KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

Poprawa testów oprogramowania stacji bazowych 5G dzięki analizie opartej na dużych modelach językowych LLM (Large Language Models)

Enhancing software testing of 5G base stations with LLM-driven analysis

Sebastian Zarębski [1,2]; Krzysztof Rusek [3]; Piotr Chołda [4]

[1] NOKIA, Kraków, sebastian.zarebski@nokia.com
[2] AGH University of Krakow, Institute of Telecommunications, Kraków, zarebski@agh.edu.pl
[3] AGH University of Krakow, Institute of Telecommunications, Kraków, krzysztof.rusek@agh.edu.pl
[4] AGH University of Krakow, Institute of Telecommunications, Kraków, piotr.cholda@agh.edu.pl

DOI: 10.15199/59.2024.4.7

Streszczenie: Duże modele językowe (LLM) umożliwiły osiąganie niespodziewanie dobrych rezultatów w generowaniu, tłumaczeniu i rozumieniu tekstu. Modele te mogą generować tekst w oparciu o ich dostosowanie do struktur ludzkiego języka i wymagania bądź żądania użytkownika. W niniejszym artykule we współpracy z firmą NOKIA zbadano wpływ specjalistycznych treści z dziedziny telekomunikacyji na model Gemma-2B. Wyniki wskazują na celowość zastosowania podejścia z oszacowaniem perpleksji w treściach o unikalnej i charakterystycznej zawartości, interesującej z punktu widzenia zastosowań telekomunikacyjnych (np. dokumentacja 3GPP). Wyniki pokazują, że odpowiednie dostrojenie modelu może zmniejszyć jego perpleksję średnio o 83% (w zależności od tematu).

Abstract: Large language models (LLMs) have transformed natural language processing, enabling remarkable achievements in text generation, translation, and comprehension. LLMs can generate text based on their intrinsic understanding of structures in human language and the user's request. This paper, reporting results obtained in collaboration with NOKIA, examines the impact of specialized telecom content on the Gemma-2B model. The findings indicate a novel approach to using perplexity assessment in the specialized field of telecommunications (3GPP-related content). It was shown that proper fine-tuning can reduce the model's perplexity by an average of 83%, depending on the topic.

Słowa kluczowe: NLP, LLM, testowanie oprogramowania, testy regresyjne

Keywords: NLP, LLM, regression testing, software testing

1. INTRODUCTION

Changing the software of 5G base stations carries the risk of introducing a bug. The tester who discovers the problem is required to provide an accurate technical description of the finding. Based on the prepared documentation, a group of developers will be selected to fix the identified bug. The quality and clarity of the description affect the understanding of the root cause of the bug and its faster repair. Understanding the content of software engineering and telecommunications is a challenge. Depending on the experience and understanding of the functionality of the selected part of the base station, the described bug contains extremely exclusive language, full of technical jargon, abbreviations, and acronyms (which appear in the 3GPP documentation). The challenge then becomes to automate the process of selecting the right group of developers based on the documentation presented [1]. An interesting option seems to be application for this task so recently popular machine learning tools, namely large language models (LLMs). These networks are trained on a corpus of a large number of texts (books, articles, web content). These models have the ability to produce text by understanding the likelihood of sequences of words within a given general context. This enables a technical process where the engineers' written content is evaluated by a chosen LLM model, which then assigns it to one of the designated programmer groups. To assess the mutual comprehension between humans and machines, that is, to determine how well the LLM can understand the structure of the provided text, the perplexity metric is employed [2]. Therefore, it is possible to examine how content not found in open data sources influences the model's performance. In partnership with NOKIA, the effect of specialized telecom content on the Gemma-2B [3] model was assessed. One of the research tasks included fine-tuning the model and comparing the outcomes.

We give the following structure to this paper. In Section 2, we describe the background of the problem and the implications associated with it. In Section 3, we describe the methods we used. In Section 4, we present our results and findings and discuss them in relation to possible future work. In Section 5, we conclude our research.

2. PROBLEM STATEMENT

The LLM model's task is to review the bug description contents and determine the correct bug topic group. These models operate on the MLM principle (masked language

model), meaning they are built to produce content based on the provided text. Each bug is formatted in a controlled style using the *outlines* library [7]. The example of such a sample is as follows:

"`product name: 5G base station \n problem summary: [eCPRI] Invalid memory reference (reason 1 'address not mapped to object') detected by kernel at address 'callproc_realtime_ue' \n label: radio heads and communication`"

Language models are designed to predict the probability distribution of the next word in a sentence based on the preceding words. This is expressed as $P(x_n|x_{n-1},..,x_1)$, where x represents a word in the sentence. These models are known as autoregressive. One characteristic of these models is that their structure aligns with a particular breakdown of the joint probability of the sequence, which is derived from the chain rule of probability and is consistently valid:

$$\prod_{n=1}^{N} P(X_n|X_1^{n-1}) \tag{1}$$

The model is trained using the conventional categorical cross-entropy loss function. Therefore, it can be interpreted as the logarithm of the joint probability of the sequence.

$$\ell_{\mathrm{CE}} = -\frac{1}{N}\sum_{n=1}^{N} \log p_\theta(x_n|x_{<n}) \tag{2}$$

Given a tokenized sequence $X = (x_1, x_2, \ldots, x_n)$, we can calculate the logarithmic probability that each word occurs within a specific context (sequence). Typically, this involves computing the negative log-likelihood, which is equivalent to maximizing the likelihood. The metric known as **perplexity** is then defined as the mean negative log-likelihood of a sequence [5].

$$\textbf{Perplexity}(X) = PPL(X) = \exp\{\ell_{\mathrm{CE}}\} \tag{3}$$

The logarithm $\log p_\theta(x_n|x_{<n})$ represents the log-likelihood of the n-th token given the previous tokens. Essentially, it evaluates the model's capability to predict consistently over the tokens defined in a corpus.

Perplexity is a metric that is highly influenced by the training data of the model. Consequently, perplexity scores cannot be directly compared across different models or datasets [6]. For instance, it is not feasible to compare the score from an original research paper with that of your own dataset. Comparisons must be made within the same corpus. Thus, the issue pertains to assessing the variation in perplexity for a specific corpus between the base model and its fine-tuned counterpart.

Cross-entropy measures the divergence between two probability distributions, indicating how much a given sequence deviates from the probability distribution modeled by the LLM. A higher cross-entropy signifies a greater difference between the given text and the model-generated text. Since the exponential function is monotonically increasing, perplexity behaves similarly. Higher perplexity indicates poorer performance of the language model in predicting the sequence. For the Gemma-2B model, a high perplexity value should be observed when compared to a corpus related to telecommunications. Fine-tuning the model is expected to reduce the final perplexity.

The Gemma-2B was selected because it is the smallest LLM available, making it attractive due to its cost-effectiveness, computational efficiency, and user-friendliness [4]. This model can operate on low- to mid-range GPU cards that support CUDA, making it viable from a business standpoint.

3. METHODS

A collection of several thousand of the latest bugs of 5G base stations software (based on their occurrence time) was utilized to examine the effect of unique content on the performance of the chosen LLM model. The origins of these bugs vary from syntax errors or code typos, through issues related to the natural expansion of existing features, to intricate and hard-to-understand causes (e.g., a communication error on the radio head interface resulting in a window freeze in the operator-side GUI). This content has been divided into thematically related groups:

1. specification of base station,
2. motherboard and its interfaces,
3. user interface and base station configuration,
4. network routing and management,
5. radio heads and communication

Every sample is tokenized using a pretrained tokenizer supplied by the model creators. The encoded sample is then quickly processed through the model to generate logits. The following step involves comparing the input tokens (referred to as 'labels') with the logits using the cross-entropy loss function. The resulting value is then averaged by the sentence length (number of tokens) and is exponentiated, as defined in Eq. (3).

The last step is to fine-tune the original model to given dataset, to move distributions of probabilities of input sentences and these that could be generated by a LLM. To make operation less demanding for the used GPU, the model is to be loaded in the quantized form. Quantization involves representing model weights using fewer bits. For instance, reducing weights from FP16 (16-bit floating-point) to 4-bit Normal Float (NF4) allows for more efficient execution. The fine-tuning process utilizes a method known as Low-Rank Adaptation (LoRA) [8]. The key aspect is to keep the pre-trained model weights fixed and incorporate trainable rank decomposition matrices into each layer of the LLM architecture, significantly lowering the number of trainable parameters. Subsequently, as each data sample is annotated, the reinforcement learning variant (RL) can be applied. Reinforcement learning from human feedback (RLHF) is a form of supervised fine-tuning (SFT) commonly used when working with LLMs [9].

4. EMPIRICAL FINDINGS

Telecommunications-related content appeared to be misunderstood by the model to varying extents. The simplest text example, with $PPL = 5.10$, was a bug summary devoid of excessive jargon or technical abbreviations: "`Customer reported serious vulnerability Deprecated SSH cryptographic settings on ssh tcp22 to be fixed`". The content of the sample pertains to general security information, which is extensively documented in open-source texts. The most challenging error was described as: "`deploy failed in step`

`monitor-installation with 'Software installation failed with exit status non-zero'"`, with $PPL = 2866$. This indicates that the first example is much closer to being generated by the chosen model than the second. Therefore, it is more likely to appear in the probability distribution known by the model. The form of writing of the second example suggests that it was created artificially, e.g. during the generation of the log of the device's operation, during which various automatic scripts compile sentence fragments into their completed form. Consequently, such samples are from a different distribution than the corpus on which the LLM was trained. This poses a significant issue when using such a model in any automated process.

Table 1: Comparison of *PPL* values for original and fine-tuned LLM for thematic groups of software bugs

Group	*Original*	*Fine-tuned*	$\Delta(\%)$
1	192.69	33.58	82.57
2	267.93	39.43	85.28
3	234.70	33.58	86.64
4	**292.06**	**61.99**	**78.77**
5	274.53	45.99	83.24

As can be seen in in Table 1. The computed perplexity metrics reveal insights into the model's inclinations and biases – by evaluating it with the initial weights, one can identify the topic that is closest to the knowledge the LLM possesses. For this study and the utilized dataset, the first topic, concerning base station specifications, comprises text samples most similar in form and meaning to the model's knowledge. Knowledge of the topic related to network routing and management turned out to be the most distant from that familiar to the model. It proved to be nearly as obscure or difficult to understand as the content related to internal product knowledge, which includes numerous hidden or internal abbreviations or acronyms. The substantial 80% reduction in perplexity due to the fine-tuning process provides direct evidence that the LLM model has successfully acquired a new token probability distribution for the given samples. The observations are backed by density functions, indicating that for each thematic group, the perplexity mass shifted to the left, approaching zero. This shift is evident in Figure 1 (group 3, with the highest change) and less in Figure 2 (group 4, with the smallest change).

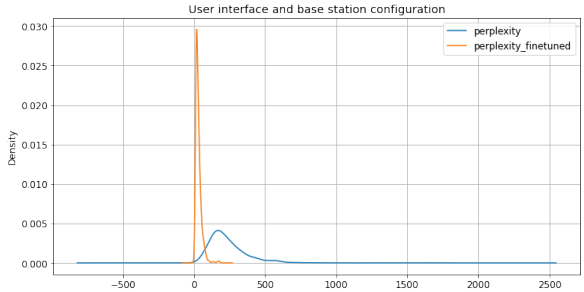


Figure 1: Density of perplexity for thematic Group 3. The highest observed change.

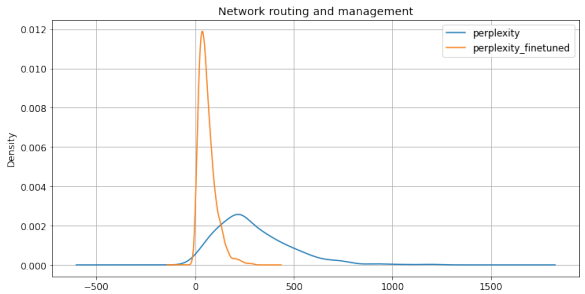


Figure 2: Density of perplexity for thematic Group 4. The lowest observed change.

As the model learns the new probability distribution of the given corpus, there is an opportunity to apply it to other tasks. One of these is to classify the given content of a software bug into a related topic. Using the original model one can see its tendency to make biased predictions, by picking mostly a topic 5 (radio heads and communication). This is noticeable in the Figure 3. The model accuracy was merely 20%. The fine-tuned model showed greater sensitivity to the context contained in the error descriptions. The rate at which individual classes are selected has also shifted, as illustrated in Figure 4. One subject is no longer preferred. Nevertheless, the sensitivity and accuracy of the identified topics were quite low, leading to an accuracy of 30% (+10%). It is evident that despite the lower perplexity, the model continues to face significant epistemic uncertainty associated with the dataset, which results in frequent classification errors.

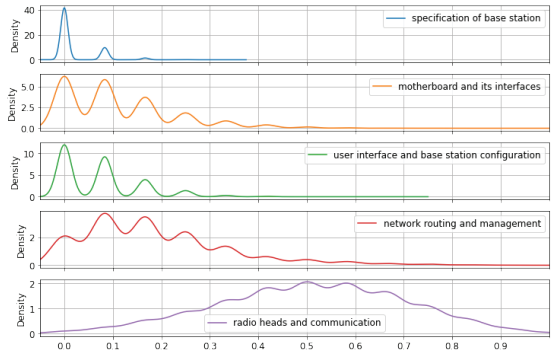


Figure 3: Density of probability of assigning each software bug to respective topic by original model

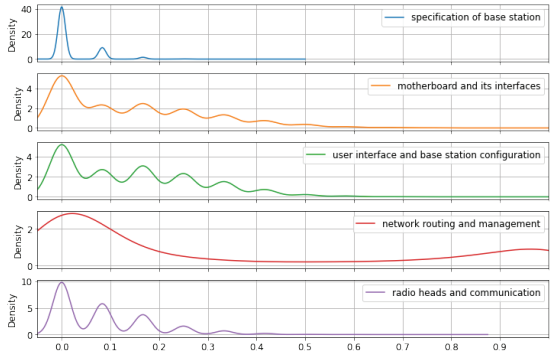


Figure 4: Density of probability of assigning each software bug to respective topic by fine-tuned model

5. CONCLUSION AND GUIDELINES

Large Language Models (LLMs) have transformed natural language processing, allowing for remarkable achievements in text generation, translation, and comprehension. However, in highly specialized contexts that fall outside of their training data, there are benefits and drawbacks that must be taken into account.

As demonstrated, LLMs can be adapted to specific datasets to enhance their performance in particular scenarios. This flexibility makes them useful for specialized applications. The vast amount of pre-training data enables LLMs to replicate human-like language patterns, even in specialized fields (as evidenced by the variation in perplexity across different topics). However, LLMs do not truly understand context. They predict the next token based on patterns in their training data, and even after fine-tuning, some token combinations remain unfamiliar to the model. Additionally, without external knowledge, LLMs might produce content that sounds plausible but is factually incorrect (e.g., when generating labels, they may alter original names into similar forms). Another point to note is that LLMs handle each prompt as an independent request. They do not retain a history of previous interactions, so each prompt related to a different software bug is treated as a separate event.

The findings indicate a novel approach to utilizing perplexity measurement in the specialized field of telecommunications. Engineers can set confidence thresholds, to find the point at which an evaluation is made as to whether a given textual content is sufficiently understood by the model. If an LLM's response shows low confidence (i.e., the perplexity of a new sample significantly exceeds the mean or another accepted threshold), the input should be reconsidered. Otherwise, external validation or alternative methods should be explored. This establishes a human-in-the-loop scenario, in which a human partially oversees the automated process, reducing the chances of LLM hallucination or sample misclassification.

BIBLIOGRAPHY

[1] Zarębski, S.; Kuzmich, A.; Sitko, S.; Rusek, K.; Chołda, P. Siamese Neural Networks on the Trail of Similarity in Bugs in 5G Mobile Network Base Stations. Electronics, 11, 3664, 2022.

[2] Miaschi, A., Brunato, D., Dell'Orletta, F. & Venturi, G. What Makes My Model Perplexed? A Linguistic Investigation on Neural Language Models Perplexity. *Proceedings Of Deep Learning Inside Out (DeeLIO): The 2nd Workshop On Knowledge Extraction And Integration For Deep Learning Architectures*. pp. 40-47, 2021.

[3] https://huggingface.co/google/gemma-2b visited at 27.06.2024

[4] Thomas Mesnard et al. "Gemma: Open Models Based on Gemini Research and Technology." arXiv preprint arXiv:2403.08295, 2024.

[5] Jelinek, Fred, Robert L. Mercer, Lalit R. Bahl, and James K. Baker. "Perplexity: A Measure of the Difficulty of Speech Recognition Tasks." The Journal of the Acoustical Society of America, 62(S1): S63, 1977.

[6] Meister, Clara, and Ryan Cotterell. "Language Model Evaluation Beyond Perplexity." In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 5328–5339. Online: Association for Computational Linguistics, 2021.

[7] Willard, Brandon T., and Rémi Louf. "Efficient Guided Generation for LLMs." arXiv preprint arXiv:2307.09702, 2023.

[8] Hu, Edward J. et al. "LoRA: Low-Rank Adaptation of Large Language Models." In Proceedings of the International Conference on Learning Representations (ICLR), 2022.

[9] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. "Deep reinforcement learning from human preferences". In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), 2017.

KONCEPCJA INTERFEJSU RADIOWEGO OPARTEGO NA UCZENIU MASZYNOWYM W PROJEKCIE MLDR

THE CONCEPT OF A MACHINE LEARNING-DRIVEN RADIO INTERFACE IN THE MLDR PROJECT

Kamil Szczęch; Szymon Szott

AGH Akademia Górniczo-Hutnicza, Kraków, {imię.nazwisko}@agh.edu.pl

DOI: 10.15199/59.2024.4.8

Streszczenie: Sieci bezprzewodowe chcąc spełnić wymagania współczesnych aplikacji, stają się niezwykle złożone co utrudnia ich konfigurację. Aby poradzić sobie z tą złożonością, można użyć technik uczenia maszynowego. Niniejsza praca opisuje koncepcję przyjętą w projekcie MLDR dotyczącą opracowania i wdrożenia interfejsu radiowego sterowanego za pomocą technik uczenia maszynowego. Opisujemy koncepcję architektury, która natywnie wspiera uczenie maszynowe, rozważane przypadki użycia, oraz przykład zastosowania opracowywanej koncepcji.

Abstract: To meet the requirements of modern applications, wireless networks are becoming extremely complex, which makes their optimal configuration challenging. Machine learning (ML) techniques can be used to address this challenge. This work describes the concept adopted in the MLDR project regarding the development and implementation of a radio interface controlled using machine learning techniques. We describe the architecture concept, which natively supports ML, the use cases considered, and an example of the application of the developed concept.

Słowa kluczowe: lokalne sieci bezprzewodowe, optymalizacja, uczenie maszynowe, zarządzanie zasobami

Keywords: machine learning, optimization, resource management, wireless local area networks

1. WSTĘP

Z dnia na dzień wymagania dotyczące łączności bezprzewodowej są coraz większe. Jest to spowodowane wprowadzaniem na rynek nowych aplikacji internetowych wymagających wysokich przepływności, małych opóźnień i dużej niezawodności. Przykładami takich rozwiązań może być Metaverse firmy Facebook czy rozwiązania holograficznej teleobecności (ang. *holographic telepresence*). Tymczasem rozwój przemysłu i Internetu rzeczy (ang. *Internet of Things*, IoT) stawia nieznane wcześniej wyzwania, którymi są inteligentne systemy, które korzystają z danych tysięcy niezależnych od siebie czujników. By sprostać takim wymaganiom, interfejsy radiowe stają się coraz bardziej skomplikowane, ponieważ korzystają z ogromnej liczby funkcjonalności i parametrów, co skutkuje problemami optymalizacyjnymi. Złożoność tych systemów sprawia, że coraz częściej brane są pod uwagę rozwiązania oparte na technikach sztucznej inteligencji i uczenia maszynowego (ang. *machine learning*, ML), które mają optymalizować parametry sieci.

Aby sprostać tym wyzwaniom, można rozważać, przez analogię do sieci sterowanych programowo, koncepcję interfejsu radiowego sterowanego przez uczenie maszynowe (ang. *machine learning-driven radio interface*, MLDR), który byłby wolny od ograniczeń dotychczasowych technik bezprzewodowych. Zadaniem takiego interfejsu jest nauczenie się komunikowania, wybór i konfiguracja protokołów komunikacyjnych i tych funkcjonalności, które najlepiej pasują do konkretnego przypadku użycia i scenariusza, spełniając w ten sposób wymagania nowych aplikacji Internetowych i wydajnie wykorzystując dostępne zasoby radiowe. Innymi słowy, interfejs MLDR powinien umożliwić wdrażanie "lokalnych sieci bezprzewodowych szytych na miarę".

Celem projektu badawczego "MLDR: Interfejs radiowy oparty na uczeniu maszynowym" jest stworzenie interfejsu radiowego sterowanego przy pomocy ML. Projekt jest finansowany przez Narodowe Centrum Nauki (NCN) w ramach międzynarodowego konkursu CHIST-ERA Call 2022 "Machine Learning-based Communication Systems, towards Wireless AI". W skład konsorcjum, oprócz AGH w Krakowie, wchodzą: lider konsorcjum – Universitat Pompeu Fabra (Hiszpania), CentraleSupélec (Francja) oraz University of Oulu (Finlandia). Wszyscy konsorcjanci pracują naukowo na styku sieci bezprzewodowych i obszarów uczenia maszynowego, dysponując uzupełniającą się wiedzą specjalistyczną. Projekt został zaplanowany na lata 2023-2026, a wstępna wersja pierwszego artykułu naukowego będąca efektem realizacji projektu jest już dostępna [1].

Celem niniejszej pracy jest przybliżenie założeń badawczych przyjętych w projekcie oraz pokazanie możliwości zastosowania opracowywanych rozwiązań. Najpierw, w rozdziale 2, opisujemy literaturę dotyczącą stosowania uczenia maszynowego do zarządzania interfejsami bezprzewodowymi. Następnie przedstawiamy koncepcję architektury interfejsu MLDR (rozdział 3) oraz, w rozdziale 4, kilka rozważanych przypadków użycia: dla sieci zorientowanych zadaniowo, ultra-szybkich łączy i adaptacyjnych sieci dostępowych. W rozdziale 5 opisujemy możliwy przykład zastosowania interfejsu MLDR, który obecnie wdrażamy. Na koniec, w rozdziale 6, podsumowujemy niniejszą pracę i wyznaczamy dalsze kierunki badań.

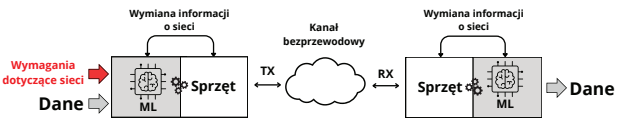
2. PRZEGLĄD LITERATURY

Idea elastycznego interfejsu radiowego sięga koncepcji radia kognitywnego (ang. *cognitive radio*, CR), które zostało

zaproponowane w roku 1999 [5]. Z czasem koncepcja CR skupiła się na efektywnym wykorzystaniu widma radiowego poprzez dynamiczny dostęp do czasowo nieużywanych pasm częstotliwości [7, 8]. Niedawny przegląd literatury z zakresu stosowania ML w CR potwierdza, że głównym problemem badawczym jest zarządzanie widmem (ang. *spectrum management*) [4]. Specyficzną cechą koncepcji MLDR jest skupienie się na dwóch dolnych warstwach (fizycznej i łącza danych). W literaturze pojawiają się też propozycje obejmujące więcej aspektów, np. MetaRadio [8], ale wydają się one trudniejsze do wdrożenia ze względu na konieczność zmiany całego systemu komunikacyjnego, a nie jedynie pojedynczych interfejsów.

Obecnie zastosowanie uczenia maszynowego w sieciach bezprzewodowych zmierza w kierunku sieci natywnie korzystających z ML, tj. zaprojektowanych z myślą o użyciu metod uczenia maszynowego [9]. Koncepcja optymalizowania parametrów sieci podobna jak w MLDR, została przedstawiona dla sieci komórkowych [2]. Również w kontekście sieci IEEE 802.11 w ostatnich latach coraz więcej badaczy zajmuje się użyciem ML do adaptacji i optymalizacji działania sieci Wi-Fi [6]. Jednym z tych podejść jest praca [3], która porusza temat optymalizacji warstwy łącza danych (MAC) w sieciach Wi-Fi. Celem pracy jest sprawdzenie możliwości zastosowania uczenia ze wzmocnieniem (ang. *reinforcement learning*, RL) do optymalizacji wielu parametrów. Autorzy zwracają uwagę, że dotychczasowe prace skupiały się na optymalizacji tylko jednego parametru, co potwierdza przegląd literatury [6]. Wybrane parametry sieci IEEE 802.11, których wartości optymalizował agent RL, to czas trwania slotu, wielkość okna współzawodnictwa (ang. *contention window*, CW) oraz poziom modulacji i kodowania. Algorytm także ustalał czy niektóre funkcjonalności IEEE 802.11 będą włączone czy wyłączone, np. wysyłanie ramek kontrolnych RTS/CTS (Request to Send/Clear to Send), wykrywanie nośnej, a także mechanizm odraczania transmisji (ang. *backoff*). Optymalizacja działania sieci odbywa się na podstawie dwóch oddzielnych kryteriów: opóźnień pakietów i zagregowanej przepływności sieci. Otrzymane wyniki porównano z wynikami sieci działającej według rozszerzenia IEEE 802.11ac. We wszystkich testach sieć korzystająca z optymalizacji MAC za pomocą agenta RL uzyskała lepsze wyniki od tradycyjnej sieci, co stanowi motywację do pracy nad bardziej ogólnymi rozwiązaniami, które optymalizują zarówno warstwę łącza danych jak i warstwę fizyczną lokalnych sieci bezprzewodowych.

3. KONCEPCJA ARCHITEKTURY



Rysunek 1: Koncepcja architektury MLDR.

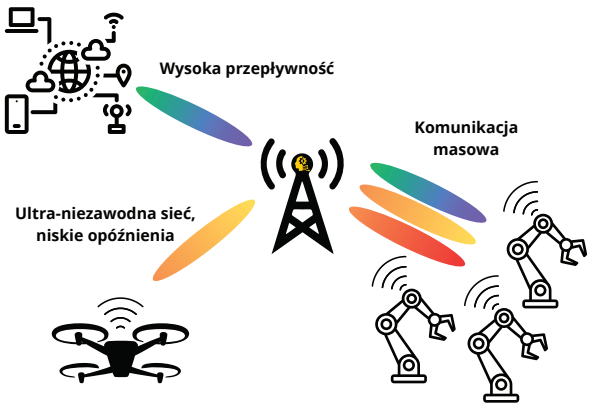
Cele stawiane przed MLDR wymagają specyficznej architektury, która różni się od tradycyjnych systemów łączności. Główną różnicą będzie możliwość adaptacji do nowych warunków poprzez użycie technik uczenia maszynowego. Dotychczasowe rozwiązania były zdane na wcześniej napisane heurystyki ze ściśle ustalonymi parametrami. W przypadku MLDR, kontrolę nad parametrami i funkcjonalnościami interfejsu ma sprawować agent RL, wykorzystujący techniki uczenia ze wzmocnieniem. Agent ma być w ciągłym kontakcie ze sprzętem by otrzymywać informacje na temat obecnego stanu sieci, ale także by przekazywać nowe parametry sieci. Koncepcja architektury systemu MLDR została przedstawiona na rysunku 1. Agenci RL stale zarządzają zarówno warstwą fizyczną jak i warstwą łącza danych na podstawie informacji ze środowiska ale także otrzymanych w procesie odbierania wiadomości (RX). Kolejną zmianą względem tradycyjnych interfejsów bezprzewodowych byłaby potrzeba zastosowania dedykowanych akceleratorów ML, które potrafią wydajniej wykonywać działania związane z uczeniem maszynowym niż standardowo używane procesory. Co więcej układy te zużywają mniej energii do obliczeń, co jest dużym atutem przy rozwiązaniach IoT.

4. PRZYPADKI UŻYCIA

W ramach projektu rozważane są trzy podstawowe przypadki użycia, w których zastosowanie interfejsu MLDR może poprawiać wydajność sieci. Dotyczą one sieci zorientowanych na zadanie, ultra-szybkich łączy oraz adaptacyjnych sieci dostępowych. Poniżej pokrótce omawiamy te trzy przypadki.

4.1 Sieć zorientowana na zadanie

Sieci zorientowane na zadanie stawiają nacisk na optymalizację w zakresie szczególnego celu. Przykładem może być sieć przemysłowa. Sieć taka może zostać wyspecjalizowana w kierunku niskich opóźnień, wysokich przepływności lub dużej liczby użytkowników. Każde z tych wymagań potrzebuje użycia innego zbioru funkcjonalności i parametrów. Dodatkowo wymagania na przestrzeni czasu mogą ewoluować. W takim scenariuszu zadaniem interfejsu MLDR byłaby adaptacja sieci do nowych wymagań użytkownika. Większa elastyczność sieci i prostota konfiguracji mogłaby znacząco zmniejszyć koszty utrzymania sieci przemysłowej, a także pozwoliłaby na wprowadzenie innowacyjnych zastosowań tak ważnych w przemyśle 4.0. Rysunek 2 przedstawia graficzną koncepcję takiego rozwiązania.

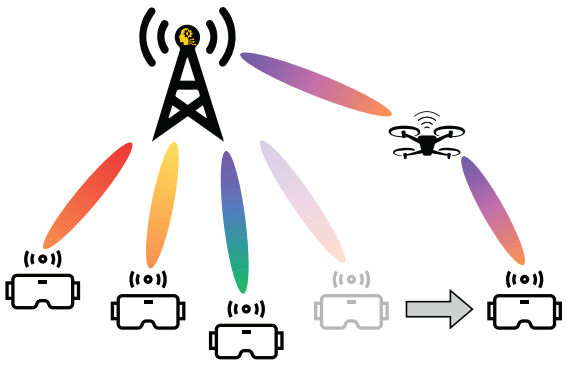


Rysunek 2: Przykładowa sieć zorientowana na zadanie.

