dźwiękowe (audio), wizualne (obraz), wizualne (wideo), oraz multimodalne.



Rys. 2. Schemat blokowy - klasyfikacja Syntionu wg. modalności funkcjonalnej.

Źródło: opracowanie własne.

W modalności tekstowej uwzględniono zarówno treści generowane przez wielkoskalowe modele językowe (LLM), jak również komponenty narracyjne integrowane w immersyjnych środowiskach multimodalnych. Modalność dźwiękowa obejmuje syntetyczny głos, rekonstrukcje mowy, struktury muzyczne oraz efekty akustyczne, generowane przy użyciu systemów TTS oraz modeli dyfuzyjnych dedykowanych przetwarzaniu audio. Modalność wizualna klasyfikowana jest w dwóch odrębnych kategoriach: obrazów statycznych oraz materiałów wideo. W przypadku modalności multimodalnej identyfikowane są złożone konfiguracje integrujące tekst, obraz, dźwięk oraz komponenty interaktywne, charakterystyczne m.in. dla środowisk VR/AR, wirtualnych postaci czy awatarów.

3.3. Klasyfikacja Syntionu według dziedziny zastosowania

Opracowana została również klasyfikacja syntionów ze względu na dziedziny ich zastosowania, mająca na celu uporządkowanie oraz ujednolicenie terminologiczne w obszarze dynamicznie rozwijających się technologii cyfrowo-syntetycznych. W klasyfikacji uwzględniono zarówno zastosowania praktyczne, jak i koncepcyjne, obejmujące różne domeny życia społecznego, nauki, przemysłu i kultury.

Rozrywka, kultura i media: syntiony wykorzystywane są do tworzenia filmów, gier, koncertów i cyfrowych postaci. Przykłady obejmują generatywne materiały wideo (Runway, Sora), wirtualnych aktorów (hologram Tupaca Shakura - Coachella), syntetyczny dubbing (Respeecher), koncerty w VR (Travis Scott), czy cyfrowych influencerów (Lil Miquela).

Medycyna i zdrowie: syntiony wspierają tworzenie danych obrazowych (MRI, CT), symulacje chirurgiczne (FundamentalVR), rekonstrukcje głosu (ElevenLabs), edukację anatomiczną (Visible Body), generatywne modele farmakologiczne (Insilico), a także personalizowane komunikaty TTS.

Edukacja i szkolenia: umożliwiają personalizację nauki: awatary wykładowców (Synthesia), wirtualne laboratoria (Labster), podręczniki ilustrowane (Midjourney, DALL•E), immersyjne kursy językowe (Mondly VR), czy symulacje przemysłowe (Tecnomatix).

Człowiek i relacje społeczne: wspierają interakcje i dobrostan psychiczny: AI-partnerzy (Replika), syntetyczne awatary w metawersum (Zepeto), emocjonalni coachowie (Woebot), generatywne rytuały w aplikacjach wellbeing (Innerworld), oraz audiowizualne medytacje (Calm AI).

Prawo, polityka i bezpieczeństwo: syntiony służą do tworzenia syntetycznych świadków w rozprawach (Stanford), realistycznych treningów służb, detekcji manipulacji (Microsoft Video Authenticator), syntetycznych przemówień politycznych oraz testowania dowodów np. audiowizualnych.

Zastosowania komercyjne i przemysłowe: obejmują cyfrowych doradców klienta, personalizowane kampanie (Jasper), wirtualne postacie w brandingu (np. Coca-Cola AI Video), reklamy (Synthesia, Pika), a także katalogi produktowe i sesje zdjęciowe generowane przez AI (Deep Agency).

Informatyka i nowe technologie: narzędzia badań: generowanie danych (StyleGAN, SynthCity), benchmarki (MMBench), testy architektur (Transformers, Diffusion Models), integracja danych modalnych (CLIP), rozwój API (OpenAI, Replicate) oraz analiza odporności modeli.

Wysoka złożoność funkcjonalna syntionów wymaga ścisłego dostosowania systemów kontroli, przejrzystości i etykietowania – zwłaszcza w medycynie i prawie, gdzie obowiązują najwyższe standardy etyczne i prawne.

3.4. Studium przypadku - zastosowanie trójwymiarowej klasyfikacji syntionów

W celu zaprezentowania użyteczności trójwymiarowego systemu klasyfikacji syntionów, przedstawiono trzy przypadki, w których możliwe jest jednoznaczne określenie typu syntionu, jego modalności funkcjonalnej oraz dziedziny zastosowania.

Przypadek 1: Syntetyczny głos Luke'a Skywalkera w serialu „The Mandalorian"

W produkcji The Mandalorian (sezon 2 i 3), głos Luke'a Skywalkera został zrekonstruowany na podstawie archiwalnych nagrań Marka Hamilla z lat 70. i 80. XX w. Za realizację odpowiada firma Respeecher, stosując zaawansowane modele uczenia maszynowego do stworzenia głębokiej replikacji głosu. Głos został w pełni zsynchronizowany z wizualną reprezentacją postaci, sam proces nie obejmował generowania mowy od podstaw – wykorzystano realne dane fonetyczne jako bazę treningową [6].

Typ syntionu: HybridSyntion.
Modalność funkcjonalna: Dźwiękowa.
Dziedzina zastosowania: Rozrywka, kultura i media.

Przypadek 2: Wirtualny pacjent w systemie FundamentalVR

Platforma FundamentalVR umożliwia lekarzom wykonywanie wirtualnych zabiegów chirurgicznych w realistycznych warunkach operacyjnych. Pacjent – to przestrzenna, interaktywna reprezentacja anatomiczna oparta na danych rzeczywistych, modelach biomechanicznych i skryptach reakcji fizjologicznych. Technologia ta wykorzystuje złożone środowisko VR oraz sensoryczne sprzężenie zwrotne, nie bazując na samodzielnie generowanych modelach AI [3].

Typ syntionu: TraditionalSyntion.
Modalność funkcjonalna: Multimodalna.
Dziedzina zastosowania: Medycyna i zdrowie.

Przypadek 3: Syntetyczne przemówienie polityczne Baracka Obamy

W 2022 r. organizacje medialne i badawcze, w tym The Guardian, opublikowały eksperymentalne syntetyczne przemówienie wygenerowane na podstawie wypowiedzi byłego prezydenta USA - Baracka Obamy. Głos, mimika i tekst wypowiedzi zostały wygenerowane z użyciem głębokich sieci neuronowych. W badaniu analizowano potencjał dezinformacyjny deepfake'ów oraz reakcje społeczne na syntetyczne treści polityczne. To klasyczny przykład syntionu wygenerowanego całkowicie przez AI w celach polityczno-symulacyjnych [8].

Typ syntionu: DeepSyntion.
Modalność funkcjonalna: Wizualna – wideo.
Dziedzina zastosowania: Prawo, polityka i bezpieczeństwo.

4. WNIOSKI

W referacie zaprezentowany został trójwymiarowy system klasyfikacyjny, w ramach którego termin Syntion pełni funkcję nadrzędnej jednostki taksonomicznej.

W zaproponowanym podziale uwzględnione zostały trzy niezależne wymiary: metody tworzenia, modalności funkcjonalne oraz dziedziny zastosowania. Nowa taksonomia syntionów może być traktowana jako struktura odwzorowująca mapę interakcji między syntetycznymi jednostkami kierującymi, a odbiorczymi systemami autonomicznymi.

Uwzględnienie tych zależności umożliwia bardziej zaawansowane modelowanie procesów informacyjnych i regulacyjnych, a także pozwala na formułowanie precyzyjnych kryteriów detekcji, oceny ryzyka i projektowania interfejsów syntetyczności zgodnych z zasadami etycznej cybernetyki.

Praca badawcza została sfinansowana w ramach projektu Wojskowej Akademii Technicznej pt.: "Nowe architektury sieci neuronowych do przetwarzania sygnałów i danych w radiokomunikacji i multimediach."
Numer projektu: UGB 22-054/2025

LITERATURA:

[1] Chesney, Robert M., Danielle K. Citron. 2019. „Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security". *California Law Review* 107 (6): 1753–1819. https://doi.org/10.2139/ssrn.3213954
[2] Digital Services Act. 2022. Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2022/2065 z dnia 19 października 2022 r. w sprawie jednolitego rynku usług cyfrowych (akt o usługach cyfrowych) i zmiany dyrektywy 2000/31/WE. https://eur-lex.europa.eu/legal-content/PL/TXT/?uri=CELEX%3A32022R2065
[3] FundamentalVR. 2023. Immersive VR Surgical Training Platform. https://www.fundamentalvr.com
[4] Mazur, Marian. 1966. *Cybernetyczna teoria układów samodzielnych.* PWN.
[5] OECD. 2023. *Artificial Intelligence in Society: Emerging Trends and Policy Implications.* Paris: OECD Publishing. https://doi.org/10.1787/1c6108b1-en
[6] Respeecher. 2022. Voice Cloning for Lucasfilm: Bringing Back the Voice of Luke Skywalker. https://www.respeecher.com/blog/luke-skywalker-voice
[7] The Guardian. 2022. Deepfake Obama: Video Shows Fake Obama Delivering Real Message. https://www.theguardian.com/technology/video/2022/aug/16/deepfake-obama-barack-obama-video
[8] U.S. Senate. 2019. Protecting Americans from Deep Fakes Act, Senate Bill S.2065, 116th Cong., 1st sess. https://www.congress.gov/bill/116th-congress/senate-bill/2065

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

Underwater Image Enhancement with Deep Learning: A MOS-Guided, Multi-Criteria Loss Function Approach

Ulepszanie obrazów podwodnych z użyciem głębokiego uczenia: podejście z wielokryterialna funkcja strat sterowana wynikami MOS

Mehrunnisa[1]; Lallouani Bouchakour[2]; Dawid Juszka[3]; Yi Zhang[4]; Mikołaj Leszczuk[5]

[1] AGH University of Krakow, Kraków mehrunnisa@agh.edu.pl
[2] Universite des Sciences et de la Technologie Houari Boumediene (USTHB and CRSTDLA), Algeria lbouchakour@usthb.dz
[3] AGH University of Krakow, Kraków dawid.juszka@agh.edu.pl
[4] Xi'an Jiaotong University, Xi'an, China, yi.zhang.osu@xjtu.edu.cn
[5] AGH University of Krakow, Kraków, mikolaj.leszczuk@agh.edu.pl

DOI: 10.15199/59.2025.4.115

Streszczenie: Artykuł przedstawia oparty na głębokim uczeniu framework do ulepszania obrazów podwodnych. Zastosowano generator typu enkoder–dekoder trenowany przy użyciu złożonej, wielokryterialnej funkcji strat. Konwencjonalny pikselowy błąd średniokwadratowy (MSE) czesto prowadzi do rozmytych rezultatów i niskiej jakości percepcyjnej. Dlatego zaproponowana funkcja strat łaczy MSE z percepcyjnymi miarami podobieństwa LPIPS i SSIM. Wagi poszczególnych składników sa dobierane na podstawie średnich ocen subiektywnych (MOS), dzieki czemu proces optymalizacji lepiej odzwierciedla ludzka percepcje wzrokowa. Badania eksperymentalne pokazuja, że proponowane rozwiazanie istotnie poprawia zarówno wierność wizualna, jak i dokładność strukturalna; model osiagnał średni SSIM równy 0,95, co potwierdza jego skuteczność w zadaniu ulepszania obrazów podwodnych.

Abstract: This study proposes a deep learning-based framework for underwater image enhancement. We utilized an encoder-decoder generator optimized through a composite loss function. Conventional pixel-wise loss functions, such as Mean Squared Error (MSE), often result in blurred output and suboptimal perceptual quality. To address these limitations, we have introduced a multi-criteria loss function that integrates MSE with perceptual similarity measures such as the Learned Perceptual Image Patch Similarity (LPIPS) and the Structural Similarity Index Measure (SSIM). The relative weights of each component are systematically adjusted using Mean Opinion Scores (MOS) to better align the optimization process with human visual perception. Empirical evaluations show that the proposed method significantly improves visual fidelity and structural accuracy. Our model achieves a high SSIM score of 0.95, validating the effectiveness of the model in underwater image enhancement.

Słowa kluczowe: Średni błąd kwadratowy (MSE), Uczona percepcyjna miara podobieństwa łatek obrazu (LPIPS), Strukturalny indeks podobieństwa (SSIM), Średnia ocena subiektywna (MOS)

Keywords: Mean Squared Error (MSE), Learned Perceptual Image Patch Similarity (LPIPS), Structural Similarity Index Measure (SSIM), Mean opinion scores (MOS).

1. INTRODUCTION

In recent years, deep learning models, especially encoder-decoder architectures and generative adversarial networks (GANs), have made significant advances in image generation and enhancement tasks. An essential factor in the success of these models is designing effective loss functions that guide the learning process and shape the quality of the output. This consideration becomes especially important in domains such as underwater image enhancement, where typical visual issues such as color fading and contrast loss pose ongoing and complex challenges [7]. In image reconstruction, the Mean Squared Error (MSE) is a popular loss function that evaluates the differences at the pixel level between the generated and reference images. Although it is simple and effective for reducing absolute reconstruction errors, it sometimes struggles to capture the perceptual quality that humans see and appreciate. Images optimized exclusively with MSE can look overly smooth and lack fine details, particularly in high-resolution tasks. This usually happens because MSE ignores spatial relationships and high-level image features.To overcome these limitations, our work introduces a composite loss function that combines perceptual metrics into the training process. In particular, we blend MSE with two perceptual similarity measures: Learned Perceptual Image Patch Similarity (LPIPS) and Structural Similarity Index Measure (SSIM). LPIPS measures similarity using deep feature representations from

pre-trained neural networks, capturing high-level perceptual details [1, 2, 5]. Meanwhile, SSIM focuses on structural accuracy, brightness, and contrast, providing a more human-centered view of image quality [6]. The total loss is expressed as a weighted sum of these three components, where the weights $(\lambda_1, \lambda_2, \lambda_3)$ are optimized using the mean opinion scores (MOS) of human evaluators. This approach ensures that the training process is aligned with subjective human perception, enhancing the practical applicability of the model.Through extensive experiments, we show that our proposed loss function significantly improves both pixel-level accuracy and perceptual quality. The enhanced underwater images are much sharper, preserve more details and closely match the ground truth, reflected by an SSIM score of 0.95. This multi-criteria loss design not only bridges the gap between numerical performance and visual appeal but also lays a solid foundation for future advancements in perceptually driven image generation and enhancement.

2. METHODOLOGY

In this study, we propose a deep learning-based method for enhancing underwater images using a generator network built on an encoder-decoder architecture. The encoder extracts key visual features from the input image, while the decoder reconstructs a clearer, more visually appealing version. This architecture effectively preserves the original structure of the image, while significantly improving its perceptual quality.

2.1 Loss Function Design

The main challenge addressed in this work is to improve the quality of underwater images that was achieved by designing an optimized loss function. To do this, we introduce a combined loss function that brings together three components: Mean squared error (MSE), learned perceptual image patch similarity (LPIPS), and structural similarity index measure (SSIM). Together, these elements work best to guide the model toward generating more clearer and more perceptually accurate images. The complete loss function is defined as:

$$L_{total} = \lambda_1 L_{MSE} + \lambda_2(1 - SSIM(x,y)) + \lambda_3 LPIPS(x,y) \quad (1)$$

where:

- L_{MSE} measures pixel-level accuracy.
- L_{SSIM} evaluates structural similarity between the generated and reference images.
- L_{LPIPS} captures perceptual similarity based on deep feature representations extracted from a pre-trained neural network.
- λ_1, λ_2, and λ_3 are tunable coefficients optimized through a validation process guided by Mean Opinion Scores (MOS).

2.2 Optimization Process

The weighting coefficients $(\lambda_1, \lambda_2, \lambda_3)$ are optimized to minimize total loss while maintaining a high correlation with human perceptual quality, as determined by MOS. The optimization involves:

- Calculating the total loss and individual losses (MSE, SSIM, LPIPS) for different combinations of weighting coefficients.
- Select the combination that achieves the best correlation between total loss and MOS scores.

2.3 Experimental Setup

The model is trained on raw images from the SAUD dataset [3], which includes a wide range of underwater environments, objects, plants, fish, statues, shipwrecks, and species. The data set is divided into 70% for training and 30% for validation. Data loaders are configured with batch sizes of 8, 4, and 1 for training, validation, and test sets, respectively, using random shuffling for training and none for validation or testing, to ensure efficient loading. Training iterates over various combinations of parameters to achieve optimal visual quality while preserving structural accuracy.

2.4 Validation and Evaluation

Enhanced images are evaluated using MSE, SSIM, and LPIPS metrics, selecting the best model based on visual realism and high SSIM. The total loss, a weighted sum of MSE, LPIPS and SSIM with tunable coefficients $(\lambda_1, \lambda_2, \lambda_3)$ is optimized by testing multiple weight combinations on a validation set and computing both total and individual losses. We then analyze their correlation with mean opinion scores (MOS) to identify the most effective weights for the next training epoch. This iterative approach ensures high pixel-level accuracy and perceptual quality, making it well suited for real-world underwater image enhancement.

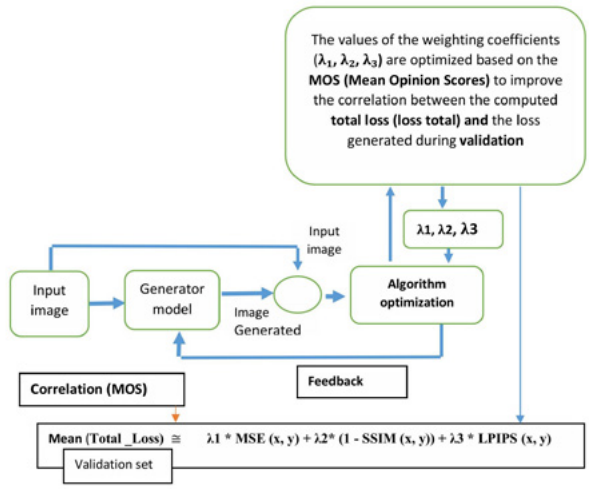


Figure 1: Loss function criteria used in image generators

3. RESULTS

In this study, we have evaluated the quality of generated images using three key metrics: MSE, SSIM, and LPIPS. These metrics assess both the accuracy of the pixel level and the perceptual quality of the underwater images produced. To enhance the generator model's performance, we explore different combinations of the weighting parameters $(\lambda_1, \lambda_2, \lambda_3)$ within a multi-criteria loss function.

3.1 Evaluation Method

For each $(\lambda_1, \lambda_2, \lambda_3)$ combination, we have calculated the total and individual losses (MSE, SSIM, LPIPS) in the validation set. The main objective is to find the combination that best correlates with subjective quality, measured by the Mean Opinion Score (MOS). The set that performs best is then used in the next training epoch, gradually improving the performance of the model.

3.2 Correlation Analysis

To find the best combination of parameters, we measure the correlation between the loss values and the MOS where the results show that a smaller λ_1, λ_3, paired with a moderate λ_2, give the highest correlation, striking the best balance between perceptual quality and overall loss.

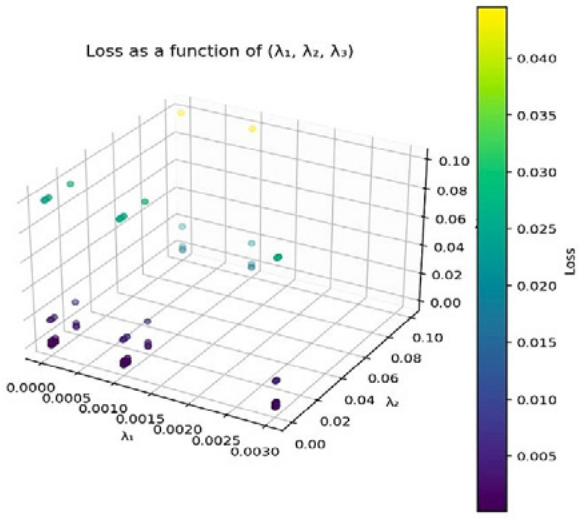


Figure 2: Visualization of loss values corresponding to various combinations of the parameters $(\lambda_1, \lambda_2, \lambda_3)$

The figure illustrates the relationship between the resulting loss values and the various combinations of $(\lambda_1, \lambda_2, \lambda_3)$. The chosen combination not only maximizes the correlation but also maintains low loss, which confirms its suitability for training.

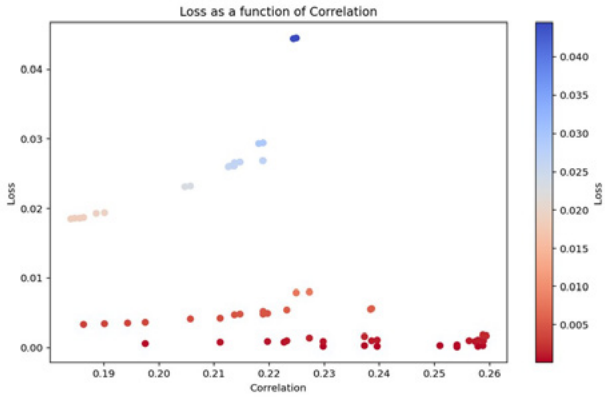


Figure 3: Relationship between total loss and correlation

Analysis proofs an inverse relationship between the quality correlation and total loss as correlation with actual quality rises but the total loss drops. Lower correlations often mean higher loss, highlighting the importance of loss minimization for strong structural similarity and perceptual quality. However, some deviations indicate model complexity and regularization also influence the results. In general, minimizing loss remains significant to generate accurate and visually coherent images. SSIM results in 60 underwater images further demonstrate the model's ability to preserve structural similarity.

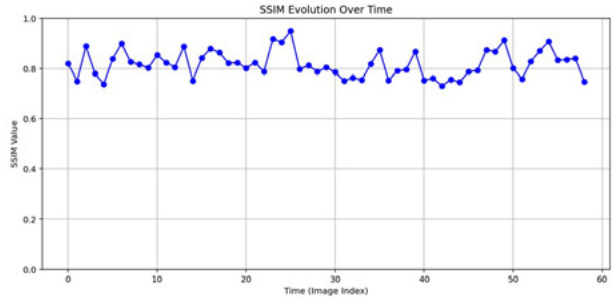


Figure 4: Distribution of SSIM values for test data samples

The distribution plot shows that images with SSIM above 0.9 preserve the structure well, which means the model retains visual detail and sharpness. For example, image 25 with an SSIM of 0.95, closely resembles the reference, which shows that the model can produce high-quality results. However, some images have lower SSIM scores due to visual complexity or loss of detail, which highlights the need for further optimization in more challenging cases.

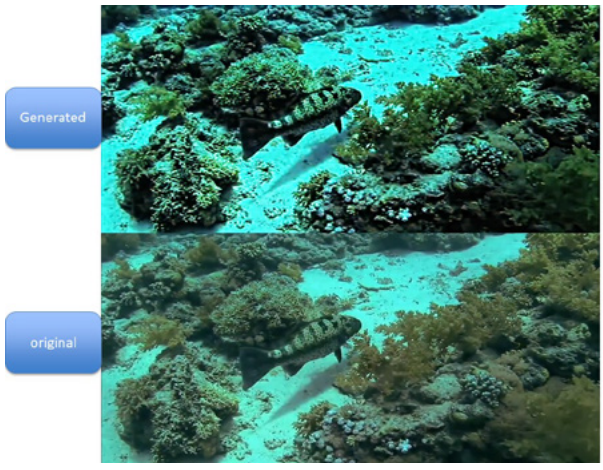


Figure 5: Final output: Original image transformed into more clear and refined version

Figure 5 shows the transformation of the original image into a more refined and clearer version. The enhancements are clearly visible in the generated image, showing increased sharpness, reduced noise, and better preservation of structural details. This transformation not only demonstrates the effectiveness of the techniques used, but also underscores their potential to achieve high-quality image restoration, even under challenging conditions.

3.3 Comparative Analysis of the Base Model and Our Proposed Model for Image Enhancement

This section compares the baseline model with our proposed enhanced model to evaluate improvements in image enhancement performance and perceptual quality.

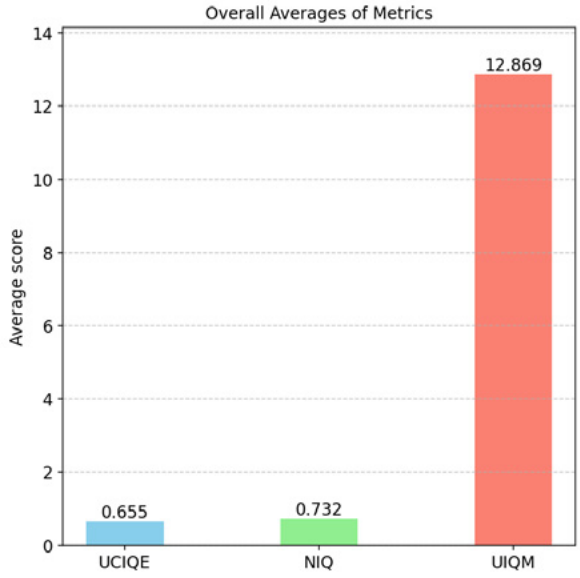


Figure 6: Baseline model architecture trained with MSE loss for underwater image enhancement

Figure 6 presents a lightweight CNN trained end-to-end on raw RGB images from the same dataset [3] without high-resolution ground truth. It uses three convolutional layers: large 9×9 kernels in the first and last layers to capture global features, and a 3×3 kernel in the middle for detail refinement. ReLU activations follow the first two layers, with a Tanh-scaled output between -1 and 1. The model is optimized with MSE loss using the Adam optimizer, achieving effective enhancement at low computational cost.