4.2 Ultra szybkie łącza

Główną cechą definiującym sieci o ultra szybkich łączach są wysokie przepływności osiągane poprzez szerokie pasmo. Sieci te nazywane często jako sieci 6G wykorzystują

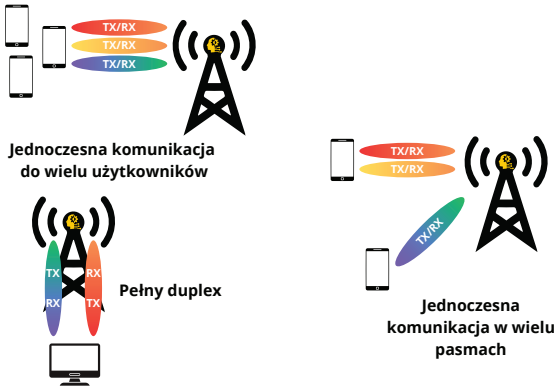
fale THz. Dodatkowo wykorzystują odbiorniki i nadajniki składające się z ogromnej liczby małych anten. Głównym problemem w przypadku tego scenariusza jest mały zasięg sieci oraz alokacja zasobów dla wielu użytkowników. Odpowiedzią na te problemy może być kształtowanie wiązki, a także użycie tak zwanych metapowierzchni (ang. *metasurface*). Nadajnik/odbiornik wspierany algorytmami ML mógłby nauczyć się jak lepiej kształtować wiązkę i uzyskać transmisję z innym użytkownikiem, pomimo braku warunków bezpośredniej widoczności (ang. *line of sight*). Ponadto, ML mógłby wspierać proces przetwarzania zarówno nadawanego jak i odbieranego sygnału, zmniejszając tym samym narzut sieci. Przykład takiego scenariusza użycia został przedstawiony na rysunku 3.



Rysunek 3: Przykład sieci o ultra szybkich łączach, koniecznych do realizacji usług wirtualnej i rozszerzonej rzeczywistości.

4.3 Adaptacyjne sieci dostępowe

Współczesne bezprzewodowe sieci lokalne wdrażane są tak, aby punkty dostępowe (ang. *access points*) były blisko użytkownika, co skutkuje silniejszym sygnałem, a zatem możliwość uzyskania wyższych prędkości transmisji. Niestety jednak prowadzi to do problemu interferencji pomiędzy sieciami. To wyzwanie mogłoby być także rozwiązane z użyciem ML. Interfejs MLDR mógłby w procesie nauki adaptować się do obecnych warunków i w sposób dynamiczny dobierać funkcjonalności, takie jak jednoczesna komunikacja wielu użytkowników (ang. *multi-user communication*) czy jednoczesne komunikacja w wielu pasmach (ang. *multi-link operation*), w celu minimalizacji zakłóceń. Przykład sieci adaptacyjnej przedstawiono na rysunku 4.



Rysunek 4: Przykład adaptacyjnej sieci dostępowej.

5. PRZYKŁAD ZASTOSOWANIA MLDR

Możliwość zastosowania koncepcji MLDR opisujemy na przykładzie związanym z sieciami zorientowanymi na zadanie (rozdział 2). Zakładamy, że operator sieci przemysłowej posiada wymagania wobec wdrażanej sieci bezprzewodowej, które może opisać jako punkt na „trójkącie MLDR" (rysunek 5). Następnie agent RL będzie konfigurował interfejsy MLDR w sieci w oparciu o te wymagania. Tabela 1 przedstawia przykładowe wymagania dla skrajnych rogów trójkąta. Każdy ze scenariuszy zakłada maksymalizację jednego z parametrów kosztem innych. Przykładowo, dla scenariusza masowej komunikacji najważniejsze dla operatora jest maksymalizacja liczby obsłużonych urządzeń kosztem przepustowości sieci i opóźnień. Przykładem takiego rozwiązania mogą być sieci sensorowe, gdzie występuje setki urządzeń, które posiadają wymagania co do dostępu do kanału radiowego, lecz uzyskanie wysokiej przepływności nie jest istotne. Pozostałe scenariusze umieszczone w rogach trójkąta także zakładają maksymalizację wyznacznika sieci takiego jak wysoka przepustowość kosztem liczby użytkowników i opóźnień w sieci. Wszystkie scenariusze zakładają, że sieć jest quasi-stacjonarna co oznacza niezmienność położenia czy wielkości generowanego ruchu. Ponadto zakładamy, że ruch jest homogeniczny tzn. każda stacja generuje ruch o podobnych parametrach. Końcowe wyniki planujemy porównać z wynikami osiąganymi przez sieć pracującą w aktualnym standardzie IEEE 802.11.

Rysunek 5: Usługi w sieci bezprzewodowej z interfejsami MLDR opisane na planie trójkąta inspirowanego standardem ITU-R Rec. M-2083. W scenariuszu MLDR dla sieci zorientowanych na zadanie zakładamy, że operator definiuje wymagania poprzez wyznaczenie punktu w obrębie trójkąta, co daje odpowiednią wagę dla każdej z metryk powiązanych z usługami.

Agent RL, zarządzający interfejsami MLDR, będzie optymalizował parametry interfejsów aby spełnić stawiane wymagania. Podobnie do [3], w pierwszym podejściu ograniczamy podzbiór parametrów, które mogą być zmieniane, do następujących parametrów i funkcjonalności typowych dla lokalnych sieci bezprzewodowych standardu IEEE 802.11:

- wartość okna współzawodnictwa CW,
- włączenie lub wyłączenie mechanizmu ochrony ramek RTS/CTS,
- włączenie lub wyłączenie wykrywania nośnej,
- włączenie lub wyłączenie agregacji ramek.

Tabela 2 zawiera przykładowe wartości, jakich możemy spodziewać się w skrajnych przypadkach (na rogach trójkąta), tj. zakładamy, że agent MLDR dzięki uczeniu on-

Tabela 1: Charakterystyka trzech usług MLDR oraz główne metryki będące celem uczenia agenta RL.

Usługa MLDR	**Użytkownicy**	**Ruch oferowany**	**Opóźnienie**	**Cel uczenia RL**
Zwiększona przepływność	1-5	Maksymalny	Brak wymagań	Przepustowość
Ultra-niezawodność, niskie opóźnienia	1-10	Niski	Bardzo niskie	Opóźnienie, współczynnik straconych pakietów
Obsługa ogromnej liczby urządzeń	>100	Niski	Brak wymagań	Liczba obsłużonych urządzeń

Tabela 2: Parametry zarządzane przez agenta RL, ich przewidywane ustawienia dla trzech usług MLDR oraz przestrzeń możliwych do skonfigurowania wartości.

Parametr	**Usługa MLDR**			**Wartości**
	Wysoka przepływność	Ultra-niezawodna sieć, niskie opóźnienia	Komunikacja masowa	
Okno współzawodnictwa CW	Niskie	Zerowe	Wysokie	2^{a-1}, $a \in \{0 \dots 10\}$
Mechanizm RTS/CTS	Wyłączony	Wyłączony	Włączony	Binarne
Wykrywanie nośnej	Włączone	Wyłączone (ALOHA)	Włączone	Binarne
Agregacja ramek	Włączona	Włączona/wyłączona	Wyłączona	Binarne

line nauczy się, że (prawdopodobnie) takie wartości będą najlepiej spełniały wymagane kryteria.

Prócz skrajnych przypadków, projekt zakłada także scenariusze, gdzie operator będzie mógł wybrać dowolny punkt w obszarze trójkąta. Współrzędne punktu będą opisane w trójwymiarowym systemie, gdzie suma składowych wynosi jeden. Każda współrzędna będzie opisywać jak bardzo agent ma skupiać się na optymalizowaniu danego współczynnika jakości sieci. Przykładowo, wymaganiami użytkownika byłaby sieć o znacznej liczbie użytkowników oraz o niskich opóźnienia i wysokiej niezawodności. Przepustowość tej sieci miałaby drugorzędne znaczenie. W takim przypadku, współrzędne mogłyby przybrać następujące wartości:

- 0,45 współczynnika masowości sieci,
- 0,45 współczynnika niezawodności sieci,
- 0,1 współczynnika przepustowości.

Największym wyzwaniem stojącym na drodze do budowy tak elastycznego interfejsu jest odpowiednia normalizacja tych współrzędnych względem wymagań stawianych przed agentem. Obecnie trwają prace nad wdrożeniem proponowanego rozwiązania w symulatorze i przy użyciu otwartoźródłowej biblioteki Reinforced-lib [10], co umożliwi walidację proponowanego rozwiązania.

6. PODSUMOWANIE

Kolejnym krokiem w zastosowaniu uczenia maszynowego dla lokalnych sieci bezprzewodowych jest tworzenie rozwiązań z wbudowanym wsparciem dla ML. W ramach projektu MLDR jest opracowywany interfejs, który może nauczyć się odpowiedniego doboru mechanizmów warstwy fizycznej i łącza danych, umożliwiając koncepcję „lokalnych sieci bezprzewodowych szytych na miarę". W niniejszej pracy przedstawiliśmy koncepcję architektury MLDR, przypadki użycia oraz omówiliśmy wdrażane obecnie przykładowe rozwiązanie. W toku dalszych prac planujemy doprecyzować opisane przypadki użycia oraz architekturę interfejsu, jak również dokonać abstrakcji podstawowych bloków komunikacyjnych. Ostatecznym efektem projektu będzie implementacja w formie *proof-of-concept* oraz określenie wydajności takiego rozwiązania.

PODZIĘKOWANIA

Praca powstała w wyniku realizacji projektu badawczego o nr DEC-2023/05/Y/ST7/00004 finansowanego ze środków Narodowego Centrum Nauki.

LITERATURA

[1] Bellalta B., Kosek-Szott K., Szott S., Wilhelmi F. 2024. „Towards an AI/ML-defined Radio for Wi-Fi: Overview, Challenges, and Roadmap". arXiv preprint arXiv:2405.12675 .

[2] Hoydis J., Ait Aoudia F., Valcarce A., Viswanathan H. 2021. „Toward a 6G AI-native air interface." IEEE Communications Magazine 59, no. 5:76-81.

[3] Keshtiarast N., Petrova M. 2024. „ML Framework for Wireless MAC Protocol Design." IEEE International Conference on Machine Learning for Communication and Networking 5–8 maja 2024, Sztokholm, Szwecja.

[4] Khalek N. et al. 2023. „Advances in Machine Learning-Driven Cognitive Radio for Wireless Networks: A Survey". IEEE Commun. Surveys & Tuts. 26(2):1201–1237.

[5] Mitola J., Maguire G. 1999. „Cognitive radio: making software radios more personal". IEEE Personal Communications 6(4):13–18.

[6] Szott S. et al. 2022. „Wi-Fi Meets ML: A Survey on Improving IEEE 802.11 Performance With Machine Learning". IEEE Commun. Surveys & Tuts. 24(3): 1843–1893.

[7] Wasilewska M., Bogucka H., Vincent Poor H. 2023. „Secure Federated Learning for Cognitive Radio Sensing". IEEE Communications Magazine 61(3):68–73.

[8] Wei G. et al. 2023. „MetaRadio: Bridging Wireless Communications Between Real and Virtual Spaces". IEEE Communications Magazine 61(6):140–146.

[9] Wilhelmi F., Szott S., Kosek-Szott K., Bellalta B. 2024. „Machine Learning & Wi-Fi: Unveiling the Path Towards AI/ML-Native IEEE 802.11 Networks." arXiv preprint arXiv:2405.11504.

[10] Wojnar M., Szott S., Rusek K., Ciezobka W. 2024. „Reinforced-lib: Rapid prototyping of reinforcement learning solutions." SoftwareX 26: 101706.

METODA PREDYKCJI ZAJĘTOŚCI PASMA W KANALE RADIOWYM W OPARCIU O GŁĘBOKIE SIECI NEURONOWE

METHOD FOR PREDICTION OF BAND OCCUPATION IN A RADIO CHANNEL BASED ON DEEP NEURAL NETWORKS)

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: 10.15199/59.2024.4.9

Streszczenie: **W artykule zaprezentowano nową metodę predykcji zajętości pasma w kanale radiowym stosującą głębokie sieci neuronowe. Metoda ta wykorzystuje strukturę autoenkodera do wieleopozoiomowej kompresji spektrogramów sygnału radiowego, co pozwala na efektywne kodowanie i dekodowanie sygnału radiowego. Implementacja prostych warstaw liniowych umożliwia szybką generację sygnału przewidywanego (predykcję spektrogramu), co jest istotne dla dynamicznie zmieniających się warunków w kanale radiowym. Dzięki zastosowaniu opisanej w artykule metodzie możliwe jest precyzyjne przewidywanie zakresów zajętości pasma radiowego dla ustalonych okresów predykcji.**

Abstract: **The article presents a new prediction of bandwidth occupancy in radio channels, which are integrated into a neural network. This method utilizes the structure of an autoencoder for multilevel compression of the radio signal's spectrogram, allowing for efficient encoding and decoding of the radio signal. The implementation of simple linear layers enables fast generation of the predicted signal (spectrogram prediction), which is crucial for dynamically changing conditions in radio channels. Thanks to the application of the method described in the article, it is possible to predict the ranges of radio bandwidth occupancy.**

Słowa kluczowe: **predykcja pasma, sieci neuronowe, sztuczna inteligencja, zajętość pasma radiowego**

Keywords: **band prediction, neural networks, artificial intelligence, radio band occupancy**

1. WSTĘP

Predykcja zajętości pasma radiowego to proces prognozowania, czy dane pasmo radiowe będzie wykorzystywane w określonym okresie czasu. Predykcja jest istotna w celu zarządzania spektrum radiowym, szczególnie w kontekstach, gdzie pasmo radiowe jest ograniczone i musi być efektywnie przydzielane pomiędzy różne usługi i aplikacje, takie jak telekomunikacja, radio, telewizja czy systemy ratunkowe. W publikacjach [1, 3, 4, 6, 8, 10] opisano zarówno teoretyczne podejścia, jak i praktyczne zastosowania różnych technik predykcji spektrum, w tym modeli Markova, analizy regresji i uczenia maszynowego.

W artykule przedstawiono metodę predykcji zajętości pasma w kanale radiowym, wykorzystującą autoenkodery [2] i proste warstwy liniowe, co stanowi istotny krok naprzód w dziedzinie komunikacji bezprzewodowej. Model dzięki zastosowaniu głębokiej kompresji umożliwia wydobycie ze spektrogramów informacji zarówno z teraźniejszości, jak i odległej przeszłości. Ta informacja następnie używana jest do predykcji zajętości pasma w przyszłości. Rozwiązanie to opiera się na analizie spektrogramu, co pozwala na precyzyjne rozpoznanie wzorców i charakterystyk sygnałów radiowych w różnych zakresach częstotliwości. Kontrastuje z innymi powszechnie stosowanymi metodami, takimi jak transformery [7] czy tradycyjne sieci rekurencyjne [5], które mimo swojej skuteczności w niektórych zastosowaniach, napotykają specyficzne ograniczenia.

Transformery, choć doskonale radzą sobie w przetwarzaniu języka naturalnego, są obliczeniochłonne, co ogranicza ich użyteczność w aplikacjach wymagających szybkiej inferencji. Ponadto, ich skomplikowana architektura i potrzeba dużych ilości danych do trenowania czynią je mniej optymalnymi w scenariuszach, gdzie dostęp do danych jest ograniczony lub gdzie wymagana jest natychmiastowa reakcja systemu.

Z kolei sieci rekurencyjne, chociaż użyteczne w modelowaniu danych sekwencyjnych, często borykają się z problemami skalowalności oraz wydłużonym czasem uczenia i inferencji, co jest niepożądane w dynamicznie zmieniających się środowiskach radiowych. Ich tendencja do zapominania odległych czasowo zależności w danych również ogranicza ich efektywność w długoterminowej predykcji zajętości pasma.

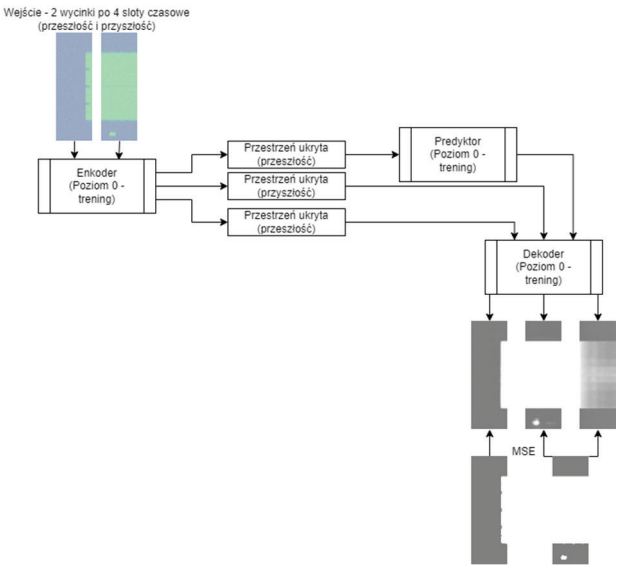
Zaproponowana metoda, opierając się na autoenkoderach, adresuje te ograniczenia poprzez wielopoziomową kompresję danych, która umożliwia efektywne i szybkie kodowanie oraz dekodowanie informacji o stanie pasma. Autoenkodery, działając w hierarchii, są w stanie ekstrahować i rekonstruować złożone wzorce w sygnałach radiowych, co jest kluczowe zarówno w krótko-, jak i długoterminowych predykcjach. Dodatkowo, implementacja prostych warstw liniowych pozwala na błyskawiczne generowanie predykcji, co jest istotne w aplikacjach czasu rzeczywistego.

2. OPIS METODY

Proponowana metoda opiera się na autoenkoderach i prostych, w pełni połączonych warstwach. Metoda ta wykorzystuje autoenkodery tylko do wielopoziomowej kompresji danych i rekonstrukcji mapy prawdopodobieństwa, podczas gdy warstwy liniowe są odpowiedzialne za predykcję. Aby reprezentacja uzyskana przez autoenkodery była regularna (chronologiczna) i pozwalała na prawidłowe generowanie map prawdopodobieństwa, należy te cechy wziąć pod uwagę w procesie treningowym. W rezultacie rekonstrukcja rzeczywistych sygnałów jest przeprowadzana w tej samej iteracji co predykcja. Ponieważ model jest wielopoziomowy, konieczne jest również zdefiniowanie podstawowej, minimalnej długości czasowej sygnału, który powinien przewidywać. Obecnie, jako dowód i wprowadzenie koncepcji, jest to spektrogram sygnału radiowego stanowiący obraz o wymiarach 192x64 podzielony na 4 panele, przy czym przestrzeń ukryta składa się z 4 wektorów o długości 256. Każdy z tych czterech wektorów reprezentuje krok czasowy z jaką jest realizowana predykcja, tak więc na najniższym poziomie są 4 wektory, a predykcja jest wykonywana dla kolejnych 4 wektorów. Liczby wektorów mogą podlegać modyfikacji w procesie treningowym, jednakże wymaga to eksperymentów skuteczności.

Pierwsze dwa poziomy zostaną przedstawione w celu zilustrowania schematu treningu. Na wczesnych etapach treningu funkcja kosztu może opierać się na porównaniu danych wejściowych, ale w potencjalnym długoterminowym przewidywaniu, dla większej liczby poziomów, ze względu na wykładniczy wzrost danych, może być wymagane skonstruowanie funkcji kosztu oparte na przestrzeni ukrytej wytrenowanych już koderów niższych poziomów.

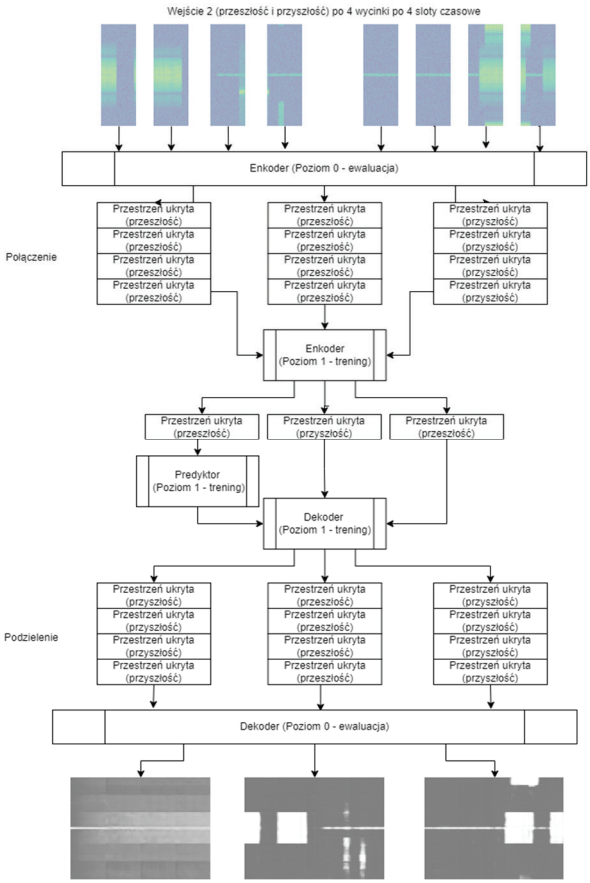
Poziom 0 jest trenowany tylko na podstawie spektrogramu. Wszystkie komponenty pokazane na Rys. 1 są trenowane w pętli.



Rys. 1. Poziom 0

Proces uczenia ostatniego modelu rozpoczyna się od Poziomu 0, który jest uczony wyłącznie na podstawie danych pochodzących ze spektrogramu. Autoenkoder na poziomie 0 kompresuje dane spektrogramu do przestrzeni ukrytej o niższym wymiarze, służącej jako skompresowana reprezentacja danych wejściowych. Ta ukryta przestrzeń pozwala na wydajne przechowywanie i przetwarzanie danych przy jednoczesnym zachowaniu ich podstawowych cech. Dekoder na poziomie 0 następnie rekonstruuje mapę prawdopodobieństwa zajętości pasma z przestrzeni ukrytej, a błąd rekonstrukcji jest wykorzystywany jako funkcja kosztu do uczenia.

Po wytrenowaniu Poziomu 0, skompresowana reprezentacja przestrzeni ukrytej jest przekazywana do następnego poziomu - Poziomu 1 (Rys. 2). Na Poziomie 1 stosowany jest kolejny autoenkoder w celu kompresji reprezentacji przestrzeni ukrytej z Poziomu 0. Koder na poziomie 1 przyjmuje reprezentację przestrzeni ukrytej z poziomu 0 jako dane wejściowe i tworzy jeszcze mniej wymiarową reprezentację przestrzeni ukrytej. Podobnie, dekoder na Poziomie 1 rekonstruuje reprezentację przestrzeni ukrytej z powrotem do przestrzeni ukrytej Poziomu 0, a błąd rekonstrukcji jest wykorzystywany jako funkcja kosztu do uczenia Poziomu 1.



Rys. 2. Poziom 1

Ten proces używania autoenkoderów do kompresji reprezentacji przestrzeni ukrytej i rekonstrukcji jej na każdym poziomie trwa aż do osiągnięcia pożądanej liczby poziomów. Każdy poziom obejmuje różne poziomy abstrakcji i zależności czasowych w danych, umożliwiając hierarchiczną reprezentację informacji o sygnale.

Po wytrenowaniu wszystkich poziomów, skompresowane reprezentacje przestrzeni ukrytej mogą być wyko-

rzystane do przewidywania. Czasy predykcji sygnału zależą od użytkownika, jeżeli potrzebuje on predykcji krótko terminowej, to może użyć małej liczby poziomów nawet bez konieczności przechowywania przestrzeni ukrytej poprzednich poziomów. Jeżeli potrzebuje predykcji długoterminowej z małym krokiem predykcji (np. predykuje 256 okna czasowe w przód, a potrzebuje aktualizacji co jedno okno czasowe), konieczne może być przechowywanie przestrzeni ukrytych z poprzednich iteracji. Warstwy liniowe na szczycie autoenkoderów generują predykcję w oparciu o skompresowane reprezentacje przestrzeni ukrytej. Sygnały z enkoderów są wykorzystywane jako dane wejściowe do warstw liniowych, które generują przewidywane sygnały. Takie podejście pozwala na wydajne i dokładne przewidywanie sygnałów przy jednoczesnym zachowaniu dużej szybkości wnioskowania, ponieważ autoenkodery służą jako etap kompresji danych, który zmniejsza wymiarowość danych wejściowych przed przekazaniem ich do warstw liniowych w celu przewidywania.

Warto zauważyć, że na wczesnych etapach szkolenia funkcja kosztu może opierać się na porównywaniu danych wejściowych w celu zapewnienia dokładnej rekonstrukcji. Jednakże, gdy horyzont predykcji rozszerza się do predykcji długoterminowej, wykładniczy wzrost danych może wymagać dodatkowych kroków wstępnego przetwarzania w oparciu o wstępnie wytrenowane kodery na poprzednich poziomach, aby skutecznie poradzić sobie ze zwiększoną ilością danych.

Rekonstrukcja danych nie tylko aproksymowanych przez predyktor jest niezbędna do prawidłowego podziału zadań między komponenty modelu. W ten sposób predyktor jest wyłącznie odpowiedzialny za generowanie prognoz, podczas gdy inne komponenty nie przejmują całego jego zadania, umożliwiając hierarchiczne szkolenie. Model dostosowuje się do dostarczonych danych szkoleniowych i odpowiednio generuje mapy prawdopodobieństwa.

Dodatkową kluczową cechą prezentowanej metody jest jej adaptacyjność, zwłaszcza w kontekście przetwarzania kolejnych próbek danych w dynamicznych środowiskach. W tradycyjnych podejściach, gdzie przewidywanie często wymaga pełnej inferencji za pomocą całego modelu, nasza metoda wyróżnia się możliwością wykorzystania już obliczonych "warstw" informacji dla nowych danych bez potrzeby ponownego przetwarzania od początku.

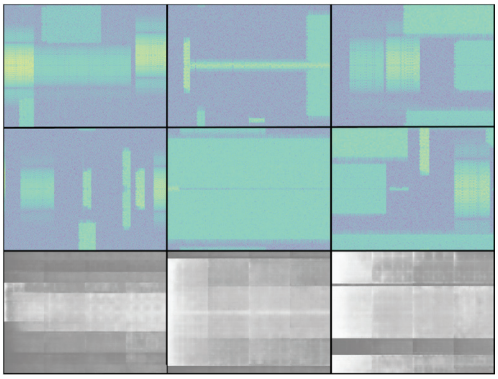
Taka adaptacyjność jest możliwa dzięki hierarchicznej i modularnej naturze modelu, gdzie skompresowane reprezentacje przestrzeni ukrytej, wygenerowane na każdym poziomie autoenkodera, mogą być efektywnie reaplikowane lub aktualizowane z mniejszym nakładem obliczeniowym Zdolność do adaptacji i wykorzystania wcześniej obliczonych danych w nowych iteracjach nie tylko zwiększa efektywność obliczeniową, ale także poprawia ogólną wydajność systemu, redukując opóźnienia w przewidywaniach i zwiększając precyzję modelu w dynamicznie zmieniających się środowiskach.

3. REZULTATY

Do eksperymentów wykorzystano zbiór danych [9] wygenerowanych w sposób sztuczny, co pozwoliło na weryfikację działania modeli w kontrolowanym środowisku. Zajętość pasma określano na podstawie występowania w tym miejscu sygnału bądź nie. Należy jednak podkreślić, że prezentowane dane nie odzwierciedlają w pełni rzeczywistych warunków, dla których model został zaprojektowany, i służą przede wszystkim ilustracji potencjału metody. Sztuczna natura danych nie pozwala na pełne wykorzystanie możliwości modelu, ze względu na brak typowych dla prognoz charakterystyk danych, takich jak okresowość i trend w zmienności zajętości pasma.

Na rysunku 3 zaprezentowano trzy wybrane próbki danych wyjściowych modelu. W wynikach można zaobserwować stopniowanie mapy prawdopodobieństwa zajętości pasma, które zmienia się wraz z upływem czasu. Stopniowanie to jest zgodne z oczekiwaniami, potwierdzając zdolność modelu do generowania prawdopodobieństw zajętości pasma w różnych przedziałach czasowych. Efekt ten świadczy o zdolności modelu do rozróżniania i adaptacji do dynamiki danych.

Mapa prawdopodobieństwa zajętości pasma, przedstawiona na rysunku w dolnym wierszu, jest kluczowym elementem wizualizacyjnym naszego modelu. Zastosowanie takiej mapy umożliwia intuicyjne przedstawienie prognoz modelu dotyczących przyszłego wykorzystania pasma. Biały kolor na tej mapie oznacza wysokie prawdopodobieństwo zajętości danego fragmentu pasma w przyszłości. Im jaśniejsze odcienie szarości, tym większa pewność modelu co do zajętości pasma w tych obszarach. Taka reprezentacja umożliwia użytkownikowi szybkie i efektywne odczytywanie informacji o potencjalnym obciążeniu pasma.



Rys. 3. Przykładowe wyniki (wiersz nr 1 - próbka teraźniejsza, wiersz nr 2 - próbka przyszła, wiersz nr 3 - mapa prawdopodobieństwa zajętości pasma określona na podstawie próbki teraźniejszej)

W celu dalszego rozwijania i potwierdzania skuteczności zaproponowanej metody niezbędne jest przeprowadzenie badań z wykorzystaniem rzeczywistych próbek danych radiowych. Rzeczywiste dane pozwolą na lepsze uchwycenie potencjalnych okresowości oraz trendów występujących w środowiskach RF. Trenowanie modelu na takich danych umożliwi badanie jego zdolności do długoterminowej predykcji, co jest kluczowe dla realnych zastosowań w dynamicznych środowiskach radiowych.

Dokładność systemu zależy jedynie od zbioru treningowego oraz jego podobieństwa do warunków rzeczywistych. Na zbiorze nierzeczywistym, losowym dokładność predykcji krótkoterminowej zachowuje się zgodnie z przewidywaniami – maleje wraz z czasem, a także przykłada większe prawdopobieństwo pojawienia się sygnału dla obszarów gdzie z reguły sygnał występuje (w naszym przypadku – środek spektrogramu). Natomiast bez odpowiedniego treningu na dużym zbiorze danych rzeczywistych, nie jest możliwe jednoznaczne określenie zachowania modelu w predykcji długoterminowej – być może model będzie w stanie wynajdywać również wzorce okresowe, z którymi klasyczne modele sobie nie radzą.

4. WNIOSKI

Wykorzystanie autoenkoderów i prostych, w pełni połączonych warstw liniowych do przewidywania zajętości pasma, jak opisano w powyższej metodzie, stanowi postęp w tej dziedzinie. Metoda ta stanowi krok naprzód w zastosowaniu technik uczenia maszynowego i głębokiego uczenia do zadań predykcyjnych, zwłaszcza w dziedzinie komunikacji bezprzewodowej i przewidywania sygnału.

Rdzeniem opisanej metody predykcji sygnału radiowego jest wykorzystanie autoenkoderów do wielopoziomowej kompresji danych i rekonstrukcji mapy prawdopodobieństwa, z warstwami liniowymi odpowiedzialnymi za przewidywanie. Takie rozdzielenie zadań pozwala temu podejściu na efektywne radzenie sobie z wysokowymiarowymi danymi, kompresując je do możliwych do zarządzania wektorów przestrzeni ukrytej, z których każdy reprezentuje krok czasowy w predykcji. To efektywne wykorzystanie danych i zasobów obliczeniowych umożliwia modelowi dostarczanie dużych szybkości wnioskowania - co jest niezbędnym wymogiem dla zadań predykcyjnych w czasie rzeczywistym.