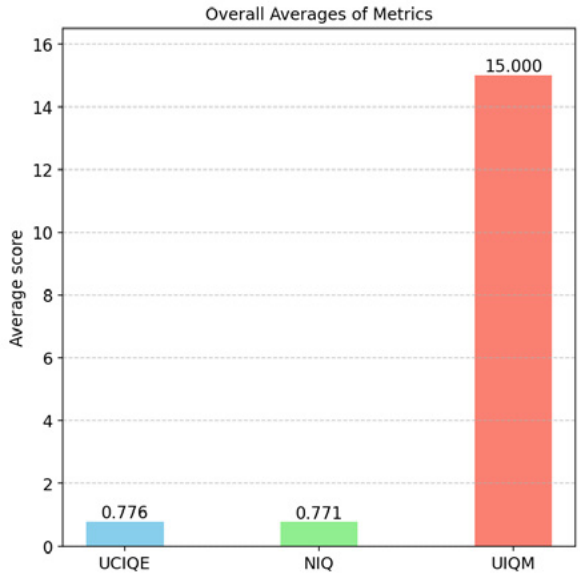


Figure 7: Enhanced model leveraging perceptual losses (LPIPS, SSIM) achieves improved visual and structural quality

Figure 7 shows the performance of the enhanced model, which extends the baseline with LPIPS and SSIM alongside MSE, with component weights tuned via MOS to match human perception. The model achieves consistent gains in non-reference metrics, with the Underwater Color Image Quality Evaluation (UCIQE), Naturalness Image Quality (NIQ) averaging 0.77, and the Underwater Image Quality Measure (UIQM) reaching 15, confirming the improved perceptual quality in underwater and general scenes[4]. This shows that the proposed model achieves superior visual fidelity and structural precision through perceptual metrics, validated by SSIM scores up to 0.95 with consistently high UCIQE, NIQ, and UIQM scores across 60 images.

4. CONCLUSION

This study proposes an enhanced framework for underwater image generation using a generator network trained with a combined loss function (MSE, LPIPS, SSIM) to incorporate perceptual and structural cues aligned with human vision. The method improves image realism and structural fidelity, achieving SSIM scores of up to 0.95, with MOS evaluations that confirm a realistic appearance. Optimal results came from weighting SSIM higher and MSE/LPIPS lower, aligning with human perception. No-reference metrics (UCIQE, NIQ, UIQM) further validated the method, showing consistently high scores across 60 generated images.

5. ACKNOWLEDGMENT

This work was supported by the Polish Ministry of Science and Higher Education with the subsidy funds of the Faculty of Computer Science, Electronics and Telecommunications of AGH University.

REFERENCES

[1] Sara Ghazanfari, Siddharth Garg, Prashanth Krishnamurthy, Farshad Khorrami, and Alexandre Araujo. R-LPIPS: An Adversarially Robust Perceptual Similarity Metric, 2023.

[2] Dawei Guan, Xinping Lin, Haoyi Zhang, and Hang Zhou. VariGAN: Enhancing Image Style Transfer via UNet Generator, Depthwise Discriminator, and LPIPS Loss in Adversarial Learning Framework. *Sensors*, 25(9), 2025.

[3] Qiuping Jiang, Yuese Gu, Chongyi Li, Runmin Cong, and Feng Shao. Underwater Image Enhancement Quality Evaluation: Benchmark Dataset and Objective Metric. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):5959–5974, 2022.

[4] Wenxia Li, Chi Lin, Ting Luo, Hong Li, Haiyong Xu, and Lihong Wang. Subjective and Objective Quality Evaluation for Underwater Image Enhancement and Restoration. *Symmetry*, 14(3), 2022.

[5] Jim Nilsson and Tomas Akenine-Möller. Understanding SSIM, 2020.

[6] Yuriy Reznik. Another look at SSIM image quality metric. *Electronic Imaging*, 35(8):305–1–305–1, 2023.

[7] Bo Xu, Dong Zhou, and Weijing Li. Image Enhancement Algorithm Based on GAN Neural Network. *IEEE Access*, 10:36766–36777, 2022.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

INSPEKCJA WIZYJNA Z WYKORZYSTANIEM SZTUCZNEJ INTELIGENCJI W CELU POPRAWY BEZPIECZEŃSTWA OBSZARU LOTNISKA

VISION INSPECTION USING ARTIFICIAL INTELLIGENCE TO IMPROVE SAFETY OF AERODROME

Jakub Suder

Zakład Układów Elektronicznych i Przetwarzania Sygnałów, Instytut Automatyki i Robotyki, Wydział Automatyki, Robotyki i Elektrotechniki, Politechnika Poznańska, Poznań, jakub.suder@put.poznan.pl

DOI: 10.15199/59.2025.4.116

Streszczenie: **Artykuł przedstawia podsumowanie badań nad wykorzystaniem sztucznej inteligencji w inspekcji wizyjnej portów lotniczych w celu zwiększenia bezpieczeństwa operacyjnego, zgodnie z wytycznymi EASA. Opracowano metody oparte na sieciach neuronowych i przetwarzaniu obrazu, wdrażane w systemach wbudowanych, wspomagające automatyzację inspekcji. Badania objęły wykrywanie obiektów FOD, oznaczeń poziomych oraz klasyfikację lamp oświetlenia nawigacyjnego, z zastosowaniem algorytmów CNN i autorskich baz danych (PUT, PLAVS1, PLAVS2). Projekt zrealizowano we współpracy Politechniki Poznańskiej z Portem Lotniczym Poznań–Ławica.**

Abstract: **The paper presents research on the application of artificial intelligence in vision-based inspection of airport infrastructure to enhance operational safety, in line with EASA guidelines. The study proposes methods based on convolutional neural networks and image processing techniques, implemented on embedded systems to support inspection automation. The research focused on three key areas: detection of Foreign Object Debris (FOD), recognition of horizontal airport markings, and classification of in-pavement airfield lighting, utilizing proprietary datasets (PUT, PLAVS1, PLAVS2) and advanced CNN architectures. The project was conducted in collaboration between Poznan University of Technology and Poznań–Ławica Airport.**

Słowa kluczowe: **bezpieczeństwo lotnicze, przetwarzanie obrazu, przetwarzanie brzegowe, uczenie maszynowe, sztuczne sieci neuronowe**

Keywords: **artificial neural networks, aviation safety, edge computing, image processing, machine learning**

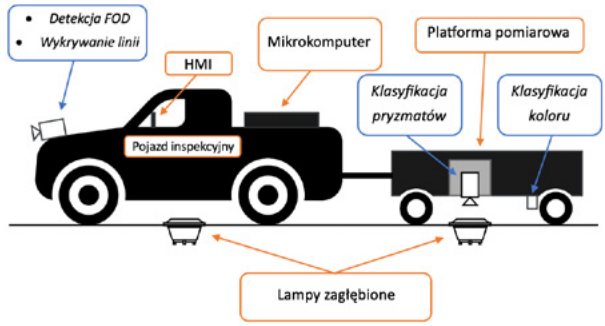
1. WSTĘP

Wraz z rosnącym natężeniem ruchu lotniczego i zwiększonym zapotrzebowaniem na niezawodność infrastruktury portów lotniczych, problematyka automatyzacji inspekcji technicznej płaszczyzn manewrowych zyskała na znaczeniu. W pracy [1] podjęto próbę opracowania systemów wspomagających detekcję i klasyfikację zagrożeń z wykorzystaniem sztucznej inteligencji, systemów wizyjnych oraz algorytmów przetwarzania obrazu, zgodnie z wytycznymi EASA (Agencja Unii Europejskiej ds. Bezpieczeństwa Lotniczego), ICAO (Organizacja Międzynarodowego Lotnictwa Cywilnego) i FAA (Federalna Administracja Lotnictwa) [2], [3].

W szczególności analizie poddano trzy kluczowe obszary inspekcji:

- wykrywanie obiektów obcych typu FOD (Foreign Object Debris)
- detekcję poziomych oznaczeń lotniskowych
- klasyfikację zagłębionych lamp oświetlenia nawigacyjnego

Zaprezentowano autorskie podejścia, opracowano bazy danych oraz implementację algorytmów z wykorzystaniem systemów wbudowanych [4], [5]. Koncepcje systemu zaprezentowano na Rys. 1.



Rys. 1. Koncepcja systemu pomiarowego

Płaszczyzny manewrowe portu lotniczego (drogi startowe, kołowania i płyty postojowe) wymagają regularnej i dokładnej inspekcji, aby zapewnić bezpieczeństwo operacji lotniczych oraz ciągłość działania infrastruktury. Dotychczasowe procedury, głównie manualne, cechuje ograniczona skuteczność i podatność na błędy ludzkie. Nowoczesne rekomendacje – w tym Artificial Intelligence Roadmap 2.0 [6] przygotowany przez EASA – podkreślają konieczność integracji narzędzi sztucznej inteligencji w procesy inspekcyjne, w tym głębokich sieci neuronowych (DNN), algorytmów decyzyjnych oraz technik komputerowego rozpoznawania obrazu.

Foreign Object Debris (FOD) stanowią jedno z najpoważniejszych zagrożeń dla operacji lotniczych. Ich obecność może prowadzić do poważnych uszkodzeń kadłuba, silników, podwozia, a w skrajnych przypadkach – katastrof lotniczych (np. Concorde Flight 4590). W pracy dokonano klasyfikacji FOD ze względu na ich pochodzenie (części samolotów, odpady lotnicze, zanieczyszczenia naturalne i meteorologiczne) oraz przeanalizowano dostępne międzynarodowe bazy danych, takie jak FOD-A, a także utworzono własny zbiór danych PUT dataset.

Zastosowanie algorytmów segmentacji i ekstrakcji cech obrazu umożliwiło detekcję FOD przy użyciu klasycznych metod (m.in. k-means) oraz modeli CNN, takich jak GoogLeNet i YOLOv5. Badania wykazały istotny wpływ rozdzielczości obrazu na skuteczność detekcji, przy czym modele głębokiego uczenia osiągnęły wyższą precyzję klasyfikacji niż podejścia klasyczne. Ostateczna architektura została zaimplementowana na platformie wbudowanej (Nvidia Jetson), zapewniając przetwarzanie obrazu w czasie rzeczywistym.

Oznaczenia poziome pełnią funkcję nawigacyjną i organizacyjną – są podstawą do bezpiecznego przemieszczania się statków powietrznych oraz pojazdów naziemnych. Z uwagi na regulacje ICAO oraz lokalnych władz lotniczych, linie te muszą być rozmieszczone zgodnie z określonymi specyfikacjami geometrycznymi, co umożliwia ich wykorzystanie do wspierania nawigacji urządzeń pomiarowych.

W pracy przygotowano bazę PLAVS1, zawierającą materiał wideo z rzeczywistych warunków lotniskowych. Opracowano i przetestowano metody wykrywania linii, w tym transformację Hougha, filtry Canny'ego oraz metody morfologiczne. Badano skuteczność tych algorytmów w kontekście wdrożenia na systemach wbudowanych. Uwzględniono także zmienne warunki oświetleniowe oraz zakłócenia obrazu powstające w wyniku opadów lub zabrudzeń.

Oświetlenie nawigacyjne wbudowane w nawierzchnię drogową (AGL – Airfield Ground Lighting) jest kluczowe dla operacji w warunkach ograniczonej widoczności. W ramach pracy opracowano system detekcji i klasyfikacji lamp na podstawie analizy stanu pryzmatów i chromatyczności emitowanego światła.

Wykorzystując autorską bazę danych PLAVS2, zastosowano sieci GoogLeNet, VGG-19 i ResNet do klasyfikacji obrazu lamp oraz czujniki koloru zgodne z przestrzenią barwną CIE 1931, umożliwiające pomiar i ocenę barwy zgodnie z normami ICAO Annex 14 [3]. Zaproponowane podejście pozwala na wczesne wykrywanie uszkodzeń lamp i automatyczne sygnalizowanie konieczności ich wymiany lub konserwacji.

W realizacji opisanych zadań wykorzystano narzędzia takie jak OpenCV, MATLAB oraz biblioteki deep learningowe TensorFlow, Keras i Deep Learning Toolbox, omawiając ich możliwości w zakresie przetwarzania obrazów w czasie rzeczywistym, klasyfikacji danych wizualnych oraz integracji z systemami wbudowanymi.

2. DETEKCJA OBIEKTÓW OBCYCH (FOD)

Autor zaprezentował nowatorski system wykrywania obiektów FOD, który opiera się na integracji systemu wizyjnego z platformą wbudowaną zainstalowaną na pojeździe serwisowym poruszającym się po płaszczyznach manewrowych lotniska. Proponowana architektura zakłada realizację czterech etapów: akwizycji obrazu, wstępnego przetwarzania, detekcji i klasyfikacji obiektów oraz generowania alertów dla operatorów w czasie rzeczywistym.

System stanowi odpowiedź na ograniczenia tradycyjnych metod, takich jak inspekcje piesze i radarowe, oferując rozwiązanie mobilne, skalowalne i zdolne do działania w dynamicznych warunkach eksploatacyjnych. Jego testy w środowisku rzeczywistym zrealizowano na terenie Portu Lotniczego Poznań–Ławica, gdzie przetestowano funkcjonalność zaproponowanego rozwiązania.

W celu trenowania i walidacji algorytmów wykorzystano dwa zestawy danych:

- FOD-A [7] – zestaw publiczny, obejmujący ponad 30 000 zdjęć wysokiej rozdzielczości z różnych portów lotniczych na świecie, zawierający bogaty zakres klas obiektów oraz metadanych (lokalizacja, czas, warunki atmosferyczne). Mimo swojej użyteczności, zestaw ten nie spełnia wszystkich wytycznych FAA [8], zwłaszcza w kontekście proporcji liczby zdjęć do reprezentowanych obiektów.
- PUT Dataset – autorski zbiór danych utworzony w oparciu o wymagania FAA oraz w konsultacji z ekspertami z portu Poznań–Ławica. Składa się z 1480 zdjęć zawierających 180 rzeczywistych obiektów FOD, podzielonych na cztery klasy materiałowe (metal, plastik, beton, drewno), z przewagą metali (ok. 60%) zgodnie z raportami FAA. Obrazy wykonano z odległości 2 m przy kącie 45°, w rozdzielczości 4272×2848.

2.1. Klasyczne metody przetwarzania obrazu

Początkowo zastosowano klasyczne podejście oparte na klasteryzacji k-średnich (k-means), które zakładało wektorową transformację obrazu, segmentację na podstawie różnic kolorystycznych, wykrywanie konturów oraz analizę morfologiczną. Mimo relatywnej prostoty, skuteczność tej metody okazała się niewystarczająca – najwyższa dokładność (73,94%) osiągnięta została dla obrazów o rozdzielczości 400×400, przy czym dalsze zwiększanie rozdzielczości skutkowało spadkiem efektywności, co sugeruje istnienie granicy optymalnej dla tej techniki.

Wnioski z tej analizy jednoznacznie wskazały na potrzebę wdrożenia bardziej zaawansowanych metod uczenia głębokiego, zdolnych do rozpoznawania wzorców nieliniowych w danych o dużym stopniu złożoności.

2.2. Implementacja sieci neuronowych – GoogLeNet i YO-LOv5

Sieć GoogLeNet, oparta na architekturze Inception, została przystosowana do klasyfikacji obrazów FOD w środowisku MATLAB. Model zoptymalizowano do przyjmowania wejść o wymiarach 224×224 piksele i trenowano na zbiorze FOD-A, wykorzystując techniki augmentacji. Architektura ta, dzięki zastosowaniu m.in. warstw 1×1 convolutions i global average pooling, umożliwiła skuteczne ekstrakcje cech przy jednoczesnym ograniczeniu liczby parametrów modelu.

GoogLeNet uzyskał dokładność 95,73% na zbiorze testowym FOD-A, co potwierdziło jego wysoką skuteczność w warunkach referencyjnych. Jednakże ograniczenia w szybkości działania oraz wysokie wymagania obliczeniowe sprawiły, że model nie był optymalny dla aplikacji czasu rzeczywistego w systemach wbudowanych [9].

Sieć YOLOv5, zaimplementowana z użyciem platformy Google Colab i Roboflow, okazała się bardziej odpowiednia dla zastosowań operacyjnych. Jej architektura – składająca się z warstw konwolucyjnych, CSP (Cross Stage Partial), upsamplingu oraz detektorów wieloskalowych – umożliwia jednoczesną detekcję i klasyfikację obiektów w czasie rzeczywistym [10]. Model trenowano zarówno na zbiorze FOD-A, jak i PUT.

Wyniki przedstawiono w Tabeli 1.Wysoka wartość mAP50 wskazuje na bardzo dobrą lokalizację i klasyfikację obiektów, natomiast niższe mAP50-95 potwierdza

wyzwania związane z wykrywaniem FOD o małej powierzchni lub zlewających się z tłem.

Tab. 1. Porównanie metryk dla baz FOD-A i PUT:

Baza danych	Rozdzielczość obrazu	Precyzja [%]	Recall [%]	mAP@50 [%]	mAP@50–95 [%]
FOD-A	400×400	98,6	99,0	98,9	87,4
PUT	600×600	99,3	100,0	99,5	69,5

3. WYKRYWANIE POZIOMYCH OZNACZEŃ LOTNISKOWYCH

W tym rozdziale przedstawiono metody wykrywania poziomych oznaczeń lotniskowych, kluczowych dla poprawnego pozycjonowania platform pomiarowych i zapewnienia precyzyjnej inspekcji oświetlenia nawigacyjnego. Z uwagi na istotne różnice między drogami publicznymi a infrastrukturą lotniskową (w tym skala i układ oznaczeń), opracowano i przetestowano dedykowane algorytmy, a także przeprowadzono ich walidację na systemach wbudowanych, takich jak rodzina Nvidia Jetson.

3.1. Zbiór danych PLAVS1

Zbiór danych PLAVS1 (Poznań Ławica Video Set 1) został utworzony na podstawie nagrań wideo wykonanych na terenie portu lotniczego Poznań–Ławica. Zebrano ponad 98 GB danych obejmujących drogi startowe, drogi kołowania i inne obszary operacyjne, zarejestrowane za pomocą dwóch modeli kamer GoPro. Zróżnicowanie lokalizacji i warunków rejestracji zapewnia reprezentatywność danych do testowania algorytmów wykrywania oznaczeń poziomych.

System [11] oparto na pojeździe pomiarowym z zamontowaną kamerą szerokokątną, umożliwiającą objęcie całej szerokości pasa w jednym kadrze. Kluczowym wyzwaniem było zapewnienie dokładnej detekcji linii centralnych i bocznych w warunkach zmiennego oświetlenia, zanieczyszczeń i ścierania farby. Szczególną trudność stanowi brak jednoznacznych punktów odniesienia przy szerokich pasach startowych oraz występowanie zakłóceń wynikających z cieni i odbić światła.

3.2. Opracowane algorytmy wykrywania linii

Pierwsze podejście bazowało na konwersji obrazu do odcieni szarości, zastosowaniu detekcji krawędzi (maski Scharra) oraz transformacji Hougha do detekcji linii prostych. Chociaż metoda ta osiągała wysoką skuteczność (ponad 90%) przy rozdzielczościach WXGA i wyższych, jej precyzja malała wraz ze zmniejszaniem rozdzielczości, a detekcja dalekich linii była utrudniona przez szum tła i nieregularności nawierzchni.

Drugie rozwiązanie wykorzystywało podejście analityczne oparte na macierzach progowych i intensywności światła. Efektywność tego algorytmu była najniższa spośród trzech badanych – zaledwie 40–45%, przy czym jego największą zaletą była szybkość działania przy niskiej rozdzielczości obrazu. Ze względu na niską skuteczność w detekcji linii żółtych i niestabilność przy zmiennym oświetleniu, metoda ta została uznana za niewystarczającą w warunkach rzeczywistych.

Trzecie, najbardziej zaawansowane podejście, opierało się na segmentacji obrazu w przestrzeni barw HSV. Algorytm wykrywał linie białe i żółte poprzez zastosowanie filtrów zakresu kolorystycznego (dla białego: [0, 0, 150] – [60, 40, 255]; dla żółtego: [10, 60, 140] – [30, 255, 255]), następnie przeprowadzano detekcję krawędzi (Canny) i probabilistyczną transformację Hougha (HoughLinesP). Wyniki wykazały niemal 100% skuteczność w wykrywaniu linii, nawet w przypadku ich częściowego wytarcia lub zakrycia zanieczyszczeniami.

Algorytmy przetestowano na platformach: Raspberry Pi 4B, Jetson Nano, Jetson Xavier AGX i Jetson Orin AGX. Przeprowadzono testy przy sześciu różnych rozdzielczościach (od 320×180 do 1920×1080). Najlepsze rezultaty osiągnął algorytm 3 – utrzymywał pełną detekcję wszystkich linii przy wysokiej wydajności (klatki/s), szczególnie na platformach Jetson Xavier i Orin. Analizowano również pobór mocy oraz stabilność przetwarzania w czasie rzeczywistym, co potwierdziło przydatność rozwiązania do zastosowań mobilnych w trudnych warunkach lotniskowych

4. AUTOMATYCZNA KLASYFIKACJA JAKOŚCIOWA LOTNISKOWEGO OŚWIETLENIA NAWIGACYJNEGO

Oświetlenie nawigacyjne zagłębione w nawierzchnię lotnisk (ang. in-pavement airfield ground lighting) stanowi kluczowy element infrastruktury lotniczej, mający bezpośredni wpływ na bezpieczeństwo operacji w warunkach ograniczonej widoczności. Zaprezentowano kompleksowe podejście do automatyzacji klasyfikacji jakości tych lamp, obejmujące zarówno ocenę zużycia pryzmatów, jak i analizę chromatyczności światła. Opracowany system opiera się na technikach przetwarzania obrazu, klasyfikacji z użyciem głębokich sieci neuronowych oraz analizie spektralnej w standardzie CIE 1931.

4.1. Klasyfikacja jakości pryzmatów lamp

Wizualna inspekcja pryzmatów stanowi mobilną alternatywę dla stacjonarnych metod goniometrycznych. Zaproponowany system składa się z kamery zamontowanej na platformie, wykonującej zdjęcia lamp z góry. Dzięki temu możliwa jest inspekcja bez konieczności ich demontażu, co istotnie skraca czas zajęcia drogi startowej i redukuje zakłócenia operacyjne.

W ramach badań utworzono zbiór danych PLAVS2, zawierający 1004 zdjęcia lamp w różnych stanach eksploatacyjnych, wykonanych przed i po konserwacji [12], [13]. Obrazy rejestrowano przy różnych warunkach oświetleniowych oraz w zróżnicowanym tle (asfalt, beton, oznaczenia poziome) w rozdzielczości 4928×3264 px i 24-bitowej głębi barw. Podział klas obejmował:

- Klasa I – pryzmaty w dobrym stanie,
- Klasa II – pryzmaty z uszkodzeniami,
- Klasa III – obrazy niejednoznaczne (np. z błędną segmentacją).

Wykorzystano algorytm bazujący na transformacie Hougha do lokalizacji lampy, a następnie operacje morfologiczne (maski Sobela, progowanie) do wykrycia pryzmatu. Obraz rotowano i skalowano w celu ujednolicenia

orientacji i zmniejszenia obciążenia obliczeniowego. Przeprowadzono dwie serie eksperymentów:

- Sieć dwuwarstwowa typu feed-forward trenowana z użyciem backpropagacji ze skalowanym gradientem sprzężonym. Osiągnięto dokładność 81,4% w klasyfikacji trójklasowej.
- Porównując GoogLeNet, VGG-19 i ResNet: najlepsze wyniki uzyskano dla GoogLeNet (88,37% dokładności), co potwierdziło zasadność jej zastosowania w aplikacjach operacyjnych. Dla porównania, VGG-19 osiągnęła 67,07%, a ResNet 75,29%.

Zastosowanie GoogLeNet pozwoliło na znaczącą poprawę względem wcześniejszych podejść i stanowi solidną podstawę do implementacji systemu w pojazdach inspekcyjnych.

4.1. Ocena chromatyczności światła lamp

Ocenę chromatyczności zaprojektowano jako osobny moduł pomiarowy [14], integrujący:

- identyfikację typu lampy na podstawie barwy światła,
- pomiar natężenia światła,
- konwersję współrzędnych barwy do przestrzeni CIE 1931 (x, y) [15],
- porównanie z normami EASA CS-ADR-DSN.

Zbadano pięć typów lamp:

- TDZ (Touchdown Zone),
- RCL_White / RCL_Red (Runway Centre Line),
- TAXI_YELLOW / TAXI_GREEN (Taxiway Line).

Pomiary wykonywano w odległościach od 0,5 do 3 m z wykorzystaniem czujnika TCS3430. Wyniki przeliczono do współrzędnych chromatyczności (x, y), znormalizowanych zgodnie z modelem CIE 1931. Wykresy chromatyczności ujawniły odchylenia niektórych lamp od wymaganych norm – szczególnie w przypadku starszych źródeł światła i zużytych filtrów barwnych.

Szczególną uwagę zwrócono na zjawiska nasycenia czujników w przypadku silnych źródeł (np. APP_Red), co wymaga kalibracji dystansu i odpowiedniego doboru parametrów akwizycji.

5. PODSUMOWANIE

Artykuł przedstawia podsumowanie prac poruszonych w rozprawie doktorskiej Autora. Praca odnosi się do zaleceń Europejskiej Agencji Bezpieczeństwa Lotniczego (EASA) w zakresie wdrażania sztucznej inteligencji do inspekcji bezpieczeństwa na lotniskach i norm ICAO oraz FAA. Opracowano autorskie systemy wspomagające detekcję FOD (obiektów obcych), oznaczeń poziomych oraz klasyfikację lamp nawigacyjnych. Wykorzystano własne bazy danych (PUT, PLAVS1, PLAVS2) oraz modele CNN (GoogLeNet, YOLOv5), osiągając skuteczność wykrywania do 99,3%. Segmentacja HSV umożliwiła niezależne od rozdzielczości wykrywanie linii z dokładnością 99,9%. W ocenie zużycia pryzmatów lamp GoogLeNet osiągnął 88,37% skuteczności. Pomiar chromatyczności światła oparto o sensor TCS3430 i przestrzeń barw CIE 1931, umożliwiając walidację zgodności z normami EASA. Przeprowadzono także analizę efektywności energetycznej algorytmów na platformach SBC (Jetson Xavier/Orin, Raspberry Pi).

Proponowane rozwiązania cechują się modułowością, mobilnością i możliwością równoczesnej realizacji wielu zadań inspekcyjnych. Wyniki badań potwierdzają tezę, że przetwarzanie wizyjne wspierane przez AI w systemach wbudowanych istotnie usprawnia i automatyzuje procesy inspekcji portów lotniczych.

LITERATURA

[1] J. Suder, „Vision inspection using artificial intelligence to improve safety of aerodrome", Ph. D. disseration, supervisor: Prof. Adam Dąbrowski, Ph.D., D.Sc., Poznań University of Technology, Poznań. [Online]. Dostępne na: https://bip.put.poznan.pl/attachments/8520/download

[2] European Union Aviation Safety Agency, „Certification Specifications and Guidance Material for Aerodrome Design (CS-ADR-DSN)". 29 marzec 2022. Dostęp: 30 czerwiec 2022. [Online]. Dostępne na: https://www.easa.europa.eu/en/downloads/136283/en

[3] INTERNATIONAL CIVIL AVIATION ORGANIZATION, *AERODROMES: aerodromes design and operations*. ICAO, 2018. [Online]. Dostępne na: https://www.iacm.gov.mz/app/uploads/2018/12/an_14_v1_Aerodromes_8ed._2018_rev.14_01.07.18.pdf

[4] J. Suder, P. Maciejewski, K. Podbucki, T. Marciniak, i A. Dąbrowski, „Measuring Platform for Quality Testing of Airport Lamps", *Pomiary Autom. Robot.*, t. 23, nr 2, s. 5–13, cze. 2019, doi: 10.14313/PAR_232/5.

[5] A. Dąbrowski, T. Marciniak, K. Podbucki, i J. Suder, „Method for measuring the quality of operation of airport lamps and a measuring set for testing the quality of operation of recessed airport lamps", P.442569

[6] „EASA Artificial Intelligence Roadmap 2.0". uropean Union Aviation Safety Agency, 10 maj 2023. Dostęp: 10 czerwiec 2024. [Online]. Dostępne na: https://www.easa.europa.eu/en/downloads/137919/en

[7] T. Munyer, P.-C. Huang, C. Huang, i X. Zhong, „FOD-A: A Dataset for Foreign Object Debris in Airports", 2021, doi: 10.48550/ARXIV.2110.03072.

[8] „Information Paper on French Study on Automatic FOD Detection Systems – Workshop EUROCONTROL, 9-10 June 2008".

[9] J. Suder i T. Marciniak, „Foreign Object Debris detection system using GoogLeNet", *Przegląd Elektrotechniczny*, t. 99, nr 11, s. 251–254, lis. 2023, doi: 10.15199/48.2023.11.47.

[10] J. Suder i T. Marciniak, „Foreign Object Debris detection at aerodromes using YOLOv5", w *2024 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, Poznan, Poland: IEEE, wrz. 2024, s. 66–71. doi: 10.23919/SPA61993.2024.10715612.

[11] J. Suder, K. Podbucki, T. Marciniak, i A. Dąbrowski, „Low Complexity Lane Detection Methods for Light Photometry System", *Electronics*, t. 10, nr 14, 2021, doi: 10.3390/electronics10141665.

[12] J. Suder, K. Podbucki, T. Marciniak, i A. Dąbrowski, „Intelligent vision system for quality classification of airport lamp prisms", w *2022 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, 2022, s. 151–154. doi: 10.23919/SPA53010.2022.9927908.

[13] J. Suder i T. Marciniak, „Fast Prototyping of In-Pavement Airport Navigation Lamp Prism Classification", w *2023 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, Poznan, Poland: IEEE, wrz. 2023, s. 95–99. doi: 10.23919/SPA59660.2023.10274434.

[14] J. Suder, K. Podbucki, i T. Marciniak, „Chromaticity measurement of airport navigation lighting using integrated colour sensor", *Opto-Electronics Review*, t. 31, nr 4. Polish Academy of Sciences (under the auspices of the Committee on Electronics and Telecommunication) and Association of Polish Electrical Engineers in cooperation with Military University of Technology, s. e147040, 2023. doi: 10.24425/opelre.2023.147040.

[15] „CIE 1931 2-deg xyz chromaticity coordinates". The Colour & Vision Research laboratory, University College London.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

ZASTOSOWANIE SIECI RESNET DO WYBORU PODZIAŁU JEDNOSTKI CTU W HEVC ALL INTRA

RESNET NETWORK FOR THE CTU UNIT PARTITIONING ALGORITHM IN ALL INTRA HEVC

Jakub Kwaśniak[1]; Mateusz Majtka[2]; Mateusz Lorkiewicz[3]; Tomasz Grajek[4]
[1] Politechnika Poznańska, Poznań, jakub.kwasniak@student.put.poznan.pl
[2] Politechnika Poznańska, Poznań, mateusz.majtka@student.put.poznan.pl
[3] Politechnika Poznańska, Poznań, mateusz.lorkiewicz@put.poznan.pl
[4] Politechnika Poznańska, Poznań, tomasz.grajek@put.poznan.pl

DOI: 10.15199/59.2025.4.117

Streszczenie: **Jednym z zastosowań sztucznych sieci neuronowych jest przyspieszanie istniejących koderów wizyjnych. W niniejszej pracy pokazano zastosowanie zaawansowanej architektury typu ResNet do przyspieszenia algorytmu wyboru podziału jednostek CTU w koderze HEVC pracującym w trybie All Intra. Przedstawione eksperymenty pokazują opłacalność zaproponowanych architektur przy uwzględnieniu sprawności kompresji oraz redukcji czasu kodowania.**

Abstract: **One application of artificial neural networks is to accelerate existing video encoders. This paper demonstrates the use of ResNet type architectures to accelerate the CTU unit partitioning algorithm in HEVC in All Intra mode. The experiments presented show the cost-effectiveness of the proposed architectures considering compression efficiency and coding time reduction.**

Słowa kluczowe: **HEVC, kodowanie wizji, ResNet, sztuczne sieci neuronowe.**

Keywords: **artificial neural networks, HEVC, ResNet, video encoding.**

1. WSTĘP

Sygnał wizyjny – obrazy zarówno statyczne, jak i ruchome – stanowi coraz większy udział w globalnym ruchu internetowym [17]. Do głównych źródeł tego ruchu należą telewizja, wideo na żądanie (Video on Demand), media społecznościowe oraz platformy streamingowe. Zwiększająca się liczba źródeł treści wizyjnych wiąże się ze wzrostem liczby urządzeń odpowiedzialnych za ich kompresję. Znaczną część z nich stanowią urządzenia mobilne, charakteryzujące się ograniczonymi zasobami obliczeniowymi i energetycznymi. W rezultacie rośnie zapotrzebowanie na nowoczesne, energooszczędne metody kompresji sygnału wizyjnego.

Jedną z najnowszych technik stosowanych w tym celu jest High Efficiency Video Coding (HEVC) [10], wykorzystywana m.in. w telewizji cyfrowej. Pomimo przyjęcia normy w 2013 roku, nadal rozwijane są nowe algorytmy sterujące pracą koderów, umożliwiające znaczące przyspieszenie procesu kompresji.

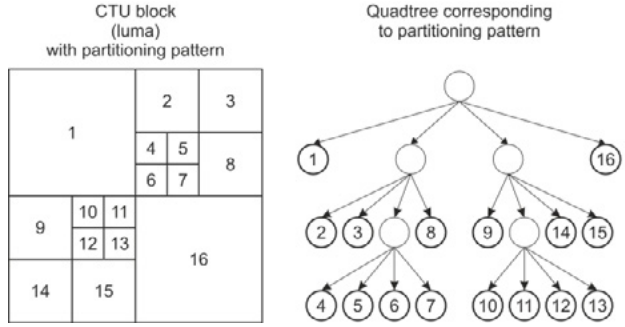
Wysoka efektywność kompresji w HEVC wynika m.in. z hierarchicznego podziału obrazu na mniejsze bloki. W pierwszym etapie obraz dzielony jest na bloki o stałej wielkości, tzw. Coding Tree Units (CTU), które następnie są dzielone na mniejsze jednostki – Coding Units (CU), poddawane właściwej kompresji. Proces ten, określany jako partycjonowanie, polega na wyszukiwaniu optymalnego podziału bloku CTU na jednostki CU. Jest to najbardziej czasochłonny etap kodowania [4].

W literaturze opisano liczne metody mające na celu przyspieszenie procesu partycjonowania, spośród których najskuteczniejsze wykorzystują sztuczne sieci neuronowe. Dotychczas stosowano głównie proste architektury, takie jak AlexNet [13] czy w pełni splotowe sieci neuronowe [7]. W ostatnich latach zaproponowano jednak bardziej zaawansowane modele, takie jak ResNet [8], które osiągają lepsze wyniki w zadaniach uczenia maszynowego.

W niniejszej pracy przedstawiono zastosowanie architektury ResNet w algorytmie partycjonowania kodera HEVC działającego w trybie All Intra.

2. PRZEGLĄD LITERATURY

Jednym z kluczowych elementów techniki HEVC [10] jest podział bloków CTU na mniejsze jednostki kodowania. Blok CTU najczęściej ma rozmiar 64×64 próbek składowej lumy. Podstawową jednostką kodowania jest Coding Unit (CU), która określa sposób kompresji próbek w obrębie bloku. Podział CTU realizowany jest zgodnie z drzewem czwórkowym (quadtree). Przykładowy schemat podziału bloku CTU na jednostki CU przedstawiono na Rysunku 1.



Rys. 1. Przykładowy podział bloku CTU na jednostki CU z odpowiadającym drzewem czwórkowym

Informacja o podziale bloku CTU prezentowana jest najczęściej na dwa sposoby. W strumieniu bitowym sposób podziału sygnalizowany jest za pomocą flag podziału. W implementacjach koderów lub dekoderów często używa się reprezentacji przy pomocy macierzy podziału - kwadratowej macierzy o rozmiarze 16×16 (dla CTU

64×64). Wartości w macierzy podziału określają głębokość jednostki CU w drzewie czwórkowym.

W literaturze zaproponowano wiele algorytmów szybkiego wyznaczania podziału CTU na jednostki CU. Celem tych algorytmów jest znaczące skrócenie czasu kodowania przy możliwie niewielkim pogorszeniu efektywności kompresji. Kompromis ten ocenia się najczęściej poprzez porównanie z koderem referencyjnym HEVC, oznaczanym akronimem HM [9].

Wśród istniejących podejść znaleźć można algorytmy, które podejmują decyzje o podziale na podstawie różnych cech obrazu, takich jak tekstura [18], cechy statystyczne [16], czy cechy heurystyczne [12]. Rozwiązania te najczęściej służą do rekursywnego wyznaczania flag podziału.

Wraz z rozwojem metod uczenia maszynowego pojawiły się algorytmy oparte na sztucznych sieciach neuronowych. Najczęściej stosowane są proste architektury splotowe, takie jak AlexNet, wykorzystywane do predykcji flag podziału [2, 6]. Zaproponowano również ich modyfikacje umożliwiające jednoczesne wyznaczanie wszystkich flag dla całego bloku CTU [5] lub bardziej zaawansowane architektury, takie jak ResNet [14].

Wraz z popularyzacją sieci w pełni splotowych, zaproponowano modele umożliwiające bezpośrednie wyznaczanie całej macierzy podziału [15]. Algorytmy tego typu, oparte na sieciach splotowych, należą obecnie do najskuteczniejszych metod szybkiego podziału bloków CTU, oferując korzystny kompromis pomiędzy czasem kodowania a efektywnością kompresji. Zazwyczaj opierają się one na prostych architekturach sieci w pełni splotowych.

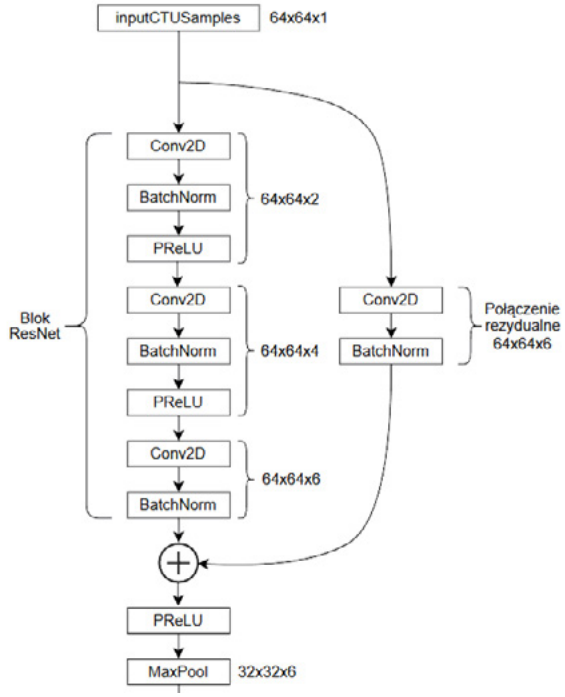
3. PROPONOWANE ROZWIĄZANIE

Niniejsza praca koncentruje się na algorytmach szybkiego podziału bloków CTU, których celem jest bezpośrednie wyznaczanie całej macierzy podziału w koderze HEVC działającym w trybie All Intra. Głównym zagadnieniem badawczym jest wykorzystanie sieci neuronowej typu ResNet w procesie predykcji struktury podziału bloku CTU.

W badaniach wykorzystano zmodyfikowaną wersję kodera HEVC, udostępnioną jako część publikacji [15]. W związku z tym architektura wejścia i wyjścia projektowanych sieci została dostosowana do wymagań zgodnych z tym oprogramowaniem.

Sieć typu ResNet [8] jest szczególnym przykładem sieci w pełni splotowej. Jej charakterystyczną cechą jest zastosowanie połączeń rezydualnych pomiędzy wejściem a wyjściem bloku warstw splotowych (kilka połączonych ze sobą warstw splotowych). Dzięki takiemu połączeniu do kolejnego bloku przekazywana jest mapa cech będąca sumą wejścia i wyjścia z danego bloku warstw splotowych. Architektura ta umożliwia skuteczniejsze i szybsze uczenie modelu, często pozwalając na osiągnięcie lepszych rezultatów niż klasyczne sieci w pełni splotowe.

Zaproponowane w niniejszej pracy sieci neuronowe oparto na architekturze ResNet, składającej się z kilku bloków rezydualnych (Rys. 2). Na wyjściu każdej sieci zastosowano funkcję aktywacji SoftMax. Szczegółowe parametry użytych architektur przedstawiono w Tabeli 1.



Rys. 2. Blok rezydualny wykorzystywany do budowy proponowanych sieci typu ResNet

Danymi wejściowymi dla sieci są próbki luminancji (luma) bloku CTU o rozmiarze 64×64. Wyjście stanowi tensor o wymiarach 4×4×4, reprezentujący prawdopodobieństwa przypisania jednej z czterech możliwych głębokości podziału do każdego z 16 podobszarów CTU. Wymiar 4×4 odpowiada zredukowanej wersji standardowej macierzy podziału (16×16), wynikającej z minimalnego rozmiaru jednostki CU (8×8). Trzeci wymiar tensora wynika z maksymalnej głębokości drzewa czwórkowego, która dla CTU 64×64 wynosi 4.

Aby wykorzystać wyjście sieci w koderze HEVC, konieczna jest konwersja tensora do macierzy o rozmiarze 16×16. Proces ten obejmuje trzy etapy:

- **Wyznaczenie głębokości podziałów**
  Zastosowanie funkcji ArgMax wzdłuż trzeciego wymiaru tensora pozwala uzyskać macierz 4×4 zawierającą estymowane głębokości podziałów.
- **Weryfikacja poprawności podziału**
  Ponieważ tylko część macierzy 4×4 odpowiada dopuszczalnym strukturom kodera HEVC, stosowana jest heurystyczna procedura korekcyjna. Jeśli liczba komórek o głębokości 0 przekracza 8, cała macierz ustawiana jest na 0. Następnie analizowane są mniejsze regiony: jeśli suma wartości w danym obszarze przekracza określony próg, jego głębokość ustalana jest na 1; w przeciwnym razie – na 2 (z pominięciem wartości 3 jako niezgodnych).
- **Odwzorowanie do formatu 16×16**
  Skorygowana macierz 4×4 jest skalowana przestrzennie do formatu 16×16 przez zwielokrotnienie jej wartości w odpowiednich regionach zgodnie ze składnią kodera HEVC.

Tab. 1. Testowane konfiguracje sieci ResNet

ResNet	Opis modyfikacji sieci.	Liczba bloków	Liczba warstw w bloku
Mod 0	Bazowa architektura sieci.	7	8
Mod 1	Zastąpienie funkcji MaxPool2D funkcją AvgPool2D.	7	8
Mod 2	Podwójny bloku ResNet przed warstwą aktywacji PReLu oraz warstwą MaxPool.	7	8
Mod 3	Zmniejszenie liczby warstw w bloku rezydualnym do pięciu.	7	5
Mod 4	Zwiększenie dwukrotnie liczby warstw w każdym bloku.	7	14
Mod 5	Zwiększenie o cztery warstwy w każdym bloku.	7	11
Mod 6	Potrójny blok ResNet przed warstwą aktywacji PReLu oraz warstwą MaxPool.	7	8
Mod 7	Wykorzystanie ReLU jako funkcji aktywacji.	7	8
Mod 8	Zastosowanie ośmiu bloków rezydualnych.	8	8
Mod 9	Zastosowanie sześciu zamiast rezydualnych.	6	8
Mod 10	Szybsze zwiększanie liczby wyjściowych map cech w kolejnych blokach rezydualnych.	7	8

4. METODOLOGIA BADAŃ

Do trenowania zaproponowanych architektur wykorzystano zbiór DIV2k [1]. Obrazy zostały zakodowane za pomocą referencyjnego kodera HM [9], a następnie podzielone na bloki CTU z przypisanymi podziałami wyznaczonymi przez HM. Zbiór obejmuje 589 589 przykładów treningowych oraz 66 650 przykładów walidacyjnych. Dla każdej wartości parametru kwantyzacji $QP = \{22, 27, 32, 37\}$, zgodnie z Common Test Conditions (CTC) dla HEVC [11], przygotowano osobny podzbiór danych uczących.

Sieci trenowano w środowisku TensorFlow, wykorzystując entropię krzyżową jako funkcję straty oraz optymalizator ADAM. Model uczono przez 50 epok z resetowaniem optymalizatora co 5 epok, korzystając z akceleracji GPU.

Ewaluację przeprowadzono za pomocą zmodyfikowanego kodera HEVC [15] po uprzedniej konwersji modeli do formatu kompatybilnego z LibTorch. Obliczenia podczas testów wykonywano wyłącznie na CPU. Do oceny wykorzystano zbiór danych testowych JCT-VC opisany w CTC, stosując standardową procedurę ewaluacji HEVC [11].

Wydajność sieci oceniono na podstawie: liczby operacji dodawania i mnożenia (w milionach [M]), liczby parametrów (w tysiącach [k]), zmiany strumienia bitowego względem kodera referencyjnego (delta Bjøntegaard'a [3]) oraz redukcji czasu kodowania względem kodera referencyjnego (w procentach).

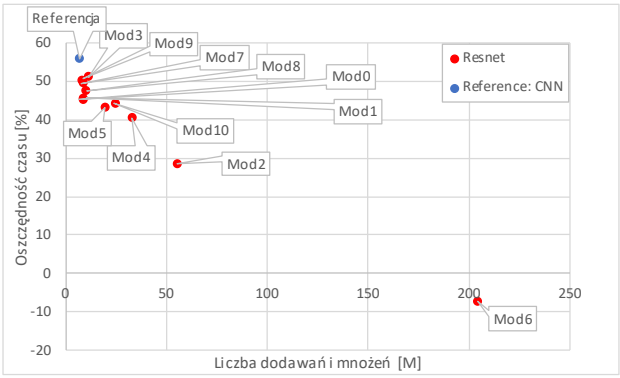
5. WYNIKI EKSPERYMENTALNE

W Tabeli 2 przedstawiono wyniki eksperymentalne dla zaproponowanych architektur sieci typu ResNet. Dla porównania przedstawiono wyniki i parametry dla jednej z najlepszych sieci w pełni splotowej dostępnych w literaturze [15]. Na Rysunkach 3-5 przedstawiono wyniki w formie graficznej, zestawiając ze sobą wybrane rezultaty.

Na podstawie Rysunku 3 obserwuje się, że zaproponowane sieci typu ResNet, mimo porównywalnej liczby dodawań i mnożeń (np. Mod9, Mod3, Mod7), wykazują mniejszą redukcję czasu kodowania niż rozwiązanie referencyjne. Wyjątkiem jest Mod6, którego złożoność powoduje wydłużenie czasu kodowania względem kodera referencyjnego. Podobne wnioski płyną z analizy oszczędności czasu kodowania (Rysunek 4). Przyczyną może być większa liczba warstw w sieciach ResNet, co skutkuje większą serializacją obliczeń i wolniejszym przetwarzaniem, mimo zbliżonej liczby operacji i parametrów.

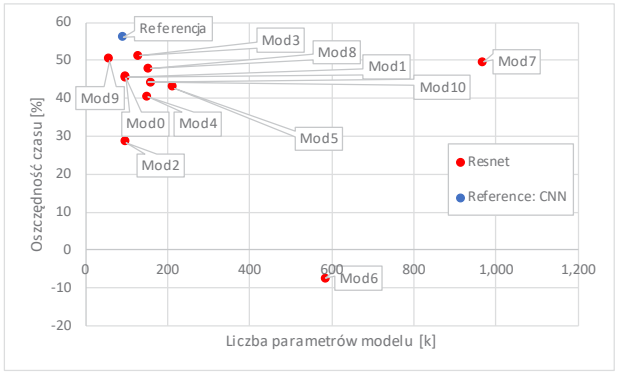
Tab. 2. Wyniki eksperymentalne

ResNet	Liczba dodawań i mnożeń [M]	Liczba parametrów modelu [k]	BD-Rate [%]	Oszczędność czasu kodowania [%]
Mod0	8,74	97,202	1,88	45,41
Mod1	8,74	97,202	1,81	45,74
Mod2	55,00	97,148	1,70	28,54
Mod3	11,13	128,188	1,87	51,27
Mod4	32,66	150,992	1,77	40,47
Mod5	19,44	212,878	1,81	43,22
Mod6	204,08	583,968	1,71	-7,46
Mod7	8,74	96,852	1,91	49,59
Mod8	9,62	152,960	1,88	47,67
Mod9	8,10	57,086	1,80	50,36
Mod10	24,49	160,262	1,77	44,28
[15]	6,76	91,600	1,87	56,08

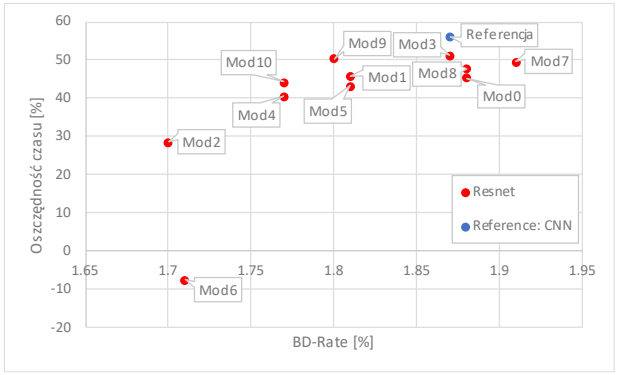


Rys. 3. Oszczędność czasu kodowania w zależności od liczby dodawań i mnożeń dla proponowanych sieci

Jednakże sieci ResNet przewyższają rozwiązanie referencyjne pod względem zmniejszenia wielkości strumienia bitowego (Rysunek 5). Dla większości modeli strumień jest mniejszy o około 0,1 p.p. względem sieci CNN z [15]. Model Mod9 osiąga zbliżoną oszczędność czasu kodowania przy wyraźnie lepszej kompresji w porównaniu do sieci referencyjnej, co potwierdza wyższą efektywność ResNet w zadaniach klasyfikacji.



Rys. 4. Wykres oszczędności czasu kodowania w zależności od liczby parametrów dla proponowanych sieci



Rys. 5. Wykres oszczędności czasu kodowania w zależności od zmiany wielkości strumienia bitowego dla proponowanych sieci.

6. WNIOSKI

Niniejsza praca przedstawia wykorzystanie sztucznych sieci neuronowych typu ResNet jako algorytmu wyznaczania podziału dla bloku CTU w koderze HEVC w trybie All Intra. W pracy przedstawiono wyniki ewaluacji dla kilku zaproponowanych wariantów sieci typu ResNet.

W porównaniu do referencyjnego rozwiązania z zastosowaniem prostej, w pełni splotowej architektury, zaproponowane sieci typu ResNet okazały się mniej wydajne pod względem oferowanej redukcji czasu kodowania. Takie wyniki otrzymano pomimo zastosowania sieci o porównywalnej liczbie parametrów i ilości obliczeń.

W artykule pokazano, ze pomimo słabszej redukcji czasu kodowania, zastosowanie sieci typu ResNet pozwala na osiągnięcie wyraźnie lepszych wyników kodowania pod względem wielkości strumienia bitowego, w porównaniu do referencyjnej sieci w pełni splotowej. Wyniki te wskazują na dalszą konieczność prowadzenia badań nad wykorzystaniem bardziej zaawansowanych architektur sieci neuronowych w opisywanym zastosowaniu.

LITERATURA

[1] Agustsson E., Timofte R.. 2017. „NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study". *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*. 1122-1131.

[2] Amna M., Imen W., Ezahra S. F.. 2021. „LeNet5-Based approach for fast intra coding". *10th Int. Symp. on Signal, Image, Video and Communications*. 1-4.

[3] Bjøntegaard G.. 2001. „Calculation of average PSNR differences between RD curves", *ITU-T SG16/Q6, dokument. VCEG-M33*.

[4] Bossen F., Bross B., Suhring K., Flynn D., 2012. „HEVC Complexity and Implementation Analysis". *IEEE Transactions on Circuits and Systems for Video Technology*. 22 (12) : 1685-1696.

[5] Chen Z., Shi J., Li W.. 2020 „Learned Fast HEVC Intra Coding". *IEEE Transactions on Image Processing*. 29 :. 5431-5446.

[6] Feng Z., Liu P., Jia K., Duan K.. 2018. „Fast Intra CTU Depth Decision for HEVC". *IEEE Access*. 6 : 45262-45269.

[7] Goodfellow I., Bengio Y., Courville Aaron. 2016. „Deep Learning". *MIT Press*.

[8] He K., Zhang X., Ren S., Sun J.. 2016. „Deep Residual Learning for Image Recognition". *IEEE Conf. on Computer Vision and Pattern Recognition*. 770-778.

[9] HHI. „HEVC reference software". https://hevc.hhi.fraunhofer.de.

[10] ISO/IEC. 2013. „Information technology — High efficiency coding and media delivery in heterogeneous environments — Part 2: High Efficiency Video Coding". *ISO/IEC IS 23008-2, także jako ITU-T Rec. H.265*.

[11] Joint Collaborative Team Video Coding (JCT-VC). 2013. „Common Test Conditions and Software Reference Configurations", *dokument JCTVC-L1100*.

[12] Kim N., Jeon S., Shim H. J., Jeon B., Lim S.-Ch., Ko H.. 2016. „Adaptive keypoint-based CU depth decision for HEVC intra coding*". IEEE Int. Symp. Broadband Multimedia Syst. Broadcast*. 1–3.