Siła tego modelu polega także na jego solidnej elastyczności w dostosowywaniu się do złożoności danych. Gdy horyzont predykcji rozszerza się do przewidywania długoterminowego, skutecznie radzi sobie z wykładniczym wzrostem danych poprzez wykorzystanie danych uzyskanych z poprzednich predykcji. Pokazuje to zdolność modelu do skalowania i adaptacji, czyniąc go wyjątkowo solidnym i odpowiednim do praktycznych zastosowań w świecie rzeczywistym.

Podsumowując, proponowana metoda przewidywania zajętości pasma oparta na autoenkoderach i warstwach liniowych stanowi wysoce wydajne podejście do przewidywania sygnału. Jej wielopoziomowa, hierarchiczna reprezentacja zapewnia szybkie, dokładne przewidywania przy jednoczesnym zachowaniu dużej szybkości wnioskowania, co stanowi znaczący krok naprzód w dziedzinie metod przewidywania sygnału.

Istnieje jednak potencjalne ryzyko, które należy wziąć pod uwagę. Na niższych poziomach modelu niektóre cechy (mogące mieć kluczowe znaczenie dla długoterminowego przewidywania) mogą zostać odrzucone podczas procesu uczenia na wyższych poziomach. Ponadto poleganie na funkcji kosztu opartej na przestrzeni ukrytej może być niedoskonałe. Staranna analiza i monitorowanie procesu treningowego jest ważne, aby rozwiązać te potencjalne problemy.

LITERATURA

[1] Akbar IA, Tranter WH (2007) Dynamic spectrum allocation in cognitive radio using hidden Markov models: Poisson distributed case. In: Proceedings 2007 IEEE SoutheastCon. pp 196–201

[2] Bank D, Koenigstein N, Giryes R (2021) Autoencoders

[3] Li Y, Dong Y, Zhang H, Zhao H, Shi H, Zhao X (2010) Spectrum Usage Prediction Based on High-order Markov Model for Cognitive Radio Networks. In: 2010 10th IEEE International Conference on Computer and Information Technology. pp 2784–2788

[4] Lin Z, Jiang X, Huang L, Yao Y (2009) A Energy Prediction Based Spectrum Sensing Approach for Cognitive Radio Networks. In: 2009 5th International Conference on Wireless Communications, Networking and Mobile Computing. pp 1–4

[5] Sherstinsky A (2020) Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. Phys Nonlinear Phenom 404:132306. doi: 10.1016/j.physd.2019.132306

[6] Tumuluru VK, Wang P, Niyato D (2010) A Neural Network Based Spectrum Prediction Scheme for Cognitive Radio. In: 2010 IEEE International Conference on Communications. pp 1–5

[7] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention Is All You Need

[8] Wang L, Hu J, Jiang R, Chen Z (2024) A Deep Long-Term Joint Temporal–Spectral Network for Spectrum Prediction. Sensors 24:1498. doi: 10.3390/s24051498

[9] Wicht J, Wetzker U, Jain V (2022) Spectrogram Data Set for Deep-Learning-Based RF Frame Detection. Data 7:168. doi: 10.3390/data7120168

[10] Yarkan S, Arslan H (2007) Binary Time Series Approach to Spectrum Prediction for Cognitive Radio. In: 2007 IEEE 66th Vehicular Technology Conference. pp 1563–1567

ROZPROSZONE WYKRYWANIE ZAJĘTOŚCI WIDMA OPARTE NA UCZENIU FEDERACYJNYM

DISTRIBUTED SPECTRUM OCCUPANCY DETECTION BASED ON FEDERATED LEARNING

Łukasz Kułacz[1], Adrian Kliks[2]

[1] Politechnika Poznańska, Poznań, lukasz.kulacz@put.poznan.pl
[2] Politechnika Poznańska, Poznań, adrian.kliks@put.poznan.pl

DOI: 10.15199/59.2024.4.10

Streszczenie: Wykrywanie zajętości widma jest kluczowym zagadnieniem umożliwiającym dynamiczny dostęp do widma. Współcześnie w celu polepszenia detekcji popularne są rozwiązania z obszaru uczenia maszynowego, w tym uczenia federacyjnego (FL). Głównym wyzwaniem w tym kontekście jest ograniczony dostęp do danych treningowych. W pracy przedstawiono podejście rozproszone FL, skupiając się na węzłach pozbawionych dostępu do danych uczących. Omówiono wyniki eksperymentu sprzętowego polegającego na wykrywaniu sygnału DVB-T.

Abstract: Spectrum occupancy detection is a key enabler for dynamic spectrum access, where machine learning algorithms are successfully utilized for detection improvement. However, the main challenge is limited access to labeled data about users' transmission presence needed in supervised learning models. We present a distributed federated learning approach that addresses this challenge for sensors without access to learning data. The paper discusses the results of the conducted hardware experiment, where FL has been applied for DVB-T signal detection.

Słowa kluczowe: detekcja zajętości widma, dynamiczny dostęp do widma, uczenie federacyjne, uczenie maszynowe

Keywords: spectrum occupancy detection, dynamic spectrum access, federated learning, machine learning

1. WSTĘP

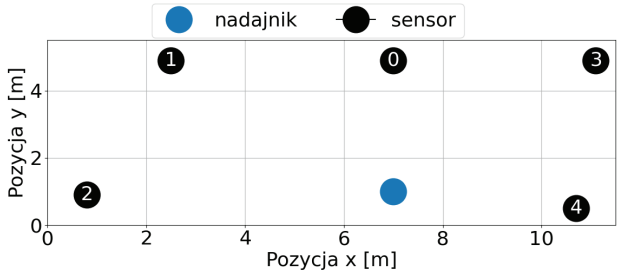
Statycznie i przydzielone na wyłączność widmo gwarantuje proste zarządzanie zasobami częstotliwości, ale jednocześnie jest dalekie od optymalnego. Liczne prace dowiodły, że takie podejście może prowadzić do tylko częściowego wykorzystania dostępnych zasobów widmowych i nie rozwiązuje problemu nieefektywnego wykorzystania widma. Dynamiczny dostęp do widma (ang. Dynamic Spectrum Access, DSA) jest jednym z obiecujących rozwiązań tego problemu [2, 5, 3]. Wykorzystując właściwość elastyczności w dostępie do zasobów widmowych, DSA pozwala on na wykorzystanie niezajętego widma do celów innych niż te, dla których zostało ono pierwotnie przeznaczone i zarezerwowane.
Zagadnienie poprawnej identyfikacji niezajętych fragmentów widma częstotliwościowego, choć koncepcyjnie proste, jest trudne w realizacji. Wiele rozwiązań zostało zaproponowanych w dziedzinie wykrywania widma (ang. Spectrum Sensing, SS) [4, 8], gdzie ostatnio rola sztucznej inteligencji (ang. Artificial Intelligence, AI) staje się coraz bardziej dominująca [7]. Jednak w przypadku zastosowania uczenia maszynowego, istotnym wyzwaniem staje się sam proces uczenia, trenowania i weryfikacji modeli AI. Jednym z możliwych rozwiązań tego problemu przekazywanie pomiędzy węzłami w systemie SS informacji o parametrach modelu. Rozwiązanie takie jest możliwe np. w uczeniu federacyjnym (ang. Federated Learning, FL) czy tzw. transfer learning [9, 1].
W [6] skupiliśmy się na badaniu możliwości zastosowania FL w wersji scentralizowanej do algorytmów SS. Jednak takie podejście wiąże się z potrzebą scentralizowanej koordynacji procesu wykrywania widma. Jednak nie zawsze okazuje się to możliwe. W tym artykule przedstawiamy wyniki eksperymentu sprzętowego, w którym rozproszone uczenie maszynowe oparte na FL zostało zastosowane do ulepszenia algorytmów SS. W szczególności skupiliśmy się na sytuacji, w której jeden z węzłów całego systemu nie ma dostępu do danych treningowych.

2. DEFINICJA PROBLEMU

Celem przeprowadzanego eksperymentu sprzętowego była praktyczna weryfikacja wydajności rozproszonego algorytmu detekcji obecności sygnału SS, w którym węzły współdzielą ze sobą parametry modeli AI, aby poprawić prawdopodobieństwo wykrycia zajętości widma. W szczególności rozważono przypadek, gdy pojedynczy węzeł jest wrażliwy na błędy detekcji ponieważ znajduje się w obszarze mierzonych niskich wartości stosunku mocy sygnału do mocy szumu (ang. Signal to Noise Ratio, SNR). Zgodnie z koncepcją rozproszonego uczenia federacyjnego, poprawa wydajności wykrywania obecności sygnału jest możliwa właśnie poprzez współdzielenie wyuczonych modeli AI, prowadząc do zwiększenia informacji o wykorzystaniu widma dzięki wiedzy od węzłów sąsiadujących. W przeprowadzonym eksperymencie laboratoryjnym każdy węzeł przeprowadza detekcję sygnału szerokopasmowego telewizji cyfrowej DVB-T, stosując w tym celu algorytm oparty na wartościach własnych.

2.1 Gromadzenie danych

Próbki sygnału oraz szumu wykorzystane w tej pracy zostały zebrane w laboratorium, gdzie rozmieszczono pojedynczy nadajnik i pięć czujników tak jak to pokazano na Rys. 1.



Rysunek 1: Topologia sieci używana podczas pomiarów.

Rolę każdego czujnika pełnił komputer stacjonarny z oprogramowaniem GNU Radio zainstalowanym w systemie operacyjnym Linux; jako stopień wejściowy odbiornika wykorzystano popularne układy radia programowalnego Universal Software Radio Peripheral (USRP) w wersji B210. Każdy odbiornik zbierał 10 milionów próbek IQ w trakcie pojedynczej sekwencji pomiarowej. W celu zapewnienia poprawnych danych do treningu modeli AI, każdy czujnik został uruchomiony 50 razy podczas transmisji sygnału DVB-T oraz 50 razy, gdy transmisja sygnału była wyłączona. Sygnał DVB-T był transmitowany przez generator sygnału Rohde&Schwarz SMBV100A, dla założonej częstotliwości środkowej 2,0 GHz i mocy transmisyjnej zmienianej w zakresie od -40 dBm do 10 dBm z krokiem o 1 dBm. Zebrane w ten sposób próbki zostały dalej wykorzystane w algorytmie detekcji zajętości widma.

2.2 Przetwarzanie danych

Przygotowanie próbek IQ zebranych przez czujniki zostało rozpoczęte od przefiltrowania pasma 8 MHz, w którym powinien być odbierany oczekiwany sygnał DVB-T. Następnie dla zbioru $L = 8$ kolejnych wektorów próbek o rozmiarze $N = 32768$ obliczono średnią odbieraną moc μ; wyznaczono także macierz korelacji $\mathbf{R}$, dla której wyznaczono wartości własne γ_i, gdzie $i = 0, 1, ...L-1$, a L to liczba wartości własnych. Jako metrykę decyzyjną przyjęto stosunek maksymalnej do minimalnej wartości własnej $T = \frac{\gamma_{\max}}{\gamma_{\min}}$ oraz wartość autokorelacji. W rezultacie otrzymano zestaw wartości wykorzystywanych dalej jako dane wejściowe w procesie uczenia maszynowego. Dodatkowo dane te musiały zostać znormalizowane, dlatego w przypadku autokorelacji przeprowadzono normalizację min-max, a w przypadku wartości własnych i metryki decyzyjnej przeprowadzono normalizację z-score. Przetestowano wiele podejść i kombinacji, aby wybrać odpowiednie metody normalizacji dla poszczególnych danych w celu wygenerowania najlepszych wyników.

3. PLAN EKSPERYMENTU

Podczas eksperymentu czujniki podejmują próbę detekcji obecności sygnału DVB-T. W scenariuszu referencyjnym każdy czujnik wykorzystuje jedynie próbki zebrane przez siebie w procesie tworzenia modelu uczenia maszynowego. Założono, że każdy czujnik tworzy własny model sieci neuronowej z dwiema warstwami (cztery węzły w każdej warstwie), a proces jest powtarzany dziesięć razy z różnymi podziałami danych treningowych i testowych (przy użyciu podejścia Stratified K-Fold). Przeprowadzone zostały również czasochłonne analizy, typowe dla algorytmów uczenia maszynowego, których celem było określenie liczby warstw i węzłów. Dwa główne wnioski z takich analiz to: brak zauważalnej poprawy dokładności modeli przy dodaniu trzeciej warstwy do modelu; zastosowanie większej liczby węzłów (około 20) w każdej z warstw nieznacznie poprawia dokładność modeli. Z uwagi na nieznaczną poprawę dokładności (maksymalnie kilka procent) oraz fakt wykorzystania uczenia federacyjnego, co wiąże się z koniecznością wymiany współczynników modeli, dobrana została celowo względnie mała liczba węzłów, która jednak zapewnia dość dobre wyniki dokładności modeli. Zwracając uwagę na główne wyzwanie w procesie detekcji zajętości widma, którym jest dostęp do danych treningowych, symulowana była sytuacja, w której jeden z czujników nie jest w stanie trenować modelu (nie ma dostępu do danych treningowych). W takiej sytuacji czujnik nieposiadający dostępu do danych treningowych wykorzystuje model utworzony z domyślnymi współczynnikami i nie zmienia ich podczas symulacji.

3.1 Uczenie federacyjne

Aby umożliwić czujnikowi nieposiadającemu dostępu do danych treningowych adaptację modelu, rozważone zostało federacyjne podejście do uczenia, w którym współczynniki modelu AI są wymieniane z sąsiednimi czujnikami. W każdej iteracji procesu uczenia federacyjnego obliczane są wagi współczynników sąsiednich czujników, aby określić, który z czujników może mieć lepszy model. W tym przypadku wzięte zostało uwagę kilka aspektów:

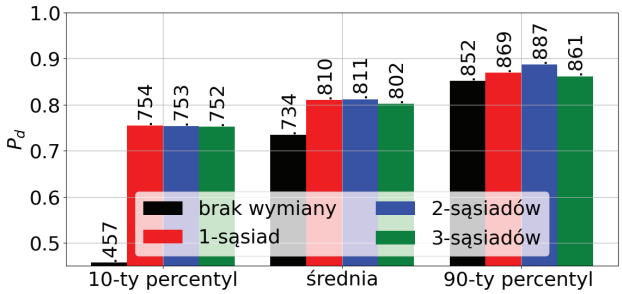
- wpływ czasu uczenia - czujnik zaczyna z wagą 0,8 dla własnych współczynników i zwiększa tę wagę do prawie 1,0 pod koniec procesu uczenia;
- wynik dokładności ostatniego kroku uczenia - waga własnych współczynników jest skalowana liniowo z ostatnim wynikiem dokładności, aby nadać priorytet etapom uczenia, gdy osiągnięty zostanie wysoki wynik dokładności własnego modelu lub nadać priorytet sąsiednim modelom, gdy osiągnięty zostanie niski wynik dokładności własnego modelu;
- w przypadku uwzględnienia wielu sąsiednich modeli ich wagi są obliczane przy użyciu ważenia odwrotnej odległości (ang. Inverse Distance Weighting , IDW).

W pracy rozważono różne współczynniki algorytmu ważenia odwrotnej odległości (dokładnie: 0, 1, 2, 3).

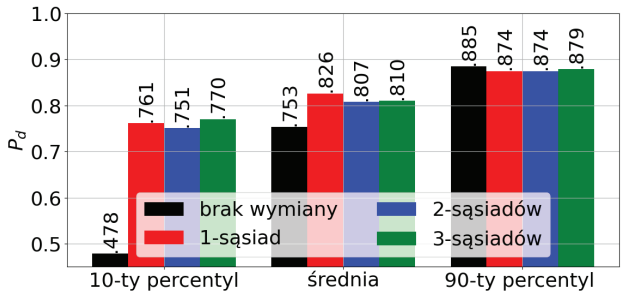
4. WYNIKI EKSPERYMENTU

Uzyskane wyniki przedstawiono z dwóch perspektyw. Pierwsza dotyczy oceny jakości procesu uczenia maszynowego, a dokładniej oceny finalnego, wytrenowanego modelu w każdym z czujników. Podstawową metryką do stosowaną do tej oceny jest miara dokładności (ang. accuracy score), jednakże określona ona tylko liczbę poprawnych predykcji w stosunku do wszystkich prób. Najczęściej nie jest to wystarczająca miara i tak w przypadku detekcji zajętości widma szczególnie interesujące są błędy predykcji,

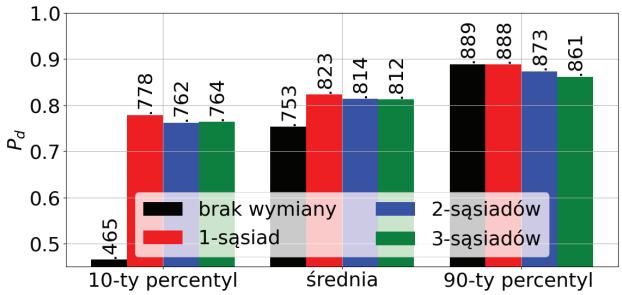
gdyż fałszywy alarm nie powinien być traktowy na równi z sytuacją nie wykrycia obecnego sygnału. Z tego powodu do oceny finalnego modelu wykorzystana została metryka F1 (ang. F1 score), która z jednakową wagą uwzględnia w ocenie precyzję (ang. precision) oraz czułość (ang. recall).



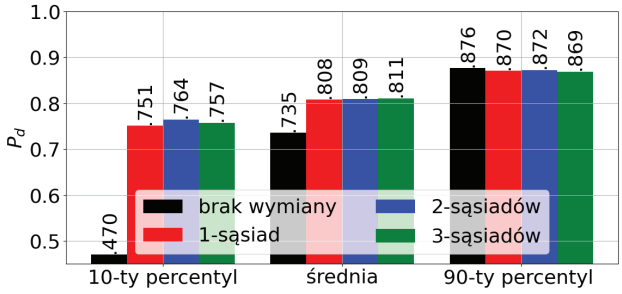
Rysunek 2: Prawdopodobieństwo detekcji (współczynnik IDW=0)



Rysunek 3: Prawdopodobieństwo detekcji (współczynnik IDW=1)



Rysunek 4: Prawdopodobieństwo detekcji (współczynnik IDW=2)



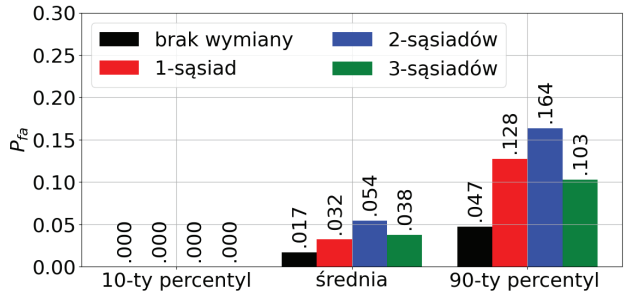
Rysunek 5: Prawdopodobieństwo detekcji (współczynnik IDW=3)

Precyzja modelu określa jaka część próbek z sygnałem udało się wykryć prawidłowo w drodze predykcji (liczba prawidłowych detekcji podzielona przez liczbę prawidłowych detekcji powiększoną przez liczbę nieprawidłowych detekcji), a czułość określa jaką część próbek sygnału pominięto w drodze predykcji (liczba prawidłowych detekcji podzielona przez liczbę prawidłowych detekcji powiększoną przez liczbę nieprawidłowych braków detekcji).

Druga przyjęta perspektywa dotyczy oceny skuteczności detekcji zajętości widma. W tym przypadku pod uwagę wzięto prawdopodobieństwo detekcji sygnału przez czujnik. Z uwagi na strukturę symulacji, gdy podczas każdej iteracji inny czujnik nie ma dostępu do danych treningowych, wyniki przedstawiono w postaci średniego wyniku dla czujnika oraz 10-tego i 90-tego percentyla, co obrazuje jeden z najsłabszych wyników i jeden z najlepszych wyników (celowo pomijając minimum i maksimum).

Uzyskane prawdopodobieństwo detekcji pokazane zostało na Rys. 2 - 5. Warto zwrócić uwagę, że w analizowanym scenariuszu, gdy wszystkie czujniki znajdują się w tym samym pomieszczeniu, a warunki propagacyjne między nadajnikiem a każdym z czujników bardzo zbliżone to wymiana współczynników modeli między różną liczbą sąsiadów przynosi bardzo zbliżone efekty. Konsekwencją tego jest również mały wpływ współczynnika IDW na wyniki. Najważniejszym jednak z uzyskanych rezultatów wnioskiem jest zwiększenie prawdopodobieństwa detekcji najsłabszych czujników (niemających dostępu do danych treningowych) z przedziału 0.457-0.478 do przedziału 0.751-0.778 (w zależności od współczynnika IDW oraz liczby sąsiadów rozważanych w procesie wymiany współczynników modeli) - co w praktycznym zastosowaniu wyraźnie zwiększa szanse na dokonanie poprawnej detekcji obecności sygnału.

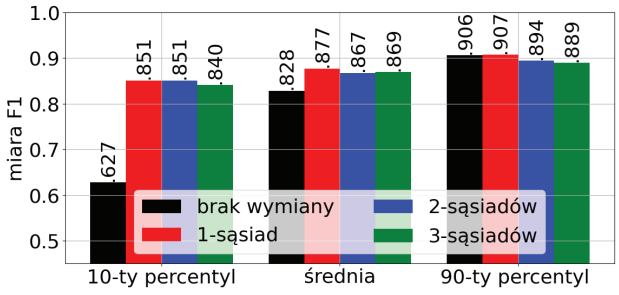
Jednocześnie warto przeanalizować wyniki prawdopodobieństwa fałszywego alarmu przedstawionego na Rys. 6 (tylko dla współczynnika IDW równego 0). Najgorsze z przypadków (tym razem reprezentowane w 90-tym percentylu wyników) zwiększają prawdopodobieństwo fałszywego alarmu z 0,047 do przedziału od 0,103 do 0,164. Warto pamiętać, że jest to wynik uzyskany jednocześnie z zwiększeniem prawdopodobieństwa detekcji z około 0,45 na około 0,75.



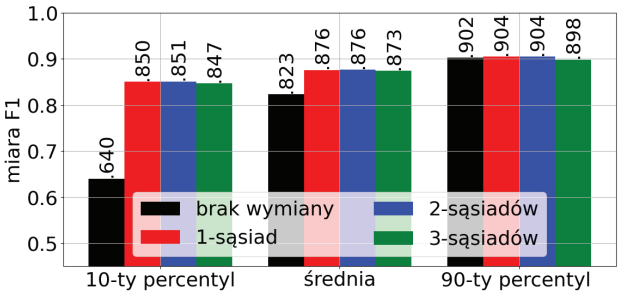
Rysunek 6: Prawdopodobieństwo fałszywego alarmu (współczynnik IDW=0)

Wyniki jakości finalnych modeli pokazane zostały na Rys. 7 - 8 w postaci uzyskanej metryki F1 (dla skrajnych wartości współczynnika IDW). Warto zauważyć, że czujniki, które nie mają możliwości podjęcia poprawnej decyzji (prawdopodobieństwo detekcji poniżej 50%) posiadają modele ocenione (miarą F1) na około 0,65. Przy ocenie średniego modelu widać zdecydowanie, że wymiana współczynników z tylko jednym sąsiadem nie działa

korzystnie i znacznie lepsze rezultaty można uzyskać poprzez wymianę z większą liczbą sąsiadów. Podobny trend odzwierciedlony jest w wynikach przedstawiających prawdopodobieństwo detekcji. Ponownie wpływ współczynnika IDW wykorzystywanego podczas tej symulacji jest pomijalny ze względu na umieszczenie wszystkich czujników w jednym pomieszczeniu. Jednakże w przypadku bardziej zróżnicowanego środowiska propagacyjnego odpowiednie kryterium doboru sąsiadów a następnie wag ich modelu byłoby kluczowe do poprawnego działania przedstawianego rozwiązania.

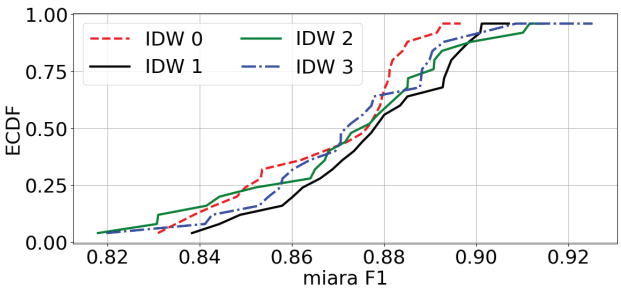


Rysunek 7: Miara F1 (współczynnik IDW=0)



Rysunek 8: Miara F1 (współczynnik IDW=3)

Wpływ współczynnika IDW na miarę F1 w scenariuszu wymiany modeli między trzema sąsiadami przedstawiono dodatkowo na Rys. 9 w formie dystrybuanty. Zauważalnym jest mniejsza rozbieżność miary F1 (między około 0,83 a 0,89) gdy każdy sąsiad ma taką samą wagę (IDW 0), a największą rozbieżność (między około 0,82 a 0,93) gdy zastosowano IDW z współczynnikiem 3. Zastosowanie większych współczynników IDW daje zatem możliwość lepszego dostosowania modelu do zastanej sytuacji, jednak wraz z zwiększonym ryzykiem uzyskania gorszego modelu. Zastosowanie mniejszych współczynników IDW powoduje większe uśrednianie modelu co minimalizuje ryzyko uzyskania słabszego modelu, ale również nie pozwala na osiągnięcie lepszych wyników lokalnie.



Rysunek 9: Dystrybuanta miary F1

5. PODSUMOWANIE

W pracy przedstawiono wyniki przeprowadzonego eksperymentu sprzętowego, w którym zastosowano rozproszone uczenie federacyjne dla poprawy detekcji obecności sygnału w obserwowanym widmie częstotliwościowym. Udało się wykazać, że w wyniku uczenia federacyjnego, które opiera się na wymianie współczynników modelu uczenia maszynowego, proces uczenia może zostać usprawniony - szczególnie w przypadku czujników z utrudnionym dostępem do danych treningowych. W ramach dalszych prac planowane jest włączenie do analizy dodatkowego czujnika, którego praca został celowo zaburzona przez niedokładne podłączenie anteny odbiorczej, a także analiza wpływu liczby wymienianych próbek (liczba węzłów w każdej z warstw modelu) na skuteczność działania całego systemu oraz porównanie wyników z systemem wykrywania zajętości widma z węzłem centralnym.

PODZIĘKOWANIE

Praca powstała w ramach projektu finansowanego przez Narodowe Centrum Nauki (nr 2021/41/N/ST7/01298).

LITERATURA

[1] Abdulrahman Sawsan, Tout Hanine, Ould-Slimane Hakima, Mourad Azzam, Talhi Chamseddine, Guizani Mohsen. 2021. "A Survey on Federated Learning: The Journey From Centralized to Distributed On-Site Learning and Beyond". IEEE Internet of Things Journal; 8(7): 5476-5497.

[2] Arjoune Youness, Kaabouch Naima. 2019. "A Comprehensive Survey on Spectrum Sensing in Cognitive Radio Networks: Recent Advances, New Challenges, and Future Research Directions". Sensors 2019; 19 (1) : 126.

[3] Dudda Torsten, Irnich Tim. 2012. "Capacity of cellular networks deployed in TV White Space". IEEE International Symposium on Dynamic Spectrum Access Networks. Bellevue USA; str. 254-265.

[4] Haykin Simon. 2005. "Cognitive radio: brain-empowered wireless communications". IEEE Journal on Selected Areas in Communications 23 (2) : 201-220.

[5] Liang Ying-Chang, Chen Kwang-Cheng, Ye Li, Mahonen Petri. 2011. "Cognitive radio networking and communications: an overview". IEEE Transactions on Vehicular Technology 60 (7) : 3386-3407.

[6] Kułacz Łukasz, Kliks Adrian. 2023. "Federated Learning-Based Spectrum Occupancy Detection". Sensors 2023; 23 (14) : 6436.

[7] Rekkas Vasileios, Sotiroudis Sotirios, Sarigiannidis Panagiotis, Wan Shaohua, Karagiannidis George, Goudos Sotirios. "Machine Learning in Beyond 5G/6G Networks—State-of-the-Art and Future Trends". Electronics. 2021; 10(22):2786.

[8] Yucek Tevfik, Arslan Huseyin. 2009. "A survey of spectrum sensing algorithms for cognitive radio applications". IEEE Communications Surveys & Tutorials 11(1) : 116-130.

[9] Wasilewska Małgorzata, Bogucka Hanna, Kliks Adrian. "Federated Learning for 5G Radio Spectrum Sensing". Sensors. 2022; 22(1):198.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

PRAKTYCZNA REALIZACJA ATAKÓW OMIJANIA SYSTEMÓW WYKRYWANIA WŁAMAŃ W SIECIACH

PRACTICAL EVASION ATTACKS ON NETWORK INTRUSION DETECTION SYSTEMS

Paweł Popiołek[1]; Jędrzej Bieniasz[1]

[1] Instytut Telekomunikacji, Politechnika Warszawska, Warszawa, Polska
pawel.popiolek.stud@pw.edu.pl
jedrzej.bieniasz@pw.edu.pl

DOI: 10.15199/59.2024.4.11

Streszczenie: **Niniejsza praca analizuje podatności systemów wykrywania włamań w sieciach (NIDS) na ataki typu przeciwstawnego. Zaproponowano metodę identyfikacji optymalnych perturbacji w ruchu sieciowym, zwiększających niewykrywalność przez NIDS. Wyniki na zbiorze danych CTU-13 wykazały, że ataki te mogą zredukować dokładność wykrywania z 99,99% do około 40% w najlepszym przypadku. Podkreśla to konieczność zwiększenia odporności systemów opartych na uczeniu maszynowym na takie zagrożenia.**

Abstract: **This study investigates the vulnerability of Network Intrusion Detection Systems (NIDS) to adversarial attacks. A prototype method was implemented to identify optimal perturbations that evade NIDS detection. The results from the CTU-13 dataset demonstrated the effectiveness of the attacks, reducing detection accuracy from 99.99% to approximately 40% in the best scenario. These findings underscore the need to enhance the resilience of machine learning-based systems against such threats.**

Słowa kluczowe: **Ataki Omijania, Bezpieczeństwo Uczenia Maszynowego, Przeciwstawne Uczenie Maszynowe, Systemy Wykrywania Włamań w Sieciach**

Keywords: **Adversarial Machine Learning, Evasion Attacks, Network Intrusion Detection Systems, Security of Machine Learning**

1. WSTĘP

W ostatnich latach rosnące zainteresowanie metodami uczenia maszynowego (ang. *machine learning*, ML) objęło różnorodne dziedziny, w tym również cyberbezpieczeństwo i systemy stosowane w tej dziedzinie [6]. Popularnym obszarem, w którym takie rozwiązania znajdują zastosowanie i okazują się skuteczne, jest wykrywanie włamań w sieciach komputerowych. Rozwiązania, które wykorzystują dane sieciowe, są określane jako Systemy Wykrywania Włamań Sieciowych (ang. *Network Intrusion Detection Systems*, NIDS). Drugą komplementarną kategorią systemów wykrywania włamań są te operujące na logach systemowych (lub szerzej, na danych hosta), znanych jako Systemy Wykrywania Włamań na Hoście (ang. *Host Intrusion Detection Systems*, HIDS), NIDS działają na danych pochodzących z monitorowania zachowań sieci w różnych jej punktach. Głównym wyzwaniem takiego podejścia do wykrywania ataków jest to, że dane sieciowe są często zaszyfrowane, a jedyny dostęp do nieszyfrowanych danych z perspektywy systemu monitorowania komunikacji ogranicza się do nagłówków przesyłanych pakietów. Analiza zagrożeń związanych z tego typu danymi jest zatem trudna i często wymaga implementacji rozwiązań budujących pewną warstwę abstrakcji między obserwowanymi pakietami sieciowymi a analizowanymi próbkami [3].

Dominujący trend w implementacji takich rozwiązań polega na analizie ruchu pod kątem przepływów sieciowych (ang. *network traffic flows*) zachodzących między hostami w sieci. Analiza ta polega na grupowaniu pojawiających się pakietów na podstawie adresów źródłowych i docelowych oraz portów, a następnie badaniu rozmiarów kolejnych pakietów, flag obecnych w komunikacji, czy przerw czasowych pomiędzy kolejnymi pakietami w przepływie. Aktualne rozwiązania zazwyczaj opierają się na odrębnej analizie przepływów dwukierunkowych (ze źródłowego do docelowego hosta i odwrotnie), gdyż wskazano, że takie podejście zwiększa skuteczność wykrywania zagrożeń cybernetycznych [4].

Jedną z pożądanych cech każdego NIDS jest odporność i wytrzymałość na próby oszukania przez atakujących. Dodatkowo, NIDS oparte na ML mają na celu zapewnienie długoterminowej efektywności logiki wykrywania w porównaniu do systemów opartych na sygnaturach (m.in. dzięki zdolnościom do wykrywania wcześniej nieznanych ataków). Jednak komponent ML w systemie może wprowadzać słabości, które mogą wykorzystać potencjalni atakujący, znajdując sposoby na obejście takiej warstwy obrony.

Wielu inżynierów systemów uczenia maszynowego często pomija odporność swoich systemów na ataki celujące bezpośrednio w algorytmy, co często pozostawia te systemy podatnymi na ataki [1]. Biorąc pod uwagę stale rosnącą popularność stosowania takich algorytmów w NIDS, oraz uznając, że te systemy często reprezentują jedną z głównych i krytycznych warstw chroniących zasoby przed atakującymi, w tym artykule postawiono istotne pytanie badawcze: *W jakim stopniu takie ataki są możliwe do przeprowadzenia przeciwko systemom NIDS w rzeczywistych scenariuszach?*

2. ATRYBUTY OPARTE NA ANALIZIE PRZEPŁYWÓW

Dokonując przeglądu literatury i znanych projektów, zidentyfikowano ogólny zestaw atrybutów wspólnych dla różnych rozwiązań analizujących przepływy sie-

ciowe. Ze względu na ich niską wzajemną korelację w analizie pakietów sieciowych, atrybuty te zostały podzielone na dwa rodzaje grup często powtarzających się w różnych rozwiązaniach: objętościowe oraz czasowe [8].

Atrybuty objętościowe obejmują takie parametry, jak liczba bajtów lub pakietów transmitowanych w trakcie przepływu. Dodatkowo, atrybuty objętościowe uwzględniają także występowanie flag w nagłówkach pakietów dla danego przepływu.

Atrybuty czasowe obejmują głównie całkowity czas trwania komunikacji (przepływu) oraz czas między przybyciem kolejnych pakietów (ang. *Packet Inter-Arrival Times*, PIATs).

Obie z grup atrybutów są mierzone zarówno kierunkowo, z rozbiciem na wskazanie nadawcy i odbiorcy w danym przepływie, jak i całościowo dla całego przepływu.

3. PROPONOWANE ROZWIĄZANIE

3.1 Techniki Perturbacji

W niniejszym podrozdziale omówiono różne metody mające na celu perturbację parametrów w pakietach sieciowych.

Kilka omawianych tutaj technik było już wcześniej badanych w literaturze, podczas gdy inne są nowe i, o ile nam wiadomo, nie były wcześniej testowane. Podstawową zasadą tych technik jest minimalizacja wpływu na rzeczywistą komunikację przy jednoczesnym potencjalnym umożliwieniu unikania systemów wykrywania. Oto niektóre z kluczowych technik:

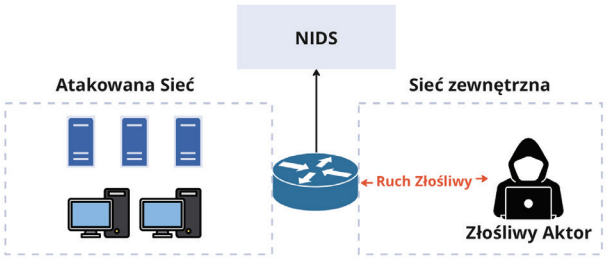
- **Uzupełnianie Zawartości (ang. *Padding*):** Metoda ta dodaje dodatkowe bajty do pakietów, nie zakłócając ich funkcjonalności sieciowej, a tym samym nie zakłócając działania ukrywanych złośliwych technik.
- **Manipulacja Czasami Delta:** Często stosowana przez frameworki Command & Control do ukrywania komunikacji [5], ta strategia wprowadza zmienność w odstępach czasowych między kolejnymi transmisjami pakietów, wpływając tym samym na czas między przybyciem pakietów (PIAT).
- **Fragmentacja Pakietów:** Poprzez modyfikację Maksymalnej Jednostki Transmisyjnej (ang. *Maximum Transmit Unit*, MTU), technika ta prowadzi do fragmentacji pakietów, aby spełnić określone ograniczenia wielkości, zmniejszając rozmiary poszczególnych pakietów (oraz generując więcej mniejszych pakietów mieszczących się w zakresie ustalonego MTU)

Kluczowym aspektem jest ustalenie, w jaki sposób poszczególne techniki będą stosowane do pakietów w danym przepływie. Rozważanie każdego pakietu indywidualnie jest niepraktyczne ze względu na konieczność ich zbiorczej analizy w ramach przepływu, dlatego techniki powinny być wdrażane równocześnie w całym przepływie (a nie na poziomie każdego pakietu osobno). Dla każdego rodzaju perturbacji zostaną określone parametry, które precyzyjnie określą, jak takie modyfikacje będą przeprowadzane. Będzie to polegać na mnożeniu kolejnych parametrów danego przepływu (np. rozmiarów następujących po sobie pakietów) przez współczynnik wyznaczony z parametrów rozkładu stanowiącego opis danej techniki perturbacji.

3.2 Model Zagrożenia

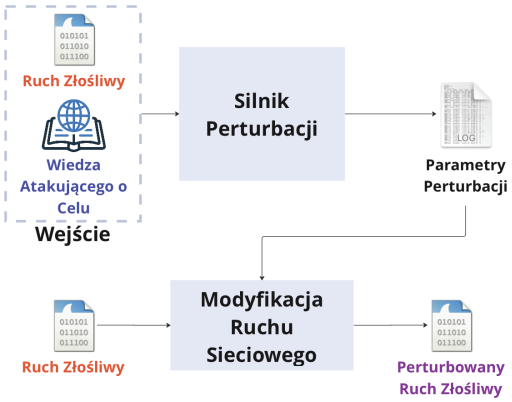
Rys. 1 ilustruje przyjęty model zagrożenia dla rozpatrywanego w artykule ataku. Chroniona sieć jest wyposażona w router brzegowy na granicy atakowanej sieci, który widzi cały ruch sieciowy. Monitorowany ruch jest następnie przetwarzany i poddawany dalszej analizie oraz wykrywaniu przez odpowiednie algorytmy w NIDS.

Typowo, atakujący uzyskuje dostęp do sieci poprzez zewnętrzny kanał, omijając router brzegowy, na przykład za pomocą socjotechnik i phishingu. Taki dostęp umożliwia atakującemu przeprowadzenie wstępnego rozpoznania, co pozwala na zebranie istotnych informacji z docelowej sieci. W omawianym modelu, atakujący może podsłuchiwać część ruchu sieciowego i wyodrębniać podstawowe parametry związane z atrybutami omówionymi w Rozdziale 2.



Rysunek 1: Model zagrożenia przyjęty w proponowanym ataku.

3.3 Wysokopoziomowa logika



Rysunek 2: Schemat wysokopoziomowy proponowanego rozwiązania.

Schemat proponowanego rozwiązania przedstawia Rys. 2. Adaptuje ono się do złośliwego ruchu sieciowego oraz wiedzy atakującego w zależności od scenariusza (cechy ruchu normalnego albo atakowany model docelowy), oblicza odpowiednie parametry, a na ich podstawie wprowadza perturbacje do rzeczywistego złośliwego ruchu.

Poszczególne komponenty przedstawione na Rys. 2 działają następująco:

- **Silnik Perturbacji**: Jego zadaniem jest rozwiązanie następującego problemu optymalizacji:

$$\min_{\mathbf{p}} \|f(\mathbf{x}_{zlosliwy} + \mathbf{p}) - f(\mathbf{x}_{normalny})\| \quad (1)$$

gdzie $\mathbf{p}$ oznacza zastosowane perturbacje, $\mathbf{x}_{zlosliwy}$ to złośliwy ruch, $\mathbf{x}_{normalny}$ to normalny ruch, a $f(\cdot)$ jest funkcją reprezentującą statystyczne cechy ruchu (przepływów sieciowych).

Do tego celu wybrano algorytm genetyczny ze względu na jego wszechstronność, efektywność w dynamicznych środowiskach oraz wysokie zdolności eksploracyjne, pozwalające unikać lokalnych minimów.

Wyzwanie związane z funkcją f polega na jej precyzyjnym zdefiniowaniu, a konkretnie na sposobie oceny podobieństwa przepływów. W przypadku naszego algorytmu genetycznego, to równanie odpowiada funkcji dopasowania (ang. *fitness*). Po gruntownej analizie stwierdzono, że funkcja zdolna do aproksymacji takich nieliniowości i uchwycenia różnych subtelnych zależności będzie miała strukturę płytkiego drzewa. Wybrano losowy las płytkich drzew (maksymalna głębokość 3) ze względu zdolności regresyjne takiego rozwiązania. Takie podejście pozwala na modelowanie złożonych wzorców przy zachowaniu efektywności obliczeniowej.

- **Parametry Perturbacji**: Algorytm genetyczny przeszukuje serię kroków, z odpowiednimi parametrami, w celu optymalizacji unikania detekcji dla każdej próbki. Parametry te, obliczone przez algorytm, tworzą listę perturbacji do zastosowania dla danej próbki.

4. BADANIA PROPONOWANEJ TECHNIKI

4.1 Metodologia

Atakowany model. Ze względu na charakterystykę modeli detekcji w kontekście wykrywania sieciowego i danych tabelarycznych, nie ma istotnych różnic w możliwościach pomiędzy różnymi typami modeli [7]. W związku z tym zastosowano model, który zapewnia solidne wyniki przy minimalnym dostrajaniu parametrów: CatBoost[1].

Zbiory danych. Do przeprowadzenia badania wykorzystano zbiór danych CTU-13 [2], który dostarcza mieszanki autentycznego ruchu złośliwego oraz normalnego, z celem umożliwienia wykrycia ruchu botnetowego. W analizowanym scenariuszu wykorzystano wszystkie oryginalne podzbiory CTU-13 w ich odpowiednich konfiguracjach, aby jak najbliżej odwzorować warunki rzeczywiste. Liczba scenariuszy poddanych testowaniu będzie odpowiadała liczbie złośliwych scenariuszy dostarczonych przez autorów zbioru. Atakowany model będzie trenowany na 80% całości ruchu normalnego oraz wszystkich scenariuszach złośliwych z wyjątkiem tego, który jest aktualnie testowany pod kątem odporności na ataki przeciwstawne. W związku z tym zostanie wytrenowanych n modeli docelowych, każdy odpowiadający innemu scenariuszowi. Takie podejście ma na celu jak najbliższe odwzorowanie sytuacji rzeczywistej, w której model docelowy jest trenowany na wielu próbkach normalnego i złośliwego ruchu, ale nie miał bezpośredniego kontaktu z próbkami ruchu, dla których dokonuje detekcji w rzeczywistych warunkach.

[1]https://catboost.ai

[2]https://github.com/ppopiolek/CyberAML

Scenariusze treningowe i testowe Zostanie wykorzystane niestandardowe narzędzie do analizy przepływów[2]. Dane zostaną poddane wyodrębnieniu kluczowych atrybutów istotnych dla narzędzi analizy przepływów, jak opisano w Rozdziale 2. Atrybuty te obejmują metryki czasowe i objętościowe, rozważane zarówno kierunkowo (źródło-do-celu i celdo-źródła), jak i całościowo (agregując cały przepływ w pojedynczy zestaw danych). We wszystkich scenariuszach, uwzględniane będą jedynie atrybuty związane z analizą statystyczną rozmiarów pakietów oraz czasów między nadejściami kolejnych pakietów danego przepływu (PIATs).

W kontekście scenariuszy testowych analizowane będą dwa przypadki obejmujące traktowanie atakowanego modelu jako ukryty lub dostępny dla atakującego. Oznacza to, że algorytm genetyczny do dopasowania wykorzysta albo bezpośrednio atakowany model, albo funkcję aproksymacyjną przygotowaną na podstawie wiedzy zdobytej przez atakującego (las losowy).

4.2 Wyniki

4.2.1 Wyniki dla oryginalnych danych

Tabela 1 opisuje wyniki predykcji próbek z różnych scenariuszy dla klasyfikatorów dla poszczególnych przypadków.

Tabela 1: Wyniki dla oryginalnych danych przed wprowadzeniem perturbacji.

Zbiór danych	*Dokładność*
20110811-neris	98,68%
20110812-rbot	99,88%
20110815-fast-flux-2	99,99%
20110815-fast-flux-2	99,99%
20110815-fast-flux	99,99%
20110815-rbot-dos	99,99%
20110816-donbot	99,99%
20110816-qvod	99,78%
20110816-sogou	99,99%
20110817-bot	99,99%
20110818-bot-2	99,99%
20110819-bot	99,09%

4.2.2 Wyniki dla próbek perturbowanych

Scenariusz ataku obejmuje wyniki, w których złośliwy ruch jest odpowiednio modyfikowany na poziomie surowego ruchu sieciowego w plikach pcap. Silnik Perturbacji oblicza odpowiednie atrybuty, które następnie są wykorzystywane przez moduł Zmiany Ruchu do przeprowadzenia rzeczywistych i celowanych perturbacji. Ostateczna analiza przepływu jest przeprowadzana za pomocą niestandardowej implementacji, która ma na celu możliwie ścisłe odwzorowanie wymagań protokołów sieciowych. Oznacza to, że atakujący w takim scenariuszu ma dostęp do narzędzia, które analizuje przepływ w docelowym systemie. Tabela 2 przedstawia szczegółowy opis wyników dla każdego zrzutu.

Tabela 2: Wyniki po wprowadzeniu perturbacji do plików pcap.

Zbiór danych	*Model docelowy*	*Własna funkcja dopasowania*
20110811-neris	80,99%	92,93%
20110812-rbot	42,34%	93,50%
20110815-fast-flux-2	75,12%	87,98%
20110815-fast-flux	74,98%	90,53%
20110815-rbot-dos	39,80%	61,69%
20110816-donbot	88,79%	92,82%
20110816-qvod	74,35%	87,03%
20110816-sogou	85,74%	94,04%
20110817-bot	83,89%	97,18%
20110818-bot-2	72,64%	79,94%
20110819-bot	75,09%	90,45%

Obserwujemy, że prezentowane techniki mogą mieć realny wpływ na zdolności wykrywania przez algorytmy detekcji, w najlepszych wypadku osiągając uniknięcie wykrywania o około 60%. Średnio 17,84% dla przypadku, w którym atakujący ma dostęp do odpytywania atakowanego systemu oraz średnio 11,99%, gdy atakujący zmuszony jest wykorzystywać ograniczoną wiedzę.

5. PODSUMOWANIE

Praca bada możliwości ataków przeciwstawnych na systemy NIDS oparte na analizie przepływów sieciowych za pomocą metod uczenia maszynowego. Wraz z rosnącym zastosowaniem tych metod w cyberbezpieczeństwie, zrozumienie zagrożeń związanych z ich wykorzystaniem jest kluczowe.

Wyniki przedstawione w artykule pokazują, że proponowane techniki mogą znacząco obniżyć skuteczność wykrywania przez NIDS. Ocena na publicznym zbiorze danych CTU-13 wykazała skuteczność ataków omijania opartych na celowych perturbacjach ruchu sieciowego, redukując dokładność wykrywania z 99,99% do około 40% w najlepszym przypadku. Podkreśla to konieczność zwiększenia odporności systemów opartych na uczeniu maszynowym na takie zagrożenia.

Przyszłe badania bazujące na wynikach tej pracy mogą obejmować głównie następujące obszary:

- Ocena proponowanych technik w rzeczywistych środowiskach NIDS opartych na analizie przepływów.
- Opracowanie techniki, która będzie efektywnie działać w środowiskach czasu rzeczywistego.
- Propozycja mniej naiwnego i potencjalnie szybszego silnika perturbacji w porównaniu do obecnie proponowanego algorytmu genetycznego.

Dalsze badania w tych obszarach zwiększą praktyczność i efektywność omawianych technik, przyczyniając się do rozwoju środków cyberbezpieczeństwa w dynamicznych i złożonych środowiskach sieciowych.

UWAGI

Badania były finansowane ze środków Politechniki Warszawskiej w ramach Programu Inicjatywa Doskonałości – Uczelnia Badawcza (IDUB).

LITERATURA

[1] Apruzzese, Giovanni, Hyrum S Anderson, Savino Dambra, David Freeman, Fabio Pierazzi, Kevin Roundy. 2023. „"Real Attackers Don't Compute Gradients": Bridging the Gap Between Adversarial ML Research and Practice". 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), 339–364. IEEE.

[2] Garcia, S, V Uhlir. 2014. „The CTU-13 dataset. A labeled dataset with botnet, normal and background traffic". Stratosphere Lab.

[3] Khraisat, Ansam, Iqbal Gondal, Peter Vamplew, and Joarder Kamruzzaman. 2019. „Survey of intrusion detection systems: techniques, datasets and challenges". Cybersecurity, 2(1): 1–22. Springer.

[4] Li, Bingdong, Jeff Springer, George Bebis, Mehmet Hadi Gunes. 2013. „A survey of network flow applications". Journal of Network and Computer Applications 36 (2) : 567–581. Elsevier.

[5] Liu, Yali. 2010. „Hide-and-seek: Concealment and detection of sensitive data exfiltration in network traffic". University of California, Davis.

[6] Sevilla, Jaime, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, Pablo Villalobos. 2022. „Compute trends across three eras of machine learning". 2022 International Joint Conference on Neural Networks (IJCNN), 1–8. IEEE.

[7] Thapa, Niraj, Zhipeng Liu, Dukka B. Kc, Balakrishna Gokaraju, and Kaushik Roy. 2020. „Comparison of machine learning and deep learning models for network intrusion detection systems". Future Internet, 12(10): 167. MDPI.

[8] Umer, Muhammad Fahad, Muhammad Sher, and Yaxin Bi. 2017. „Flow-based intrusion detection: Techniques and challenges". Computers Security, 70: 238–254. Elsevier.

ROZWÓJ I ZASTOSOWANIE SYSTEMU MONITORINGU URZĄDZEŃ AUTOMATYKI PRZEMYSŁOWEJ SMUAP

DEVELOPMENT AND APPLICATION OF INDUSTRIAL AUTOMATION DEVICE MONITORING SYSTEM - SMUAP

Franciszek Rakowski[1]; Piotr Ostapowicz[2]; Maciej Stróżyk[3]; Maciej Głowiak[4]; Marek Smolik[5]

[1] ICsec S.A., Poznań, franciszek.rakowski@icsec.pl
[2] Poznańskie Centrum Superkomputerowo-Sieciowe afiliowane przy Instytucie Chemii Bioorganicznej PAN, Poznań, thewasp@man.poznan.pl
[3] Poznańskie Centrum Superkomputerowo-Sieciowe afiliowane przy Instytucie Chemii Bioorganicznej PAN, Poznań, mackostr@man.poznan.pl
[4] Poznańskie Centrum Superkomputerowo-Sieciowe afiliowane przy Instytucie Chemii Bioorganicznej PAN, Poznań, mac@man.poznan.pl
[5] ICsec S.A., Poznań, marek.smolik@icsec.pl

DOI: 10.15199/59.2024.4.12

Streszczenie: **W odpowiedzi na rosnące wyzwania stawiane przedsiębiorstwom produkcyjnym przez rewolucję przemysłową 4.0, w tym zagrożenia atakami cybernetycznymi, opracowany został System Monitoringu Urządzeń Automatyki Przemysłowej (SMUAP). Architektura SMUAP oparta jest na pasywnym monitorowaniu ruchu sieciowego oraz analizie komunikacji między urządzeniami. Moduł analityczny systemu oparty jest głównie na sztucznej inteligencji i uczeniu maszynowym, natomiast narzędziem do monitorowania sieci jest pasywny sniffer (w kilku opcjach, włączając technologię FPGA).**

Abstract: **In response to the growing challenges faced by manufacturing enterprises due to the Industry 4.0 revolution, including the threat of cyber attacks, the Industrial Automation Monitoring System (SMUAP) has been developed. The architecture of SMUAP is based on passive monitoring of network traffic and analysis of communication between devices. The analytical module of the system is mostly based on artificial intelligence and machine learning, while the network monitoring tool is passive sniffer (in several options, including FPGA technology).**

Słowa kluczowe: **cyberbezpieczeństwo, detekcja, malware, sieci przemysłowe, system IDS, FPGA**

Keywords: **cybersecurity, detection, malware, industrial networks, intrusion detection systems, FPGA**

1. WSTĘP

W ramach odpowiedzi na rosnące wyzwania, które stawia przed przedsiębiorstwami produkcyjnymi rewolucja przemysłowa 4.0, opracowany został zestaw funkcjonalności zwiększających poziom cyberbezpieczeństwa - System Monitoringu Urządzeń Automatyki Przemysłowej (w skrócie SMUAP). System SMUAP obejmuje w pełni funkcjonalne urządzenie do pasywnego sniffowania sieci przemysłowej (sniffer w technologii FPGA oraz w formie softwarowej) oraz innowacyjny serwer analityczny z szeregiem unikalnych funkcjonalności. Synergia między tymi dwoma komponentami umożliwia monitorowanie sieci i urządzeń automatyki i wykrywanie anomalii oraz podatności w sieciach przemysłowych. SMUAP jest systemem rozszerzającym funkcjonalności istniejącego już rozwiązania SCADvance XP w obszarze automatycznej detekcji i charakterystyki elementów sieci (urządzeń końcowych, serwerów, urządzeń sieciowych) pod kątem ich podatności na zagrożenia oraz umożliwiający szybką i automatyczną inspekcję pakietów komunikacji sieciowej w celu wykrycia anomalii w ruchu sieciowym i zagrożeń z nich wynikających. Dlatego też nowe funkcjonalności SMUAP zostały zintegrowane w SCADvance XP tworząc nowej generacji, kompleksowe rozwiązanie klasy IDS (*Intrusion Detection System*). System SCADvance XP umożliwia monitorowanie w trybie rzeczywistym, egzaminowanie i niemal natychmiastowe reagowanie na sytuacje niepożądane w działaniu sieci przemysłowej (włączając ataki typu *Zero-Day*). Rozwiązanie te komercjalizowane jest przez spółkę ICsec S.A. i implementowane w infrastrukturze przemysłowej klientów, w tym w infrastrukturze krytycznej, np. w przedsiębiorstwach wodno-kanalizacyjnych lub wydobywczych. Początki systemu sięgają projektu badawczo-rozwojowego o akronimie SMUAP (POIR.01.01.01-00-0756/19, faza badań przemysłowych i prac rozwojowych), który był współfinansowany przez Narodowe Centrum Badań i Rozwoju w Polsce. Celem tego artykułu jest przedstawienie najważniejszych wyników badawczych projektu, które stanowią istotny wkład w aktualny stan wiedzy w tym zakresie.

2. Architektura i funkcjonalności Systemu Monitoringu Urządzeń Automatyki Przemysłowej

System SMUAP stanowi dedykowane przemysłowi kompleksowe rozwiązanie do cyberochrony sieci przemysłowych, które oparte jest na analizie komunikacji po-

między urządzeniami. Działanie systemu jest uruchamiane na serwerze centralnym, do którego przekazywane są dane zbierane przez sondy nasłuchujące, w tym sondę USS. Nieprzerwany monitoring sieci może dostarczyć wartościowych informacji dla profesjonalistów odpowiedzialnych za ochronę sieci OT (*operational technology*).

System stanowi część sprzętowa, na którą składa się Uniwersalny Sniffer Sieciowy (USS) wraz z firmware oraz część softwarowa, na którą składa się innowacyjny serwer analityczny oparty na metodach sztucznej inteligencji i uczenia maszynowego. Ponadto, w odpowiedzi na pojawiające się potrzeby rozwiązań sieciowych, w ramach projektu opracowany został Hybrydowy Sniffer Sieciowy (HSS). HSS to oprogramowanie typu sniffer, które można instalować na dowolnych urządzeniach używanych w sieciach przemysłowych opartych na procesorze x86. HSS realizuje wszystkie funkcjonalności USS, natomiast ich wykonywanie w pełnym lub ograniczonym zakresie zależy od parametrów sprzętu, na którym zostanie zainstalowany. W związku z tym, iż HSS realizuje wszystkie funkcjonalności USS w dalszej części artykułu uwaga skupiona zostanie głównie na funkcjonalnościach serwera analitycznego oraz USS.

Interfejs użytkownika systemu zaprojektowany został w sposób modularny, umożliwiając jego manualną konfigurację oraz obserwację struktury sieci i jej dynamiki w czasie rzeczywistym. Użytkownik systemu ma możliwość zobaczenia podstawowych informacji o sieci: liczbę węzłów, natężenie komunikacji, graf sieci itp., jak również dokonać manualnej konfiguracji systemu, tj. ustalić liczbę uruchamianych modułów, czas pracy i ich precyzję. Ważnym atutem tego systemu są moduły do automatycznego wykrywania zagrożeń. Złożoność systemu powoduje także, iż istnieje również znaczna liczba szczegółowych parametrów konfiguracyjnych, które w istotny sposób mogą wpływać na działanie algorytmów. Parametry są konfigurowane automatycznie poprzez specyficzne dla modułów procedury. Funkcjonalność automatycznej - wystarczająco dobrej - konfiguracji systemu jest podstawowym wymaganiem stawianym systemom klasy IDS, mającym obsługiwać duże instalacje przemysłowe.

System SMUAP w oparciu o pasywny monitoring ruchu sieciowego realizuje szereg funkcjonalności kluczowych dla realizacji polityk bezpieczeństwa sieci OT. Najistotniejsze przedstawione zostały poniżej. Pierwszą jest **moduł *Device Discovery* (DDM)**, który służy do rozpoznawania szczegółów urządzeń w sieci OT. Kolejną funkcjonalnością jest zaawansowany zestaw narzędzi uczenia maszynowego ***Configurable Artificial Intelligence* - CAI** (tzw. system CAI) umożliwiający automatyczną konfigurację hiperparametrów sterujących pracą obecnych w oprogramowaniu algorytmów sztucznej inteligencji służących wykrywaniu anomalii w ruchu sieciowym, mogących świadczyć o przeprowadzanym właśnie ataku hakerskim lub innej niechcianej ingerencji zewnętrznej w stan i dynamikę sieci przemysłowej. System CAI składa się z zestawu modułów, które śledzą poszczególne cechy dynamiki komunikacji, nie tylko na poziomie dozwolonych zakresów, ale także na poziomie morfologii sygnału. Optymalne wykorzystanie tych modułów można otrzymać dla połączeń, dla których występują wyraźne charakterystyki zmienności cech od prostych typu sinusoidalnego czy prostokątnego do bardziej złożonych, ale z wyraźnymi powtarzającymi się motywami w ich morfologii. Wykorzystanie tych narzędzi wiąże się z zaangażowaniem znaczniejszych zasobów pamięciowych i obliczeniowych oraz z procedurą wdrożenia na poziomie zaawansowanym. Innym modułem SMUAP jest ***Vunerability Checker* (VC)**. Służy on do analizy podatności urządzeń w systemie na podstawie bazy danych o urządzeniach i podatnościach: CPE (*Common Platform Enumeration*), CVE (*Common Vulnerability Enumeration*) i CWE (*Common Weakness Enumeration*). Moduł VC pozwala na porównanie listy urządzeń w sieci z bazami podatności i na tej podstawie znajduje podatne na ataki elementy. Kolejną funkcjonalnością jest moduł ***Thread Traffic Analyzer* (TTA)**, który służy do monitorowania sieci bazując na sygnaturach znanych ataków. Jest to system wykrywania anomalii i włamań, który jest oparty o zdefiniowane i znane reguły ruchu i zachowania. TTA używa zestawów znanych reguł do monitorowania ruchu sieciowego i wyzwala alerty za każdym razem, gdy wystąpią podejrzane zdarzenia. Prawdopodobieństwo anomalii jest oparte o poziom zagrożenia danej reguły.