[13] Krizhevsky A., Sutskever I., Hinton G. E. 2012. „ImageNet classification with deep convolutional neural networks", *25th Int. Conf. on Neural Information Processing Systems*, 1097–1105.

[14] Li Y., Li L., Zhuang Z., Fang Y., Yang Y.. 2021. „ResNet Approach for Coding Unit Fast Splitting Decision of HEVC Intra Coding". *IEEE Sixth Int. Conf. on Data Science in Cyberspace*. 130-135.

[15] Lorkiewicz M., Stankiewicz O., Domański M., Hang H.-M., Peng W.-H.. 2024. „Complexity-Efficiency Control With ANN-Based CTU Partitioning for Video Encoding". *IEEE Access*. 12 : 102536-102551.

[16] Lu Y., Huang X., Liu H., Zhou Y., Yin H., Shen L.. 2020. „Hierarchical classification for complexity reduction in HEVC inter coding". *IEEE Access*. 8 : 41690–41704.

[17] Sandvine. 2023. „Sandvine's 2023 Global Internet Phenomena Report Shows 24% Jump in Video Traffic, with Netflix Volume Overtaking YouTube".

[18] Zhang Q., Wang Y., Huang L., Jiang B.. 2020. „Fast CU partition and intra mode decision method for H.266/VVC". *IEEE Access*. 8 : 117539–117550.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

DETEKCJA POCIĄGÓW PRZY UŻYCIU SYGNAŁÓW OKAZJONALNYCH SIECI LTE

TRAINS DETECTION USING LTE SIGNALS OF OPPORTUNITY

Olga Błaszkiewicz; Krzysztof K. Cwalina; Alicja Olejniczak; Piotr Rajchowski; Jarosław Sadowski

Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki
email: {olga.blaszkiewicz, krzysztof.cwalina, alicja.olejniczak, piotr.rajchowski, jarsadow1}@pg.edu.pl

DOI: 10.15199/59.2025.4.118

Streszczenie: **Referat przedstawia wyniki eksperymentu polegającego na powiązaniu zmian w rejestrowanych symbolach referencyjnych w sygnałach odbieranych ze stacji bazowej LTE przez odbiornik umieszczony w pobliżu czynnej linii kolejowej z przejazdem pociągów. W odróżnieniu od znanych rozwiązań bazujących na koncepcji radaru pasywnego, w zaprezentowanym stanowisku do detekcji obecności pociągów użyto pojedynczego odbiornika z anteną dookólną.**

Abstract: **The paper presents the results of an experiment based on linking changes in reference symbols in signals received from LTE base station by a receiver placed near an active railway line with the passage of trains. Unlike known solutions based on the concept of passive radar, in the presented experiment a single receiver with an omnidirectional antenna was used to detect the presence of trains.**

Słowa kluczowe: **detekcja obiektów, detekcja pociągów, LTE, symbole referencyjne.**

Keywords: **cell-specific reference signals, LTE, object detection, train detection.**

1. WSTĘP

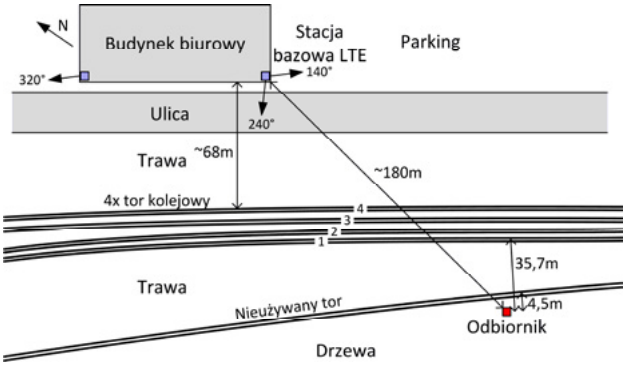
Zmiany w geometrii i właściwościach środowiska propagacji fal elektromagnetycznych w łączach radiowych mogą w istotny sposób wpływać na parametry odbieranych sygnałów. Zmienność warunków odbioru sygnałów radiowych spowodowana obecnością budynków, budowli, pojazdów czy ludzi w środowisku jest w większości przypadków traktowana jako zjawisko niekorzystne, które należy ograniczać lub kompensować przy użyciu odpowiednich technik odbioru i przetwarzania sygnałów. Ale z drugiej strony, zmienność parametrów odbieranych sygnałów radiowych pozwala w pewnym zakresie na pozyskiwanie informacji o środowisku propagacji fal radiowych, w tym także przy wykorzystaniu sygnałów radiowych już występujących w danym środowisku, czyli na zasadzie sygnałów okazjonalnych. W literaturze można spotkać przykłady eksperymentów polegających na detekcji obecności obiektów poruszających się, np. samochodów czy pociągów, na podstawie analizy różnic pomiędzy sygnałami odbieranymi z istniejących nadajników radiowych w sposób bezpośredni oraz sygnałów odbitych od monitorowanych obiektów, czyli na zasadzie tzw. radaru pasywnego. Przykładowo w [4] wykorzystano odbiór sygnałów łącza w dół ze stacji bazowej w systemie LTE (ang. Long Term Evolution) nie tylko do wykrywania obecności pojazdów ale także do ich klasyfikacji, chociaż warunki realizacji pomiarów, zaprezentowane w tej publikacji, były bardzo rygorystyczne. Podobną zasadę pomiaru, bazującą na koncepcji radaru pasywnego, zastosowano w [2] do detekcji obecności pojazdów na drodze przy wykorzystaniu sygnału telewizji cyfrowej DVB-T (ang. Digital Video Broadcasting-Terrestrial). W obu przytoczonych przykładach jako sygnały przypadkowe użyte były sygnały wieloczęstotliwościowe OFDM (ang. Orthogonal Frequency Division Multiplexing), których szerokie pasmo oraz możliwość obserwacji zjawisk selektywnych częstotliwościowo sprzyjają ich użyciu do sondowania stanu środowiska. Ale koncepcja radaru pasywnego wymaga stosowania przynajmniej dwóch torów odbiorczych i anten kierunkowych do rozróżnienia sygnałów bezpośrednich i sygnałów odbitych od monitorowanych obiektów. Zamiast stosowania podwójnego toru odbiorczego, w niniejszym artykule opisano możliwość pozyskania informacji o obecności pociągów na torach na podstawie analizy zmiany zależności fazowych pomiędzy symbolami referencyjnymi w sygnale łącza w dół ze stacji bazowej LTE, obserwowanych przy użyciu odbiornika z pojedynczym torem odbiorczym i dookólną anteną, zainstalowanego w pobliżu linii kolejowej.

2. REJESTRACJA SYGNAŁÓW

Do realizacji pomiarów wybrano torowisko linii kolejowej przebiegającej na granicy między gdańskimi dzielnicami Aniołki i Młyniska. W pobliżu tego torowiska, na dachu budynku biurowego, znajduje się stacja bazowa telefonii komórkowej pracująca m.in. w standardzie LTE (FDD) na częstotliwości 1844,9 MHz z szerokością kanału 10 MHz. Stanowisko odbiorcze zostało umieszczone po drugiej stronie torowiska w miejscu wskazanym na schematycznej mapie na rys. 1. Najbliższy tor, znajdujący się zaledwie 4,5 m od miejsca rejestracji sygnałów LTE, nie był wykorzystywany do ruchu kolejowego w trakcie kampanii pomiarowej. Dwa kolejne tory, oznaczone na rys. 1 jako ‘1’ i ‘2’ służą do ruchu regionalnego i dalekobieżnego, podczas gdy tory ‘3’ i ‘4’ obsługują ruch metropolitalny.

Do rejestracji sygnałów użyto urządzeń radia programowalnego USRP B200 mini z anteną dookólną Cobham OA2-0.3-10V zainstalowaną na niskim maszcie około 0,5 m nad poziomem torowiska. Wzajemne roz-

mieszczenie anten stacji bazowej i odbiornika pozwalało na zachowanie warunków propagacji LOS (ang. Line of Sight) ponad dachami przejeżdżających pociągów, więc ich obecność nie blokowała całkowicie propagacji sygnałów radiowych. Sygnały łącza w dół ze stacji bazowej LTE były próbkowane z częstotliwością 15,36 MHz i zapisywane na dysku do późniejszego przetwarzania.



Rys. 1. Położenie stanowiska odbiorczego względem torów i stacji bazowej LTE.

Podczas 35 minut rejestracji zarejestrowano przejazd dwunastu pociągów po wszystkich czterech torach w obu kierunkach. Spis pociągów, czasów przejazdu i używanych torów zawiera tabela 1. Warunki rejestracji sygnałów, w tym stosowanie anteny dookólnej oraz odbiornika taktowanego typowym oscylatorem kwarcowym, o deklarowanej względnej stabilności częstotliwości rzędu 10^{-7}, odpowiadają warunkom odbioru sygnałów, możliwym do osiągnięcia w typowych terminalach użytkowników (telefony, modemy) sieci komórkowych.

Tab. 1. Zarejestrowane przejazdy pociągów.

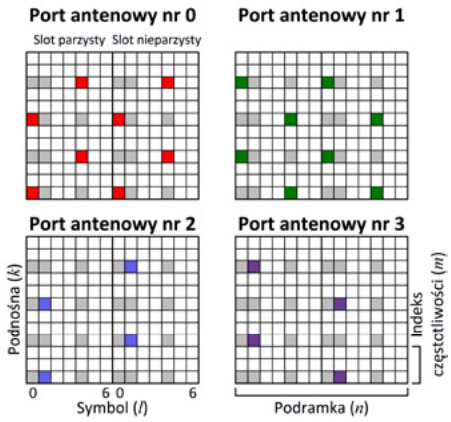
Lp.	Czas obserwacji [s]	Numer toru i typ pociągu
1	300	4 – metropolitalny
2	405	3 – metropolitalny
3	508	2 – regionalny
4	758	1 – regionalny
5	822	2 – regionalny
6	919	4 – metropolitalny
7	1001	3 – metropolitalny
8	1493	4 – metropolitalny
9	1631	3 – metropolitalny
10	1664	1 – regionalny
11	1898	2 – dalekobieżny
12	2000	1 – dalekobieżny

3. PRZETWARZANIE SYGNAŁÓW

Zastosowane stanowisko odbiorcze z jedną anteną, w dodatku o charakterystyce dookólnej, nie pozwala na skuteczną detekcję obecności/przejazdu pociągów metodą pasywnego radaru bistatycznego, z uwagi na wysoką trudność w separacji zmiennego echa spowodowanego ruchem pociągów od stałego echa budynków i budowli w środowisku propagacji sygnału. Z kolei bezpośrednia obserwacja bardzo małych zmian fazy poszczególnych podnośnych sygnału OFDM, spowodowanych zmianami w środowisku propagacyjnym, jest praktycznie niemożliwa na tle znacznie większych wahań fazy będących efektem ograniczonej stabilności oscylatora lokalnego przemiany częstotliwości w zastosowanym odbiorniku i generatora taktującego próbkowanie sygnałów. Zamiast tego zastosowano porównywanie parametrów sygnałów odbieranych z różnych portów antenowych stacji bazowej LTE.

Nadajnik LTE, którego sygnały były rejestrowane w pobliżu torowiska, pracował w konfiguracji czteroportowej i podłączony był do anteny dwupolaryzacyjnej (+45°/–45°), prawdopodobnie wyposażonej w cztery niezależne zestawy promienników. Pojęcie portu antenowego w standardzie LTE nie jest w pełni tożsame z fizycznymi złączami antenowymi urządzeń nadawczo-odbiorczych, jest to koncepcja warstwy przetwarzania sygnałów, z którą skojarzone są unikatowe symbole referencyjne, umożliwiające estymację parametrów odbieranych sygnałów. Porty antenowe są mapowane do portów fizycznych urządzeń radiowych, ale nie zawsze jest to przypisanie „jeden do jednego". Jednakże w przypadku portów antenowych o numerach od 0 do 3 najczęściej spotyka się ich mapowanie bezpośrednio do oddzielnych elementów antenowych przy dywersyfikacji przestrzennej lub polaryzacyjnej albo oddzielnych wiązek przy stosowaniu szyków anten adaptacyjnych, a standard definiuje unikatowe sekwencje symboli referencyjnych oznaczanych jako CRS (Cell-specific Reference Signal) nadawanych oddzielnie przez każdy z tych portów antenowych w sposób umożliwiający ich niezależny odbiór nawet przy stosowaniu pojedynczego toru odbiorczego w terminalu użytkownika.

Postać symboli referencyjnych CRS i ich rozmieszczenie w siatce zasobów czasowo-częstotliwościowych w sygnale LTE jest jednoznacznie określona wartością identyfikatora komórki (N_{ID}^{cell}) [1]. Wartości zespolonej sekwencji pseudoprzypadkowej $r_{l,n_s}(m)$ są zdefiniowane w sekcji 6.10.1.1 specyfikacji [1], gdzie l to numer symbolu a n_s to numer slotu. Następnie wykonywane jest mapowanie sekwencji r na docelowe symbole CRS a_{k,l,n_s}^{p}, gdzie k to numer podnośnej OFDM natomiast p to numer portu antenowego. Dla analizowanego sygnału LTE o szerokości pasma 10 MHz i normalnego (krótkiego) przedrostka cyklicznego k przyjmuje wartości od 0 do 599, l może mieć tylko wartości 0, 1 lub 4, n_s zmienia się od 0 do 19 a numer portu p zawiera się w zakresie od 0 do 3. Przykładowe rozmieszczenie symboli CRS w siatce zasobów w sygnale LTE przedstawiono symbolicznie na rys. 2.



Rys. 2. Przykładowe rozmieszczenie symboli CRS w siatce zasobów LTE dla emisji sygnałów przy użyciu czterech portów antenowych.

Znając nominalne wartości symboli referencyjnych a^p_{k,l,n_s} oraz ich estymaty z zarejestrowanych sygnałów $\hat{a}^p_{k,l,n_s}$ obliczane jest przybliżenie transmitancji kanału radiowego jako:

$$\hat{h}^p_{k,l,n_s} = \frac{\hat{a}^p_{k,l,n_s}}{a^p_{k,l,n_s}} \quad (1)$$

Dla portów antenowych nr 0 i 1 nadawane są dwa symbole referencyjne na danej podnośnej k w czasie podramki, po jednym w slocie o numerze parzystym i nieparzystym, podczas gdy dla pozostałych dwóch portów występuje tylko jeden symbol CRS na podramkę. Ponieważ dokładna wartość częstotliwości oraz numeru symbolu, w którym występują symbole referencyjne nie jest istotna w dalszej części analizy, dla uproszczenia zapisu używany będzie tylko numer podramki n:

$$n = \left\lfloor \frac{n_s}{2} \right\rfloor \quad (2)$$

obejmujący 14 kolejnych symboli oraz indeks częstotliwości:

$$m = \left\lfloor \frac{k}{3} \right\rfloor \quad (3)$$

obejmujący trzy podnośne. Dla portów antenowych nr 0 i 1 dla każdej kombinacji indeksów (m, n) występują dwa symbole referencyjne, więc stosowane jest uśrednianie ich wartości:

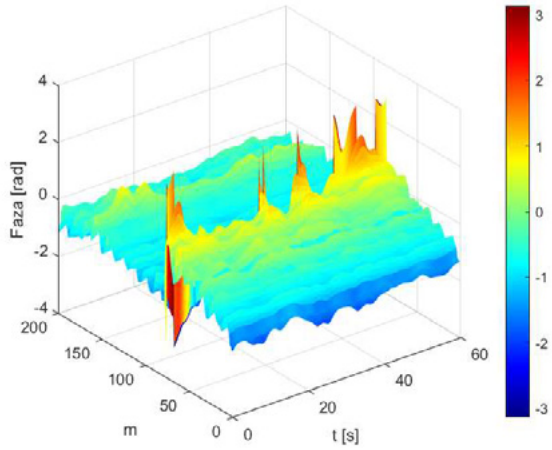
$$\hat{h}^p_m(n) = \frac{\hat{h}^p_{k,l,2n} + \hat{h}^p_{k,l,(2n+1)}}{2} \quad (4)$$

Dla portów nr 2 i 3 dla dowolnej kombinacji (m, n) występuje tylko jedna estymata funkcji przejścia kanału (1) więc uśrednianie $\hat{h}^p_m(n)$ według (4) nie jest wymagane.

Następnym krokiem przetwarzania sygnałów było wykrycie różnic pomiędzy funkcjami przejścia kanałów dla różnych par portów antenowych (p_1, p_2) w ramach tego samego indeksu częstotliwości m, ale ponieważ należało oczekiwać, że największe różnice będą widoczne przy analizie faz estymat funkcji przejścia (4), zaproponowano analizę wartości fazy wyrażenia:

$$u^{p_1,p_2}_m(n) = \hat{h}^{p_2}_m(n) \cdot \left(\hat{h}^{p_1}_m(n)\right)^* \quad (5)$$

Niestety chociaż w ramach danego nadajnika LTE stosowana jest zasada, że w momentach emisji symboli CRS przez jeden port antenowy emisje sygnałów ze wszystkich pozostałych portów są wstrzymane, położenie symboli CRS w siatce zasobów zależy m.in. od identyfikatora N^{cell}_{ID}. Oznacza to, że nie tylko inne stacje bazowe w tej samej sieci jednoczęstotliwościowej ale nawet nadajniki sąsiednich sektorów w tej samej stacji bazowej używają tych elementów siatki zasobów czasowo-częstotliwościowych do transmisji innych danych użytecznych lub sterujących, generując wzajemnie interferencje wewnątrzsystemowe. Ponieważ jednak obliczanie (1) dla sygnału z właściwej stacji bazowej powinno zwracać wyniki o dość dużej korelacji czasowej, a więc z dominacją składowych niskoczęstotliwościowych, podczas gdy sygnały interferujące z innych sektorów i stacji bazowych dają wyniki o niskim stopniu korelacji, zastosowana została filtracja dolnoprzepustowa (5) w dziedzinie czasu przy użyciu filtru FIR o bardzo niskiej częstotliwości granicznej wynoszącej zaledwie 0,5 Hz, przy czym filtrację realizowano oddzielnie dla każdego indeksu częstotliwości m. W efekcie uzyskano dwuwymiarową funkcję czasu i częstotliwości dla każdej z sześciu możliwych kombinacji numerów portów antenowych (p_1, p_2). Przykładowy wykres wartości faz wyrażenia (5) po filtracji dolnoprzepustowej, uzyskany dla pary portów antenowych nr 1 i 2, przedstawiono na rys. 3. Wykres został ograniczony do okresu 60 sekund dla zwiększenia czytelności.



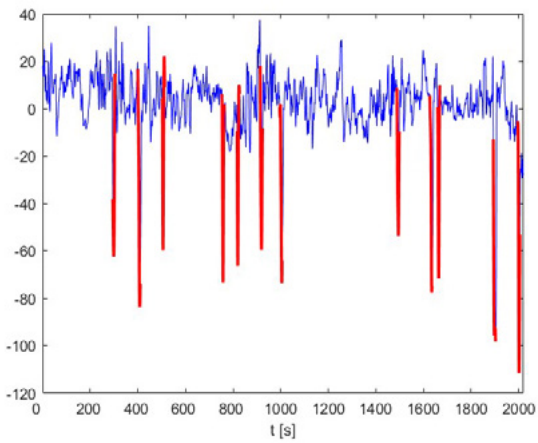
Rys. 3. Przykładowy wykres fazy różnic symboli CRS po filtracji dolnoprzepustowej w sygnale odebranym ze stacji bazowej LTE dla pary portów antenowych nr 1 i 2.

4. DETEKCJA POCIĄGÓW

Relatywnie szerokie pasmo rejestrowanych sygnałów (około 9 MHz dla nominalnego pasma LTE 10 MHz) oraz geometria środowiska propagacyjnego sprawiają, że w analizowanym eksperymencie kanał radiowy powinien być selektywny częstotliwościowo. Stąd obecność (przejazd) pociągu na torach może wpływać tylko na część podnośnych LTE, ale ogólnie duża liczba stosowanych podnośnych (symbole CRS występują na 200 podnośnych z 600 dostępnych w paśmie 10 MHz) może sugerować, że wpływ obecności pociągów może być widoczny na więcej niż jednej podnośnej i powinna istnieć możliwość wskazania grupy podnośnych LTE o widocznych skorelowanych efektach spowodowanych przejazdem pociągu. Weryfikację korelacji danych wykonano przy użyciu analizy głównych składowych (PCA, ang. Principal Component Analysis). PCA to analiza statystyczna zbioru obserwacji bazująca na transformacji wielowymiarowej przestrzeni danych, zdefiniowanej danymi z kolejnych obserwacji, tak by następowało maksymalizowanie wariancji wyników względem poszczególnych współrzędnych przestrzeni po transformacji wymiarów [3]. W ten sposób zdefiniowane współrzędne nowej przestrzeni danych definiują wektory bazowe objaśniające najbardziej istotną statystycznie zmienność danych pomiarowych.

Danymi wejściowymi do analizy PCA były unormowane wartości fazy wyrażenia (5) po filtracji dolnoprzepustowej, dla każdej z sześciu możliwych kombinacji portów antenowych, co daje łącznie 1200 wektorów danych będących funkcją czasu. Po obliczeniu sum ważonych tych wektorów, z wagami zdefiniowanymi postaciami wektorów własnych skojarzonymi z największymi wartościami własnymi macierzy korelacji wektorów wejściowych okazało się, że wyniki odpowiadające kilku składnikom PCA wykazują istotną korelację z obecnością pociągów na torach. Wykres przedstawiony na rys. 4 zawiera liniową kombinację ważoną unormowanych różnic faz symboli referencyjnych na różnych podnośnych, gdzie wartości wag do kombinacji liniowej wynikają z drugiego i trzeciego wektora własnego wyniku analizy PCA. Kolorem czerwonym na tym wykresie zaznaczono przebieg w przedziałach czasu od ośmiu sekund przed zapisanym nominalnym czasem przejazdu pociągu po torach w bezpośredniej bliskości stanowiska odbiorczego do ośmiu sekund po tym czasie. Korelacja występowania lokalnych minimów przebiegu z obecnością pociągów na torach jest bardzo wyraźnie widoczna, niemal identyczne wyraźne minima występują w momentach przejazdu pociągów po każdym z czterech czynnych torów i w obu kierunkach. Przy tak dużej różnicy pomiędzy przebiegiem wykresu na rys. 4 w momentach przejazdu pociągów, względem sytuacji pustego torowiska, detekcję obecności pociągów można zrealizować stosunkowo prostym detektorem progowym.

Druga wartość własna wyników PCA stanowiła 13,4% a trzecia 8,45% sumy wszystkich wartości własnych, co oznacza, że wykres z rys. 4 tłumaczy blisko 22% zmienności różnic faz symboli referencyjnych dla wszystkich możliwych par portów antenowych w stacji bazowej i wszystkich indeksów częstotliwości (1200 wektorów danych będących funkcją czasu). Zatem wpływ pociągów na parametry łącza radiowego, jakimi w analizowanym przypadku była różnica faz sygnałów nadawanych przez wieloportowy nadajnik stacji bazowej telefonii komórkowej LTE, można uznać za dość znaczny.



Rys. 4. Liniowa kombinacja ważonych różnic faz symboli referencyjnych CRS z wagami zdefiniowanymi drugim wektorem własnym wyniku analizy PCA (bez jednostki). Kolor czerwony – przedziały czasu obecności pociągów na torach.

Analiza PCA nie może być wprost użyta do detekcji obecności pociągów na torach, gdyż po pierwsze dla analizowanego przypadku wykresy podobne do rys. 4 uzyskano dla kilku zestawów wag z kilku wektorów własnych wyniku analizy PCA, co wskazuje, że wykres z rys. 4 nie wyjaśnia całej zmienności wartości różnic faz symboli CRS wynikającej z obecności pociągów. Po drugie zaś analiza PCA nie wskazuje które wektory własne tłumaczą jakie zależności danych wejściowych, co oznacza, że nawet powtarzając pomiary w podobnych warunkach wybór wyników z analizy PCA optymalnych do detekcji obecności pociągów może być odmienny. Zagadnienie optymalnego sumowania ważonego wyników filtracji wyrażenia (5) do celów detekcji obecności pociągów na torach zasługuje więc na dalszą analizę.

5. PODSUMOWANIE

Zależność parametrów odbieranych sygnałów od geometrii środowiska propagacji fal radiowych, która w radiokomunikacji często utrudnia zachowanie wysokiej jakości odbioru, pozwoliła na uzyskanie powiązania obecności (przejazdu) pociągów po torach w pobliżu miejsca rejestracji sygnałów z mierzalnymi parametrami tych sygnałów. Taka detekcja istotnych zmian w łączu radiowym, chociaż nie spełnia wymogów stawianych jakimkolwiek systemom detekcji zajętości linii kolejowych np. do celów zarządzania ruchem, może być użyta np. do „śledzenia" pociągu przez wiązki anten adaptacyjnych czy modyfikacji parametrów pracy algorytmów przydziału zasobów w stacjach bazowych.

Wybór miejsca rejestracji sygnałów stacji bazowej LTE był dość przypadkowy i wynikał z dostępności miejsca do wygodnej instalacji stanowiska odbiorczego. Na podstawie uzyskanych wyników można domniemywać, że wyraźniejszą zależność zmienności parametrów odbieranych sygnałów LTE powinno się obserwować w warunkach propagacji LOS, a weryfikacja wpływu miejsca i warunków rejestracji sygnałów na detekcję pociągów może być jednym z kierunków dalszych prac.

LITERATURA

[1] 3GPP. 2024. *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation*, TS 36.211, V18.0.1

[2] Capria Amerigo, Dario Petri, Marco Martorella, Michele Conti, Enzo Mese, Fabrizio Berizzi. 2010. *DVB-T passive radar for vehicles detection in urban environment*, 2010 IEEE International Geoscience and Remote Sensing Symposium, Honolulu, HI, USA, 2010, str. 3917-3920, doi: 10.1109/IGARSS.2010.5649675.

[3] Sanguansat Parinya (ed). 2012. *Principal Component Analysis*. InTech. doi:10.5772/2340.

[4] Sardar Santu, Amit K. Mishra, Mohammed. Z. A. Khan. 2019. *Vehicle detection and classification using LTE-CommSense*, IET Radar Sonar Navig., vol. 13, str. 850-857, 2019, doi: 10.1049/iet-rsn.2018.5231.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

NIEBINARNE KODOWANIE LDPC TYPU RAPTOR O KRÓTKIEJ DŁUGOŚCI BLOKU

SHORT BLOCKLENGTH NONBINARY RAPTOR-LIKE LDPC CODING

Wojciech Sułek[1]; Jakub Hyla[2]; Marcin Kucharczyk[3];

[1] Politechnika Śląska, Gliwice, wsulek@polsl.pl
[2] TKH Technology, Leszno, jakubhyla93@gmail.com
[3] Politechnika Śląska, Gliwice, mkucharczyk@polsl.pl

DOI: 10.15199/59.2025.4.119

Streszczenie: Przedstawiono skuteczny, adaptacyjny schemat korekcji błędów transmisji oparty o niebinarne kodowanie LDPC typu Raptor (RL), o strukturze quasi-cyklicznej (QC). Dane zakodowane kodem o wysokiej sprawności (HR) są uzupełniane o symbole redundancji inkrementalnej (IR), generowane przez niebinarne równania parzystości w ciele GF(2^q). Zaprezentowano koncepcję procedury konstrukcji systemu NB QC-RL-LDPC, metodologię eksperymentów numerycznych oraz wyniki pokazujące skuteczność rozwiązania dla krótkich bloków.

Abstract: An efficient, adaptive error correction scheme is presented, based on nonbinary Raptor-like (RL) LDPC coding with a quasi-cyclic (QC) structure. Data encoded with a high-rate (HR) code is supplemented with incremental redundancy (IR) symbols, generated by nonbinary parity-check equations over the Galois Field GF(2^q). The concept of the NB QC-RL-LDPC system construction procedure is presented, along with the methodology of numerical experiments and results demonstrating the effectiveness of the solution for short block lengths.

Słowa kluczowe: korekcja błędów, kodowanie korekcyjne, kody LDPC, kody niebinarne.

Keywords: Error Correction Coding, LDPC, NB-LDPC.

1. WSTĘP

Kody LDPC (Low-Density Parity-Check) [4] to nowoczesne kody korekcyjne, które charakteryzują się wysoką wydajnością i są szeroko stosowane w systemach komunikacyjnych, takich jak Wi-Fi, DVB-S2 czy 5G. Ich struktura pozwala na skuteczną detekcję i korekcję błędów przy stosunkowo niskim narzucie obliczeniowym, szczególnie w przypadku kodów o dużej długości bloku.

Rozszerzenie kodów LDPC na niebinarne ciała Galois (GF(2^q)) pozwala uzyskać lepsze właściwości korekcji błędów, szczególnie dla krótkich bloków danych, co czyni je atrakcyjnymi dla transmisji o niskim opóźnieniu. Pomimo większej złożoności implementacyjnej, nowoczesne architektury sprzętowe umożliwiają efektywną realizację niebinarnych kodów LDPC. Szczególnie interesującą klasą są kody quasi-cykliczne, które umożliwiają efektywne kodowanie i dekodowanie dzięki uporządkowanej strukturze macierzy kontrolnej.

Prace badawcze prowadzone w zespole Politechniki Śląskiej [6] dotyczą między innymi niebinarnego kodowania LDPC typu Raptor-Like (RL), o strukturze QC, łączącego zalety kodowania o wysokiej przepustowości z możliwością adaptacyjnego dodawania redundancji (Incremental Redundancy – IR). Dane są najpierw przesyłane z użyciem kodu HR, a w przypadku problemów z dekodowaniem – uzupełniane o dodatkowe symbole korekcyjne. Zamiast projektowania kodów na podstawie analizy asymptotycznej, autorzy zaproponowali podejście oparte na bezpośredniej generacji kodów o małym rozmiarze bloku. Metoda oparta jest na tworzeniu grafu o pożądanych własnościach i symulacjach Monte Carlo. Krótkie kody NB-RL-LDPC są znakomicie dostosowane do transmisji z niskimi opóźnieniami, co jest zgodne z kierunkiem rozwoju kodowania w 5G i 6G dla uzyskania własności URLLC (*Ultra Reliable Low Latency Communications*) [7].

W niniejszym tekście przedstawiono metodologię prac badawczych nad systemami transmisji adaptacyjnej wykorzystującej niebinarne kodowanie RL oraz ogólny opis metody konstrukcji kodów QC NB-RL-LDPC, z zastosowaniem grafowej optymalizacji kodu bazowego oraz symulacji Monte Carlo. Zaproponowany system transmisyjny z redundancją inkrementalną zakłada możliwość zmiany modulacji podczas retransmisji. Eksperymenty pokazują, że taka kombinacja kodowania niebinarnego i adaptacyjnej modulacji pozwala na poprawę wydajności transmisji, zwłaszcza w środowisku o niskim stosunku sygnału do szumu (SNR), przewyższając efektywność rozwiązań binarnych stosowanych m.in. w 5G.

2. PODSTAWY

Niebinarne kody LDPC (NB-LDPC) to klasa blokowych liniowych kodów korekcyjnych, definiowanych nad ciałem Galois GF(2^q) przy $q > 1$. Proces kodowania dla kodu (N, K) o sprawności $R = K/N$ polega na dodaniu $M = N - K$ elementów nadmiarowych do wektora informacyjnego $u = [u_1, u_2, \ldots, u_K]^T$, co tworzy wektor kodowy $c = [c_1, c_2, \ldots, c_N]^T$, którego elementy należą do ciała GF(2^q).

Kod NB-LDPC jest opisany przez rzadką macierz

kontroli parzystości $\mathbf{H}_{M\times N}$ z elementami w ciele GF(2^q). Wektor $\hat{\mathbf{c}}$ jest uznawany za poprawne słowo kodowe, jeśli spełnia równanie $\mathbf{H}\hat{\mathbf{c}} = \mathbf{0}$ w arytmetyce GF(2^q). Każdy wiersz $\mathbf{H}$ reprezentuje jedno równanie kontrolne, które łącznie są wykorzystywane w iteracyjnym algorytmie dekodowania korekcyjnego [2, 4].

W rozważanej podklasie QC kodów NB-LDPC, macierz kontrolna $\mathbf{H}$ przedstawiona w (1), składa się z kwadratowych podmacierzy cyklicznych permutacji (CPM) $\mathbf{P}^s$, otrzymanych przez cykliczne przesunięcie w prawo wierszy macierzy jednostkowej $\mathbf{I}_{P\times P}$ o ($s \mod P$) pozycji,

$$\mathbf{H} = \begin{bmatrix} z_{1,1}\mathbf{P}^{s_{1,1}} & z_{1,2}\mathbf{P}^{s_{1,2}} & \cdots & z_{1,J}\mathbf{P}^{s_{1,J}} \\ z_{2,1}\mathbf{P}^{s_{2,1}} & z_{2,2}\mathbf{P}^{s_{2,2}} & \cdots & z_{2,J}\mathbf{P}^{s_{2,J}} \\ \vdots & \vdots & \ddots & \vdots \\ z_{I,1}\mathbf{P}^{s_{I,1}} & z_{I,2}\mathbf{P}^{s_{I,2}} & \cdots & z_{I,J}\mathbf{P}^{s_{I,J}} \end{bmatrix}, \quad (1)$$

gdzie $s_{i,j} \in 0, 1, \ldots, P-1$ określa wartości przesunięć cyklicznych (CSV), a $z_{i,j}$ wskazuje pozycje niezerowych podmacierzy w macierzy $\mathbf{H}$ oraz określa współczynniki w ciele Galois GF(2^q) w bloku równań kontrolnych. Wszystkie współczynniki $z_{i,j}$ można zebrać razem, tworząc tzw. macierz bazową kodu $\mathbf{Z}$ o rozmiarze $I \times J$.

Każda macierz kontrolna może być reprezentowana przez graf Tannera. Zawiera on węzły symbolowe odpowiadające symbolom kodowym (lub kolumnom macierzy $\mathbf{H}$), węzły kontrolne odpowiadające równaniom parzystości (lub wierszom $\mathbf{H}$) oraz krawędzie wskazujące pozycje niezerowych elementów w $\mathbf{H}$. Podobnie można zdefiniować graf bazowy odpowiadający macierzy bazowej $\mathbf{Z}$. Skuteczność korekcyjna kodu o krótkiej długości bloku zależy bezpośrednio od kluczowych właściwości strukturalnych jego grafu. W szczególności obecność krótkich cykli w grafie kodu pogarsza skuteczność odpowiadającego mu kodu LDPC.

Konkatenacja binarnego kodu LDPC o wysokiej sprawności HR z kodem o rzadkiej macierzy generującej do redundancji inkrementalnej IR, jest powszechnie znana jako kodowanie typu Raptor-Like LDPC (RL-LDPC) [1]. W niniejszej pracy badane jest uogólnienie tego podejścia na przypadek kodowania niebinarnego (NB).

Niebinarne kody QC-RL-LDPC, będące połączeniem niebinarnego kodu HR i niebinarnego kodu IR, można zdefiniować przy użyciu połączonej macierzy kontrolnej w następującej postaci:

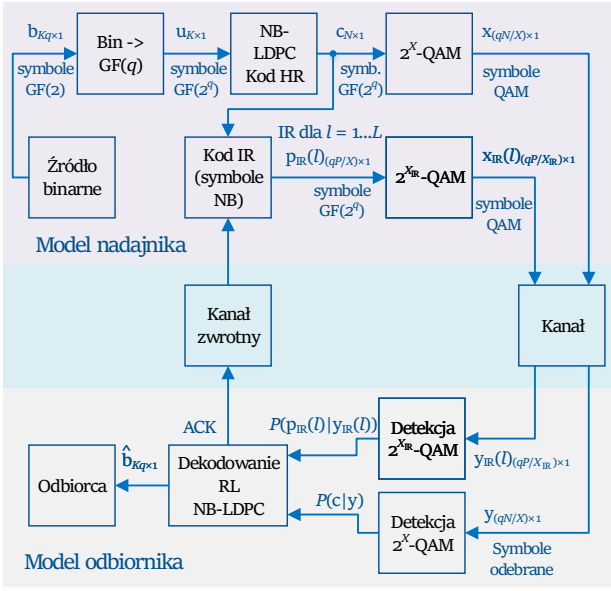
$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{\mathrm{HR}} & \mathbf{0} \\ \mathbf{H}_{\mathrm{IR}} & \mathbf{I} \end{bmatrix} \quad (2)$$

- $\mathbf{I}$ to macierz jednostkowa, a $\mathbf{0}$ macierz zerowa,
- $\mathbf{H}_{\mathrm{HR}}$ to macierz kontrolna QC NB-LDPC nad GF(2^q) o wymiarach $M \times N$, definiująca kod HR, zbudowana z podmacierzy $P\times P$, jak w (1), o strukturze dwu-diagonalnej, umożliwiającej efektywne kodowanie,
- $\mathbf{H}_{\mathrm{IR}}$ to niebinarna macierz kontrolna kodu IR nad GF(2^q) o wymiarach $LP\times N$, również podzielona na podmacierze $P\times P$, analogicznie do (1), lecz złożona z L makrowierszy. Daje to łącznie LP równań kontrolnych IR. Kod IR generuje dodatkowe LP symboli parzystości, które można obliczyć za pomocą macierzy generującej: $\mathbf{G} = [\mathbf{I}, \mathbf{H}_{\mathrm{IR}}^T]$ nad GF(2^q), co oznacza, że jest to kod LDGM. Symboloe kontrolne IR są kodowane w L porcjach oznaczonych jako $\mathbf{p}_{\mathrm{IR}}(l)$, gdzie $l = 1, \ldots, L$, poprzez podział $\mathbf{H}_{\mathrm{IR}}$ na L makrowierszy, po P wierszy każdy.

3. METODOLOGIA

Model analizowanego systemu komunikacyjnego z niebinarnym kodowaniem i redundancją inkrementalną, wykorzystujący transmisję QAM, przedstawiono na rys. 1. Po stronie nadajnika zakodowany wektor HR $\mathbf{c}$ o rozmiarze $N \times 1$ nad GF(2^q) jest mapowany na wektor $\mathbf{x}$ zawierający symbole 2^X-QAM, które są transmitowane przez kanał z zakłóceniami.

Jeśli przeprowadzone w odbiorniku dekodowanie początkowo wysłanego bloku kodowego HR nie zakończy się sukcesem po maksymalnej liczbie iteracji, nadajnik generuje i wysyła kolejne porcje symboli IR $\mathbf{p}_{\mathrm{IR}}(l)$, $l = 1, \ldots, L$, aż do momentu otrzymania potwierdzenia poprawnego dekodowania (ACK). Modulacja bloków IR wykorzystuje niezależny modulator, $2^{X_{\mathrm{IR}}}$-QAM. Każdy wektor IR, $\mathbf{p}_{\mathrm{IR}}(l)$ ma w podstawowej wersji długość P i pochodzi z jednej makrowarstwy macierzy $\mathbf{H}_{\mathrm{IR}}$, ale możliwe jest także grupowanie makrowarstw w celu zwiększenia bloków IR. Kanał zwrotny kontroluje retransmisje poprzez sygnalizację ACK/NACK.



Rysunek 1: Model systemu transmisyjnego z kodowaniem NB RL-LDPC.

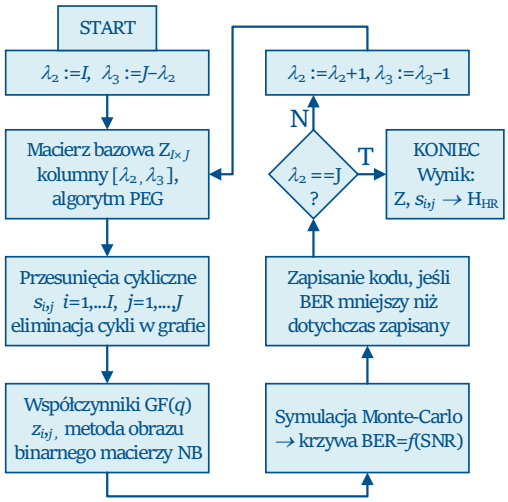
Po stronie odbiornika, zmodulowane symbole x_i, $i = 1, \ldots, qN/X$ są zniekształcane przez zakłócenia kanałowe, przy czym eksperymenty numeryczne prowadzono dla modelu kanału AWGN. Odebrane wartości $\mathbf{y} = [y_1, \ldots, y_{qN/X}]^T$ są wykorzystywane do obliczenia prawdopodobieństw $P(c_n = a|\mathbf{y})$, dla $n = 1, \ldots, N$ oraz dla każdego $a \in$ GF(2^q), które inicjalizują proces dekodowania. Algorytm dekodowania wykorzystuje metodę BP (*Belief Propagtion*) w wesji FFT-BP [2], uzupełnioną o rozszerzanie danych wejściowych oraz struktury grafu o kolejne fragmenty skojarzone z kolejnymi transmisjami IR. Każda iteracja dekodowania składa się z trzech etapów: aktualizacji wiadomości z węzłów kontrolnych do symbolowych $\{Q_{cn\to sn}\}$, następnie aktualizacji wiadomości z węzłów symbolowych do kontrolnych $\{Q_{sn\to cn}\}$, a na końcu próbnego dekodowania. Proces ten jest kończony, jak tylko w wyniku dekodowania uzyskany zostanie wektor kodowy $\hat{\mathbf{c}}$ spełniający wszystkie równania kontrolne, z którego ekstrahowany jest wektor informacyjny $\hat{\mathbf{b}}$.

Propagowane wiadomości $Q^a_{cn_m \to sn_n}$ określają przekonania, że m-te równanie parzystości jest spełnione, jeśli n-ty element wektora kodowego (c_n) ma wartość $a \in \mathrm{GF}(2^q)$; natomiast wiadomości $Q^a_{sn_n \to cn_m}$ reprezentują przekonania, że element c_n wektora kodowego przyjmuje wartość a, przy uwzględnieniu informacji pochodzących od pozostałych węzłów oraz początkowych prawdopodobieństw $P(\mathbf{c}|\mathbf{y})$. Przekonania mogą reprezentować szacunki prawdopodobieństw, logarytmów prawdopodobieństw lub logarytmów ilorazów prawdopodobieństw (LLR) i mogą być obliczane za pomocą jednego z dobrze znanych algorytmów, takich jak BP, FFT-BP, Min-Max lub Min-Sum. Szczegóły dotyczące algorytmów propagacji przekonań można znaleźć np. w [2, 6].

Jeśli limit iteracji dla dekodowania kodu HR zostanie osiągnięty bez uzyskania poprawnego wyniku dekodowania wstępnego, proces jest zatrzymywany i wznawiany po odebraniu nowej porcji danych IR. Dla każdego kolejnego odebranego fragmentu, zbiory wiadomości $\{Q_{cn\to sn}\}$ i $\{Q_{sn\to cn}\}$ są rozszerzane o dodatkowe elementy odpowiadające części macierzy $\mathbf{H}_{\mathrm{IR}}$, które są odpowiednio inicjalizowane, po czym wykonywane są kolejne iteracje dekodowania.

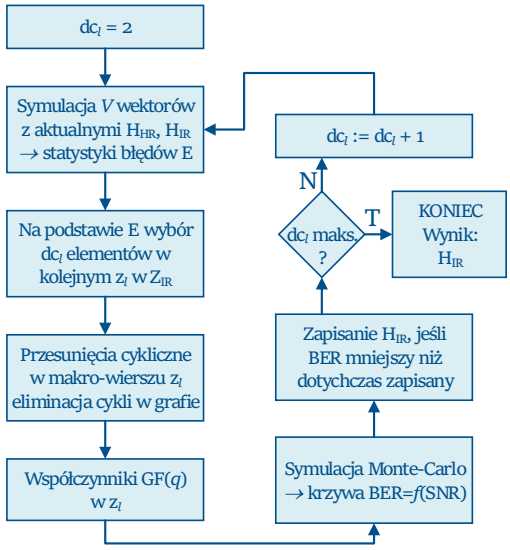
4. KONSTRUKCJA KODÓW NB RL-LDPC

Opracowany sposób projektowania kodów NB RL-LDPC oparty jest na algorytmicznym poszukiwaniu rozwiązania, osobno dla części HR i IR kodu RL-LDPC o strukturze QC. Oba składniki są projektowane niezależnie, z uwzględnieniem maksymalizacji wydajności korekcji błędów na każdym etapie redundancji inkrementalnej. Proces projektowania wykorzystuje odpowiednią konstrukcję grafu Tannera. Cechy grafu, takie jak liczba krótkich cykli, mają bezpośredni wpływ na zdolności korekcyjne kodu oraz skuteczność dekodowania. Dodatkowo projektowanie zarówno części HR, jak i IR, jest wspierane eksperymentami symulacyjnymi. Taki proces pozwala na weryfikację przyjętych modeli, w tym algorytmu dekodowania, co zapewnia, że ostateczny system działa zgodnie z oczekiwaniami fazy projektu.



Rysunek 2: Procedura konstrukcji kodu HR

Pierwszym etapem projektowania kodu RL-LDPC jest konstrukcja macierzy $\mathbf{H}_{\mathrm{HR}}$ – ogólna koncepcja jest przedstawiona na rys. 2. Optymalne rozkłady wag kolumn macierzy kontrolnej, oznaczone $[\lambda_2, \lambda_3]$ są wynikiem działania algorytmu, uzyskanym jako produkt uboczny optymalizacji kodu HR – wykorzystano tylko kolumny rzędu 2 i 3, co jest skutecznym rozwiązaniem w przypadku kodów niebinarnych (quasi-regularnych). Procedura rozpoczyna się od $\lambda_2 = I$, co odpowiada liczbie kolumn o wadze 2 wymaganych do uzyskania formy dualno-diagonalnej macierzy $\mathbf{H}_{\mathrm{HR}}$. Następnie λ_2 jest stopniowo zwiększane (a λ_3 zmniejszane), aż do maksymalnej wartości $\lambda_2 = J$. Każda iteracja pętli zawiera przygotowanie macierzy bazowej $\mathbf{Z}$, optymalizację przesunięć cyklicznych w kodzie QC (1), wyznaczenie współczynników w $\mathrm{GF}(2^q)$ oraz symulacje Monte-Carlo. Finalnie wybierany jest kod o najmniejszej stopie błędów dla założonego SNR w kanale AWGN.



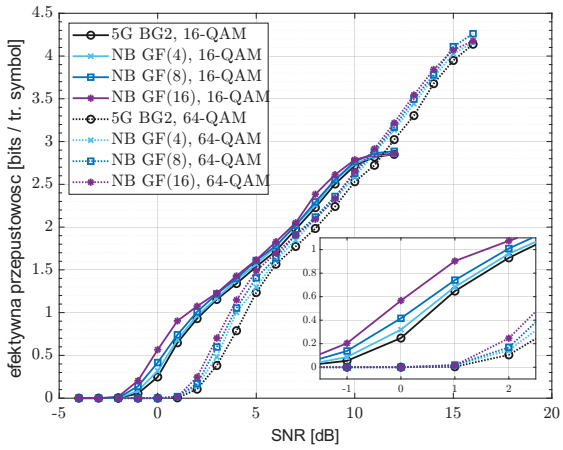
Rysunek 3: Procedura dodawania l-tego makrowiersza do $\mathbf{H}_{\mathrm{IR}}$.

W drugim etapie, projektowaniu kodu IR, każda iteracja projektowa, przedstawiona na rys. 3, polega na dodaniu makrowiersza do macierzy $\mathbf{H}_{\mathrm{IR}}$, określanego w symulacjach. Krokowe konstruowanie $\mathbf{H}_{\mathrm{IR}}$, wspierane symulacjami, uwzględnia porównania wydajności w sposób odpowiadający rzeczywistej pracy systemu, gdzie kolejne porcje IR są transmitowane w takiej kolejności, w jakiej projektowano odpowiadające im makrowiersze macierzy $\mathbf{H}_{\mathrm{IR}}$.

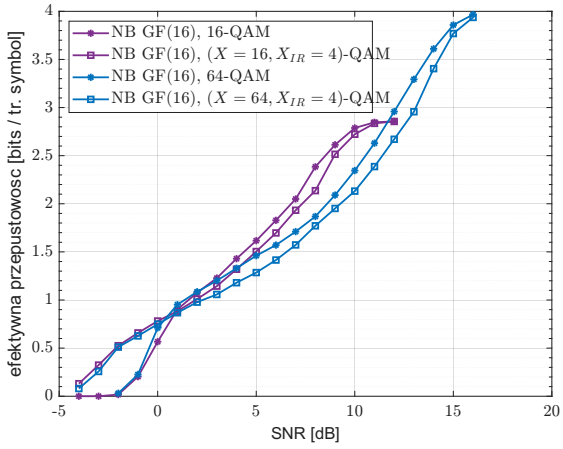
Każdy obieg pętli algorytmu generuje jeden makrowiersz w $\mathbf{H}_{\mathrm{IR}}$, wybierając symbole do równania parzystości w kolejnej transmisji bitów nadmiarowych. Kluczowym etapem jest wybór indeksów elementów niezerowych w $l$tym makrowierszu $\mathbf{z}_l$. W tym celu wykorzystano macierz błędów $\mathbf{E}_{V\times N}$, która zawiera informacje o błędnych symbolach w próbce V symulowanych transmisji, dekodowanych z użyciem $\mathbf{H}_{\mathrm{HR}}$ oraz dotychczas zaprojektowanej części $\mathbf{H}_{\mathrm{IR}}$. Większa liczba jedynek w n-tej kolumnie $\mathbf{E}$ wskazuje na większe prawdopodobieństwo, że symbol c_n zostanie błędnie zdekodowany na tym etapie. Metoda selekcji indeksów symboli polega zatem na wyborze skojarzonych z kolumnami $\mathbf{E}$ o największej wadze.

W opracowanych procedurach, współczynniki nad ciałem $\mathrm{GF}(2^q)$ są wybierane z uprzednio obliczonego zbioru zestawów współczynników, zgodnie z metodą zaproponowaną w [5].

5. WYNIKI EKSPERYMENTÓW NUMERYCZNYCH



Rysunek 4: Wyniki: transmisja IR w kanale AWGN, z wykorzystaniem kodów NB RL-LDPC oraz kodów 5G BG2, dla długości bloku informacyjnego $K = 80$ bitów i zróżnicowanych rzędów GF i modulacji QAM.



Rysunek 5: Wyniki: transmisja IR z odmiennymi rzędami modulacji QAM w części HR oraz IR, X, X_{IR}

Eksperymenty dotyczyły kodów o różnych rzędach ciała Galois oraz ich porównania z systemem IR wykorzystującym binarne kody LDPC standardu 5G. Symulacje przeprowadzono z modelem kanału AWGN przy różnych poziomach SNR, a wyniki przedstawiono w formie miary wydajności powszechnie stosowanej w badaniach nad schematami IR – efektywnej przepustowości (*transmission goodput*) [3]. Miara ta jest definiowana jako średni stosunek liczby poprawnie dostarczonych bitów informacyjnych do całkowitej liczby przesłanych symboli QAM, niezbędnych do osiągnięcia bezbłędnego dekodowania.

Na rys. 4 przedstawiono wyniki dla systemu wykorzystującego kody NB QC-RL-LDPC nad różnymi ciałami: GF(16), GF(8) oraz GF(4), zaprojektowanymi za pomocą przedstawionych algorytmów. Zaprezentowano wyniki dla konfiguracji transmisji z modulacją 16-QAM oraz 64-QAM. Wszystkie kody NB nieco przewyższają pod względem wydajności system IR oparty na binarnych kodach 5G, przy czym najlepsze rezultaty uzyskano dla kodu działającego nad ciałem o najwyższym rzędzie, GF(16), co jest zgodne z oczekiwaniami.

Potrzeba retransmisji występuje zwykle, gdy warunki propagacji sygnału w kanale są niekorzystne. Wówczas, zmiana na modulację niższego rzędu, np. QPSK, podczas retransmisji, może okazać się korzystna. Zweryfikowano to eksperymentalnie, a przykładowe wyniki przedstawiono na rysunku 5. Wyniki dotyczą schematu transmisji typu Raptor, który wykorzystuje mieszane rzędy modulacji QAM – początkową transmisję z modulacją ($X = 64$)-QAM lub ($X = 16$)-QAM oraz retransmisje z użyciem ($X_{\mathrm{IR}} = 4$)-QAM (QPSK).