W dalszej części artykułu, ze względu na zastosowanie nowatorskich metod sztucznej inteligencji i uczenia maszynowego, przybliżone zostaną dwie wspomniane wyżej funkcjonalności: *Device Discovery* oraz *Configurable Artificial Intelligence.*

3. Algorytmy wykrywania anomalii w sieciach OT bazujące na metodach AI i ML

Działanie systemu opiera się na tym, iż odczytuje i przekazuje do serwera analitycznego pakiety sieciowe wraz z nadanymi im znacznikami czasowymi. Pakiety te mogą należeć do szerokiego repertuaru protokołów komunikacji sieciowej, zarówno ethernetowych jak i szeregowych. Ze względu na wymóg uniwersalności systemu detekcji anomalii, każdemu pakietowi przypisywany jest wektor deskryptorów o metodzie tworzenia inwariantnej w stosunku do typu pakietu. Metoda ta wykorzystuje zawartość informacyjną pakietu, cechy opisujące wielkość pakietu oraz znaczniki czasu nadania pakietu. Celem działania serwera analitycznego jest wykrycie anomalii przejawiającej się w obrębie poszczególnego pakietu lub ich grupy. Określenie definicji anomalii w procesach stochastycznych nie jest jednoznaczne i zależy od kontekstu i celu detekcji. Normatywność w silnym sensie, stosowana w algorytmach CAI, nie jest oparta o przedziały rejestrowanych wielkości, ale opiera się także o morfologię sygnału. Dla sygnałów, które zawierają powtarzające się fragmenty, mikromotywy jest możliwe zdefiniowanie ich słownika - który stanowi normę dla sygnału. Słownik mikromotywów morfologii sygnału definiowany jest niejawnie poprzez zastosowanie metod autoregresyjnego uczenia maszynowego. tj. *Passive-Aggresive, MulitLayer Perceptron*, czy *RidgeRegression*. Ścisłe definicje niniejszych metod, umożliwiających praktyczne działanie zawierają także informacje o wartościach hiperparametrów wykorzystanych zarówno w trakcie procedury preprocesingu danych jak i dla ścisłego zdefiniowania samego algorytmu. Jednym z kluczowych, nieznanych apriori hiperparametrów jest wielkość okna czasowego agregacji

cech pakietów. Kolejnymi hiperparametrami są parametry definiujące architekturę poszczególnych algorytmów - przykładowo dla jednego z algorytmów wykorzystywanych w systemie CAI, MLP (*Multi-Layer Perceptron*) będzie to liczba warstw i liczność neuronów w poszczególnych warstwach. W SMUAP faza treningu zawiera zatem rozbudowane procedury wyboru optymalnych algorytmów, stanowiące jądro mechanizmu CAI. Automatycznej konfiguracji podlegają parametry: preprocesingu danych (wybór okna czasowego) oraz parametrów metod ML. W drugiej kolejności dokonywany jest wybór najskuteczniejszych zestawów parametrów w połączeniu z predefiniowanymi statystykami oryginalnych wartości (minimalna i maksymalna wartość w oknie, wartość oczekiwana, odchylenie standardowe) oraz wybór najskuteczniejszej metody ML spośród sześciu zaimplementowanych: *Passive-Agressive, Multi-Layer Perceptron, Rigde-Regression* czy *Long-Short Term Memory Neural Network*. Efektem końcowym działania systemu CAI jest praca algorytmu wykrywającego anomalie na poszczególnych połączeniach w sieci OT w reżimie automatycznie zoptymalizowanych parametrów przetwarzania danych oraz przy wybranych optymalnych (najskuteczniejszych) algorytmach AI.

Kolejnym szczególnie nowatorskim modułem systemu SMUAP jest moduł automatycznego wykrywania i charakterystyki urządzeń - *Device Discovery Module* (DDM). Moduł umożliwia automatyczne wykrywanie typu i specyfikacji urządzenia będącego węzłem monitorowanej sieci (jako producenta, modelu i wersji, nazwy oraz opisu urządzenia w zgodzie ze standardem CPE - *Common Platform Enumeration*) na podstawie głębokiej inspekcji pakietów przesyłanych przez dane urządzenie oraz algorytmów wykorzystujących AI. Techniki wykrywania urządzeń można podzielić na aktywne i pasywne. W pierwszym przypadku system aktywnie odpytuje urządzenia sieciowe w celu uzyskania informacji o ich parametrach. W drugim przypadku system analizuje obecny w sieci ruch, by wydedukować (często statystycznie), rodzaj urządzeń biorących udział w komunikacji. Obie powyższe metody charakteryzuje swoisty zestaw zalet i wad. Aktywne wykrywanie urządzeń pozwala zwykle na dużo wyższą skuteczność analizy, zarówno w kontekście trafności wyników jak i możliwości odpytywania nieaktywnych (w czasie badania) urządzeń. Jest też jednocześnie „hałaśliwe i natrętne", gdyż generuje dodatkowy ruch sieciowy i wpływa na standardową pracę urządzeń. Pasywne wykrywanie urządzeń nie wpływa negatywnie na przepustowość sieci i monitorowanych urządzeń. Wymaga jednak często skomplikowanych, celowanych (w kierunku wąskiej klasy urządzeń) i heurystycznych algorytmów oceny, oraz jest nieczułe na urządzenia będące w uśpieniu (nieaktywne). Ze względu na powyższą charakterystykę, pasywne wykrywanie urządzeń pracuje zwykle w trybie pracy ciągłej, pozwalając na monitoring w czasie rzeczywistym, co umożliwia ciągłe monitorowanie zachowań urządzeń w sieci, a tym samym szybką reakcję na zmiany i zagrożenia. Ponadto dzięki pasywnemu wykrywaniu urządzeń system działa bez ingerencji w samą sieć, co minimalizuje zakłócenia i ryzyko wpływu na działania innych systemów oraz oszczędność zasobów, ponieważ mogą działać na niewielkiej mocy obliczeniowej i zasobach, co sprawia, że są one bardziej efektywne energetycznie i mniej obciążają infrastrukturę sieciową.

Zaimplementowany w systemie SMUAP model rozpoznawania urządzeń działający w sieci Ethernet składa się z trzech podmodułów, gdzie predykcja odnośnie nazwy producenta i modelu jest wykonywana kaskadowo. Pierwszy z podmodułów jest oparty na odczytaniu i zestawieniu adresu MAC z bazą OUI i zapewnia rozpoznanie nazwy producenta. Drugi podmoduł jest oparty o protokoły rozgłoszeniowe (LLDP, CDP, CIP), które zawierają informacje zdefiniowane przez producenta o urządzeniu. Ze względu na fakt, iż nie wszystkie urządzenia w danej instalacji wysyłają pakiety rozgłoszeniowe, utworzony został trzeci podmoduł, wykorzystujący uczenie maszynowe i rozpoznający je na podstawie wcześniej wytrenowanego klasyfikatora typu Las Losowy (*Random Forest*), zawierającego niejawną bazę sygnatur producenta, modelu i wersji urządzenia lub oprogramowania.

4. Uniwersalny Sniffer Sieciowy - sprzętowa akceleracja z wykorzystaniem specjalizowanych układów cyfrowych

Jak zostało wyżej wspomniane, elementem systemu SMUAP jest Uniwersalny Sniffer Sieciowy (USS). Jest to sprzętowa sonda umożliwiająca pasywne monitorowanie sieci przemysłowych (OT) i komputerowych (IT) w czasie rzeczywistym oraz przesyłanie odczytanych danych z monitorowanych interfejsów do dalszej analizy i archiwizacji na serwerze analitycznym. Warto wspomnieć, że u podstaw wnioskowania na temat zjawisk transmisyjnych (przykładowo: wykrywanie ataków czy anomalii) leży analiza podstawowych struktur binarnych, wykorzystywanych do spełnienia założeń modelu odniesienia OSI. Całościowe przeniesienie "ciężaru" przetwarzania na klasyczne jednostki skalarne (procesory systemowe ogólnego przeznaczenia) radykalnie obniża poziom przepływności danych, ograniczając tym samym ogólną przydatność systemu. Wprowadzenie akceleracji sprzętowej w zakresie rozbioru i analizy ciągów binarnych - zwłaszcza w przypadku sieci budowanych przy wykorzystaniu interfejsów o wysokich przepustowościach (>1Gbps) staje się koniecznością, czego wyrazem jest dobrze obserwowany i dynamiczny rozwój procesorów sieciowych - układów ASIC, wyspecjalizowanych do szeroko pojętego przetwarzania danych sieciowych.

Sonda sieciowa systemu SMUAP jest zbudowana jako urządzenie elektroniczne wyposażone w jednostkę przetwarzania zintegrowaną z zasobami FPGA oraz zestaw interfejsów sieciowych (Ethernet, RS, CAN, itp.), pozwalających na podłączenie urządzenia do dowolnej sieci w celu monitorowania. Główną funkcjonalnością sondy USS jest obsługa ścieżki przetwarzania i transport danych z monitorowanych interfejsów do serwera analitycznego. Kluczowe przetwarzanie danych jest wykonywane przez jednostki sprzętowe zaimplementowane w strukturach FPGA, a cały proces jest konfigurowany

i nadzorowany z poziomu aplikacji zarządzania. Urządzenie to charakteryzuje się możliwością kompleksowego przetwarzania strumieni transmisyjnych, rozumianego przez obsługę rozmaitych interfejsów sieciowych. Pomimo wyraźnego wzrostu znaczenia standardu *ethernet* w obszarze sieci przemysłowych, bardzo istotną jest możliwość monitorowania zjawisk transmisyjnych, związanych z prostszymi wariantami, wykorzystywanymi niezmiennie do budowy zaawansowanych struktur automatyki przemysłowej, ze względu na ich skuteczną i stabilną pracę (RS485, CAN – Profibus, Modbus, Can).

Zasadniczym założeniem projektu stało się opracowanie architektury niezależnego układu cyfrowego, wyspecjalizowanego do przetwarzania danych bez względu na standard transmisyjny. Ośrodkiem implementacji sprzętowego akceleratora został tzw. obszar "logiki programowalnej" (PL) wybranego układu MPSoC (Xilinx/AMD ZU4EG). Cechą wspólną obsługiwanych transmisji - ze względu na ich zgodność z założeniami modelu odniesienia OSI, jest nadawanie i odbiór ciągów binarnych o skończonej długości, przeto normalizacja na poziomie warstwy sieciowej pozwoliła na zaprojektowanie uniwersalnej jednostki, oznaczanej skrótowo **SPU** (*Stream Processing Unit*), zorientowanej na przetwarzanie danych w ruchu.

Istotą przetwarzania, przy zastosowaniu opracowanej jednostki jest trójetapowa praca, polegająca na: a) dekompozycji ciągów na podciągi o parametryzowanej (programem przetwarzania dla SPU) długości i parametryzowanym umiejscowieniu ciągu; b) analizie logicznej wyekstrahowanych danych; c) syntezie strumienia wyjściowego. Dopuszczalna gradacja podziału strumienia oraz stopień zrównoleglenia procesu przetwarzania są parametryzowane na poziomie źródła opisu sprzętu. Praca jednostki SPU jest ustalana na podstawie programu przetwarzania strumienia i w tym zakresie widać podobieństwo do typowych jednostek skalarnych (rozkazy z listy instrukcji, układane w sekwencję), jednak brak ograniczeń wynikających z ortogonalności instrukcji procesorów RISC, oraz możliwość wyzwalania równoległych procesów, wpływającego pozytywnie na sprawność obróbki danych, wprowadza odmienną jakość procesów przetwarzania strumieni (w odniesieniu do klasycznego schematu). Krótkim przykładem, który dostatecznie dobrze obrazuje istotę różnic w schemacie przetwarzania jest selekcja niesąsiadujących ze sobą pól pakietu, reprezentowanych przez zróżnicowane typy danych, oraz ich złożenie (konkatenacja) w celu uzyskania ciągu wyjściowego, stanowiącego podstawę adresowania pamięci CAM (*Content Addresable Memory*). W przypadku klasycznej jednostki przetwarzającej, złożoność procesu wyrażana przez liczbę sekwencyjnie występujących cykli maszynowych – ze względu na konieczność prowadzenia operacji arytmetyczno-logicznych, oraz wcześniejszą organizację danych w przestrzeni adresowej, będzie wyraźnie wyższa w odniesieniu do jednostki wyspecjalizowanej do przetwarzania strumieni.

5. Podsumowanie

System SMUAP, opracowany w odpowiedzi na współczesne wyzwania cyberbezpieczeństwa, wyróżnia się jako wyjątkowe rozwiązanie na europejskim rynku oraz odpowiada na wiele aktualnych zagrożeń występujących w sieciach OTDzięki innowacyjnym funkcjonalnościom bazującym na metodach AI/ML oraz architekturze Uniwersalnego Sniffera Sieciowego stanowi niezaprzeczalne wzmocnienie bezpieczeństwa infrastruktury OT.

Unikalną cechą systemu SMUAP jest wykorzystanie inteligentnych algorytmów wykrywania anomalii opartych na metodach sztucznej inteligencji i uczenia maszynowego, co pozwala na automatyczną analizę ruchu sieciowego i wykrywanie zagrożeń. Ponadto, kluczowym elementem systemu jest USS, który jest rozwiązaniem sprzętowym i umożliwia monitorowanie sieci przemysłowych w czasie rzeczywistym. Wykorzystuje on zaawansowane technologie przetwarzania strumieniowego i sprzętowej akceleracji, co pozwala na kompleksową analizę danych w ruchu. Zaprojektowany został jako urządzenie elektroniczne wyposażone w układ FPGA i szereg interfejsów sieciowych, umożliwiających połączenie z sieciami w różnych technologiach w celu ich monitorowania.

LITERATURA

[1] AMD-Xilinx. 2019. UG585 Zynq-7000 SoC Technical Reference Manual v1.12.2.;
[2] AMD-Xilinx. 2020. UG1085 Zynq UltraScale+ Device Technical Reference Manual.
[3] Baer J.-L. 2009. “Microprocessor Architecture: From Simple Pipelines to Chip Multiprocessors”, Cambridge University Press.
[4] Briggs, F.A. et al., 1977, "Organization of semicoductor memories for parallel-pipelined processors." IEEE Trans. Comput.
[5] Hajda J., Jakuszewski R. and Ogonowski S. 2021. “Security challenges in industry 4.0 PLC systems”, Applied Sciences, 11(21), 9785.
[6] Kim B., Alawami M.A., Kim E., Oh S., Park J. and Kim H. 2023. “A comparative study of time series anomaly detection models for industrial control systems”, Sensors, 23(3), 1310.
[7] Kotu V., Deshpande B. 2019. “Time Series Forecasting, in Data Science Morgan Kaufmann”, [Chapter 12].
[8] Kumar S., Dalal S., Dixit V. 2014. “The OSI Model: overview on the seven layers of computer networks”, ISSN 2348-1196
[9] Pratomo A. B., Burnap P., Theodorakopoulos G. 2020. "BLATTA: Early Exploit Detection on Network Traffic with Recurrent Neural Networks", Security and Communication Networks.
[10] Twardawa M. G., Smolik M., Rakowski F., Kwiatkowski J., Meyer N. 2024. “SCADvance XP – an intelligent Polish system for threat detection and monitoring of industrial networks”, Security and Defence Quarterly.
[11] Zadnik M., Korenek J., 2005, "Netflow probe intended for high-speed networks"

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

WPŁYW PARAMETRÓW SIECI TYPU TREE PARITY MACHINE NA BEZPIECZEŃSTWO ALGORYTMU WYRÓWNUJĄCEGO WAGI

THE IMPACT OF TREE PARITY MACHINE NETWORK PARAMETERS ON THE SECURITY OF THE WEIGHT EQUALIZATION ALGORITHM

Miłosz Stypiński

AGH Akademia Górniczo-Hutnicza,
Instytut Telekomunikacji,
al. Mickiewicza 30, 30-059 Kraków

DOI: 10.15199/59.2024.4.13

Streszczenie: Znaczący postęp w technologi obliczeń z wykorzystaniem komputerów kwantowych jest ogromnym zagrożeniem dla obecnie wykorzystywanej kryptografii asymetrycznej. Sieci neuronowe typu Tree Parity Machine są alternatywnym algorytmem uzgodnienia klucza kryptograficznego. W ostatnim czasie zaproponowano szereg rozwiązań usprawniających właściwości klucz kryptograficznego otrzymanego z wykorzystaniem wspomnianego rozwiązania. Celem tego artykułu jest zbadanie zależności pomiędzy bezpieczeństwem usprawnień sieci Tree Parity Machine, a jej parametrami i rozmiarem.

Abstract: Significant advancements in quantum computing technology pose a major threat to currently used asymmetric cryptography. Tree Parity Machines are an alternative algorithm for cryptographic key agreement. Recently, several solutions have been proposed to enhance the properties of cryptographic keys obtained using this approach. The aim of this article is to investigate the relationship between the security of Tree Parity Machine network improvements and its parameters and size.

Słowa kluczowe: uzgodnienie klucza kryptograficznego, uczenie wzajemne, sztuczne sieci neuronowe, kryptografia postkwantowa

Keywords: key agreement, mutual learning, artificial neural networks, postquantum cryptography

1. WSTĘP

Komputery kwantowe stanowią potencjalną rewolucję w dziedzinie zaawansowanych obliczeń. Dzięki zastosowaniu zasad mechaniki kwantowej, takich jak superpozycja i splątanie kwantowe, komputer kwantowy pozwala na przyspieszenie rozwiązywania problemów należących do klasy Bounded-error Quantum Polynomial Time (BQP) [2].

Aktualnie wykorzystywane algorytmy kryptografii asymetrycznej korzystają z funkcji algebraicznych, które są na dzień dzisiejszy niezwykle trudne do odwrócenia. Niestety algorytm Shora pozwala na rozwiązanie problemów takich jak faktoryzacja dużych liczb, bądź znalezienie logarytmu dyskretnego w ciele skończonym [14]. W wyniku wspomniane problemy również należą do klasy BQP, a w rezultacie komputer kwantowy o odpowiedniej liczbie kubitów będzie w stanie zagrozić bezpieczeństwu wykorzystywanej kryptografii asymetrycznej.

Znalezienie algorytmów kryptografii postkwantowej, czyli odpornych na działanie komputera kwantowego, jest niezwykle ważne dla utrzymania porównywalnego poziomu bezpieczeństwa w telekomunikacji. Znane instytucje standaryzacyjne prowadzą badania nad wspomnianymi algorytmami. Znaczącymi osiągnięciami w tej kwestii jest rozstrzygnięcie konkursu prowadzonego przez National Institute of Standards and Technology (NIST) [1], gdzie wyłoniono trzech finalistów. Ponadto, rozwijane są biblioteki programistyczne implementujące algorytmy odporne na działanie komputera kwantowego. Znaczącym osiągnięciem w tej tematyce jest projekt Open Quantum Safe [15].

Alternatywnym protokołem uzgodnienia klucza kryptograficznego, będącym tematem tej publikacji, jest wzajemna synchronizacja sieci neuronowych typu Tree Parity Machine (TPM). W [5, 7, 10] została po raz pierwszy zaproponowana struktura sieci typu TPM. Ponadto we wspomnianej publikacji zostało udowodnione, że wzajemne uczenie tych sieci neuronowych skutkuje pełną synchronizacją tych sieci, czyli odpowiadające sobie wagi w obu sieciach posiadają dokładnie takie same wartości. Ponadto autorzy . Do sieci neuronowych typu TPM zaproponowane wiele usprawnień. W [11] i [12] zostały przedstawione usprawnienia polegające na wykorzystaniu odpowiednio algorytmów genetycznych oraz inspirowanego naturą algorytmu optymalizacyjnego. Zmiana sposobu uczenia sieci typu TPM jest oddzielną rodziną usprawnień. Zaproponowane usprawnienia z tej rodziny zostały opisane w [3, 4, 17], gdzie autorzy zaproponowali wykorzystanie odpowiednio wartości zespolonych, wektorów oraz liczb niebinarnych jako wartości wektora wejściowego podczas synchronizacji omawianych sieci neuronowych. Ponadto sieci TPM zostały poddane różnym testom bezpieczeństwa [6, 9, 13]. Wśród przeprowadzonych testów nie zdefiniowano żadnego algorytmu, który umożliwiałby przyspieszenie ataku przy użyciu komputera kwantowego.

Sieci neuronowe typu TPM wykorzystujące niebinarne wektory wejściowe są przedmiotem badań tej publikacji (Nonbinary Tree Parity Machine (NBTPM)).

Wykorzystanie niebinarnych wektorów wejściowych skutkuje przyspieszeniem procesu uczenia sieci neuronowych, a w rezultacie skraca czas wymagany do uzgodnienia klucza kryptograficznego. Negatywnym efektem wy-

korzystania niebinarnych wartości wejściowych jest szybsza synchronizacja jednostki wykonującej atak typu man-in-the-middle [16] oraz większa podatność sieci NBTPM na odzyskanie klucza kryptograficznego w wyniku podsłuchiwania kanału komunikacyjnego. Dalsza ewaluacja bezpieczeństwa sieci neuronowych NBTPM została przeprowadzona w [18]. W celu zmniejszenia negatywnych skutków wynikających z zastosowania niebinarnych wektorów wejściowych, w [19] zaproponowano algorytm, który sprowadza otrzymany klucz kryptograficzny do rozkładu jednorodnego. W tej publikacji zostanie przebadany wpływ rozmiaru sieci NBTPM na skuteczność wspomnianego algorytmu.

2. OPIS MODELU

Struktura TPM to dwuwarstwowa sieć neuronowa o binarnej odpowiedzi. W zależności od rozważanego wariantu wartości uczące mogą przybierać różne wartości. W pierwowzorze sieci TPM wykorzystywane są wartości binarne, jednak w tym artykule rozważany jest wariant sieci NBTPM. Dla tego modelu wektor uczący składa się z liczb całkowitych x_{kn} przybierających wartości od $-M$ do M. W pierwszej warstwie sieci neuronowej typu NBTPM znajduje się K neuronów. Każdy z nich posiada N dedykowanych wejść. Z każdym wejściem rozważanego modelu istnieje powiązana waga w_{kn}, która może przyjmować wartości całkowite z przedziału od $-L$ do L. Funkcja aktywacji w każdym neuronie pierwszej warstwy to zmodyfikowana funkcja signum, która zamiast przybierać wartość równą 0 zwraca -1 albo 1. Wartość y_k, będąca odpowiedzią k-tego neuronu, została przedstawiona w równaniu (1). Wszystkie wartości wynikowe neuronów pierwszej warstwy przekazywane są bezpośrednio do drugiej warstwy, będącej jednocześnie ostatnią. Funkcja aktywacji wspomnianego pojedynczego neuronu jest równa iloczynowi wejść, co stanowi ostateczną, binarną odpowiedź sieci. Model sieci TPM został przedstawiony na rysunku 1.

$$y_k = f_{signum}(\sum_{n=1}^{N} x_{kn} \cdot w_{kn}) \tag{1}$$

Proces uczenia, dalej nazywany synchronizacją lub uczeniem wzajemnym, nie jest zgodny ze standardowym algorytmem propagacji wstecznej ze względu na nieciągłość wartości jakie mogą przyjmować wagi modelu sieci NBTPM. W celu wzajemnego nauczenia dwóch sieci, czyli doprowadzeniu ich do posiadania identycznych wag, wykonywany jest dedykowany algorytm, którego skutkiem jest pełna synchronizacja, co jest tożsame z tym, że obie sieci neuronowe są identyczne. Algorytm uczenia jest przedstawiony poniżej [8].

1. Członkowie synchronizacji uzgadniają zestaw parametrów synchronizowanej sieci NBTPM, czyli wartości dla parametrów (K, L, M, N) oraz niezależnie losują wstępne wartości wag z przedziału liczb całkowitych od $-L$ do L.
2. Losowy uczestnik generuje wektor losowy X i udostępnia go przez publiczny kanał drugiemu uczestnikowi synchronizacji. Wektor składa się z wartości x_{kn}, gdzie x_{kn} jest liczbą całkowitą z przedziału od $-M$ do M
3. Uczestnicy obliczają odpowiedź swoich sieci NBTPM i dzielą się nią wzajemnie
4. W przypadku, gdy odpowiedzi sieci są sobie równe uczestnicy stosują wybraną wcześniej metodę uczenia. Poniżej przedstawiono dostępne metody uczenia, gdzie t oznacza numer iteracji algorytmu, a funkcja Θ to funkcja przyjmująca wartość 1, jeżeli wszystkie jej argumenty są sobie równe; w przeciwnym wypadku przyjmuje wartość 0.
   - *Hebbian*
     $$w_{kn}(t+1) = w_{kn}(t) + O(t)x_{kn}(t)\Theta(y_k(t), O(t)),$$
   - *Anti-Hebbian*
     $$w_{kn}(t+1) = w_{kn}(t) - O(t)x_{kn}(t)\Theta(y_k(t), O(t)),$$
   - *błądzenie losowe*
     $$w_{kn}(t+1) = w_{kn}(t) + x_{kn}(t)\Theta(y_k(t), O(t)).$$
5. Kroki 2-4 wykonywane są aż do osiągnięcia pełnej synchronizacji obu sieci neuronowych.

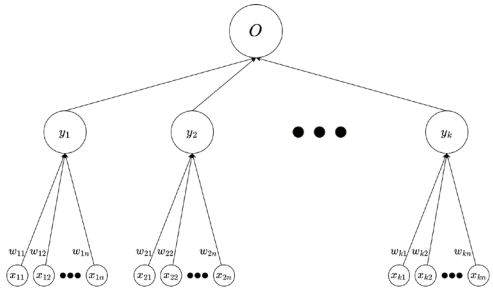
Po wykonaniu algorytmu obie sieci neuronowe są identyczne, czyli odpowiadające sobie wagi posiadają równe sobie wartości. Wektor W stworzony z wag w_{kn} stanowi uzgodniony klucz kryptograficzny, który może następnie zostać wykorzystany w utworzeniu bezpiecznego kanału komunikacyjnego.

Poza pozytywnymi skutkami zastosowania wariantu NBTPM istnieją też skutki negatywne, takiej jak nierównomierny rozkład wartości w wektorze W zwany efektem wartości krańcowych [17]. Jako rozwiązanie problemu tego problemu zaproponowano algorytm, który sprowadza rozkład wag do rozkładu jednostajnego [19]. Działanie algorytmu składa się z trzech etapów przedstawionych poniżej.

- *Wyrównywanie wag* – w tej fazie algorytm analizuje wektor wag W element po elemencie, jednocześnie mierząc częstotliwość występowania konkretnej wartości. Jeżeli podczas danej iteracji rozważana wartość jest najczęściej występującą to jest podmieniana na wartość dotychczas występującą najrzadziej;
- *Odrzucenie wag* – w wyniku nierównomiernego występowania wag, entropia każdego elementu jest mniejsza od entropii hipotetycznego rozkładu jednostajnego. Aby pozbyć się informacji nadmiarowych, pojedyncze elementy wektora W są odrzucane w sposób deterministyczny;
- *Podmiana wartości* – w celu poprawienia losowości otrzymanego wektora, zostaje on zamieniony na wektor binarny, a następnie podzielony na bloki równej długości. Następnie tak przygotowane bloki są podmieniane przez wartość zwracaną przez wcześniej wybraną funkcje haszującą. Dodatkowo, wybór funkcji haszującej implikuje długość bloku na jaką jest dzielona binarna forma wektora W.

Dokładne działanie algorytmu zostało przedstawione w [19].

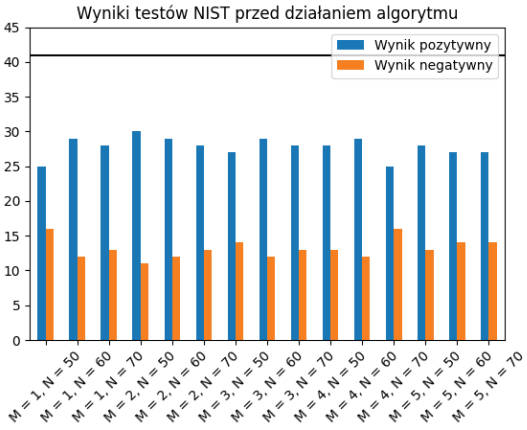
W dalszej części publikacji przedstawione zostaną pogłębione badania nad opisany algorytmem. Celem badań będzie sprawdzenie zależności pomiędzy parametrami (rozmiarem) sieci neuronowej typu NBTPM, a jakością uzyskanego klucza kryptograficznego.



Rysunek 1: Architektura modelu sieci typu Tree Parity Machine

3. WYNIKI

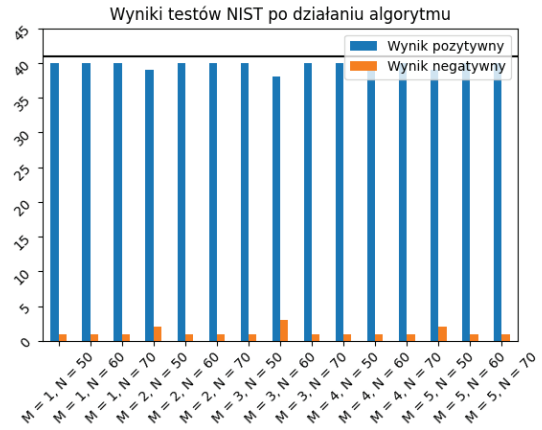
W celu analizy zależności pomiędzy rozmiarem sieci NBTPM, a jakością otrzymanego klucza kryptograficznego przeprowadzone zostały badania symulacyjne. W trackie symulacji dwie sieci NBTPM synchronizowały się ze sobą czego efektem było uzyskanie klucza kryptograficznego. Tak utworzony klucz kryptograficzny był następnie przetwarzany przez algorytm sprowadzający wagi do rozkładu jednostajnego. Następnie obie wersje otrzymanego klucza kryptograficznego zostały poddawane testom opisanym w standardzie NIST 800-22 [1], których celem jest badanie jakości generowanych liczb losowych do zastosowań kryptograficznych. Zestaw testów NIST 800-22 składa się z 15 różnych testów, z których niektóre są parametryzowane, co łącznie daje 41 wyników. Wykorzystywane testy badają różne właściwości rozważanego ciągu losowego, m. in.: częstotliwość występowania wartości 0 i 1 w całym badanym ciągu binarnym jak i jego podciągach, widmo badanego klucza kryptograficznego, częstotliwość występowania wzorców oraz entropię. Podczas symulacji rozważano następujące zestawy parametrów (rozmiary) sieci NBTPM $K = 3, L = 8, M = 1; 2; 3; 4; 5, N = 50, 60, 70$, a każdy scenariusz został powtórzony 100 razy w celu otrzymania dłuższego klucza kryptograficznego do badań.



Rysunek 2: Wyniki testów przed wykorzystaniem algorytmu

Przed zastosowaniem algorytmu równoważącego wartości w otrzymanym wektorze wag zestaw testów NIST 800-22 oscyluje od 11 do 16 z 41 testów zakończonych niepowodzeniem. Następujący fakt jest spowodowany nierównomiernym rozkładem wag, gdzie wartości graniczne lub bliskie zeru występują częściej w zależności od wykorzystanej reguł uczącej [8]. Wyniki przed wykonaniem algorytmu rozważanego w tym artykułu przedstawione są na rysunku 2.

Po zastosowaniu algorytmu wyrównującego wagi liczba testów zakończonych niepowodzeniem znacząco spada, bez znaczących zależności pomiędzy parametrami sieci, a liczbą niepowodzeń. Liczba testów zakończonych niepowodzeniem mieści się w przedziale od 1 do 3, gdzie maksimum ma miejsce dla $M = 3, N = 60$. Pojedyncze testy nie dające wyniku pozytywnego spowodowane są probabilistyczną naturą testów. Stosunek testów zakończonych powodzeniem do testów zakończonych wynikiem negatywnym został przestawiony na rysunku 3.