Zysk wydajności wynikający z tego podejścia jest wyraźnie widoczny w zakresie niskiego stosunku sygnału do szumu (SNR). Wraz ze spadkiem SNR dochodzi się do wartości, dla której zwiększona efektywność widmowa modulacji wyższych rzędów przestaje kompensować rosnący poziom szumu. Bardziej efektywna staje się modulacja QPSK, co widoczne jest jako punkt przecięcia krzywych na wykresie.

6. PODSUMOWANIE

W niniejszej pracy przedstawiono ramy numerycznej konstrukcji oraz eksperymentalnej weryfikacji wydajności niebinarnego rozszerzenia schematu kodowania RL-LDPC z przyrostową redundancją. Zaprezentowano model systemu transmisji, w którym kodowanie NB jest połączone z modulacją QAM o dowolnym rzędzie, a także opisano iteracyjny algorytm dekodowania. Opracowano również algorytm konstrukcji kodów NB QC-RL-LDPC. Przeprowadzone eksperymenty wykazały, że niebinarne kodowanie typu Raptor oferuje lepsze osiągi w porównaniu do binarnych kodów z systemu 5G. Dodatkowo, zróżnicowanie rzędów modulacji – poprzez zastosowanie modulacji wyższego rzędu dla transmisji początkowej oraz niższego rzędu dla retransmisji IR – może przynieść poprawę możliwości transmisji w warunkach niskiego SNR.

LITERATURA

[1] Chen T. Y., Vakilinia K., Divsalar D., Wesel R. D. 2015. „Protograph-Based Raptor-Like LDPC Codes". IEEE Trans. Commun., 63 (5): 1522-1532.

[2] Declercq D., Fossorier M. 2007. „Decoding Algorithms for Nonbinary LDPC Codes Over GF(q)".IEEE Trans. Commun., 55 (4): 633–643.

[3] Jayasooriya S., Shirvanimoghaddam M., Ong L., Johnson S. J. 2018. „Raptor Codes for Higher-Order Modulation Using a Multi-Edge Framework". IEEE Wireless Commun. Lett., 7 (3): 110-113.

[4] MacKay D. J. C. 1999. „Good Error-Correcting Codes Based on Very Sparse Matrices". IEEE Trans. Inf. Theory, 45 (3): 399-431.

[5] Poulliat C., Fossorier M., Declercq D.. 2008. „Design of Regular (2,dc)-LDPC Codes over GF(q) Using Their Binary Images". IEEE Trans. Commun., 56 (10): 1626-1635.

[6] Hyla Jakub, Sułek Wojciech. 2025. „Short Blocklength Nonbinary Raptor-Like LDPC Coding Systems Design and Simulation". IEEE Access, 13: 3849-3863.

[7] Wang F., Jiao J., Zhang K., Wu S., Zhang Q.. 2022. „Self-Adaptive Ordered Statistics Decoder for Finite Block Length Raptor Codes Toward URLLC". IEEE Internet Things Journal, 9 (5): 3282-3297.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

PROJEKT I REALIZACJA SYSTEMU ROZPROSZONEGO Z WYKORZYSTANIEM TECHNOLOGII LoRa

DESIGN AND EVALUATION OF DISTRIBUTED LoRa SYSTEM

Jakub Browarski; Daniel Niepla; Robert Niepla, Jerzy Demkowicz[1]

[1] Politechnika Gdańska, Gdańsk, demjot@eti.pg.edu.pl

DOI: 10.15199/59.2025.4.120

Streszczenie: **W artykule przedstawiono kompletny projekt, implementację i testy rozproszonego, energooszczędnego systemu monitorowania środowiska opartego na bezprzewodowej komunikacji LoRa. Celem było umożliwienie pozyskiwania danych z odległych czujników zdalnych przy minimalnym zużyciu energii oraz przy użyciu zasilanych energią słoneczną. Projekt ma charakter rozwojowy, pełna rozproszoność wymaga jeszcze implementacji protokołów.**

Abstract: **The article presents the complete design, implementation and testing of a distributed, energy-saving environmental monitoring system based on wireless LoRa communication. The goal was to enable data collection from remote sensors with minimal power consumption and the use of solar energy.**

Słowa kluczowe: **Lora, IoT, system rozproszony, komunikacja szerokopasmowa.**

Keywords: **Lora, IoT, distributed system, broadband communication.**

1. WSTĘP

Bezprzewodowe sieci czujników mają szerokie zastosowanie w wielu dziedzinach związanych z ochroną środowiska. Wykorzystuje się je m.in. do monitorowania siedlisk przyrodniczych, śledzenia migracji zwierząt, wczesnego wykrywania pożarów lasów, wspierania rozwoju rolnictwa precyzyjnego, jak również w zarządzaniu sytuacjami kryzysowymi związanymi z klęskami żywiołowymi [1]. Dzięki nim możliwe jest uzyskanie danych w czasie rzeczywistym, co pozwala na szybszą i dokładniejszą reakcję na zmiany w środowisku. Modulacja chirp, wykorzystywana w LoRa (ang. *Long Range*) to technika modulacji oparta na transmisji szerokopasmowej (ang. *Spread Spectrum*), która zapewnia wysoką odporność na zakłócenia oraz umożliwia transmisję danych na dużą odległość przy minimalnym zużyciu energii. LoRa wykorzystuje modulację Chirp Spread Spectrum (CSS). W tej metodzie częstotliwość sygnału jest zmieniana w czasie, co pozwala na uzyskanie odporności na interferencje oraz umożliwia duży zasięg transmisji, w porównaniu do tradycyjnych metod modulacji. CSS pozwala na uzyskanie znacznie lepszego zasięgu i odporności na szumy przy niższym poborze mocy. Odbiornik LoRa może dekodować transmisje 19,5 dB poniżej poziomu szumów [5]. Modulacja chirp umożliwia dostosowanie parametrów transmisji, takich jak szerokość pasma, czas trwania chirpów oraz moc sygnału nadajnika, co pozwala na optymalizację zarówno zasięgu, jak i szybkości transmisji. Istnieją cztery główne parametry konfiguracyjne, które wpływają na wydajność modulacji chirp [4]:

- Carrier Frequency (CF), częstotliwość nośna: Określa częstotliwość środkową, wokół której będzie odbywało się przesunięcie sygnału,
- Spreading Factor (SF), współczynnik rozproszenia: Określa szybkość zmiany częstotliwości, który wpływa na szybkość transmisji i zasięg. Im wyższy spreading factor (np. SF12), tym większy zasięg, ale mniejsza szybkość transmisji. Standardowe wartości SF to od 6 do 12, (w projekcie 12),
- Bandwidth (BW), pasmo częstotliwości: Określa szerokość pasma transmisji, która wpływa na szybkość transmisji i odporność na zakłócenia. Węższe pasmo zmniejsza szybkość transmisji, (62.5kHz),
- Coding Rate (CR), współczynnik kodowania: Określa stopień korekcji błędów. Wyższy współczynnik korekcji (np. CR4/5) zwiększa odporność na błędy transmisji, ale zmniejsza efektywną szybkość transmisji.

Zasięg transmisji w technologii LoRa może wynosić nawet do 30 km w terenie otwartym (nad wodą) i kilka kilometrów w miastach [8], co czyni ją bardzo dobrym rozwiązaniem do zastosowań, gdzie duży zasięg i niskie zużycie energii jest kluczowe. Alternatywne rozwiązania, takie jak systemy satelitarne, mimo (teoretycznie) pełnego pokrycia powierzchni Ziemi, wiążą się z wysokimi kosztami eksploatacji i wymaganiami związanymi z instalacją skomplikowanych anten. Systemy oparte na GSM wymagają kosztownej infrastruktury telekomunikacyjnej, podczas gdy sieci Wi-Fi oferują bardzo ograniczony zasięg. Analizę porównawczą dostępnych standardów przedstawiono w pracy Emerging Wireless Technologies In The Internet Of Things: A Comparative Study [2]. W

ramach analizy zostały porównane takie technologie jak Wi-Fi, Bluetooth, Zigbee, 6LOWPAN, WIMAX, Z Wave, 4G i LoRa. Wybrane cechy tych standardów przedstawiono w tabeli 1.

Tab. 1. Porównanie standardów komunikacyjnych [2]

Technologia	*Szybkość*	*Przybliżony zasięg*	*Zużycie energii*
WiFi	do kilku Gb/s	do 100 m	Niskie do Wysokie
Bluetooth	2,1 Mbps	<100m	Niskie
ZigBee	250 kb/s	10 - 20m	Bardzo niskie
6LOWPAN	-200 kb/s	10 - 20m	Bardzo niskie e
WiMAX	do 100 Mbps	50km	Wysokie
Zwave	40 kb/s	30m	Bardzo niskie
LTE	Do 1 Gb/s	kilka km	Wysokie
LoRa	0,3 do 100 kb/s	3 - 5km	Bardzo niskie

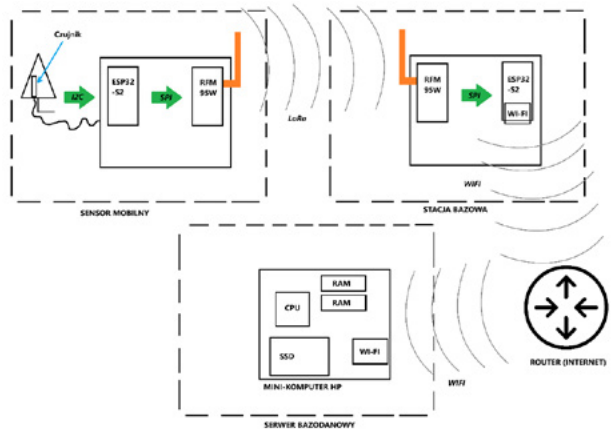
Jednym z głównych wymagań dla technologii komunikacji między sensorem mobilnym a stacją bazowa jest niskie zużycie energii. Spełniające ten warunek technologie to: ZigBee, 6LOWPAN, Zwave oraz LoRa. Kolejnym istotnym wymaganiem jest duży zasięg komunikacji; w tym zakresie najlepsze wyniki osiągają technologię WiMAX, LTE oraz LoRa. Podsumowując, LoRa wydaje się jedyną technologią, która jednocześnie spełnia oba kluczowe wymagania – bardzo niskie zużycie energii oraz duży zasięg transmisji. W związku z tym została wybrana jako podstawowa technologia komunikacyjna w projektowanym systemie.

2. OPIS SYSTEMU I METODOLOGII

W niniejszym rozdziale przedstawiono opis realizacji systemu do zbierania danych pomiarowych.

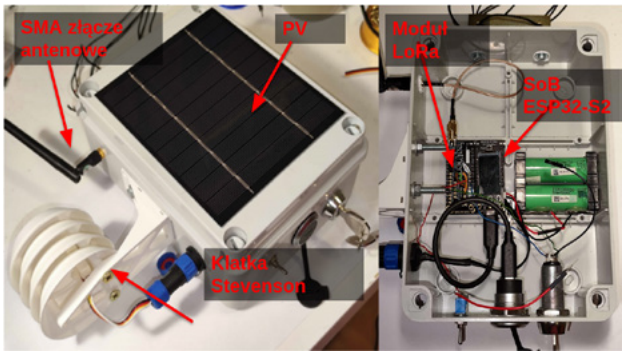
2.1. Architektura systemu

System składa się z sensora mobilnego, który pełni rolę urządzenia zbierającego dane z otoczenia. Składa się on z mikrokontrolera ESP32-S2, który odpowiada za przetwarzanie danych, oraz modułu komunikacyjnego RFM95W, umożliwiającego bezprzewodowe przesyłanie danych do Stacji Bazowej. Stacja Bazowa odbiera dane przesyłane z sensorów mobilnych za pomocą modułu RFM95W. Następnie dane są przesyłane do mikrokontrolera ESP32-S2, który realizuje funkcję bramy komunikacyjnej, przesyłając dane za pomocą Wi-Fi do serwera bazodanowego. Serwer bazodanowy oparty na kompaktowym komputerze x86 odpowiada za przechowywanie danych w bazie danych oraz ich udostępnianie. Serwer bazodanowy powstał na bazie minikomputera HP EliteDesk 600 G1 DM. Stacja Bazowa realizuje proces komunikacji poprzez stałe nasłuchiwanie kanału transmisyjnego za pomocą modemu radiowego RFM96W. Po odebraniu pakietu danych jest on identyfikowany i przetwarzany. Komunikacja odbywa się za pomocą protokołu TCP. Używany do tego jest modem WiFi wbudowany w mikrokontroler ESP32-S2 oraz sieć Wi-Fi, do której podłączona jest stacja bazowa.



Rys. 2. Schemat blokowy systemu

Sensor mobilny odpowiada za odczyt danych z czujników oraz przekazanie ich poprzez modem LoRa do Stacji Bazowej. Wyposażony jest, podobnie jak stacja bazowa w moduł Adafruit ESP32-S2 Reverse TFT Feather, moduł LoRa RFM96W, złącze SMA i zewnętrzną antenę 433 MHz oraz akumulator. Elementy wewnętrzne zamknięte są w szczelnej obudowie IP65. Klatka Stevensona z czujnikiem Unit ENV IV, przymocowana jest do obudowy (zob. Rys. 3).



Rys. 3. Sensor mobilny zamknięty w klatce Stevensona

Czujnik środowiskowy z układami Sensirion SHT40 i Bosh BMP280, umożliwia pomiar temperatury, wilgotności i ciśnienia atmosferycznego. Z powodu wysokich wymagań dotyczących niskiego zużycia energii, zarówno sensor mobilny, jak i stacja bazowa, wykorzystują ten sam moduł Adafruit ESP32-S2 Reverse TFT Feather oraz moduł Adafruit RFM96W. W tabeli 2 porównane zostały pobory prądu dla ESP32. Tryby oszczędzania energii Light Sleep i Deep Sleep, pozwalają na minimalizację zużycia prądu poprzez selektywne spowolnienie lub całkowite wyłączenie modułów takich jak CPU, podtrzymanie pamięci czy moduł szyfrujący.

Układ RFM96W to moduł radiowy oparty na układzie SX1278 firmy Semtech. Moduły serii RFM9XX działają w paśmie ISM (Industrial, Scientific and Medical – pasmo dostępne bez licencji), wspiera różne częstotliwości, takie jak 433 MHz, 868 MHz oraz 915 MHz. Dzięki temu można dostosować moduł do różnych pasm częstotliwościowych.

Tab. 2. Porównanie poboru prądu w różnych stanach mikrokontrolera ESP32-S2 [3]

Tryb	*Opis*	*Pobór prądu [mA]*
Standardowy (włączony modem Wi-Fi)	CPU działające na częst. 240 MHz	<310
Standardowy (włączony modem Wi-Fi)	CPU 240 MHz	32
	CPU 80 MHz	20
Light-Sleep	CPU wyłączone, podtrzymanie RAM	0.75
Deep-Sleep	CPU wyłączone, tylko obsługa RTC	0.02
Wyłączony	Układ jest wyłączony	0.001

Wspiera modulację: FSK, GFSK, MSK, GMSK, LoRa i OOK [6]. Układ ma maksymalną moc wyjściową 20 dBm (100 mW) i oferuje szybść transmisji w zakresie od 0,018 kb/s do 37,5 kb/s, w zależności od konfiguracji. Układ charakteryzuje się niskim poborem mocy, wynoszącym około 12 mA podczas odbioru oraz 90 mA podczas transmisji z mocą 17 dBm. W trybie uśpienia zużywa poniżej 1 µA. RFM96W obsługuje szerokość pasma od 7,8 kHz do 500 kHz oraz Spreading Factor (SF) od 6 do 12.

Układ MAX17048 wykorzystuje zaawansowany algorytm modelowania baterii litowo-jonowych ModelGauge™ do ciągłego śledzenia względnego stanu naładowania baterii (SOC – State Of Charge). Algorytm ModelGauge eliminuje konieczność stosowania rezystora pomiaru prądu oraz cykli uczenia się baterii wymaganych w tradycyjnych wskaźnikach naładowania. Dodatkowo wykorzystano układ MCP73831 zarządzający ładowaniem akumulatorów litowo-jonowych (Rys.3). W systemie wykorzystano anteny SMA 433 MHz o impedancji wynoszącej 50 Ω. Ponadto system wyposażony jest w przełącznik bistabilny dźwigniowy pozwala na zmianę źródła zasilania układów regulacji napięcia RT9080 oraz panel słoneczny monokrystaliczny o mocy maksymalnej 2,5 W.

3. TESTY SYSTEMU

Testy zostały podzielone na cztery fazy. Testy funkcjonalne, testy pracy systemu w warunkach rzeczywistych, testy zasięgowe oraz testów wydajności energetycznej.

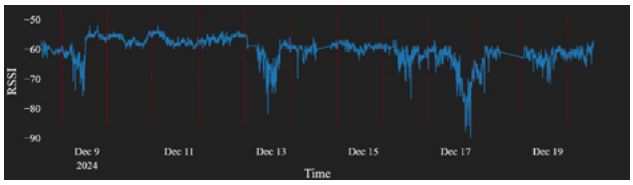
3.1. Testy funkcjonalne

Podczas fazy pierwszej, zostały przetestowane połączenia elektryczne, zweryfikowana komunikacja i możliwość programowania mikrokontrolerów ESP32-S2. Sprawdzono poprawność komunikacji mikrokontrolerów, testowe odczyty czujników, sensorów. Po pierwszym udanym uruchomieniu systemu rozpoczęto kilkudniowe testy mające sprawdzić stabilność działania systemu. Trwały one około 4 dni. W tym czasie udało się uzyskać około 15 tysięcy pomiarów. Testowane były różne interwały transmisji - 10 sekund oraz 300 sekund. Już w trakcie pomiarów został dodany pomiar naładowania baterii, poziomu ładowania/rozładowywania.

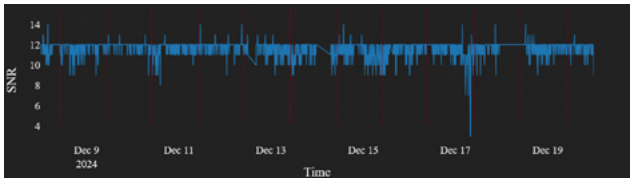
3.2. Testy w warunkach rzeczywistych

System został wystawiony na warunki atmosferyczne na okres 11 dni (8 grudzień – 19 grudzień). Przez cały ten okres stacja pracował na własnym zasilaniu (fotowoltaicznym i akumulatorowym). System zapisuje do bazy danych kilka wartości pomiarowych i telemetrycznych, są to: temperatura, wilgotność, ciśnienie, RSSI (ang. *Received Signal Strength Indication* - wskaźnik mocy odbieranego sygnału) [dBm], SNR (Signal to Noise Ratio – stosunek sygnału do szumu) [dB], stopień naładowania akumulatora LiPo oraz prędkość ładowania Chargerate.

Rysunki 4 i 5 przedstawiają kolejno wartości RSSI oraz SNR. RSSI jest to poziom mocy odbieranego sygnału radiowego wyrażony w jednostce dBm (moc odniesiona do mocy 1 mW). SNR to stosunek mocy sygnału użytecznego do mocy szumu wyrażony w decybelach (dB). Obie wartości mierzone są przez modem stacji bazowej podczas odbioru danych od sensora mobilnego.



Rys. 4. Wykres RSSI



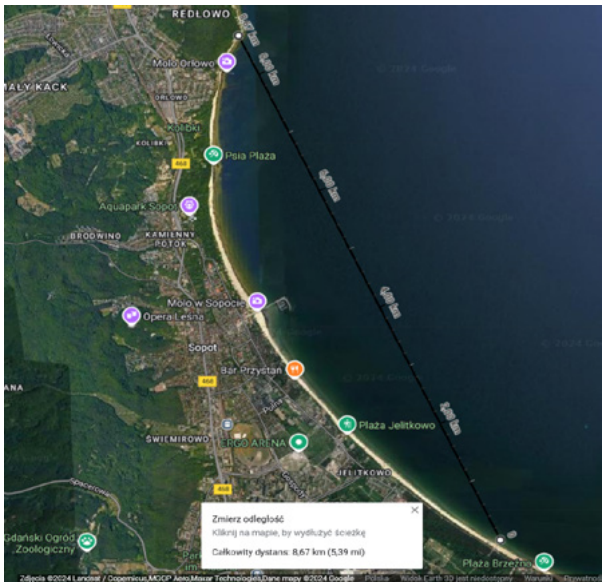
Rys. 5. Wykres SNR

Stacja bazowa znajdowała się w niewielkiej odległości od sensora w warunkach widoczności. Sensor mobilny cały czas znajdował się w jednym miejscu, jednak stacja bazowa nieco się przemieszczała, była zmieniana orientacja anteny. Wynik tych niewielkich zmian możemy zaobserwować na Rys. 4 i Rys. 5.

3.3. Testy zasięgowe

W ramach trzeciej fazy testów wykonane zostały dwa testy zasięgowe systemu w warunkach Line of Sight (LoS) – w zasięgu wzroku. Pierwszy test polegał na próbie zestawienia połączenia na maksymalnym dostępnym zasięgu w warunkach zasięgu wzroku. Wybrane do tego zadania została para miejsc molo Brzeźno dla stacji bazowej oraz Klif Orłowski

(wysokość około 30 metrów) dla sensora mobilnego. Jest to odległość około 8670 metrów w linii prostej. Pierwsza strefa Fresnela dla odległości 8.7km oraz częstotliwości 434 MHz wynosi 38.8 metra. Na tak dużych odległościach wysokość stacji nadawczej i odbiorczej ma więc ogromne znaczenie (zob. Rys. 6).



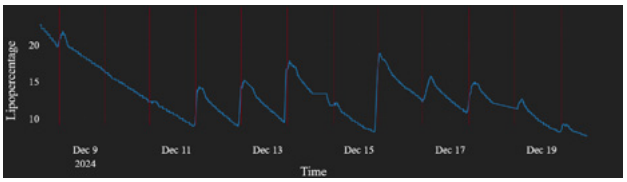
Rys. 6. Odległość pomiędzy poszczególnymi elementami systemu

Drugie podejście do testu zasięgowego zostało podjęte pomiędzy mostem drogowym na rzece Nogat w Malborku oraz jednym z wałów przeciwpowodziowych wspomnianej rzeki we wsi Grobelno. Odległość między urządzeniami wyniosła około 2800 metrów.

W ramach testów udało się nawiązać komunikację pomiędzy testowanymi urządzeniami. W momencie, gdy sensor znajdował się na nabrzeżu, poziom RSSI wyniósł -106 dBm (± 2 dB) a SNR (signal-to-noise ratio - stosunek sygnału do szumu) wyniósł 3 dB (± 1 dB). Biorąc pod uwagę deklaracje producenta układu RFM96W [8], układ ten jest w stanie odbierać, przy bieżących ustawieniach sygnał o sile RSSI -128 dBm oraz wartości SNR -10 dB. Z tego wynika, że maksymalny zasięg LoS systemu przy dobrych warunkach może wynosić nawet ponad 4 kilometry.

3.4. Wydajność energetyczna systemu

Rys. 7 przedstawia wykres procentowego naładowania akumulatora Li-Po. Testy rozpoczęto z naładowaniem akumulatora na poziomie około 23%. Jak można zauważyć, pierwsze trzy dni ładowanie akumulatora panelem słonecznym prawie nie miało miejsca. Spowodowane to jest wcześniej wspomnianym bardzo dużym zachmurzeniem. Jednak w następnych dniach - Słońce dziennie pojawiało się średnio 1-2 godziny. Podczas testów procent ładowania/ rozładowania akumulatora w dzień umiarkowanie słoneczny +10 p.p., w dzień pochmurny ~0 p.p., i w dzień o całkowitym zachmurzeniu/mgle -5 p.p. na jeden dzień. Należy zaznaczyć, że pomiary zostały wykonane w grudniu. W całym teście procent naładowania akumulatora spadł z 22.7% na 7.6%. Spadek o 15.1 pp.



Rys. 7. Wykres procentowego zużycia baterii LiPo

procentowych. Daje to średnio spadek w miesiącach zimowych 1.37 punktu procentowego na 24 godziny. Przy pełnym naładowaniu urządzenie mogłoby działać w takich warunkach ~73 dni. Spadek naładowania akumulatora bez doładowywania wynosi ~5.7 punktów procentowych na 24 godziny.

4. WNIOSKI

Technika LoRa ma zastosowanie m.in. w rolnictwie, inteligentnych miastach,, logistyce, IIoT, medycynie itp. W ramach niniejszej pracy udało się zaprojektować i skonstruować rozproszony energooszczędny system monitoringu środowiska, wykorzystujący technikę LoRa. Istnieje wiele kierunków rozwoju, które mogą poszerzyć możliwości systemu oraz zwiększyć jego funkcjonalność i efektywność. Dalszy rozwój/badanie może obejmować optymalizację komunikacji, protokół komunikacyjny zwiększający możliwości sensorów rozproszonych, dedykowana aplikacja na telefon, badania koegzystencji z innymi systemami w paśmie.

LITERATURA

[1] Stojmenović I. 2005. Handbook of Sensor Networks: Algorithms and Architectures. John Wiley & Sons:Hoboken, NJ, USA, Rozdz. 1, s. 6-8.

[2] Elkhodr M., Shahrestani S., Cheung H. 2016. "Emerging Wireless Technologies in the Internet of Things: A Comparative Study". International Journal of Wireless & Mobile Networks Vol. 8, No. 5, Październik, s. 67-82

[3] ESP32-S2 Series Datasheet v1.7, Espressif Systems: https://www.espressif.com/sites /default/files/documentation/esp32-s2_datasheet_en. pdf - data dostępu 15.12.2024 r.

[4] Bor M., Vidler J., Roedig U. 2016. "LoRa for the Internet of Things". EWSN '16 Proceedings of the International Conference on Embedded Wireless Systems and Networks. Junction Publishing, AUT, pp. 361-366. ISBN 9780994988607

[5] RT9080 Datasheet v8, Richtek Technology Corporation, data dostępu 15.12.2024: https://www.richtek.com/CustomerDownload?cDow nloadItem=%7bB7137095-D669-4BB9-B606-3EFE 206229A8%7d .

[6] RFM95W/96W/98W - Low Power Long Range Transceiver Module V2.0, HOPE MICROELECTRONICS, data dostępu 28.12.2024: https://www.hoperf.com/modules/lora/RFM96W.ht ml .