Rysunek 3: Wyniki testów po wykorzystaniu algorytmu

4. PODSUMOWANIE

W niniejszym artykule przedstawiono postęp prac nad rozwojem kryptografii z wykorzystaniem sieci neuronowych. Zaprezentowana została zasada działania sieci TPM, zaproponowane usprawnienie w postaci wariantu NBTPM oraz korzyści i ryzyka płynące z jego zastosowania. Ponadto omówiono algorytm, którego celem jest równoważenie rozkładu wag i zostały przeprowadzone dalsze badania, które weryfikowały działanie algorytmu dla różnych wariantów sieci NBTPM.

Wykonane badania wykazały, że zastosowanie algorytmu sprowadzającego rozkład otrzymanych wag do rozkładu jednostajnego pozytywnie wpływa na jakość klucza kryptograficznego otrzymanego w skutek uczenia wzajemnego. Ponadto działanie algorytmu jest niezależne od parametrów wykorzystywanej sieci NBTPM.

Kryptografia wykorzystująca sieci neuronowe umożliwia bezpieczne uzgodnienie klucza kryptograficznego, odpornego na ataki komputerów kwantowych. Jednak w celu osiągnięcia wymaganego bezpieczeństwa kluczowe jest odpowiednie dobranie parametrów sieci oraz zapobieganie wpływowi efektu wartości krańcowych. Konieczne są także dalsze badania, aby zweryfikować tę metodę w środowiskach podatnych na ataki kryptograficzne.

LITERATURA

[1] Bassham, L. E., Rukhin, A. L., Soto, J., Nechvatal, J. R., Smid, M. E., Barker, E. B., Leigh, S. D., Levenson, M., Vangel, M., Banks, D. L., Heckert, N. A., Dray, J. F., and Vo, S. Sp 800-22 rev. 1a. a statistical test suite for random and pseudorandom number generators for cryptographic applications. Tech. rep., Gaithersburg, MD, USA, 2010.

[2] Cao, Y., Zhao, Y., Wang, Q., Zhang, J., Ng, S. X., and Hanzo, L. The evolution of quantum key distribution networks: On the road to the qinternet. *IEEE Communications Surveys Tutorials 24*, 2 (2022), 839–894.

[3] Dong, T., and Huang, T. Neural cryptography based on complex-valued neural network. *IEEE Transactions on Neural Networks and Learning Systems 31*, 11 (2020), 4999–5004.

[4] Jeong, S., Park, C., Hong, D., Seo, C., and Jho, N. Neural cryptography based on generalized tree parity machine for real-life systems. *Security and Communication Networks 2021* (Feb 2021), 6680782.

[5] Kinzel, W., Metzler, R., and Kanter, I. Dynamics of interacting neural networks. *Journal of Physics A: Mathematical and General 33*, 14 (apr 2000), L141.

[6] Klimov, A., Mityagin, A., and Shamir, A. Analysis of neural cryptography. In *Advances in Cryptology — ASIACRYPT 2002* (Berlin, Heidelberg, 2002), Y. Zheng, Ed., Springer Berlin Heidelberg, pp. 288–298.

[7] Metzler, R., Kinzel, W., and Kanter, I. Interacting neural networks. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 62*, 2 Pt B (2000), 2555–2565.

[8] Ruttor, A. *Neural Synchronization and Cryptography.* PhD thesis, University of Würzburg, 2007.

[9] Ruttor, A., Kinzel, W., Naeh, R., and Kanter, I. Genetic attack on neural cryptography. *Phys. Rev. E 73* (Mar 2006), 036121.

[10] Ruttor, A., Reents, G., and Kinzel, W. Synchronization of random walks with reflecting boundaries. *Journal of Physics A: Mathematical and General 37*, 36 (aug 2004), 8609.

[11] Santhanalakshmi, S., K., S., and Patra, G. K. Analysis of neural synchronization using genetic approach for secure key generation. In *Security in Computing and Communications* (Cham, 2015), J. H. Abawajy, S. Mukherjea, S. M. Thampi, and A. Ruiz-Martínez, Eds., Springer International Publishing, pp. 207–216.

[12] Sarkar, A., Khan, M. Z., Singh, M. M., Noorwali, A., Chakraborty, C., and Pani, S. K. Artificial neural synchronization using nature inspired whale optimization. *IEEE Access 9* (2021), 16435–16447.

[13] Shacham, L. N., Klein, E., Mislovaty, R., Kanter, I., and Kinzel, W. Cooperating attackers in neural cryptography. *Phys. Rev. E 69* (Jun 2004), 066137.

[14] Shor, P. W. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM Journal on Computing 26*, 5 (1997).

[15] Stebila, D., and Mosca, M. Post-quantum key exchange for the internet and the open quantum safe project. In *Selected Areas in Cryptography – SAC 2016* (Cham, 2017), R. Avanzi and H. Heys, Eds., Springer International Publishing, pp. 14–37.

[16] Stypiński, M., and Niemiec, M. Impact of nonbinary input vectors on security of tree parity machine. In *Multimedia Communications, Services and Security* (Cham, 2022), A. Dziech, W. Mees, and M. Niemiec, Eds., Springer International Publishing, pp. 94–103.

[17] Stypiński, M., and Niemiec, M. Synchronization of tree parity machines using nonbinary input vectors. *IEEE Transactions on Neural Networks and Learning Systems* (2022), 1–7.

[18] Stypiński, M., and Niemiec, M. Security of neural network-based key agreement protocol for smart grids. *Energies 16*, 10 (2023).

[19] Stypiński, M., and Niemiec, M. Weight equalization algorithm for tree parity machines, 2024.

Wybrane metody wykrywania i przeciwdziałania atakom na infrastrukturę dostępową i aplikacje sieci 5G

SELECTED METHODS FOR DETECTING AND COUNTERING ATTACKS ON 5G NETWORK ACCESS INFRASTRUCTURE AND APPLICATIONS

Sławomir Hanczewski[1]; Piotr Remlein[1];

[1] Wydział Informatyki i Telekomunikacji, Politechnika Poznańska, Poznań,
slawomir.hanczewski@put.poznan.pl, piotr.remlein@put.poznan.pl

DOI: 10.15199/59.2024.4.14

Streszczenie: **W artykule przedstawiono charakterystykę wybranych ataków możliwych do przeprowadzenia na usługi sieci 5G. Opisano również w jaki sposób można wykorzystać algorytmy sztucznej inteligencji (ang. artificial intelligence, AI) do wykrywania ataków na sieci 5G. Ponadto przedstawiono słabości tych algorytmów, które mogą być wykorzystane przez intruza do naruszenia ich bezpieczeństwa. Artykuł zawiera również niewielką część wyników uzyskanych w ramach prowadzonych badań nad identyfikacją ataków w sieciach pakietowych. Przykładem takiej sieci jest sieć 5G. Za pomocą algorytmów SI dokonano analizy danych dostępnych w zbiorze danych CIC-IDS-2017.**

Słowa kluczowe: **ataki na sieci 5G, algorytmy SI, wykrywanie anomalii, bezpieczeństwo 5G.**

Keywords: **5G network attacks, AI algorithms, anomaly detection, 5G security .**

1. WSTĘP

Jedną z najbardziej rozwijających się współcześnie technologii jest sztuczna inteligencja (SI). Rozwiązania bazujące na SI wykorzystywane są niemal w każdym obszarze naszego życia. Również w obszarze sieci i systemów komputerowych oraz ich bezpieczeństwa. Rozwiązania tego typu są lub będą również szeroko stosowane w sieciach 5G. Należy o tym fakcie pamiętać projektując rozwiązania wspierające proces zabezpieczania systemów informatycznych. Mechanizmy wykorzystujące rozwiązania SI mogące wykrywać zagrożenia, same mogą być również źródłem podatności. Do wykrycia nietypowego ruchu w sieciach komputerowych konieczna jest analiza całego ruchu sieciowego oraz podjęcie decyzji o tym czy jest on typowy czy też odbiega od normy. Mechanizm SI podejmuje poprawne decyzje w przypadku gdy ruch typowy zostanie zakwalifikowany jako typowy (True Positive) a ruch nietypowy jako nietypowy (True Negative). Błędy występują wtedy gdy ruch typowy zostaje oceniony jako nietypowy (False Positive) lub gdy ruchu nietypowy zostanie zakwalifikowany jako typowy (False Negative). O ile zakwalifikowanie typowego ruchu sieciowego jako potencjalnie niebezpiecznego i w konsekwencji jego zablokowanie sprawi, że w pewnym okresie czasu, użytkownicy nie będą mogli korzystać z usług oferowanych przez system to popełnienie błędu typu False Negative może mieć poważne konsekwencje dla systemu.

Idealnie działające rozwiązanie wspomagające bezpieczeństwo systemu informatycznego podejmuje tylko poprawne decyzje. Ponieważ jednak, dysponujemy rozwiązaniami, którym do ideału może brakować bardzo wiele, zdefiniowanych zostało szereg parametrów pozwalających ilościową i jakościową ocenę mechanizmów wspierających bezpieczeństwo systemów informatycznych. Należą do nich dokładność (ang. accuracy), precyzja (ang. precision), czułość (ang. recall, sensitivity) oraz tzw. miara F1 (ang. F1-score).

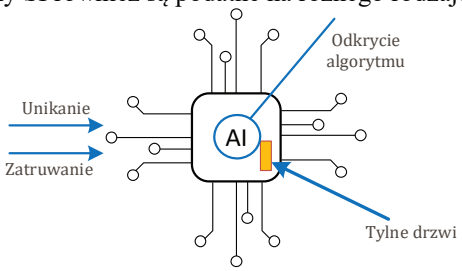
2. ATAKI NA USŁUGI SIECI 5G

Komunikacja we współczesnych sieciach opiera się na transmisji pakietowej wykorzystującej protokoły IP w wersji 4 oraz 6. Wiąże się to z akceptacją całego bagażu pozytywnych i negatywnych aspektów z wiązanych z transmisją pakietową. W literaturze dotyczącej problematyki bezpieczeństwa we współczesnych sieciach teleinformatycznych, ataki dzieli się na wewnętrzne i zewnętrzne. W przypadku ataków zewnętrznych atakujący znajduje się po za siecią, która jest celem ataku (urządzenia w niej działające). Ataki tego typu wymagają od atakującego w pierwszej kolejności przełamania zabezpieczeń związanych z dostępem do danej sieci a następnie w dalszych krokach eskalację działań już wewnątrz sieci. Z punktu widzenia sieci 5G i oferowanych w niej usług, realizacja tego typu ataku będzie związana przełamaniem zabezpieczeń działających w warstwie fizycznej i łącza danych (zgodnie z modelem OSI). Jest to dodatkowe utrudnienie dla atakującego, choć nie jest to niemożliwe. Ataki wewnętrzne są łatwiejsze do wykonania. Wiąże się to z faktem, że atakujący ma już dostęp do sieci, która (lub wybrane urządzenie) jest celem ataku. Bardzo często jako osoba posiadająca pewne uprawnienia w sieci (np. dostęp do urządzeń, znajomość danych logowania), co niestety utrudnia szybkie wykrycie a tym samym przeciwdziałanie takim atakom. Analizą systemów po wykryciu ataku (niezależnie od jego typu) zajmuje się tzw. kryminalistyka cyfrowa (ang. digital forensics) pozwalająca na podstawie cyfrowych śladów określić cel, przebieg i sprawcę ataku. Niezależnie od usytuowania atakującego, do najczęstszych ataków w sieciach teleinformatycznych należą:

- Ataki odmowy dostępu do usługi DoS (ang. denial of service) wraz z wariantem "rozproszonym" DDoS (ang. distributed denial of service);
- Ataki typu „człowiek pośrodku" MitM (ang. man in the middle);
- Przejmowanie kontroli i eskalacja uprawnień;
- Ataki na algorytmy nauczania maszynowego (ang. machine learning, ML).

3. ATAKI NA ALGORYTMY ML

Nauczanie maszynowe (należące do rozwiązań SI) z powodzeniem wykorzystywane jest do, zarówno zapewnienia bezpieczeństwa systemom komputerowym, jak i zarządzania ich zasobami. Możliwości „nauczenia" algorytmu typowych zachowań jest bardzo atrakcyjnym rozwiązaniem pozwalającym wykrywać anomalie ruchu sieciowego. Nie mniej jednak należy pamiętać, że same algorytmy SI również są podatne na różnego rodzaju ataki.



Rysunek 1. Ataki na mechanizmy ML

Do najczęściej rozważanych zaliczane są (Rysunek 1):

- Unikanie (ang. evasion)

Jest to bardzo ciekawy atak ponieważ aby go przeprowadzić konieczne jest przygotowanie odpowiednich danych, które są przesyłane do przetworzenia w atakowanym mechanizmie wykorzystującym ML. Załóżmy, że wykorzystywane algorytm jest przygotowany do pracy (z powodzeniem zakończyła się faza ucznia i algorytm z powodzeniem przeszedł testy potwierdzające poprawność jego działania). Zatem podając dane każdorazowo algorytm powinien zwrócić poprawną odpowiedź. Jednak jak wskazują badania [1] możliwe jest takie spreparowanie danych, które na pierwszy rzut oka nie wydają się być zmanipulowane a ich podanie na wejście algorytmu powoduje że zwracane są błędne wyniki. Przykładem takiego ataku jest „stealth T-shirt" pozwalająca na ukrywanie osób przed algorytmami wykrywającymi obiekty w sekwencjach wideo [2]. W ramach przeprowadzonego eksperymentu opracowano wzór, którego umieszczenie na koszulce powodowało błąd w działaniu algorytmy. Dla algorytmu osoba w koszulce z tym osobliwym wzorem jest ukryta choć wszyscy dokoła mogą ją przecież zobaczyć. Jest to przykład niezwiązany bezpośrednio z sieciami 5G ale pokazuje jakie mogą być konsekwencje tego typu ataków.

- Zatruwanie (ang. poisoning)

Zatruwanie polega na spreparowaniu fałszywych danych wykorzystywanych w procesie uczenia algorytmu. Atakujący musi mieć dostęp do algorytmu w czasie jego uczenia co może być wystarczającym zabezpieczeniem przed tym atakiem. Nie mniej jednak, spreparowane dane są przygotowane w taki sposób aby wydawałoby się, że algorytm działa poprawnie a tylko na wybrane dane będzie reagować w inny zaplanowany przez atakującego sposób [3].

- Tylne drzwi (ang. backdoor)

Podobnie jak tradycyjne programy, w modelach ML można osadzać ukryte, dodatkowe funkcjonalności w postaci tzw. tylnych drzwi. Tylko osoba, która umieściła backdoor w algorytmie wie, jak je wywołać. Dla pozostałych użytkowników funkcjonalność ta będzie ukryta. W przeciwieństwie do tradycyjnych programów, model sieci neuronowej składa się tylko z zestawu parametrów, bez kodu źródłowego. Dlatego backdoory w modelach ML są trudniejsze do wykrycia niż w tradycyjnych programach. Te tylne drzwi są zwykle wszczepiane przez dodanie jakiś konkretny neuron do modelu sieci neuronowej. Model z tylnymi drzwiami reaguje w taki sam sposób jak oryginalny model na normalnym wejściu, ale na określonym wejściu, odpowiedzi są kontrolowane przez backdoora. W pracy [4] zaproponowano metodę osadzania backdoorów w modelach ML. Backdoory mogą być uruchamiane tylko wtedy, gdy wejście obraz zawiera określony wzór i z modelu trudno go znaleźć, a nawet wykryć, czy taki backdoor istnieje. Większość z tych ataków ma miejsce podczas generowania lub przesyłania modelu.

- Odkrycie algorytmu (ang. model extraction)

Atak polega na odkryciu zasad działania algorytmu ML. Jest to możliwe, gdy atakujący ma dostęp do urządzeń, na których algorytm jest testowany oraz do danych wejściowych i wyjściowych algorytmu (np. w czasie jego uczenia). Pozwala to na jego uruchomienie w innych niż pierwotnie zakładano urządzeniach. Gotowe do pracy algorytmy są niezwykle cenne, dlatego ich kradzież może przynieść ich właścicielom szereg strat finansowych. Przykład opisany w pracy [5] pozwala na wykradanie algorytmu przez API (ang. application programming interface) usługi AIaaS (AI as a Service).

4. WYBRANE METODY WYKRYWANIA ATAKÓW

Metody wykrywania ataków zależą od wielu elementów. Bardzo skuteczną metodą jest wykrywanie anomalii ruchu w sieci. Przykładem takiego rozwiązania, które jest z powodzeniem wykorzystywane w sieciach IP są systemy IPS/IDS (Intrusion Prevention System/ Intrusion Detection System), których działanie opiera się na wykrywaniu odstępstw od dobrze znanych wzorców ruchu charakterystycznych dla różnych protokołów sieciowych. Dzięki temu możliwe jest wykrywanie nowych, jeszcze dobrze nie poznanych ataków. Ponieważ w sieciach 5G tworzone będą różne segmenty o diametralnie różnych właściwościach, dlatego też konieczne będzie określenie charakterystyk ruchu w każdym z nich. Oczywiście można przyjąć, że w przypadku segmentu pozwalającego na szerokopasmowy dostęp będący odpowiednikiem dostępowych sieci przewodowych, można w pierwszym etapie prac nad metodami wykrywania wykorzystać wzorce ruchu znane dla sieci przewodowych. W przypadku pozostałych sieci

konieczne jest przeprowadzenie odpowiednich badań. Niezależnie od typu segmentu do opisu wzorca ruchu można zaliczyć: liczbę typów przesyłanych wiadomości (w tym ich wielkość), liczbę wiadomości w jednostce czasu oraz odstępy pomiędzy nimi, sekwencję przesyłanych wiadomości oraz zawartość wiadomości (o ile to możliwe).
Pierwsze trzy właściwości zależą zasadniczo od przyjętego protokołu sieciowego, ponieważ to w specyfikacji protokołów określane są warunki poprawnego działania protokołu. Zawartość przesyłana w polach danych wiadomości zależy ściśle od zastosowania – zwłaszcza w przypadku IoT, gdy możliwości stosowania algorytmów szyfrujących są ograniczone.

Wykrywanie ataków opisanych we wcześniejszych rozdziałach można zrealizować w oparciu o obserwację poszczególnych atrybutów sieciowych (np. bajty odebrane na sekundę, kolizje (procent), straty). Można zastosować algorytm wykrywania anomalii skupiający się na liczbie bajtów ukierunkowanych na serwer, z uwzględnieniem flag RST/FIN TCP. Celem działania algorytmu może być identyfikacja anomalii w bieżącym oknie czasowym na podstawie porównania z wzorcem wyznaczonym na podstawie poprawnej niezakłócanej pracy sieci tzw. profilu normalnym. Można posłużyć się metodą wyznaczającą różnicę pomiędzy entropią normalnego ruchu i generowanego podczas próby ataku. Przy detekcji próby ataku można też brać pod uwagę częstotliwość operacji HTTP wykonywanych na serwerze.

5. ROZWIĄZANIA SI

Dzięki środowisku języka Python możliwe jest relatywnie proste przygotowanie i testowanie rozwiązań wykorzystujących nauczanie maszynowe do wykrywania anomalii w sieciach IP. Na etapie testów niezastąpione wydaje się być środowiska Anaconda oraz Jupyter Notebook. Do testowania rozwiązań AI pozwalających na wykrywanie anomalii w sieciach IP zostało wykorzystane rozwiązanie zaproponowane w [6]. Jest to zbiór skryptów języka Python pozwalających na przygotowanie i przetestowanie znanych algorytmów SI do wykrywania anomalii sieciowych. Nie jest to jedyne rozwiązanie tego typu ale zdaniem autorów raportu jego użycie jest intuicyjne i powala na dostosowanie do swoich potrzeb (zgodnie z udzielona licencją). Omawiany zbiór pozwala na testowanie między innymi następujących algorytmów: Naive Bayes classifier, Random forest, KNN (k nearest neighbours).

6. ZBIORY DANYCH

Aby mechanizmy wykorzystujące rozwiązania SI mogły skutecznie wykrywać nietypowy ruch będący efektem niepożądanych aktywności muszą być odpowiednio przygotowane. W tym celu wykorzystuje się wzorcowe dane, które pozwolą przygotowywanemu rozwiązaniu na nauczenie się czym jest typowy i nietypowy ruch w sieci. Jeżeli na tym etapie zostaną wykorzystane błędne czy może nawet niewłaściwe dane to później, w rzeczywistym środowisku, mechanizm będzie podejmował błędne decyzje. Dlatego też wybór danych wykorzystywane w procesie treningu mechanizmu jest niezwykle ważne. W literaturze nie znajdziemy wielu zbiorów danych, które można wykorzystać do budowy rozwiązań SI. Wiąże się to z faktem, że przygotowanie takich zbiorów danych wiąże się najczęściej z wysokimi kosztami ich przygotowania. Inna przyczyną braku odpowiednich zbiorów danych jest względna nowość rozwiązań sieciowych lub ich nietypowość a tym samym mała ich popularność. W przypadku braku potrzebnych danych można je przygotować samodzielnie. Wymaga to oczywiście nakładu pracy oraz przygotowanie testowej infrastruktury sieciowej z której będą pobierane dane. W ramach prowadzonych eksperymentów wykorzystany został zbiór CICIDS2017 [6]. W dalszych pracach planowane jest pozyskiwanie danych ze środowiska testowego budowanego w ramach projektu 5GSTAR.
Zestaw danych CICIDS2017 zawiera przechwycony ruch sieciowy zawierający między innymi dane związane z najbardziej popularnymi atakami w roku 2017. Zbiór ten dostępny jest w postaci plików .pcap, zawierających przechwycone w sieci pakiety. Można je otworzyć i analizować np. w programie Wireshark. Zawiera również pliki .csv będące wynikiem analizy ruchu sieciowego przy użyciu programu CICFlowMeter z oznakowanymi przepływami na podstawie znacznika czasu, źródłowego i docelowego IP, źródłowych i docelowych portów, protokołów i typu ataku.
Celem zespołu, który przygotował zbiór CICIDS2017 było przygotowanie zbioru realistycznego ruchu. Okres przechwytywania danych rozpoczął się o godzinie 9 rano w poniedziałek, 3 lipca 2017 r. i zakończył się o godzinie 17 w piątek 7 lipca 2017 r., w sumie przez 5 dni. Poniedziałek jest dniem „normalnej" pracy i obejmuje tylko ruch typowy bez anomalii. Zrealizowane ataki to Brute Force FTP, Brute Force SSH, DoS, Heartbleed, Web Attack, Infiltracja, Botnet i DDoS. Zostały one wykonane zarówno rano, jak i po południu we wtorek, środę, czwartek i piątek.

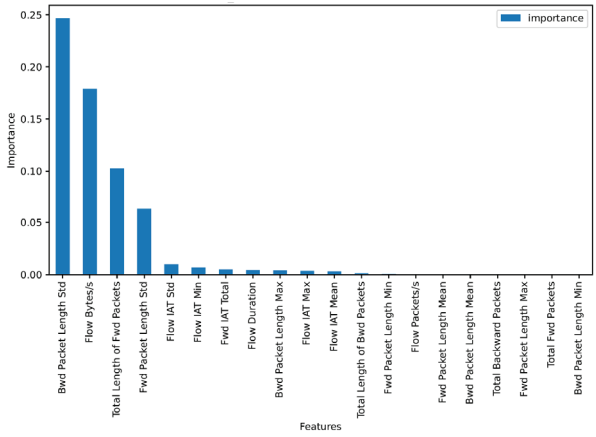
7. WŁAŚCIWOŚCI ANOMALII RUCHU

Niezależnie od tego czy ruch sieciowy jest ruchem typowym czy tez anomalią można go opisać za pomocą jego charakterystycznych cech. Część z nich będzie występować (lub przyjmować określone wartości) tylko dla ruchu typowego, część będzie występować (lub przyjmować określone wartości) tylko w przypadku strumienia pakietów będącego efektem przeprowadzonego ataku. Część cech będzie występować w każdym przypadku. Dzięki określeniu charakterystycznych cech ruchu sieciowego możliwe będzie jednoznaczne opisanie typu ruchu, co z kolei doprowadzi do możliwości wykorzystania tych danych w procesie przygotowania rozwiązań bazujących na Sztucznej Inteligencji. Określenie cech, które za pomocą można opisać ruch sieciowy to niezwykle ważny etap w procesie tworzenia rozwiązań pozwalających na detekcję anomalii w sieciach IT.
Przygotowując dane do procesu określania cech charakterystycznych strumieni danych należy sprawdzić jak są one opisane. Po identyfikacji strumieni zawartych w plikach .pcap np. za pomocą programu CICFlowMeter może się okazać, że część cech nie zostanie wykryta i do opisu zostanie wstawione puste pole zakończone „;". Z jednej

strony mamy zachowany spójny opis tzn. każdy strumień opisany jest za pomocą takiej samej liczby pól (cech) ale z drugiej część z nich może być pusta co będzie prowadziło do błędnej analizy przez rozwiązania SI. Dlatego też przed rozpoczęciem analizy cech strumieni puste pola powinny być uzupełnione wartościami neutralnymi dla danej cechy. Najczęściej będzie to wartość 0. Dysponując przygotowanymi danymi można przystąpić do właściwej analizy danych. W pierwszej kolejności należy wykryć cechy charakterystyczne strumieni: typowych oraz tych związanych z atakami sieciowymi. Znając cechy anomalii można przystąpić do procesu przygotowania rozwiązań AI. W tym celu dostępny zbiór danych dzieli się na dwa podzbiory: jedne służy do uczenia, drugi do testowania.

8. WYKRYWANIE CECH ANOMALII RUCHU

W ramach prowadzonych badań nad identyfikacją ataków w sieciach pakietowych, jaką jest niewątpliwie sieć 5G, dokonano za pomocą algorytmów SI analizy danych dostępnych w zbiorze CIC-IDS-2017. W wyniku analizy wszystkich danych związanych ze strumieniami zarejestrowanymi w czasie ataków otrzymano rezultaty przedstawione na rysunku 2. Ponieważ jednak wszystkich ataków nie można mierzyć jedną miarą, to ocenę ważności cech należy przeprowadzić dla każdego ataku oddzielnie.



Rysunek 2. Cechy określone dla wszystkich ataków łącznie

Wykryte cechy charakterystyczne wszystkich ataków, których strumienie danych zostały zawarte w zbiorze CICIDS2017 zostały przedstawione w [7].

9. PODSUMOWANIE

Sieci 5G projektowane są z uwzględnieniem potencjalnych problemów z bezpieczeństwem. Pomimo to należy założyć, że zarówno w warstwie fizyczne, łącza danych, jak i w warstwach wyższych może dojść do incydentów wpływających na bezpieczeństwo działania całej sieci. Sieć 5G oferuje szereg możliwości przesyłania danych z założonymi wymaganiami co do wartości parametrów QoS. Pomimo tej różnorodności, wykorzystując doświadczenie zdobyte w tradycyjnych sieciach IP, możliwe jest wykrywanie niepożądanych zachowań. Doświadczenie to wskazuje bowiem, że najefektywniejszą metodą wykrywania ataków jest wykrywanie anomalii w ruchu sieciowym.

Rozwiązania wykorzystujące Sztuczną Inteligencję to potężne narzędzie do walki z nietypowymi zachowaniami w sieciach teleinformatycznych. Dobrze przygotowane mechanizmy potrafią w czasie rzeczywistym analizować, informować i przeciwdziałać wszelkim anomaliom ruchowym co jest jednoznaczne z wykrywaniem i przeciwdziałaniem atakom sieciowym. W przypadku sieci 5G, gdzie w sieci szkieletowej wykorzystywana jest transmisja danych również można je wykorzystywać. Jednak przygotowanie dedykowanego rozwiązania musi się związać z pozyskaniem próbek danych z tej sieci.

Praca wykonana w ramach projektu 5gSTAR „Zaawansowane metody i techniki wykrywania i przeciwdziałania atakom na infrastrukturę dostępową i aplikacje sieci 5G" nr grantu: CYBERSECIDENT/487845/IV/ NCBR/2021

LITERATURA

[1] N. Papernot, P. McDaniel i I. Goodfellow, „Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples," Cornrll University, 24 May 2016.

[2] duoergun0729.github.io, „Stealth T-shirt," AdvBox, 16 January 2020. [Online]. Available: https://github.com/advboxes/AdvBox/blob/master/applications/StealthTshirt/README.md. [Data uzyskania dostępu: 30 sierpień 2022].

[3] C. Weng, J. Chen, Y. Yang, X. Ma i J. Liu, „Poisoning attacks and countermeasures in intelligent networks: Status quo and prospects," *Digital Communications and Networks,* tom 8, nr 2, pp. 225-234, April 2022.

[4] T. Gu, B. Dolan-Gavitt i S. Garg, „BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain," Cornell University, 22 08 2017.

[5] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter i T. Ristenpart, „Stealing Machine Learning Models via Prediction APIs," Cornell University, 09 09 2016.

[6] I. Sharafaldin, A. H. Lashkari i A. A. Ghorbani, „Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," w *4th International Conference on Information Systems Security and Privacy (ICISSP)*, 2018.