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

Inteligencja Rozproszona w migracji funkcji VNF

Swarm Intelligence in VNF function migration

Bartłomiej Mastej[1];

[1] Szkoła Doktorska Politechniki Warszawskiej,
Warszawa, bartlomiej.mastej.dokt@pw.edu.pl

DOI: 10.15199/59.2025.4.121

Streszczenie: W artykule przedstawiona została koncepcja wykorzystania Inteligencji Rozproszonej przy problemie migracji funkcji VNF. Został zaproponowany algorytm, który na podstawie lokalnych interakcji umożliwia tworzenie pola elektrycznego. Dzięki zastosowaniu pola elektrycznego węzły miałyby możliwość samodzielnego podejmowania decyzji o migracji VNF. Algorytm został poddany eksperymentom, a uzyskane wyniki wskazują na jego prawidłowe działanie i potencjał zastosowania przy migracji VNF.

Abstract: The article presents the concept of using Swarm Intelligence in the migration of VNF functions. An algorithm was proposed, which allows the creation of an electric field based on local interactions. By using the electric field, nodes would be able to make VNF migration decisions on their own. The algorithm has been subjected to experiments, and the results show its correct functioning and potential for application in VNF migration.

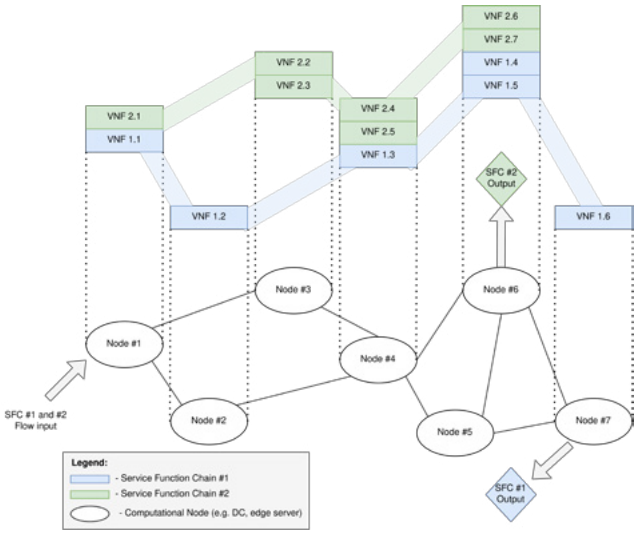
Słowa kluczowe: inteligencja rozproszona, migracja funkcji VNF, ruting pola elektrycznego, samoorganizacja

Keywords: electric-field-based routing, self-organization, swarm inteligence, VNF migration

1. WPROWADZENIE

Wraz z pracami nad siecią 6G rozwijają się technologie ją współtworzące i wspierające, nazywane z j.ang. *"enablers"*. Mają one na celu dostarczenie nowych możliwości sieci mobilnych, jednakże stawia się na możliwe wykorzystanie istniejącej już infrastruktury. Jedną z tych technologii jest Cloud-Continuum, ma ona na celu wykorzystanie wszystkich dostępnych zasobów znajdujących się w sieci. Na ich podstawie ma tworzyć spójną infrastrukturę - poczynając od chmury centralnej (cloud), przez serwery znajdujące się na krawędzi sieci (edge), aż do terminali (far edge). W artykule przyjęto nazwę węzeł (ang. node) określającą dowolny element infrastruktury obliczeniowej sieci.

Ze względu na możliwość wykorzystania dynamicznie pojawiających się i znikających zasobów infrastrukturalnych, jednym z istniejących problemów jest migracja zwirtualizowanych funkcji sieciowych (ang. *Virtualized Network Functions* - VNF). VNF'y są to wszelkiego rodzaju funkcje (np. firewall, analiza ruchu sieciowego), które połączone ze sobą tworzą usługę. Takie połączenie najczęściej wymaga ustawienia ich w określonej kolejności w celu realizacji danej usługi - VNF'y tworzą w ten sposób łańcuch (ang. *Service Function Chain* - SFC).



Rysunek 1: Przykład użycia VNF'ów tworzących dwa SFC na infrastrukturze sieciowej.

VNF'y konsumują zasoby infrastrukturalne między innymi poprzez zużycie procesora czy obciążenie łącza. Dodatkowo mogą się zdarzyć błędy, które uniemożliwiają kontynuację pracy VNF'a na danym węźle. Wykorzystanie terminali użytkowników znacznie zwiększa częstość występowania takich zdarzeń, gdyż są zasilane bateryjnie i z reguły są mobilne. W takich wypadkach konieczna jest migracja VNF'ów aby nie przerwać działania usługi.

Problem migracji VNF'ów jest najczęściej rozwiązywany z wykorzystaniem pełnej wiedzy o systemie tzw. wiedzy globalnej. Przykładowo, jak można zauważyć w artykule przeglądowym [1] najpopularniejsze metody rozwiązywania tego problemu wykorzystują różne metody optymalizacyjne takie jak: programowanie liniowe, algorytmy heurystyczne i metaheurystyczne oraz uczenie maszynowe. Jednakże wskazane w artkule metody służą do rozmieszczania VNF'ów w sieci, a nie migracji w trakcie działania. W celu przeprowadzenia migracji są one ponownie rozmieszczane. Takie podejście nie sprawdza się przy częstej zmianach w topologii sieci, gdyż wymaga za każdym razem wykonywanie kosztownych obliczeń.

W związku z tym, w artykule [2] zaproponowano politykę migracji VNF'ów, zamiast ich ponownego rozmieszczania, wykorzystującą programowanie liniowe i zmniejszającą przeszukiwaną przestrzeń przy użyciu heurystyki. Inne podejście zaprezentowano w artykule [3], gdzie zaproponowano dwa algorytmy migracji VNF'ów: pierwszy skupiony na możliwościach obliczeniowych węzłów, natomiast drugi brał pod uwagę przepustowość między węzłami. Oba algorytmy działają dzieki monitorowaniu wszystkich dostępnych zasobów na węzłach z perspektywy globalnej. VNF'y były migrowane do węzłów posiadających wolne zasoby i będących w najbliższej okolicy najkrótszej ścieżki.

Uwzględniając potencjalnie silnie dynamiczną naturę infrastruktury typu Cloud-Continuum, w artykule jest zaproponowane alternatywne podejście do migracji VNF'ów, skupiające się na perspektywie pojedynczego węzła zamiast na całości systemu. Podejście opiera się o paradygmat Inteligencji Rozproszonej (IR), którego celem jest samodzielna organizacja systemu. Jest ona uzyskana na podstawie lokalnych interakcji, dodatkowo możliwie unikane jest wykorzystanie wiedzy globanej [4]. Systemy IR najczęściej tworzy się operując na dwóch płaszczyznach: poziom marko (system) i poziom mikro (pojedynczy agent). W kolejnej części artykułu zaprezentowany jest model systemu i model pojedynczego agenta. Następnie został opisany algorytm umożliwiający samoorganizację łańcuchów SFC, wykorzystujący zasadę działania pola elektrycznego.

Zaproponowany algorytm umożliwia natychmiastowe określenie dla danego węzła wszystkich możliwych jego sąsiadów będących bliżej węzła celu. Umożliwia to samodzielne podejmowanie decyzji o migracji VNF'a bez wstrzymania pracy systemu i bez konieczności informowania całego systemu. Warto zauważyć, że algorytm samodzielnie znajduje alternatywne ścieżki i samodzielnie dostosowuje się do zmieniającej topologii.

2. MODEL SYSTEMU

Można zauważyć, że węzły oraz działające na nich VNF'y działają na dwóch warstwach abstrakcji. Każdy VNF należy do jakiegoś łańcucha SFC. Z perspektywy SFC nie jest istotne, na jakim węźle działa dana funkcja VNF. Ważnym jest jedynie, aby ciągłość trwania usługi była zachowana. Z perspektywy węzłów istotnym natomiast jest optymalne wykorzystanie zasobów - możliwie duża ich alokacja, jednakże nie dopuszczanie do ich przeciążenia. Z tego powodu za poziom makro systemu IR można uznać SFC jako całość.

W celu przetestowania zaproponowanego rozwiązania został wykorzystany model możliwie zbliżony do [3]. W związku z tym zostały przyjęte następujące założenia:

- VNF'y tworzące SFC muszą być ułożone w określonej kolejności.
- Dany SFC składa się z konkretnej liczby VNF'ów.
- VNF'y są już rozmieszczone na węzłach sieci.

Ostatnie założenie wynika z problematyki migracji VNF'ów - musi się ona odbywać w trakcie działania usługi w ramach SFC, w przeciwieństwie do zadania początkowego rozmieszczenia VNF'ów. Przyjęty scenariusz zakłada, że VNF'y zostały rozłożone na węzłach będących na najkrótszej ścieżce między źródłem a celem.

3. MODEL AGENTA (WĘZŁA)

W zaproponowanym rozwiązaniu pojedynczym agentem jest węzeł - to on podejmuje decyzję o migracji VNF'ów jeżeli następuje taka konieczność. Warstwa abstrakcji - SFC - tworzy wirtualne pole elektryczne (co zostanie przedstawione później), które jest wykorzystywane do lokalnego podejmowania decyzji. Dzięki niemu każdy węzeł, na którym pracuje VNF jest w stanie natychmiast podjąć decyzję do którego węzła należy daną funkcję zmigrować. Węzły mają możliwość komunikacji między sobą, na podstawie lokalnej komunikacji tworzone jest wirtualne pole elektryczne.

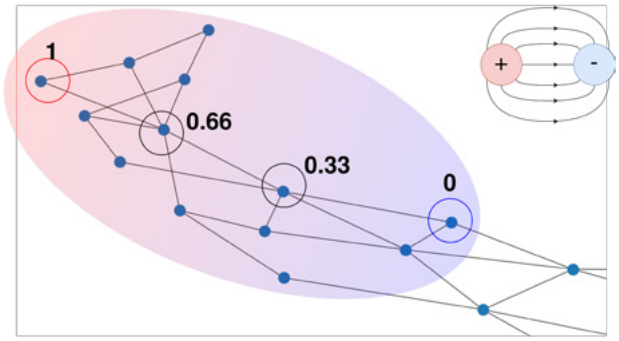
Tabela 1: Notacja

Notacja	*Objaśnienie*
q_i	wartość ładunku na węźle i
x_i	węzeł i bez określonego ładunku
n_i	węzeł i z określonym ładunkiem
N	zbiór wszystkich węzłów n_i
S_i	zbiór wszystkich sąsiadów węzła i

4. RUTING POLA ELEKTRYCZNEGO

Ruting pola elektrycznego (ang. Electric-field-based routing - EFR) zaproponowany przez [5], jak nazwa wskazuje opiera się o ideę pola elektrycznego. Dwa elementy o przeciwnych ładunkach elektrycznych tworzą pole wektorowe, określające dla każdego punktu siłę oddziaływującą na pojedynczy ładunek elektryczny.

Analogicznie w rutingu zamiast tych dwóch elementów wystepuje źródło i cel. Natomiast "siła" oddziaływująca między nimi wpływa na wszystkie węzły pośrednie określając ich "ładunek". Na Rysunku. 2 zaprezentowano takie pole oddziaływujące między węzłami w sieci. Na niebiesko zaznaczono węzeł będący źródłem, a dastynację na czerwono.



Rysunek 2: Pole "elektryczne" węzłów.

4.1 Opis algorytmu samoorganizacji SFC

Zaproponowany algorytm, w przeciwieństwie do pierwotnego wykorzystania idei pola elektrycznego w rutingu, nie ma na celu zapewnienia dwustronnej komunikacji, lecz jego celem jest stworzenie samoorganizującego się łańcucha VNF'ów dla każdego SFC. W związku z tym różnice między EFR oraz algorytmem są znaczące,

a podobieństwo opiera się na wykorzystaniu tego samego principium. Przykładowo, algorytm samoorganizacji nie wymaga znajomości przez węzeł źródłowy lokalizacji węzła docelowego.

W zaproponowanym algorytmie pole elektryczne jest obliczane osobno **dla każdego SFC**. Jak zostało zaznaczone w modelu systemu, VNF'y są już rozmieszczone na węzłach w ramach działających SFC, jeszcze przed początkiem działania algorytmu. Węzły tworzące pierwotną ścieżkę zaznaczono kółkami na Rysunku 2, źródło ma ładunek 1, natomiast cel 0. W pierwszym kroku jest wykorzystywany jedyny element wiedzy globalnej w celu przyśpieszenia działania - obliczane są wartości ładunków między źródłem a celem - ich potencjałów.

Pozostałe węzły będące w sąsiedztwie do węzłów początkowych mają wyliczane swoje ładunki. Wartość ładunku jest wyliczana na dwa sposoby i zależy jedynie od liczby sąsiadów węzła x_i, którzy mają już wyliczony ładunek. Jeżeli tylko jeden sąsiad posiada ładunek, stosowane jest równanie 1, natomiast w przypadku posiadania większej liczby sąsiadów, wartość ładunku jest średnią arytmetyczną wszystkich z nich 2. Gdy sąsiedzi zmienią wartość swojego ładunku, dany węzeł również aktualizuje swój ładunek, ponownie korzystając z równań 1 i 2. Węzły mające wyliczoną wartość najkrótszej ścieżki nie aktualizują swoich ładunków - dopóki nie dojdzie do poważniejszych zmian w topologii to pozostaną optymalną ścieżką.

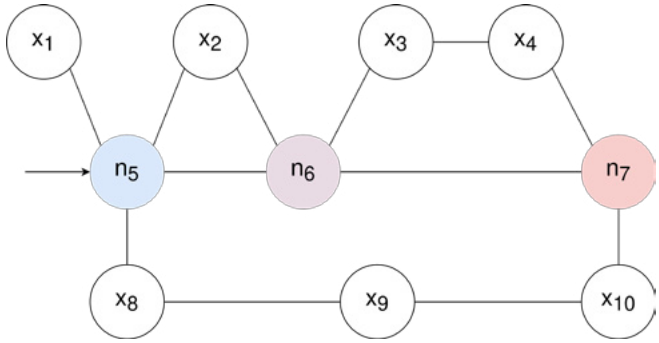
$$|S_{x_i} \in N| = 1 \implies q_{x_i} = \frac{(1 + q_{n_j})}{2}; n_j \in S_{x_i} \quad (1)$$

$$|S_{x_i} \in N| > 1 \implies q_{x_i} = \frac{\sum_j^{n_j \in S_{x_i}} q_{n_j}}{|S_{x_i} \in N|} \quad (2)$$

W przypadku zmian w topologii system samodzielnie się dostosuje, nawet zmieniając źródło czy cel. Ze względu na wymianę informacji sąsiadujące węzły będą budowały pole elektryczne, gdyż nastąpi zmiana w ich otoczeniu. W przypadku zmian w topologii najkrótszej ścieżki węzeł ją opuszczający powinien wskazać swojego zastępcę/zastępców w celu uniknięcia zbędnej wymiany informacji - system ponownie musiałby się zbalansować. Alternatywnie można wyznaczyć pewną dokładność ϵ, która ograniczy wymianę informacji między sąsiadami.

4.2 Zasada działania.

"Prąd" przepływa, gdy powstaje różnica potencjałów. Podobnie w zaproponowanym algorytmie migracja funkcji jest możliwa, gdy występuje różnica "potencjałów" - wartości ładunków. Łatwo zauważyć, że z persektywy pojedynczego wezła, gdy jego sąsiad posiada niższą wartość potencjału to musi mieć możliwość dotarcia do celu (ta zasada nie działa w drugą stronę, lecz można stosować heurezę - jeżeli wartość ładunku sąsiada jest zbliżona do watości własnego ładunku wyliczonego za pomocą równania 1 to najprawdopodobniej nie jest połączony z innym węzłem posiadającym ładunek).



Rysunek 3: Zastosowanie algorytmu rutingu na różnych przykładach topologicznych.

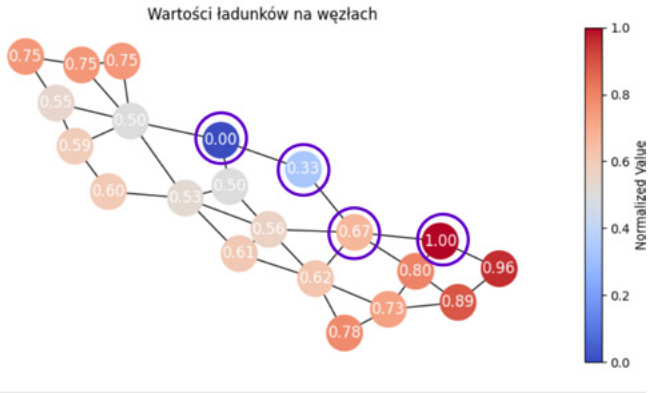
Na Rysunku 3 zaprezentowano trzy węzły mające obliczoną wartość ładunku (n_5, n_6, n_7) oraz węzły $x_1, x_2, x_3, x_4, x_8, x_9, x_{10}$, których wartość ładunku będzie obliczana. Zgodnie z algorytmem $q_{x_1} = (1 + q_{n_5})/2$, a zatem $q_{x_1} < q_{n_5}\ \forall q_{n_5} \in [0, 1)$ i równe, gdy $q_{n_5} = 1$ - węzeł źródłowy (najwyższa wartość). Drugi przypadek: $q_{x_2} = (q_{n_5} + q_{n_6})/2$, zgodnie z założeniem "prąd" płynie wtedy i tylko wtedy, gdy występuje różnica potencjałów, a dokładniej, gdy $q_{n_5} > q_{n_6}$. Z tego powodu $q_{n_5} > q_{x_2} > q_{n_6}\ \forall q_{n_5}, q_{n_6}$ Następny przypadek wymaga najpierw obliczenia wartości ładunku dla węzłów x_3 i x_4 jak dla x_1, a następnie tak samo jak dla x_2. Ostatni przypadek x_8, x_9, x_{10} również oblicza się analogicznie - najpierw oblicza się wartość ładunku dla x_8 i x_{10}, a następnie dla x_9. Następnie procedurę się powtarza (najpierw x_8 i x_{10}, następnie x_9), aż do osiągnięcia zadowalającego ϵ, w celu przyśpieszenia samodzielnego balansowania.

5. EKSPERYMENTY

Celem eksperymentów było zweryfikowanie poprawności działania algorytmu samoorganizacji opartego na polu elektrycznym oraz oszacowanie jego efektywności.

Testy zostały przeprowadzone w samodzielnie stworzonym symulatorze, w którym zostały zamodelowane węzły zgodnie z topologią ItalyNet [6]. Eksperymenty zostały przeprowadzone dla pojedynczego SFC, gdyż każde kolejne SFC jest wyliczane w taki sam sposób. Nie ma interferencji między SFC zatem w celu przetestowania algorytmu ruting wystarczającym jest zbadanie pojedynczego pola elektrycznego. Liczba VNF'ów w łańcuchy SFC należała do zakresu $[1; 8]$. Węzły będące na początku i końcu łańcucha SFC była losowane w taki sposób, aby odległość między nimi była równa liczbie VNFów. Węzły będące na początku i końcu SFC były losowane, a najkrótsza ścieżka między nimi była wyliczona na podstawie algorytmu Dijkstry (stąd maksymalna długość SFC wynosi 7).

Na Rysunku 4 można zaobserwować wartości pola elektrycznego na węzłach dla SFC składającego się z 4 węzłów. Na rysunku zaznaczona jest najkrótsza ścieżka, jednakże można zauważyć, że w przypadku niedrożności któregoś z węzłów poruszając się zgodnie z zasadą przepływu w stronę niższej wartości ładunku zawsze dotrze się do celu. Co więcej, zawsze im niższa wartość ładunku, tym krótsza ścieżka. Stąd wybór alternatywnej ścieżki względem najkrótszej zawsze sprowadza się do wybrania drugiej najniższej wartości ładunku.



Rysunek 4: Zbalansowane pole elektryczne z zaznaczoną najkrótszą ścieżką.

5.1 Ocena efektywności

Algorytm należy do samoorganizujących się, z tego powodu kluczowym wskaźnikiem wydajności jest czas potrzebny do osiągnięcia stanu stabilnego. Taki stan umożliwia rozmieszczenie łańcucha SFC z zachowaniem jego niezawodności na większej liczbie tras oraz umożliwia migrację VNFów do większej liczby węzłów (przypadek ścieżki x_8, x_9, x_{10}, która balansuje się dłużej).

Dla algorytmu wyznacznikiem stanu stabilności systemu jest brak występowania zmian w polu elektrycznym. Jeżeli pole elektryczne nie zmieni się w trakcie dwóch kolejnych iteracji oznacza, że stało się stabilne - ładunki nie są już przeliczane na żadnym węźle. Zatem wszystkie węzły mają już przeliczone wartości ładunków. Zostało przeprowadzone 1000 testów zgodnie z wcześniejszym opisem. Wyniki otrzymane na ich podstawie można zobaczyć w Tabeli 2. Czas stablizacji został określony jako średnia liczba iteracji wymiany wiadomości między węzłami.

Tabela 2: Średni czas dostosowania pola elektrycznego z podziałem na długość łańcucha SFC.

Długość łańcucha	*Średni czas stabilizacji*
2	9,811
3	8,669
4	7,176
5	5,783
6	4,509
7	4,000

Przy długości $SFC = 7$ łatwo zauważyć, że szybki czas stabilizacji wynika ze względnie szybkiego zapełnienia całej planszy, gdyż większość węzłów pierwotnie neutralnych znajduje się w pobliżu węzłów posiadających ładunek. Im mniejsza długość łańcucha, tym ciężej uzyskać stabilność, gdyż system dąży do zbalansowania wszystkich węzłów. Podobna sytuacja wystąpiłaby w przypadku udziału bardzo dużej sieci węzłów.

W związku z tym, można zredukować rozprzestrzenianie się ładunku dodając pewne ograniczenia. Przykładowo, jeżeli węzeł oblicza wartość swojego ładunku na podstawie równania 1, przekazuje o tym informację dalej. Wówczas dalsze węzły nie posiadające ładunku nie będą go rozpropagowywały. Ograniczenia mogą powodować z jednej strony wyższą wdajność, jednakże część ścieżek może z tego powodu zostać pominięta. Dlatego inną możliwością jest zmniejszanie zasad ograniczeń jeżeli obecna liczba ścieżek nie będzie wystarczająca.

6. PODSUMOWANIE

W artykule został zaprezentowany algorytm samoorganizacji działający na zasadzie wirtualnego pola elektrycznego. Algorytm opiera się o wykorzystanie lokalnych interakcji w celu samoorganizacji - tworzącym system Inteligencji Rozproszonej. Celem artykułu było zaproponowanie wykorzystania IR do rozwiązywania problemu klasycznie rozwiązywanego za pomocą wiedzy globalnej.

Zaproponowany algorytm jest elementem mającym w przyszłości umożliwić migrację funkcji VNF w infrastrukturze typu Cloud-Continuum uwzględniającym możliwości występowania dynamicznych zmian w zasobach sieciowych. W tej pracy zaproponowano algorytm samoorganizacji, który skupia się na perspektywie węzłów i traktuje SFC jako warstwę abstrakcji. Ma to na celu umożliwienie podejmowanie decyzji dotyczących migracji VNF'ów samodzielnie przez dany węzeł - bez konieczności posiadania zewnętrznego orkiestratora.

W dalszej pracy koniecznym będzie przetestowanie zdolności algorytmu do rozwiązania migracji VNF'ów oraz porównanie go z klasycznymi algorytami. Wartym sprawdzenia jest również zweryfikowanie możliwości działania algorytmu w środowisku dynamicznym, między innymi pod względem czasu uzyskania stabilności oraz porównanie wydajności w tym zadaniu z innymi algorytmami. Z całą pewnością biorąc pod uwagę możliwość wykorzystania terminali użytkowników w sieci 6G, które charakteryzując się niestabilnością i mobilnością, warto zbadać możliwości jakie daje im Inteligencja Rozproszona.

LITERATURA

[1] Zhang, B., Fan, Q., Zhang, X. et al. A survey of VNF forwarding graph embedding in B5G/6G networks. Wireless Netw 30, 3735–3758 (2024). https://doi.org/10.1007/s11276-021-02741-9

[2] V. Eramo, E. Miucci, M. Ammar and F. G. Lavacca, "An Approach for Service Function Chain Routing and Virtual Function Network Instance Migration in Network Function Virtualization Architectures," in IEEE/ACM Transactions on Networking, vol. 25, no. 4, pp. 2008-2025, Aug. 2017, doi: 10.1109/TNET.2017.2668470.

[3] B. Yi, X. Wang, M. Huang and K. Li, "Design and Implementation of Network-Aware VNF Migration Mechanism," in IEEE Access, vol. 8, pp. 44346-44358, 2020, doi: 10.1109/ACCESS.2020.2978002

[4] J. Kennedy, Swarm Intelligence. Boston, MA: Springer US, 2006, pp. 187–219. https://doi.org/10.1007/0-387-27705-6_6

[5] Nam T. Nguyen, An-I Andy Wang, Peter Reiher, and Geoff Kuenning. 2004. Electric-field-based routing: a reliable framework for routing in MANETs. SIGMOBILE Mob. Comput. Commun. Rev. 8, 2 (April 2004), 35–49. https://doi.org/10.1145/997122.997129

[6] Borylo, Piotr & Lason, Artur & Rzasa, Jacek & Szymanski, Andrzej & Jajszczyk, A.. (2016). Green Cloud Provisioning Throughout Cooperation of a WDM Wide Area Network and a Hybrid Power IT Infrastructure. Journal of Grid Computing. 14. 10.1007/s10723-015-9354-7.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2025

METODA ORKIESTRACJI WIRTUALNYCH ZAPÓR SIECIOWYCH W SIECIACH OPERATORSKICH[1]

ORCHESTRATION OF VIRTUAL FIREWALLS IN OPERATOR'S NETWORK

Bartosz Kopeć-Persiński*, Andrzej Bęben*

* Politechnika Warszawska, {bartosz.kopec-persinski; andrzej.beben}@pw.edu.pl

DOI: 10.15199/59.2025.4.122

Streszczenie: Zapewnienie bezpieczeństwa sieci jest obecnie jednym z głównych wyzwań operatorów. Rozwój technik wirtualizacji funkcji sieciowych pozwala na dostosowanie wydajności wirtualnych zapór sieciowych do aktualnego obciążenia sieci. W artykule zaproponowano nową metodę orkiestracji wirtualnymi zaporami sieciowymi, która jednocześnie skaluje liczbę instancji oraz steruje rozpływem ruchu co prowadzi do lepszego wykorzystania zasobów. Eksperymenty przeprowadzone z wykorzystaniem modelu MILP oraz danych zmierzonych w sieci operatora potwierdziły efektywność proponowanej metody.