[7] S. Hanczewski, K. Wesołowski, P. Remlein i F. Idzikowski, „Badania nad wykrywaniem ataków na sieć dostępową i aplikacje w systemie 5G, Raport wewn. projektu NCBR CYBERSECIDENT/487845/IV/ NCBR/2021 pn. Zaawansowane metody i techniki wykrywania i przeciwdziałania atakom na infrastrukturę dostępową i aplikacje," 2023.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

ESTYMACJA POŁOŻENIA I ORIENTACJI W SYSTEMIE LOKALIZACYJNYM Z CZĘŚCIOWĄ SYNCHRONIZACJĄ WĘZŁÓW REFERENCYJNYCH

POSITION AND ORIENTATION ESTIMATION IN POSITIONING SYSTEM WITH PARTIAL SYNCHRONIZATION OF REFERENCE NODES

Jarosław Sadowski[1]; Jacek Stefański[2]

[1] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, jarsad@eti.pg.edu.pl
[2] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, jstef@eti.pg.edu.pl

DOI: 10.15199/59.2024.4.15

Streszczenie: **W referacie przedstawiono koncepcję systemu lokalizacyjnego do jednoczesnej estymacji położenia i orientacji obiektów. Obiekt ruchomy jest wyposażony w zestaw zsynchronizowanych odbiorników. Nadajniki referencyjne systemu są podzielone na grupy i pracują synchronicznie tylko w obrębie swojej grupy, bez zachowania synchronizacji pomiędzy grupami nadajników.**

Abstract: **The paper presents the concept of a positioning system for simultaneous estimation of the position and orientation of objects. The moving object is equipped with a set of synchronized receivers. The system's reference transmitters are divided into groups and work synchronously only within their group, without maintaining synchronization between groups of transmitters.**

Słowa kluczowe: **radiolokalizacja, radionawigacja, systemy asynchroniczne.**

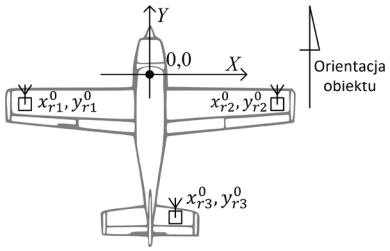
Keywords: **asynchronous systems, radiolocalization, radionavigation**

1. WSTĘP

W większości systemów radiolokalizacyjnych, w których estymacja położenia węzłów ruchomych realizowana jest na zasadzie pomiaru zależności czasowych pomiędzy odbieranymi sygnałami radiowymi, tzn. pomiar czasów nadejścia sygnałów TOA (*Time Of Arrival*) lub różnic czasów nadejścia sygnałów TDOA (*Time Difference Of Arrival*), wymaga synchronizacji pracy węzłów referencyjnych. Synchroniczna praca nadajników jest niezbędna do prawidłowego działania podstawowych usług lokalizacyjnych w sieciach komórkowych, satelitarnych systemach nawigacyjnych GNSS (*Global Navigation Satellite Systems*) i była także podstawą funkcjonowania naziemnych systemów nawigacyjnych jak np. Loran-C, a z krajowych rozwiązań AD-2 i Jemiołuszka [1]. Jak wiadomo, w rozwiązaniach synchronicznych zakłócenia pracy lub awaria układów utrzymywania synchronizacji może prowadzić do błędnych wskazań lub braku dostępu do usług lokalizacyjnych. Jako alternatywę rozważa się asynchroniczne systemy radiolokalizacyjne, w których pozycje terminali ruchomych można estymować bez konieczności synchronizacji pracy węzłów odniesienia [2]. W referacie przedstawione jest rozwiązanie pośrednie: system radiolokalizacyjny bazujący na pomiarach różnic czasów nadejścia sygnałów, w którym nieruchome węzły referencyjne, w prezentowanym rozwiązaniu będące nadajnikami (możliwe jest także rozwiązanie z referencyjnymi odbiornikami), są podzielone na grupy a synchronizacja ich pracy realizowana jest tylko w obrębie danej grupy, bez synchronizacji pomiędzy grupami. Taka organizacja systemu z synchronizacją węzłów tylko w lokalnych grupach może być wygodna przy nieregularnym rozmieszczeniu stacji odniesienia: stacje położone blisko siebie są łączone przewodowo lub bezprzewodowo łączem synchronizacyjnym, ale bez konieczności tworzenia łączy do synchronizacji węzłów w dużych odległościach od siebie. Dodatkowo, w prezentowanym systemie zaproponowano, aby lokalizowany obiekt także był wyposażony w zestaw dwóch lub więcej zsynchronizowanych urządzeń (odbiorników), podłączonych do anten rozmieszczonych w różnych miejscach danego obiektu. Jest to zatem rozszerzenie geometrii systemu lokalizacyjnego zaproponowanego w [3] przez zastąpienie pojedynczych stacji referencyjnych grupą lokalnie zsynchronizowanych nadajników i dopuszczenie stosowania większej liczby odbiorników na lokalizowanym obiekcie ruchomym. Dzięki temu uzyskuje się możliwość jednoczesnej estymacji położenia i orientacji obiektu ruchomego. Mimo synchronicznej pracy odbiorników zakłada się, że odbiorniki te nie mogą realizować detekcji kąta nadejścia sygnałów AOA (*Angle Of Arrival*), więc oba parametry lokalizacyjne (współrzędne i kąt orientacji) mierzone są metodą pomiaru różnic czasów detekcji sygnałów TDOA.

2. STRUKTURA SYSTEMU

Lokalizowany obiekt, którego położenie i orientacja mają być określone, jest wyposażony w zestaw N zsynchronizowanych odbiorników.

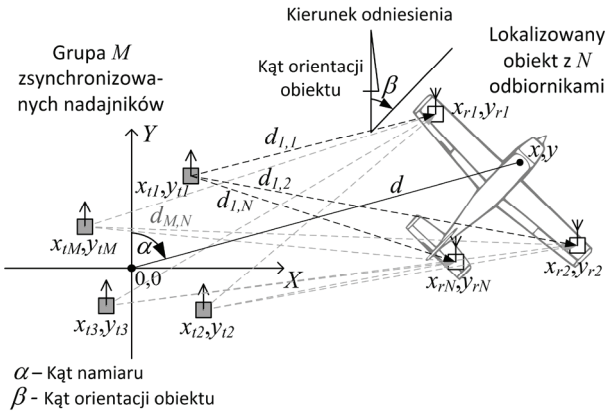


Rys. 1. Lokalny układ współrzędnych skojarzony z lokalizowanym obiektem

Obiekt ma zdefiniowany punkt odniesienia, względem którego będą przedstawiane jego współrzędne oraz kąt orientacji. Zatem współrzędne każdego n-tego odbiornika można zapisać jako x_{rn}^0, y_{rn}^0 w lokalnym układzie współrzędnych skojarzonym z lokalizowanym obiektem, o początku w punkcie odniesienia tego obiektu i osi Y zgodnej z kierunkiem orientacji tego obiektu, jak przedstawiono na rys. 1.

2.1. Jedna grupa węzłów referencyjnych

W najprostszej wersji w systemie lokalizacji dwuwymiarowej występuje tylko jedna grupa złożona z M zsynchronizowanych nadajników referencyjnych, rozmieszczonych na ograniczonym obszarze z powodu występowania ograniczeń w maksymalnej odległości pomiędzy urządzeniami, przy której można realizować synchronizację ich pracy.



Rys. 2. Geometria fragmentu systemu z jedną grupą zsynchronizowanych węzłów referencyjnych

W globalnym, nieruchomym układzie współrzędnych, współrzędne n-tego odbiornika zainstalowanego na obiekcie znajdującym się w punkcie (x, y) i obróconym o kąt β można obliczyć z lokalnych współrzędnych odbiorników x_{rn}^0, y_{rn}^0 według zależności:

$$x_{rn} = x + x_{rn}^0 \cos(\beta) + y_{rn}^0 \sin(\beta) \quad (1)$$

$$y_{rn} = y + y_{rn}^0 \cos(\beta) - x_{rn}^0 \sin(\beta) \quad (2)$$

Zestaw zsynchronizowanych odbiorników może wykonywać pomiary różnic czasów nadejścia sygnałów od wszystkich nadajników, przy czym sposób realizacji tych pomiarów wykracza poza tematykę niniejszego referatu. Oznaczając przez $t_{rm,n}$ czas odbioru sygnału z m-tego nadajnika przez n-ty odbiornik, wyniki wykonywanych pomiarów są określone jako:

$$t_{rm_2,n_2} - t_{rm_1,n_1} = \left(t_{tm_2} + \frac{d_{m_2,n_2}}{v}\right) - \left(t_{tm_1} + \frac{d_{m_1,n_1}}{v}\right), \quad (3)$$

gdzie m_2, n_2 oraz m_1, n_1 to identyfikatory nadajników (m_1, m_2), których sygnały są używane do realizacji pomiaru i identyfikatory odbiorników (n_1, n_2), które wykonują pomiar różnic czasów nadejścia tych sygnałów $\left(t_{rm_2,n_2} - t_{rm_1,n_1}\right)$. Zmienne t_{tm_1} i t_{tm_2} to czasy rozpoczęcia emisji sygnałów lokalizacyjnych odpowiednio przez nadajnik m_1 i m_2, których bezwzględne wartości nie są znane po stronie odbiorczej, jednak nadajniki w grupie pracują synchronicznie to znana jest różnica czasów emisji sygnałów $\left(t_{tm_2} - t_{tm_1}\right)$, więc może być usunięta z (3). Obliczona z równania (3) różnica odległości $\left(d_{m_2,n_2} - d_{m_1,n_1}\right)$ jest powiązana z poszukiwanymi współrzędnymi odbiorników na lokalizowanym obiekcie równaniem:

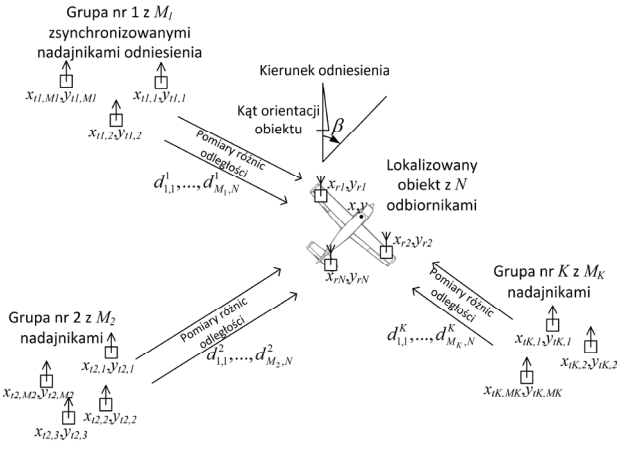
$$d_{m_2,n_2} - d_{m_1,n_1} = \sqrt{\left(x_{rn_2} - x_{tm_2}\right)^2 + \left(y_{rn_2} - y_{tm_2}\right)^2} - \sqrt{\left(x_{rn_1} - x_{tm_1}\right)^2 + \left(y_{rn_1} - y_{tm_1}\right)^2} \quad (4)$$

co po uwzględnieniu (1) i (2) pozwala na obliczenie współrzędnych punktu referencyjnego lokalizowanego obiektu (x, y) i kąta orientacji β jeśli w pomiarach uczestniczą co najmniej dwa nadajniki i dwa odbiorniki, chociaż w takim przypadku może wystąpić niejednoznaczność lustrzanego odbicia. Jeżeli lokalizowany obiekt jest wyposażony tylko w jeden odbiornik to określenie kąta orientacji nie będzie możliwe, a system staje się równoważny klasycznemu podejściu do lokalizacji na bazie TDOA.

Jakość obliczania położenia obiektu w tym rozwiązaniu w istotny sposób zależy od odległości obiektu od grupy nadajników referencyjnych, gdyż zwiększaniu odległości towarzyszy zmniejszanie kątów, pod jakimi przecinają się w tym systemie linie pozycyjne według znanych zasad propagacji błędów pomiarów w systemach pozycyjnych. Inaczej jednak zachowuje się pomiar kąta orientacji. Jeżeli zachowana jest stała jakość pomiaru różnic czasów nadejścia pakietów (3), to wyznaczanie kąta orientacji z (1) i (2) charakteryzuje się relatywnie małą zależnością od odległości, więc zakres stosowalności tego rozwiązania do pomiaru orientacji obiektu jest szerszy niż do pomiaru położenia.

2.2. Struktura ogólna systemu z wieloma grupami nadajników

Rozmieszczenie kilku grup nadajników referencyjnych w obszarze większym niż wynika to z ograniczeń możliwości synchronizacji urządzeń, prowadzi do struktury systemu lokalizacyjnego przedstawionej na rys. 3.



Rys. 3. Struktura systemu radiolokalizacyjnego z grupami lokalnie synchronizowanych węzłów

Lokalizowany obiekt jest nadal wyposażony w N odbiorników, pracujących synchronicznie. Ale infrastruktura stała systemu jest podzielona na K grup stacji referencyjnych, zsynchronizowanych tylko lokalnie wewnątrz grupy. Liczba nadajników M_k w k-tej grupie

$(k = 1 \ldots K)$ może być różna, tak samo jak różne może być przestrzenne rozmieszczenie stacji referencyjnych w poszczególnych grupach. W przypadku gdy $K = 1$, $N = 1$ uzyskuje się typowy system radiolokalizacyjny TDOA, podczas gdy konfiguracja $N = 2$ i $M_k = 1$ dla każdego $k = 1 \ldots K$ daje równania pozycyjne identyczne jak w systemie zaprezentowanym w publikacji [3].

Obliczanie położenia i kąta orientacji obiektu w systemie z wieloma grupami lokalnie synchronizowanych stacji referencyjnych realizowane jest przez zdefiniowanie układu równań (4) według tych samych zależności (1-3) co dla systemu z synchroniczną pracą wszystkich nadajników referencyjnych. Jedyną różnicą jest ograniczenie, że nie można zapisywać w tym samym równaniu (4) wyników pomiarów uzyskanych dla sygnałów pochodzących z nadajników w różnych grupach, z uwagi na nieznane zależności pomiędzy ich nominalnymi czasami emisji.

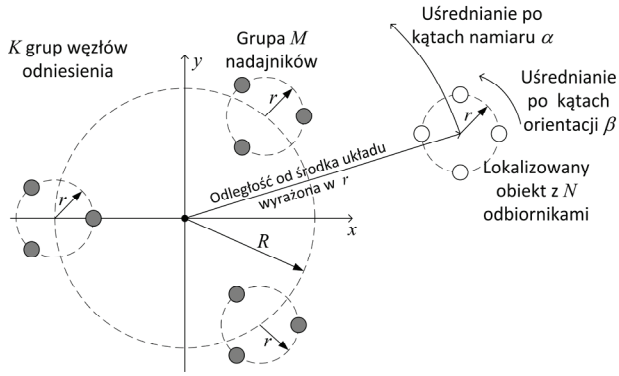
Układ równań nieliniowych (4) może być rozwiązany przy pomocy jednego ze znanych algorytmów iteracyjnych, np. algorytmu Gaussa-Newtona. Ponieważ sama metoda poszukiwania rozwiązania przy użyciu tego algorytmu nie jest specyficzna dla zaproponowanego systemu radiolokalizacyjnego, nie będzie tutaj przedstawiana. Warto jednak zwrócić uwagę na problem doboru punktu początkowego do pierwszej iteracji algorytmu Gaussa-Newtona, który w zaproponowanym systemie charakteryzuje się stosunkowo dużą skłonnością do rozbieżności. Początkową pozycję i kąt orientacji można wyznaczyć zgrubnie przez obliczenie przybliżenia kąta orientacji β i kąta namiaru α (patrz rys. 2), uzyskanych przy założeniu nieskończenie dużej odległości pomiędzy obiektem ruchomym a grupą nadajników:

$$d_{m_2,n_2} - d_{m_1,n_1} = \left[\left(x^0_{rn_2}\cos\beta + y^0_{rn_2}\sin\beta\right) - \left(x_{tm_2}\cos\alpha + y_{tm_2}\sin\alpha\right)\right] - \left[\left(x^0_{rn_1}\cos\beta + y^0_{rn_1}\sin\beta\right) - \left(x_{tm_1}\cos\alpha + y_{tm_1}\sin\alpha\right)\right] \quad (5)$$

oddzielnie dla każdej grupy nadajników. Za kąt orientacji dla dokładnych obliczeń przyjmuje się średnią wartość β otrzymaną z (5) dla każdej z grup nadajników, natomiast początkowe współrzędne (x, y) wyznacza się z kątów namiaru α zależnościami dla systemów kątowych (AOA).

3. SYMULACJE

Duża liczba możliwych kombinacji parametrów systemu, takich jak: liczba odbiorników i ich rozmieszczenie na lokalizowanym obiekcie, liczba grup nadajników referencyjnych, liczność poszczególnych grup nadajników, wzajemne rozmieszczenie grup nadajników względem siebie i rozmieszczenie nadajników w grupach, utrudnia wykonanie kompleksowego badania jakości estymacji położenia i orientacji w zaproponowanym systemie. Dla przedstawienia danych w sposób jak najbardziej uniwersalny, umożliwiający przeskalowanie wyników symulacji na inną geometrię oraz rozmiary systemu, założono, że zarówno odbiorniki zainstalowane na lokalizowanym obiekcie jak i nadajniki w każdej z grup są rozmieszczone równomiernie na okręgu o promieniu r, a wszystkie parametry błędów realizacji pomiarów różnic odległości oraz wynikowych błędów estymacji położenia będą wyrażone w krotności promienia r. Także wszystkie grupy nadajników referencyjnych były w symulacjach rozmieszczone równomiernie na okręgu o promieniu R, przy ustalonych proporcjach pomiędzy r oraz R. Geometrię systemu przyjętą na potrzeby symulacji przedstawiono na rys. 4.



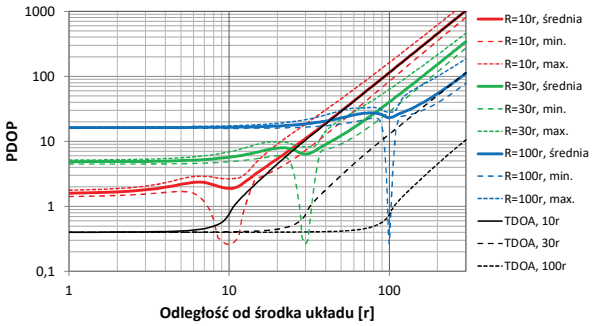
Rys. 4. Ogólna struktura systemu przyjęta do symulacji

Celem symulacji było określenie wartości współczynników rozmycia DOP (*Dilution Of Precision*), które są miarą przełożenia błędów pomiarów, wykonywanych w danym systemie lokalizacyjnym, na uzyskiwane wyniki. Współczynnik rozmycia położenia PDOP (*Position Dilution of Precision*) [4] jest uśrednioną pochodną wyznaczonego położenia obiektu po błędach realizacji pomiarów, w tym systemie pomiarów różnic czasów nadejścia sygnałów przeliczonych na różnice odległości by uzyskać wartość niemianowaną, natomiast współczynnik rozmycia kąta orientacji DDOP (*Direction Dilution Of Precision*) będzie tu rozumiany jako uśredniona pochodna błędu określania kąta orientacji w radianach po błędach pomiarów różnic odległości, unormowanych względem r.

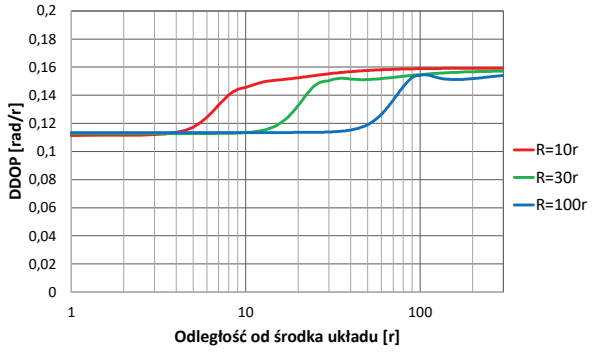
W zależności od konkretnej pozycji i kąta orientacji lokalizowanego obiektu, takie same wartości błędów realizacji pomiarów przekładać się mogą na różne błędy wyznaczania położenia i orientacji, dlatego symulacje wykonano dla różnych wartości odległości lokalizowanego obiektu od środka układu grup stacji referencyjnych, różnych wartości promienia R oraz różnych kątów orientacji i namiaru, przy czym wyniki w postaci wartości PDOP i DDOP uśredniono dla wszystkich kątów orientacji i namiaru oraz przedstawiono wartości minimalne i maksymalne z symulacji. Dodatkowo na wykresach PDOP pokazano wyniki uzyskane dla symulacji pracy klasycznego, synchronicznego systemu TDOA z pięcioma stacjami bazowymi rozmieszczonymi równomiernie na okręgu o promieniu R (czarne linie) i jednym odbiornikiem na lokalizowanym obiekcie w celu ułatwienia oceny pracy systemu z częściową synchronizacją stacji bazowych.

Na rys. 5 i rys. 6 przedstawiono odpowiednio wykresy PDOP i DDOP w funkcji odległości lokalizowanego obiektu od środka układu grup stacji bazowych, przy trzech grupach stacji bazowych ($K = 3$) z trzema nadajnikami w każdej ($M = 3$) i trzema odbiornikami na lokalizowanym obiekcie ($N = 3$). Linie ciągłe na rys. 5 prezentują wyniki uśrednione, podczas gdy liniami przerywanymi zwizualizowano rozrzut wartości minimalnych

i maksymalnych PDOP uzyskanych w symulacjach. Można zauważyć, że jakość estymacji położenia w zaproponowanym systemie jest niemal niezmienna wewnątrz obszaru ograniczonego grupami stacji bazowych, natomiast przemieszczenie obiektu poza okrąg o promieniu R (wyniki zaprezentowano dla konfiguracji $R = 10r$, $R = 30r$ i $R = 100r$) powoduje szybkie narastanie PDOP w funkcji odległości, z nachyleniem takimi samym jak w klasycznym systemie lokalizacji TDOA. Największe rozbieżności pomiędzy minimalnymi i maksymalnymi wartościami PDOP uzyskano, gdy lokalizowany obiekt znajdował się bardzo blisko okręgu R, na którym rozmieszczone były grupy nadajników, gdyż w tym przypadku w zależności od kąta namiaru α występuje największe zróżnicowanie odległości lokalizowanego obiektu od najbliższej grupy nadajników referencyjnych.



Rys. 5. Zależność wartości współczynnika PDOP od odległości obiektu od środka układu grup stacji bazowych. Parametry systemu: K=3, M=3, N=3.



Rys. 6. Zależność wartości współczynnika DDOP od odległości obiektu od środka układu grup stacji bazowych. Parametry systemu: K=3, M=3, N=3.

W porównaniu z PDOP, współczynnik DDOP określający przełożenie błędu pomiaru różnic odległości na błąd estymacji kąta orientacji obiektu β charakteryzuje się bardzo małą zależnością odległości obiektu od środka układu grup stacji bazowych. Pogorszenie jakości estymacji kąta orientacji poza okręgiem R w porównaniu z lokalizacjami wewnątrz okręgu R nie przekracza 50%, a rozbieżność pomiędzy wartościami minimalnymi i maksymalnymi, uzyskanymi podczas symulacji, była na tyle mała, że wartości skrajne DDOP nie zostały uwzględnione na rys. 6.

Poprawę jakości estymacji położenia w zaproponowanym systemie można uzyskać zarówno zwiększając liczbę odbiorników N, liczbę grup nadajników referencyjnych K jak i liczbę nadajników w każdej grupie M, ale parametry te nie mają jednakowego wpływu na właściwości systemu. Zwiększane liczby nadajników w grupach jak też liczby odbiorników na lokalizowanym obiekcie prowadzi do większej redukcji błędu estymacji położenia, niż podobnej wielkości zwiększenie liczby grup nadajników.

4. PODSUMOWANIE

Zaproponowany system radiolokalizacyjny jest uogólnieniem rozwiązania zaprezentowanego w publikacji [3]. Ponieważ problem precyzyjnej synchronizacji węzłów referencyjnych w sieciach radiowych narasta ze zwiększaniem odległości pomiędzy synchronizowanymi urządzeniami, w tym rozwiązaniu założono, że synchroniczna praca jest możliwa tylko w obrębie grup nadajników referencyjnych, rozmieszczonych na ograniczonym obszarze, podczas gdy różne grupy nadajników pracują niezależnie od siebie, bez synchronizacji.

Jednym z możliwych obszarów zastosowania zaproponowanego systemu jest lotnictwo. Samoloty pasażerskie i transportowe są wystarczająco duże do zainstalowania w nich kilku odbiorników z odpowiednio rozsuniętymi antenami, natomiast naziemne stacje referencyjne mogłyby być ulokowane np. na turbinach wiatrowych. Turbiny zwykle są budowane w grupach tworzących farmę wiatrową z infrastrukturą sterującą, pozwalającą na realizację synchronizacji pracy urządzeń na poszczególnych turbinach, podczas gdy nadajniki referencyjne na różnych farmach oddalonych o dziesiątki kilometrów, nie wymagałyby synchronicznej pracy.

LITERATURA

[1] Specht Cezary, Adam Weintrit, Mariusz Specht. 2016. “A History of Maritime Radio-Navigation Positioning Systems used in Poland”. Journal of Navigation. Vol. 69, str. 1-13

[2] Stefański Jacek. 2018. “Asynchroniczne metody radiolokalizacyjne”. Wydawnictwo Naukowe PWN

[3] Stefański Jacek, Jarosław Sadowski. 2022. “Asynchroniczna metoda jednoczesnej estymacji położenia i orientacji obiektu za pomocą dwóch nadajników”, Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne, 4/2022: 567-570

[4] Tahsin Mahdia, Sultana Sunjida, Reza Tasmia, Hossam-E-Haider Md. 2015. “Analysis of DOP and its preciseness in GNSS position estimation”. 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), str. 1-6

OGRANICZENIA WIDMOWEJ METODY ESTYMACJI CZĘSTOTLIWOŚCI W SENSORZE LOKALIZACYJNYM ZBUDOWANYM NA KOMPUTERZE JEDNOPŁYTKOWYM

LIMITATIONS OF THE SPECTRAL FREQUENCY ESTIMATION METHOD IN A LOCATION SENSOR BUILT ON SINGLE-BOARD COMPUTER

Kacper Bednarz[1]; Jarosław Wojtuń[2]; Cezary Ziółkowski[3]

[1] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, kacper.bednarz@wat.edu.pl
[2] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, jaroslaw.wojtun@wat.edu.pl
[3] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, cezary.ziolkowski@wat.edu.pl

DOI: 10.15199/59.2024.4.16

Streszczenie: **Lokalizacja mobilnych źródeł emisji radiowych wiąże się z koniecznością wykonywania wszelkich obliczeń w czasie rzeczywistym. Projektując sensor lokalizacyjny przeznaczony do wykorzystania na dronach, mamy natomiast do czynienia z ograniczonym rozmiarem ładunku. Metodą lokalizacji, która może zostać zaimplementowana na takim nośniku jest signal Doppler frequency, bazująca na estymacji częstotliwości dopplerowskiej. Częstotliwość ta może być wyznaczana na wiele sposobów. W tym artykule przebadano możliwości obliczeniowe różnych komputerów jednopłytkowych wchodzących w skład elementów sensora lokalizacyjnego, w przypadku zaimplementowania na nich widmowej metody estymacji częstotliwości. W oparciu o uzyskane wyniki określono ograniczenia związane z maksymalną szerokością pasma sygnałów emitowanych przez lokalizowane nadajniki.**

Abstract: **The location of mobile radio emission sources requires performing all calculations in real time. When designing a location sensor intended for use on drones, we are faced with a limited payload size. The location method that can be implemented on such a carrier is signal Doppler frequency, based on Doppler frequency estimation. This frequency can be determined in many ways. This article examines the computational capabilities of various single-board computers included in the location sensor components when a spectral frequency estimation method is implemented on them. Based on the results obtained, limitations related to the maximum bandwidth of signals emitted by the located transmitters were determined.**

Słowa kluczowe: **lokalizacja, odtwarzanie, częstotliwości, signal Doppler frequency (SDF), komputer jednopłytkowy (SBC).**

Keywords: **localization, frequency, recovery, signal Doppler frequency (SDF), single-board computer (SBC).**

1. WSTĘP

Zdolność do lokalizacji nadajników radiowych jest niezwykle istotna w dzisiejszych czasach zwłaszcza w zastosowaniach wojskowych. Dynamika współczesnego pola walki przekłada się na coraz mobilniejsze wykorzystanie środków radiowych. Skutkuje to potrzebą lokalizacji emiterów w czasie rzeczywistym. To natomiast wiąże się z dodatkowymi utrudnieniami związanymi przede wszystkim z ograniczonymi możliwościami obliczeniowymi wykorzystywanych komputerów.

W ramach realizowanego projektu opracowywany jest sensor lokalizacji, który docelowo zostanie umieszczony na bezzałogowym statku powietrznym (BSP). Takowa implementacja przekłada się na ograniczenia rozmiarowe i wagowe dotyczące możliwego do przeniesienia ładunku. Do przetwarzania zarejestrowanych sygnałów oraz implementacji algorytmów wybranej metody lokalizacji niezbędne jest zatem wykorzystanie niewielkich komputerów jednopłytkowych (SBC), których możliwości obliczeniowe są znacznie ograniczone.

Jedną z częstotliwościowych metod lokalizacji nadajników radiowych jest SDF (ang. signal Doppler frequency) [1]. Pozwala ona na określenie współrzędnych położenia emitera z wykorzystaniem pojedynczego, poruszającego się odbiornika [2], [3]. Zasadniczym elementem wpływającym na dokładność lokalizacji jest dokładność odtwarzania częstotliwości z zarejestrowanego sygnału. W [4] przeprowadzono badania symulacyjne pokazujące w jaki sposób błąd estymacji częstotliwości przekłada się na dokładność lokalizacji. W publikacji założono, iż emiter znajduje się wzdłuż trajektorii ruchu odbiornika, w punkcie o współrzędnych $(x, y, z) = (1{,}1, -0.1)$ km. Przyjęto, iż oscylacje częstotliwości mają rozkład normalny o różnej wartości średniej μ i odchyleniu standardowym σ. W przypadku, gdy $N(\mu, \sigma)$ Hz wynosiło $N(0.1{,}0.01)$ udawało się uzyskać błędy lokalizacji rzędu pojedynczych metrów, podczas gdy dla $N(0.1, 2.72)$ błąd ten potrafił sięgać 45 metrów. Badanie wskazuje, iż dokładność estymacji częstotliwości jest zasadniczym elementem przekładającym się na błąd lokalizacji. Rozważaną metodą estymacji częstotliwości jest analiza widmowa odebranego sygnału [5]. Pozwala ona na wyznaczenie częstotliwości zarówno dla sygnałów harmonicznych, jak i emisji zmodulowanych poprzez kluczowanie fazy sygnału [6].

Celem tego artykułu jest zatem ocena ograniczeń związanych z dopuszczalną szerokością pasma sygnałów emitowanych przez lokalizowane nadajniki, w przypadku wykorzystania widmowej metody estymacji częstotliwości zaimplementowanej na SBC. Skupiono się ponadto wyłącznie na SBC możliwych do umieszczenia na BSP.

Układ pozostałej części artykułu został zorganizowany następująco. Sekcja II wprowadza w sposób estymacji częstotliwości z wykorzystaniem analizy widmowej. Stanowisko i scenariusz pomiarowy opisano w rozdziale III. Rozdział IV przedstawia uzyskane wyniki pomiarów. Opracowanie kończy krótkie podsumowanie.

2. WIDMOWA METODA ESTYMACJI CZĘSTOTLIWOŚCI

W przypadku sygnałów harmonicznych, po przekształceniu sygnału z dziedziny czasu do dziedziny częstotliwości, możemy zaobserwować charakterystyczny prążek w widmie na częstotliwości nośnej f_0. Analiza widmowa polega zatem na wyznaczeniu maksimum widma, a częstotliwość mu odpowiadająca jest równoważna częstotliwości odebranego sygnału, co ilustruje poniższe równanie:

$$f_0 = \arg\max_{k\in\{0,1,\ldots,N-1\}}\{|X(k)|\} \quad (1)$$

gdzie $X(k)$ – widmo częstotliwościowe sygnału, N – liczba punktów transformaty Fouriera.