Abstract: Today, network security is one of the main challenges for telco operators. The development of network function virtualization techniques allows for adjusting the performance of virtual firewalls to the current network load. The paper proposes a new orchestration method for virtual firewalls, which simultaneously scales the number of instances and controls traffic distribution, leading to better resource utilization. Experiments performed using the MILP model and data set measured in the operator's network confirmed the effectiveness of the proposed method.

Słowa kluczowe: wirtualne zapory sieciowe, wirtualizacja funkcji sieciowych, orkiestracja, optymalizacja, programowanie liniowe

Keywords: wirtual firewalls, network function virtualisation, orchestration, optimization, linear programming

1. WPROWADZENIE

Zapewnienie bezpieczeństwa i integralności danych w sieciach telekomunikacyjnych jest jednym z największych wyzwań dla operatorów. Według raportu *Barometr cyberbezpieczeństwa* [6] w 2024 roku 83% badanych firm wykryło incydent bezpieczeństwa. Podstawową metodą ochrony sieci przed atakami i niechcianym ruchem jest zastosowanie inteligentnych systemów wykrywania i prewencji IDS/IPS (*Intrusion Detection Systems / Intrusion Prevention Systems*) [5]. Systemy te wykorzystują zapory sieciowe FW (*Fire walls*), instalowane w punktach napływu ruchu PoP (*Point of Presence*), aby filtrować niepożądane pakiety. Zapory sieciowe wykorzystują metody bazujące na analizie znaczników, profili ruchowych, zawartości pakietów DPI (*Deep Packet Inspection*)[7] lub metody uczenia maszynowego do wykrywania anomalii i niepożądanych wzorców ruchu [8].

Ze względu na coraz bardziej złożone sposoby ataków opracowywane są nowe algorytmy detekcji zagrożeń, wymagające większej mocy obliczeniowej do analizy przesyłanych strumieni danych [4]. Obciążenie zapór sieciowych zwykle zmienia się zgodnie z cyklem dobowym ruchu, jednakże w sytuacji skoordynowanych ataków np. DDoS (*Distributed Denial of Service*) lub awarii rutingu obciążenie to może rosnąć w sposób nieprzewidywalny. W przypadku sprzętowych zapór sieciowych FW operatorzy wymiarują urządzenia uwzględniając prognozę największego obciążenia przyjmując pewien margines uwzględniający potencjalny wzrost ruchu. Rozwój technik wirtualizacji funkcji sieciowych NFV/CNF (*Netowrk Function Virtualization / Cloud Native Function*), umożliwia wdrożenie wirtualnych zapór sieciowych (vFW), których wydajność można dostosowywać do aktualnego obciążenia skalując horyzontalnie liczbę działających instancji. Skalowanie wymaga opracowania efektywnych algorytmów orkiestracji. Algorytmy te są odpowiedzialne za uruchomienie nowych lub usuwanie nadmiarowych instancji vFW w systemie.

Orkiestracja wirtualnymi funkcjami sieciowymi VNF (*Virtual Network Function*) jest ważnym i szeroko badanym współcześnie problemem, zarówno w obszarze komercyjnym jak i akademickim. Autorzy [9] przedstawili systematyczny przegląd rozwiązań dotyczących alokacji zasobów w NFV, klasyfikując je do jednego z czterech głównych problemów: VPTR (*VNF Placement and Traffic Routing*), VNFP (*VNF Placement*), TRR (*Traffic Routing*) i VRC (*VNF Redeployment and Consolidation*). Każdy z wymienionych wariantów problemu opiera się o: 1) zagadnienie pozycjonowania i skalowania liczby instancji VNF oraz 2) znalezienia ścieżki w grafie sieci, która spełni wymagania dot. pojemności łączy oraz kolejności obsługi w ramach zdefiniowanej sekwencji usług SFC (*Service Function Chain*). Autorzy [1] przedstawili rozwiązanie problemu alokacji vFW za pomocą algorytmu MaxSMT. Założono, iż vFW może zostać utworzony na dowolnym węźle oraz włączono do rozwiązania zarządzanie politykami bezpieczeństwa.

W artykule zaproponowano nową metodę orkiestracji wirtualnych zapór sieciowych vFW, która umożliwia dostosowanie wydajności systemu vFW do aktualnego obciążenia. Metoda stanowi rozszerzenie

[1]Prace były realizowane w ramach projektu KPOD.01.18-IW.03-0009/24, pt. „Krajowe laboratorium sieci i usług 5G wraz z otoczeniem", współfinansowanego w ramach Krajowego Planu Odbudowy sfinansowanego przez Unię Europejską NextGenerationEU.

podejścia VPTR, polegającą na jednoczesnym horyzontalnym skalowaniu liczby instancji vFW uruchomionych w poszczególnych punktach PoP lub centrach danych operatora oraz sterowaniu rozpływem ruchu do uruchomionych instancji vFW. Przeprowadzone eksperymenty, z wykorzystaniem opracowanego modelu systemu IDS/IPS bazującego na metodach programowania matematycznego, wykazały, że jednoczesne uwzględnienie sterowanie rozpływem ruchu pozwala lepiej wykorzystać zasoby uruchomionych instancji vFW zmniejszając koszty działania systemu a jednocześnie zapewniając większą odporność na trudne do przewidzenia zmiany obciążenia.

W rozdziale 2 przedstawiono proponowaną metodę orkiestracji dotyczącą problemu orkiestrowania vFW opisanego w rozdziale 1. W rozdziale 3 przedstawiono wyniki badań, potwierdzających efektywność proponowanej metody względem obecnie stosowanych rozwiązań. W rozdziale 4 podsumowano prace i przedstawiono kierunki dalszych badań.

2. PROPONOWANE ROZWIĄZANIE

Proponowana metoda orkiestracji monitoruje napływ ruchu do sieci i na tej podstawie wyznacza liczbę instancji vFW oraz ich lokalizację, a także wyznacza rozpływ ruchu z węzłów PoP do miejsc uruchomienia instancji vFW. W okresie małego obciążenia sieci, orkiestrator zmniejsza liczbę instancji vFW i przekieruje część ruchu do pozostałych węzłów z uruchomionymi instancjami vFW. Analogicznie, w trakcie największego obciążenia sieci, orkiestrator zwiększa liczbę instancji vFW tak, aby nie dopuścić do przeciążenia vFW i w konsekwencji przesłania niebezpiecznego ruchu do sieci. Proponowany algorytm optymalizuje liczbę uruchomionych instancji vFW, ich lokalizację oraz rozpływ ruchu, tak aby zminimalizować zarówno koszty zużywanych zasobów obliczeniowych jak i koszty transmisji pakietów przesyłanych poza lokalny węzeł PoP. Sterowanie rozpływem ruchu pozwala zmniejszyć liczbę instancji vFW dzięki pełnemu wykorzystaniu ich zdolności obsługowej, co nie jest możliwe gdy vFW są uruchamiane wyłącznie lokalnie w PoP. Ponadto, proponowana metoda zwiększa odporność systemu na nagły wzrost ruchu, gdyż algorytm może uruchamiać nowe instancje vFW dopóki choć w jednym węźle są dostępne wymagane zasoby obliczeniowe. Jednakże, przekaz niebezpiecznego ruchu poza lokalny węzeł wymaga odizolowania go od pozostałych przepływów w sieci operatora.

2.1 Model infrastruktury

W proponowanej metodzie infrastrukturę operatora zamodelowano jako graf skierowany $G(N, E)$, gdzie N reprezentuje zbiór węzłów umożliwiających uruchamianie instancji vFW, podczas gdy zbiór E zawiera predefiniowane ścieżki przekazu pakietów pomiędzy węzłami $p \in E$. Węzły $n \in N$ zaprezentują węzły PoP do których napływa ruch oraz centra danych operatora. Umożliwiają one uruchamianie instancji vFW, przy czym zwykle węzły PoP zlokalizowane na brzegu sieci dysponują ograniczonymi zasobami obliczeniowymi w porównaniu do centrów danych. Uruchomienie pojedynczej instancji vFW wymaga użycia określonej ilości zasobów, tj. liczby rdzeni procesora CPU oraz ilości pamięci operacyjnej RAM. Zasoby te są zajęte przez cały czas działania instancji vFW i są zwalniane po jej zatrzymaniu. Proponowana metoda decyduje, czy lepiej uruchomić nową instancję vFW, czy też przekierować strumień pakietów do innego węzła na którym działają instancje vFW. Decyzja ta jest podejmowana w zależności od całkowitych kosztów rozważanego rozwiązania.

2.2 Zdolność obsługowa instancji vFW

Przyjęto założenie, iż pojedyncza instancja vFW może przetwarzać strumień pakietów o zadanej maksymalnej przepływności bitowej η. Dla uproszczenia przyjęto, że wartość ta jest niezależna od liczby analizowanych strumieni i rozmiarów przesyłanych pakietów. Zdolność obsługową (pojemność) pojedynczej instancji vFW η, oszacowano na podstawie analizy rozwiązań vFW oferowanych przez firmy Cisco ASA [2] i Fortinet Fortigate [3]. Pojedyncza instancja vFW wykorzystująca 4 rdzenie procesora i 6 GB pamięci RAM może przeprowadzać analizę zagregowanego ruchu o przepływności bitowej do 10 Gbit/s.

2.3 Napływ strumieni pakietów

Proponowana metoda orkiestracji wykorzystuje dane wolumetryczne o przepływnościach bitowych strumieni pakietów napływających do sieci. W badaniach wykorzystano dane zmierzone co godzinę w ciągu jednej doby u operatora sieci komórkowej w Polsce. Optymalizację przeprowadzono dzieląc strumienie pakietów na dziesięć klas strumieni $[F1, F10]$ względem kolejnych decyli rozkładu zmierzonej przepływności bitowej. Klasa F1 reprezentuje 10% strumieni o niższej przepływności bitowej, zaś klasa F10 uwzględnia 10% strumieni o najwyższej przepływności. Proponowany algorytm rozpływu ruchu przyjmuje jednakową wartość przepływności bitowej dla każdego strumienia należącego do danej klasy. Precyzję działania algorytmu można zwiększyć przez uwzględnienie większej liczby klas strumieni, co jednak wiąże się ze zwiększoną złożonością obliczeniową proponowanej metody.

2.4 Model optymalizacyjny

Proponowana metoda orkiestracji wykorzystuje technikę optymalizacji bazującą na modelach MILP (*Mixed Integer Linear Programming*) do wyznaczenia wymaganej liczby instancji vFW, ich lokalizacji oraz rozpływu ruchu. W modelu MILP przyjęto następujące oznaczenia: 1) $x \in X$ oznacza węzeł PoP, gdzie wpływa ruch, 2) $y \in N$ wskazuje węzeł, gdzie uruchomiono instancje vFW, 3) $c \in C$ wskazuje rozważaną klasę strumieni, a $e \in E$ wskazuje ścieżkę przekazu danych w sieci. W modelu wykorzystywane są dwie zmienne decyzyjne: 1) zmienna n_y, określa liczbę instancji vFW uruchomionych w węźle y, oraz 2) zmienna y_{xyc} określa wolumen ruchu należący do klasy c, który napływa w węźle PoP x i jest obsługiwany w węźle y. Pozostałe parametry wykorzystywane w modelu podsumowano w Tab. 1.

Tab. 1: Parametry wykorzystywane w modelu.

δ_{xc} - ruch w klasie c	β_c - przepływność bitowa w klasie c	η - wydajność instancji vFW
i_y^{max} - dostępna liczba rdzeni CPU	j_y^{max} - dostępna pamięć RAM	l_y^{max} dostępna liczba akceleratorów
i - liczba rdzeni per instancja vFW	j - ilość RAM per instancja vFW	P_{xy} - ścieżka przekazu
ξ - koszt alokacji vFW	ζ_e - koszt transmisji	κ - koszt nieanalizowanego ruchu

2.4.1 Ograniczenia i warunki brzegowe

Ograniczenie (1) zapewnia, że dla każdej kasy c, suma strumieni przydzielonych do obsługi na poszczególnych węzłach nie przekroczy wartości ruchu napływającego w węźle x. Ograniczenia (2) i (3) zapewniają, że zasoby obliczeniowe wykorzystywane przez instancje vFW uruchomione na poszczególnych węzłach nie przekroczą zasobów obliczeniowych dostępnych w rozważanym węźle. Ograniczenie (4) dotyczy zapewnienia dostępności akceleratorów sprzętowych, jeśli są one wymagane dla uruchomienia instancji vFW. Ostatnie ograniczenie (5) zapewnia, że suma strumieni skierowanych do obsługi do danego węzła nie przekroczy całkowitej zdolności obsługowej instancji vFW uruchomionych w tym węźle.

$$\sum_{y} y_{xyc} \leq \delta_{xc}, \quad \forall x \in X, \forall c \in C \tag{1}$$

$$n_y \cdot i_y \leq i_y^{max}, \quad \forall y \in Y \tag{2}$$

$$n_y \cdot j_y \leq j_y^{max}, \quad \forall y \in Y \tag{3}$$

$$n_y \leq l_y^{max}, \quad \forall y \in Y \tag{4}$$

$$\sum_{xc} y_{xyc} \cdot \beta_c \leq n_y \cdot \eta, \quad \forall_y \in Y \tag{5}$$

2.4.2 Funkcja celu

Funkcja celu w (6) obejmuje i) koszt alokacji każdej instancji vFW, ii) koszt transmisji ruchu między węzłami PoP a węzłami obliczeniowymi oraz iii) koszt ruchu, który nie jest analizowany przez system vFW (ruch stracony).

$$\begin{aligned} minimize : F(y,n) = \sum_{y} n_y \cdot \xi \\ + \sum_{c} [\beta_c \cdot \sum_{e} [\zeta_e \cdot (\sum_{x} \sum_{y} y_{xyc} | e \in E)]] \\ + \kappa \cdot \sum_{xc} \beta_c (\delta_{xc} - \sum_{y} y_{xyc}) \end{aligned} \tag{6}$$

Rozwiązując przedstawiony problem, algorytm znajduje optymalny przydział vFW, opisany wektorem n_y, który minimalizuje liczbę instancji vFW, a także znajduje optymalny sposób kierowania strumieni z węzłów PoP do miejsc uruchomienie vFW uwzględniając koszty przekazu pakietów.

3. OCENA EFEKTYWNOŚCI

W celu oceny efektywności proponowanego rozwiązania zdefiniowano 10 scenariuszy testowych odnoszących się do 3 różnych podejść do zabezpieczenia sieci przez: 1) fizyczne zapory sieciowe, 2) wirtualne zapory sieciowe vFW wykorzystujące skalowanie vFW jedynie w węzłach POP oraz 3) wirtualne zapory sieciowe vFW dodatkowo wykorzystujące mechanizm sterowania rozpływem ruchu. Eksperymenty przeprowadzono wykorzystując narzędzie CPLEX. Parametry przeprowadzonych badań zestawiono w Tab. 2. "Typ" określa, czy badana jest fizyczna zapora sieciowa, czy instancja wirtualnej zapory vFW. W przypadku badań nad fizycznymi FW, zastosowano założenie

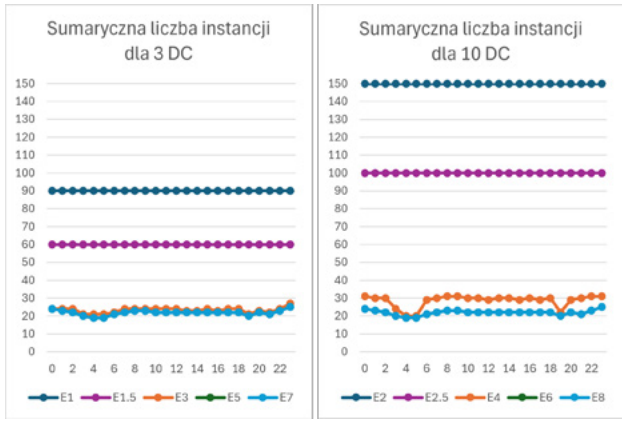
Tab. 2: Zestawienie eksperymentów.

	E1	E1.5	E2	E2.5	E3
Typ	fiz.	fiz.	fiz.	fiz.	wirt.
Skala	100%	66.67%	50%	33.34%	auto
DC	3	3	10	10	3
Dystr.	brak	brak	brak	brak	brak
Trm.	brak	brak	brak	brak	brak

	E4	E5	E6	E7	E8
Typ	wirt.	wirt.	wirt.	wirt.	wirt.
Skala	auto	auto	auto	auto	auto
DC	10	3	10	3	10
Dystr.	brak	tak	tak	tak	tak
Trm.	brak	niskie	niskie	wys.	wys.

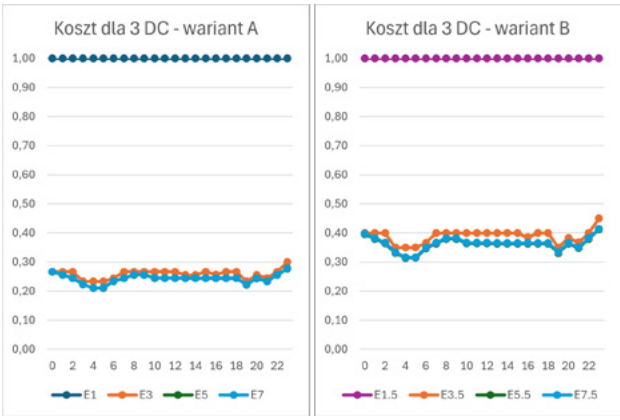
dotyczące skalowania wydajności, w odniesieniu do % całego ruchu, jaki musi obsłużyć 1 węzeł (wiersz "Skala" - przykładowo 100% oznacza, że każdy węzeł musi być w stanie obsłużyć cały ruchu w sieci). Wiersz "Dystr." oznacza czy badany scenariusz zakłada sterowanie rozpływem ruchu pomiędzy węzłami, czy statyczne skalowanie. Wiersz "Trm." informuje jakie są koszty transmisji ruchu w sieci - pod hasłem "niskie" rozumie się operatora, który posiada własną infrastrukturę transmisyjną, dla którego dodatkowa transmisja ruchu nie stanowi znacznego kosztu. Hasło "wysokie" oznacza operatora, który nie posiada własnej infrastruktury i ponosi wysokie koszta dzierżawy zasobów transmisyjnych od innych dostawców.

Ocenę efektywności badanych scenariuszy przeprowadzono na podstawie analizy porównawczej 2 parametrów. Pierwszym z nich, zaprezentowanym na Rys. 1, jest sumaryczna liczba instancji vFW uruchomionych na wszystkich węzłach sieci. W przypadku fizycznych FW, zastosowano wymiarowanie na określony w Tab. 2 procent całego ruchu w sieci. W przypadku vFW przedstawiono wyniki działania algorytmu dostosowującego liczbę instancji do napływającego ruchu sieciowego.



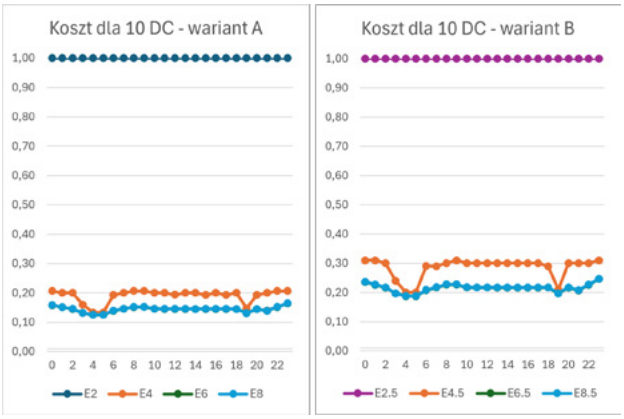
Rys. 1: Zmiana liczby instancji vFW w trakcie doby.

Drugim badanym parametrem, przedstawionym na Rys. 2 oraz Rys. 3, jest wartość funkcji celu (6) rozumianej jako koszt ponoszony przez operatora sieci. Równanie (6) zwraca wartości z zakresu $[0, 3]$, gdzie każdy składnik (koszt alokacji zasobów, koszt transmisji oraz koszt utraty ruchu) zwraca wartości w przedziale $[0, 1]$. W przypadku fizycznych zapór sieciowych, wartość została odwzorowana jako 100% kosztu alokacji oraz zerowy koszt transmisji i utraty ruchu. Wyniki przeprowadzonych analiz przedstawiono na dwóch wykresach, co umożliwia porównanie scenariuszy zakładających odpowiednio trzy oraz dziesięć węzłów DC/PoP. W obu przypadkach uwzględniono dwa warianty analizy: A oraz B.



Rys. 2: Porównanie kosztów dla 3 węzłów DC/PoP.

W wariancie A oceniano efektywność algorytmów przy założeniu wyższego poziomu przewymiarowania fizycznych zapór sieciowych, natomiast wariant B odnosił się do scenariuszy z niższym poziomem przewymiarowania. Taki podział ma na celu odzwierciedlenie różnych strategii planowania pojemności w sieciach operatorskich. Wariant A reprezentuje podejście zakładające wyższe nakłady inwestycyjne, umożliwiające przystosowanie infrastruktury do obsługi awarii wielokrotnej. Z kolei wariant B odzwierciedla strategię minimalizacji kosztów poprzez projektowanie zapór sieciowych z myślą o pojedynczym punkcie awarii. Eksperymenty oznaczone skrótami E1-E8 mają tożsame parametry z podanymi w Tab. 2, a eksperymenty oznaczone E1.5-E8.5 mają zmniejszony koszt transmisji.



Rys. 3: Porównanie kosztów dla 10 węzłów DC/PoP.

Na podstawie analizy Rys. 1–3 sformułowano następujące wnioski. Zastosowanie wirtualizacji funkcji sieciowych (vFW) w połączeniu z mechanizmem sterowania rozpływem ruchu umożliwia efektywniejsze wykorzystanie zasobów w sieci operatora. W przypadku tradycyjnych, fizycznych zapór sieciowych jest konieczne ich przewymiarowanie, aby uniknąć sytuacji przeciążenia i odrzucania ruchu w sytuacjach awaryjnych. Z kolei wariant oparty na vFW pozwala na elastyczne zarządzanie zasobami – alokowane są one dynamicznie zgodnie z bieżącym zapotrzebowaniem, a w razie nagłych zmian (np. awaria DC lub zmiany w rozkładzie ruchu) jest możliwe szybkie zwiększenie pojemności systemu. Zastosowanie mechanizmów sterowania rozpływem ruchu umożliwia dalsze obniżenie kosztów, ponieważ maksymalne przewymiarowanie jest ograniczone do jednej instancji vFW. W eksperymentach 3 i 4 zastosowano jedynie mechanizm skalowania — nowa instancja vFW była uruchamiana po wysyceniu poprzedniej. W eksperymentach 5–8 wprowadzono mechanizm sterowania rozpływem ruchu, który umożliwia przekierowanie obciążenia do węzłów z wolnymi zasobami. Takie podejście prowadzi do bardziej efektywnego wykorzystania infrastruktury oraz ogranicza liczbę wymaganych instancji vFW.

4. PODSUMOWANIE

W artykule zaproponowano nową metodę orkiestracji wirtualnych zapór sieciowych vFW, która umożliwia dostosowanie wydajności systemu vFW do aktualnego obciążenia wykorzystując mechanizm sterowania rozpływem ruchu. Przeprowadzone badania wykazały, że zastosowanie wirtualizacji zapór sieciowych może znacząco obniżyć koszty operacyjne, umożliwić współdzielenie zasobów infrastruktury wirtualnej (np. wykorzystanie wolnych zasobów nocą do realizacji kopii zapasowych) oraz zwiększyć odporność sieci na nieprzewidziane zdarzenia i ataki. Dalsze prace będą dotyczyły: 1) zwiększenia szczegółowości modelu uwzględniając m.in. wpływ rozmiarów pakietów, typu ruchu i liczby aktywnych sesji; 2) opracowania nowych modeli heurystycznych (w szczególności modeli sztucznej inteligencji), które pozwolą na uzyskanie porównywalnych wyników przy mniejszej złożoności obliczeniowej i krótszym czasie obliczeń; 3) oceny złożoności obliczeniowej metody oraz 4) eksperymentalnej weryfikacji proponowanego rozwiązania w ramach zaimplementowanego prototypu systemu.

LITERATURA

[1] D. Bringhenti, G. Marchetto, R. Sisto, F. Valenza, and J. Yusupov, *Automated optimal firewall orchestration and configuration in virtualized networks*, in NOMS 2020 - 2020 IEEE/IFIP Network Operations and Management Symposium, 2020, pp. 1–7.

[2] Cisco, *Cisco Adaptive Security Virtual Appliance (ASAv) Data Sheet*, 2021. Remote access (19.01.2025): https://www.cisco.com/c/en/us/products/collateral/security/adaptive-security-virtual-appliance-asav/adapt-security-virtual-appliance-ds.html.

[3] Fortinet, *FortiGate-VM on VMware ESXi*, 2024. Remote access (19.01.2025): https://www.fortinet.com/content/dam/fortinet/assets/data-sheets/FortiGate_VM_ESXi.pdf.

[4] V. Jain and A. Mitra, *Real-Time Threat Detection in Cybersecurity: Leveraging Machine Learning Algorithms for Enhanced Anomaly Detection*, 11 2024, pp. 315–344.

[5] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, *Survey of intrusion detection systems: techniques, datasets and challenges*, Cybersecurity, 2 (2019), p. 20.

[6] KPMG, *Barometr cyberbezpieczeństwa 2025*, tech. rep., KPMG, 2025.

[7] J. Pandey, S. Rai, and S. R, *Assessment of deep packet inspection system of network traffic and anomaly detection*, IJSRSET, (2023), pp. 680–688.

[8] R. R. Shanthi, N. K. Sasi, and P. Gouthaman, *A new era of cybersecurity: The influence of artificial intelligence*, in 2023 International Conference on Networking and Communications (ICNWC), 2023, pp. 1–4.

[9] S. Yang, F. Li, et al., *Recent advances of resource allocation in network function virtualization*, IEEE Transactions on Parallel and Distributed Systems, PP (2020), pp. 1–1.

POLITECHNIKA GDAŃSKA

WYDZIAŁ ELEKTRONIKI, TELEKOMUNIKACJI I INFORMATYKI

WWW.ETI.PG.EDU.PL

KRiT 2025

GDAŃSK,
10 - 12 WRZEŚNIA 2025

Rubinowy Sponsor

Złoty Sponsor

Sponsor „Złotych Cyborgów"

Partnerzy honorowi

Partnerzy medialni