W przypadku sygnałów zmodulowanych wyznaczenie częstotliwości w widmie jest znacznie trudniejsze ze względu na jego rozproszony charakter. Często nie zawiera ono częstotliwości charakterystycznej, na której można oprzeć estymację częstotliwości sygnału. Znaleziono rozwiązanie tego problemu dla sygnałów z kluczowaniem fazy [6]. O ile w widmie sygnału $x(t)$ z n-wartościową manipulacją fazy nie występuje częstotliwość charakterystyczna, o tyle można ją zaobserwować w widmie sygnału podniesionego do n-tej potęgi $x^n(t)$ na częstotliwości nf_0. Widmo takiego sygnału możemy oznaczyć poprzez $X_n(k)$. Chcąc zatem wyznaczyć częstotliwość nośną dla sygnału z dwuwartościową manipulacją fazy (BPSK) posłużymy się następującym równaniem:

$$f_0 = \frac{\arg\max_{k\in\{0,1,\ldots,N-1\}}\{|X_2(k)|\}}{2} \quad (2)$$

Analogicznie w przypadku zastosowania czterowartościowej manipulacji fazy (QPSK), wykorzystamy wzór przedstawiony poniżej:

$$f_0 = \frac{\arg\max_{k\in\{0,1,\ldots,N-1\}}\{|X_4(k)|\}}{4} \quad (3)$$

Należy tutaj zauważyć, iż dokładność odtwarzania częstotliwości (rozdzielczość transformaty Fouriera) Δf zależy od częstotliwości próbkowania f_s oraz liczby punktów transformaty N zgodnie z zależnością [7]:

$$\Delta f = \frac{f_s}{N} \quad (4)$$

W przypadku sygnałów kwadraturowych częstotliwość próbkowania f_s potrzebna do ich wiernego odtworzenia jest równa szerokości pasma B [8].

3. STANOWISKO I SCENARIUSZ POMIAROWY

Zasadniczy element stanowiska pomiarowego stanowił badany SBC wraz z odpowiednim oprogramowaniem. Przebadane zostały trzy różne SBC, które potencjalnie mogą zostać zamontowane na BSP. Były nimi:

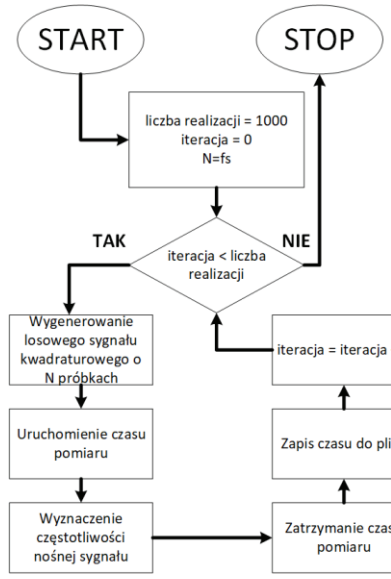
- Raspberry Pi 3B+ (RPi 3B+)
- Raspberry Pi 4B (RPi 4B)
- ODROID H3

W przypadku obu Raspberry Pi wykorzystano system operacyjny PiSDR oparty na Raspbianie [9], zaś na ODROID H3 zainstalowano Ubuntu 22.04.3 LTS.

Skrypt zawierający wszelkie niezbędne do wykonania pomiaru obliczenia napisany został w języku Python. Schemat blokowy głównego algorytmu przedstawiono na rys. 1. Badanie polegało na generacji sygnału zespolonego o N próbkach. Dla wygenerowanego sygnału mierzony był czas estymacji częstotliwości nośnej z wykorzystaniem metody widmowej opisanej w sekcji 2. Do obliczenia DFT (ang. Discrete Fourier Transform) wykorzystano algorytmy FFT (ang. Fast Fourier Transform) biblioteki NumPy. Określono czas obliczeń w przypadku analizy sygnału harmonicznego (równanie (1)), BPSK (równanie (2)), QPSK (równanie (3)). Dokładność odtwarzania częstotliwości Δf ustalona została na 1 Hz. Badanie przeprowadzono w dwóch wariantach:

- gdy N stanowiło potęgę dwójki, $N = 2^k = \{32768, 65536, 131072, 262144, 524288, 1048576, 2097152, 4194304\}$
- gdy N nie było potęgą dwójki, $N = 2^k - 1$

W przypadku, gdy czas obliczeń dla danej wartości N sięgał rzędu kilkunastu sekund, nie wykonywano badania dla większych wartości N.



Rys. 1. Główny algorytm skryptu pomiarowego

4. WYNIKI BADAŃ

W tabelach 1, 2 i 3 przedstawiono szczegółowe wyniki dotyczące czasu wykonywania obliczeń związanych z estymacją częstotliwości charakterystycznej odpowiednio dla platformy RPi 3B+, RPi 4B oraz Odroid H3 w zależności od liczby punktów transformaty Fouriera N. W poszczególnych kolumnach pokazano wartość średnią oraz odchylenie standardowe czasu niezbędnego do wyznaczenia częstotliwości charakterystycznej w przypadku analizy sygnału harmonicznego (kolumna 1), BPSK (ko-

lumna 2) oraz QPSK (kolumna 3). Wartość średnią i odchylenie standardowe każdorazowo wyznaczono dla serii zawierającej 1000 pomiarów. Na rys. 2 przedstawiono przykładowe wyniki serii pomiarów wykonanych na RPi 4B dla sygnału harmonicznego i $N = 32768$. Ze względu na brak znajomości rozkładu danych pomiarowych, do analizy statystycznej wykorzystano nierówność Czebyszewa [10], [11]:

$$P(|Y - \mu| \geq q\sigma) \leq \frac{1}{q^2} \tag{5}$$

gdzie: Y – zmienna losowa o dowolnym rozkładzie, $\mu = E(Y)$ – wartość oczekiwana, $\sigma^2 = E[(Y - \mu)^2]$ – wariancja.

Równanie (5) oznacza, że prawdopodobieństwo, iż wartość zmiennej losowej Y odchyla się od swojej średniej o więcej niż q odchyleń standardowych, jest nie większe niż $1/q^2$.

Tab. 1. Czas wykonywania obliczeń dla RPi 3B+

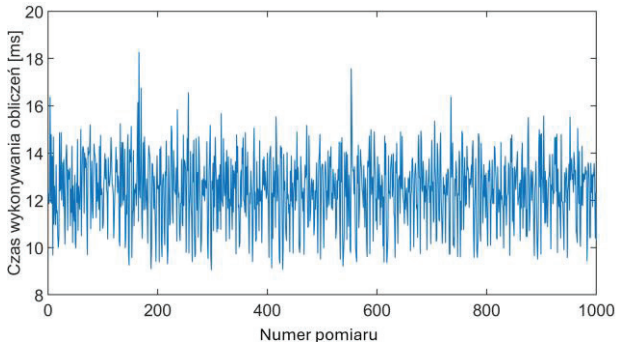
	Czas obliczeń, gdy analizowano sygnał [s]:		
N	*harmoniczny*	*BPSK*	*QPSK*
32768	0,031±0,003	0,027±0,002	0,029±0,002
65536	0,066±0,005	0,059±0,005	0,062±0,003
131072	0,131±0,008	0,120±0,002	0,123±0,002
262144	0,258±0,015	0,237±0,003	0,246±0,007
524288	0,507±0,011	0,508±0,049	0,518±0,025
1048576	1,079±0,018	1,152±0,076	1,159±0,030
2097152	2,494±0,010	2,459±0,011	2,501±0,079
32767	0,074±0,008	0,075±0,010	0,073±0,003
65535	0,304±0,002	0,266±0,003	0,257±0,002
131071	0,642±0,003	0,579±0,003	0,577±0,002
262143	0,495±0,010	0,480±0,004	0,479±0,002
524287	2,970±0,010	2,870±0,015	2,903±0,007
1048575	2,022±0,044	2,069±0,046	2,144±0,047
2097151	13,067±0,175	13,111±0,094	13,434±0,239

Tab. 2. Czas wykonywania obliczeń dla RPi 4B

	Czas obliczeń, gdy analizowano sygnał [s]:		
N	*harmoniczny*	*BPSK*	*QPSK*
32768	0,012±0,001	0,013±0,002	0,011±0,001
65536	0,029±0,003	0,023±0,002	0,024±0,002
131072	0,068±0,009	0,050±0,003	0,052±0,003
262144	0,132±0,017	0,096±0,007	0,095±0,003
524288	0,241±0,011	0,211±0,005	0,221±0,021
1048576	0,413±0,049	0,423±0,032	0,456±0,013
2097152	0,804±0,016	0,937±0,019	0,861±0,049
4194304	1,932±0,028	2,030±0,016	2,035±0,012
32767	0,023±0,002	0,023±0,002	0,026±0,003
65535	0,133±0,003	0,121±0,011	0,138±0,004
131071	0,238±0,009	0,206±0,011	0,248±0,008
262143	0,206±0,010	0,165±0,001	0,210±0,005
524287	1,013±0,011	0,895±0,038	0,975±0,004
1048575	0,768±0,034	0,825±0,004	0,822±0,022
2097151	4,404±0,219	4,507±0,063	4,543±0,046
4194303	13,540±0,153	14,207±0,078	13,357±0,044

Tab. 3. Czas wykonywania obliczeń dla ODROID H3

	Czas obliczeń, gdy analizowano sygnał [s]:		
N	*harmoniczny*	*BPSK*	*QPSK*
32768	0,005±0,001	0,005±0,001	0,005±0,001
65536	0,010±0,002	0,010±0,002	0,010±0,001
131072	0,022±0,004	0,021±0,003	0,021±0,003
262144	0,037±0,003	0,039±0,004	0,039±0,002
524288	0,066±0,004	0,078±0,008	0,075±0,004
1048576	0,108±0,003	0,150±0,009	0,148±0,005
2097152	0,215±0,004	0,257±0,026	0,258±0,007
4194304	0,544±0,062	0,466±0,006	0,537±0,003
32767	0,011±0,002	0,010±0,001	0,011±0,001
65535	0,037±0,002	0,033±0,003	0,036±0,002
131071	0,069±0,001	0,062±0,002	0,067±0,002
262143	0,063±0,003	0,060±0,002	0,065±0,002
524287	0,214±0,003	0,231±0,016	0,260±0,013
1048575	0,172±0,004	0,187±0,005	0,212±0,005
2097151	0,929±0,004	0,939±0,007	0,975±0,004
4194303	2,756±0,011	2,777±0,008	2,877±0,159



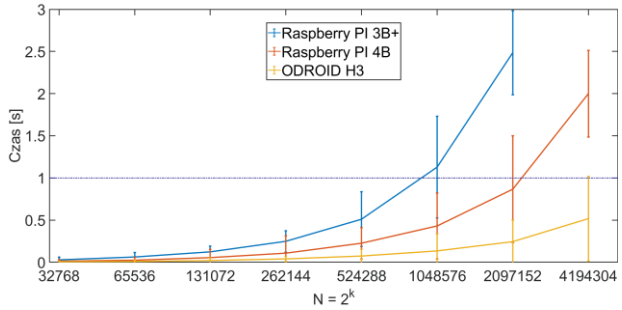
Rys. 2. Przykładowe wyniki serii pomiarów wykonanych na RPi 4B dla sygnału harmonicznego i N = 32768

Na rys. 3 i 4 przedstawiono, określone z 99% prawdopodobieństwem (q=10), czasy wykonywania obliczeń związanych z wyznaczaniem częstotliwości charakterystycznej dla badanych SBC w zależności od liczby punktów N. W tym przypadku uśredniono wyniki serii pomiarowych dotyczących sygnału harmonicznego, BPSK oraz QPSK.

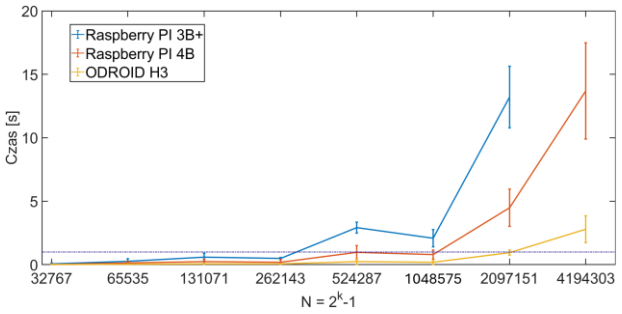
Zakładając, iż $\Delta f = 1$ Hz, częstotliwość próbkowania f_s i pasmo sygnału B, co do wartości są równe liczbie punktów N. Z rys. 3 i 4 wynika, że na szybkość wykonywania obliczeń znacząco wpływa spełnienie warunku $N=2^k$.

Zakładając, że $f_s=B=2^k$ oraz $\Delta f = 1$ Hz z wykorzystaniem RPi 3B+, możemy wyznaczyć częstotliwość charakterystyczną dla sygnałów zmodulowanych fazowo, których pasmo jest mniejsze niż 0,5 MHz. Przy założeniu, że wartość częstotliwości wyznaczana jest co sekundę, ΔT=1 s. Wykorzystując RPi 4B lub ODRID H3 pasmo analizowanego sygnału może być szersze i wynosić odpowiednio 1 MHz oraz 4 MHz, przy tych samych założeniach ($\Delta f = 1$ Hz, ΔT=1 s). Chcąc wyznaczać częstotliwość charakterystyczną z inną (mniejszą lub większą) rozdzielczością, bądź też wykonywać obliczenia w innych odstępach czasu ΔT (częściej lub rzadziej) należy

odpowiednio dobrać pasmo analizowanego sygnału. Przykładowe szerokości pasma dla różnych wartości Δf i ΔT dla RPi 4B przedstawiono w tab. 4.



Rys. 3. Czas wyznaczania częstotliwości charakterystycznej dla badanych SBC ($N = 2^k$)



Rys. 4. Czas wyznaczania częstotliwości charakterystycznej dla badanych SBC ($N = 2^k - 1$)

Tab. 4. Proponowane maksymalne szerokości pasma dla sygnałów zmodulowanych fazowo

	Δf = 1 Hz	Δf = 2 Hz	Δf = 4 Hz
ΔT=0,1 s	65536 Hz	131072 Hz	262144 Hz
ΔT=0,5 s	524255 Hz	1048576 Hz	2097152 Hz
ΔT=1 s	1048576 Hz	2097152 Hz	4194304 Hz

5. PODSUMOWANIE

Celem artykułu była ocena ograniczeń związanych z dopuszczalną szerokością pasma sygnałów emitowanych przez lokalizowane nadajniki, w przypadku wykorzystania widmowej metody estymacji częstotliwości zaimplementowanej na SBC.

Uzyskane wyniki wskazują, iż spełnienie warunku $N = 2^k$ umożliwia znacznie skrócić czas wykonywanych na SBC obliczeń. Uwzględniając wpływ dokładności estymacji częstotliwości na błąd lokalizacji metody SDF, opisany w sekcji 1, można uznać, iż rozdzielczość Δf = 2 Hz pozwoli na lokalizację emiterów z błędem nie większym niż 50 m. Przyjęta dokładność Δf oraz możliwości obliczeniowe SBC determinują natomiast dostępne do analizy szerokości pasma sygnałów emitowanych przez lokalizowane nadajniki. Przykładowo wykorzystanie RPi 4B pozwoli na lokalizację nadajników radiowych pracujących z emisją QPSK w pasmie do 2 MHz, wyznaczając częstotliwość charakterystyczną co 1 sekundę (tab. 4). Chcąc to robić częściej, np. co 0,1 sekundy, szerokość pasma analizowanego sygnału nie może w tym przypadku przekraczać 100 kHz. Do analizowania sygnałów o znacznie szerszym pasmie, lepszym rozwiązaniem okaże się ODROID H3. Jego transport będzie jednak wymagał znacznie większego BSP niż w przypadku obu Raspberry Pi, co nie zawsze jest możliwe.

Praktyczne wykorzystanie powyższych wniosków znajdzie swoje odzwierciedlenie w projektowanym sensorze lokalizacyjnym, zbudowanym w oparciu o SBC i radio programowalne. Pozwolą one na odpowiedni dobór SBC oraz parametrów przetwarzania sygnałów w zależności od stawianych wymagań projektowych, zwłaszcza tych związanych z częstością wykonywanych obliczeń.

LITERATURA

[1] P. Gajewski, C. Ziółkowski, i J. M. Kelner, „Przestrzenna dopplerowska metoda lokalizacji źródeł sygnałów radiowych”, *Biuletyn Wojskowej Akademii Technicznej*, t. 60, nr 4, s. 187–200, 2011.

[2] J. Kelner i C. Ziółkowski, „Zastosowanie metody SDF do lokalizacji źródeł emitujących sygnały z manipulacją fazy”, *Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne (ISSN: 1230-3496)*, t. 87, s. 205–208, cze. 2014.

[3] R. Szczepanik i J. M. Kelner, „Implementacja metody SDF na platformie radia programowalnego”, *Elektronika : konstrukcje, technologie, zastosowania*, t. Vol. 60, nr nr 3, 2019, doi: 10.15199/13.2019.3.7.

[4] K. Bednarz, J. Wojtuń, J. M. Kelner, i K. Różyc, „Frequency Instability Impact of Low-Cost SDRs on Doppler-Based Localization Accuracy”, *sensors*, luty 2024, doi: https://doi.org/10.3390/s24041053.

[5] J. Kelner i C. Ziółkowski, *Lokalizacja źródeł emisji na podstawie analizy widmowej odbieranego sygnału*. 2014. doi: 10.13140/RG.2.1.1082.4804.

[6] J. Kelner i C. Ziółkowski, „Zastosowanie technologii SDF do lokalizowania źródeł emisji BPSK i QPSK”, *Przeglad Elektrotechniczny*, t. 91, s. 61–65, mar. 2015, doi: 10.15199/48.2015.03.14.

[7] T. P. Zieliński, *Cyfrowe przetwarzanie sygnałów. Od teorii do zastosowań*. Warszawa: Wydawnictwa Komunikacji i Łączności, 2014. Dostęp: 22 maj 2024. [Online]. Dostępne na: https://www.wkl.com.pl/cyfrowe-przetwarzanie-sygnalow-od-teorii-do-zas,1,1,400?

[8] R. G. Lyons, *Wprowadzenie do cyfrowego przetwarzania sygnałów*. 2010. Dostęp: 22 maj 2024. [Online]. Dostępne na: https://www.wkl.com.pl/wprowadzenie-do-cyfrowego-przetwarzania-sygnalow,1,1,1322?

[9] L. Cruz, „luigifcruz/pisdr-image”. 21 maj 2024. Dostęp: 22 maj 2024. [Online]. Dostępne na: https://github.com/luigifcruz/pisdr-image

[10] R. V. Hogg, J. W. McKean, i A. T. Craig, *Introduction to mathematical statistics*, Eighth edition. Boston: Pearson, 2019.

[11] R. Routledge, „Chebyshev’s inequality”. Dostęp: 22 maj 2024. [Online]. Dostępne na: https://www.britannica.com/science/Chebyshevs-inequality

SELEKCJA WĘZŁÓW POPRZEZ MINIMALIZACJĘ BŁĘDU LOKALIZACJI W WĄSKICH PRZEJŚCIACH

ANCHOR SELECTION THROUGH POSITIONING ERROR MINIMIZATION IN NARROW PASSAGES

Marcin Kołakowski

Instytut Radioelektroniki i Technik Multimedialnych, Politechnika Warszawska, Warszawa
marcin.kolakowski@pw.edu.pl

DOI: 10.15199/59.2024.4.17

Streszczenie: W referacie przedstawiono metodę selekcji wyników pomiarów przeznaczoną dla systemów lokalizacji osób wewnątrz pomieszczeń. Metoda zakłada podział pomieszczenia na strefy, którym przypisane są zestawy węzłów systemu pozwalające na lokalizację z największa dokładnością. Najkorzystniejsze zestawy są przyporządkowywane do stref na podstawie miary uwzględniającej fizyczne ograniczenia lokalizacji użytkownika i najczęściej obierane przez niego trajektorie ruchu. Badania eksperymentalne wykazały, że zastosowanie metody pozwala na zwiększenie dokładności lokalizacji w stosunku do stosowania stałych zestawów węzłów.

Abstract: The paper presents a measurement results selection method for indoor person positioning systems. The technique assumes dividing the system area into zones, which are assigned anchor sets yielding the highest positioning accuracy. The anchor sets are assigned based on an accuracy metric considering physical limitations of user's whereabouts and typical paths he takes. The experiments have shown that the proposed method increases the positioning accuracy compared to using fixed sets of anchor nodes.

Słowa kluczowe: lokalizacja wewnątrzbudynkowa, metody adaptacyjne, UWB

Keywords: adaptive methods, indoor positioning, UWB

1. WSTĘP

W typowym scenariuszu pracy radiowego systemu lokalizacyjnego, do jednoznacznego wyznaczenia położenia użytkownika w dwóch wymiarach potrzebne są co najmniej trzy wyniki pomiarów. Jednakże, w większości implementacji, aby skutecznie pokryć całą założoną przestrzeń działaniem systemu stosowane są wielowęzłowe infrastruktury dostarczające większą liczbę wyników.

Ze względu na zjawiska występujące w kanale radiowym (tłumienie, przysłanianie, odbicia) część wyników może być obarczona zniekształceniami skutkującymi zwiększeniem błędu lokalizacji. Nadmiarowość wyników zwracanych przez infrastrukturę stwarza możliwość wybrania jedynie tych pozwalających na otrzymanie wysokiej dokładności lokalizacji [4].

Selekcja wyników nie jest zadaniem prostym [1]. W systemach o wielowęzłowej infrastrukturze, liczba możliwych kombinacji wyników pomiarów może być znacząca. Przykładowo, w systemie składającym się z sześciu węzłów, liczba możliwych kombinacji co najmniej trzech węzłów wynosi 42. Liczba ta jest znacząco większa w przypadku systemów stosujących metody różnicowe np. Time Difference of Arrival (TDOA). Ze względu na to, że wybierane są zestawy par, liczba liniowo niezależnych kombinacji wyników w analogicznym systemie to 2811.Tak duże liczby sprawiają, że ręczny wybór korzystnych zestawów wyników jest praktycznie niemożliwy i konieczne jest zastosowanie zautomatyzowanych metod.

Najpopularniejsze metody selekcji wyników oparte są na analize parametrów sygnału w celu detekcji pracy w warunkach NLOS (ang. Non Line of Sight). Przykładowo, w [8], analizowany jest poziom sygnałów nadawanych przez węzły oraz ich dostępność. Zaproponowana metoda zakłada, że sygnały o wysokim poziomie zostały nadane przez pobliskie węzły w związku z czym prawdopodobieństwo tego że sygnał propagował przez przeszkody jest mniejsze niż w przypadku bardziej oddalonych.

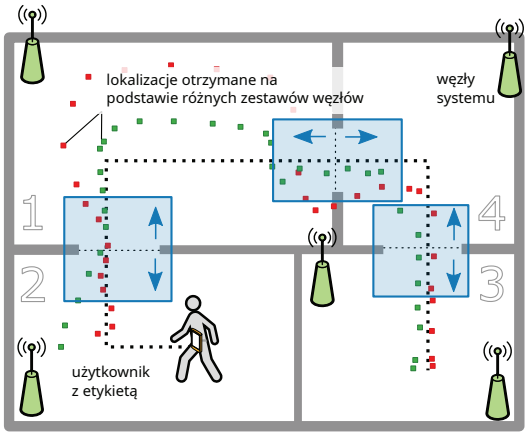
Metoda opisana w [2] oparta jest na analizie bieżącej lokalizacji użytkownika i ocenie prawdopodobieństwa wystąpienia wyników. Algorytm oblicza lokalizację użytkownika na podstawie wszystkich możliwych kombinacji wyników i wybiera wynik najbliższy predykcji.

Metody [6, 7] zakładają podział środowiska na strefy, do których przyporządkowano konkretne zestawy węzłów. W [6] korzystne zestawy dla każdej ze stref są określane poprzez minimalizację błędu lokalizacji samolokalizującego się robota. W [7] analizowana jest dokładność lokalizacji w wąskich przejściach gdzie fizyczna lokalizacja użytkownika jest ograniczona.

W referacie zaprezentowano metodę będącą kolejnym etapem rozwoju podejścia zaprezentowanego w [7]. W zaproponowanym rozwiązaniu, dobór korzystnych zestawów wyników dla stref również jest realizowany poprzez analizę dokładności lokalizacji użytkownika w wąskich przejściach. W odróżnieniu od wcześniejszego rozwiązania, zamiast odległości wyznaczonych lokalizacji od osi przejścia, analizowana jest dokładność odwzorowania trajektorii ruchu względem typowo obieranej przez użytkownika.

2. KONCEPCJA METODY

Koncepcja metody została zilustrowana na Rys. 1.

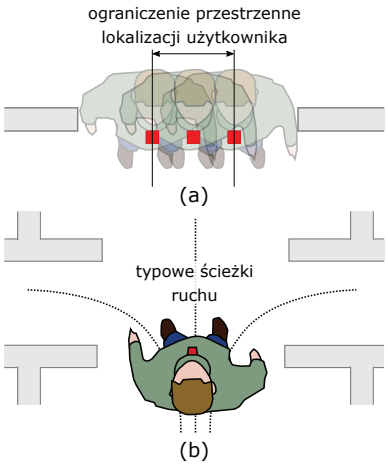


Rysunek 1: Koncepcja metody selekcji węzłów

Metoda zakłada podział pomieszczenia objętego działaniem systemu na rozłączne strefy i przyporządkowanie im zestawów węzłów (lub par węzłów), dla których osiągane są najdokładniejsze wyniki lokalizacji. Najkorzystniejsze zestawy węzłów są określane dla różnych zwrotów użytkownika tak by uwzględnić zaburzenia propagacji wynikające z przysłaniania etykiety ciałem. Wyniki, na podstawie których lokalizowany jest użytkownik są wybierane dynamicznie na podstawie jego aktualnej lokalizacji i zwrotu.

To jakie zestawy węzłów są najkorzystniejsze dla danej strefy określane jest automatycznie podczas korzystania z systemu. Po zgromadzeniu przez system reprezentatywnego zbioru wyników pomiarów (np. z kilku dni) analizowana jest dokładność lokalizacji na podstawie różnych zestawów wyników i wybierane są najkorzystniejsze.

Ustalenie najkorzystniejszych zestawów węzłów wymaga określenia dokładności lokalizacji. W przyjętym scenariuszu, rzeczywista lokalizacja użytkownika nie jest znana. Nie można więc obliczyć bezwzględnego błędu lokalizacji. Zaproponowana metoda oparta jest na założeniu, że lokalizacja użytkownika jest w niektórych miejscach fizycznie ograniczona co przedstawiono na Rys. 2.



Rysunek 2: Ograniczenia lokalizacji użytkownika: (a) ograniczenia przestrzenne; (b) typowe trajektorie ruchu użytkownika

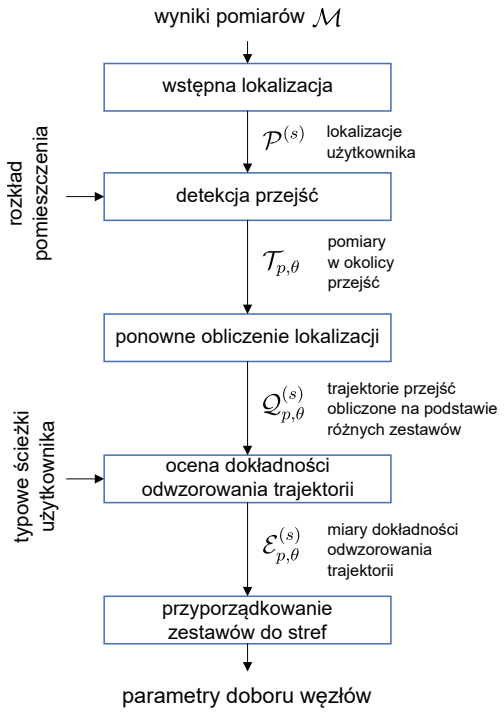
W środowisku wewnątrzbudynkowym, lokalizacja użytkownika jest ograniczona poprzez ściany i obiekty w nim występującym. Szczególnym przypadkiem są wąskie przejścia np. drzwi (Rys. 2.a). Jako, że typowe drzwi mają między 0,7 a 0,9 m szerokości a rozstaw barków typowego człowieka wynosi ok 0,5–0,6 m, lokalizacja użytkownika poprzez przechodzenia przez drzwi zawiera się w wąskim obszarze 0,3–0,4 m. Dodatkowo, po przejściu przez drzwi użytkownik zazwyczaj porusza się po jednej z typowych dla siebie trajektorii (Rys. 2.b).

Dokładność lokalizacji w przejściach można więc ocenić poprzez analizę dokładności odwzorowania trajektorii ruchu. W zaproponowanej metodzie wykorzystywana jest do tego odległość Frécheta [3]. Węzły dla danych stref są dobierane na podstawie analizy dokładności lokalizacji w przejściach leżących na ich obszarze.

Zaproponowana metoda nie jest związana z konkretnym typem systemu i może znaleźć zastosowanie w systemach wąskopasmowych jak i ultraszerokopasmowych stosujących zarówno metody oparte na pomiarze odległości jak i różnicowe np. TDOA.

3. IMPLEMENTACJA METODY

Schemat blokowy procedury przyporządkowania zestawów węzłów strefom przedstawiono na Rys. 3.



Rysunek 3: Schemat blokowy procedury przyporządkowania zestawów węzłów do stref

Pierwszym etapem procedury jest wstępna lokalizacja użytkownika na podstawie zgromadzonych przez system wyników pomiarów $\mathcal{M}$:

$$\mathcal{M} = \{z_k \mid k = 1...K\} \tag{1}$$

$$\mathcal{P}^{(s)} = \{p_k^{(s)} \mid k = 1 \dots K\} \tag{2}$$

$$p_k^{(s)} = \mathrm{LS}(z_k, s) \tag{3}$$

gdzie z_k to wyniki pomiarów a $p_k^{(s)}$ to wyniki lokalizacji za pomocą estymatora najmniejszych kwadratów (ang.

Least Squares) dla zestawu węzłów s. Wyniki wstępnej lokalizacji są analizowane w celu wykrycia detekcji przejść przez znajdujące się w strefach wąskie obszary. W tym celu analizowana jest odległość wyników lokalizacji od osi przejścia oraz to czy użytkownik pokonał całe przejście bez zawracania. Efektem jest zbiór wyników pomiarów odpowiadającym przejściom $\mathcal{T}_{p,\theta}$:

$$\mathcal{T}_{p,\theta} = \{t_{p,\theta,i} \mid i = 1 \dots N_{\text{det}}\} \tag{4}$$

$$t_{p,\theta,i} = \{z_k \mid k = m_i \dots n_i\} \tag{5}$$

gdzie p to numer przejścia, θ zwrot użytkownika podczas jego pokonywania a N_{det} jest łączną liczbą zarejestrowanych dla tych parametrów trajektorii.

Lokalizacja użytkownika w obszarze przejść jest ponownie wyznaczana na podstawie wielu zestawów węzłów.

$$\mathcal{Q}^{(s)}_{p,\theta} = \{q^{(s)}_{p,\theta,i} \mid i = 1 \dots N_{\text{det}}\} \tag{6}$$

$$q^{(s)}_{p,\theta,i} = \{\text{LS}(z, s) \mid z \in t_{p,\theta,i}\} \tag{7}$$

W przypadku poniższej implementacji, analizowane są wszystkie możliwe zestawy. Dla systemów składających się z wielu węzłów korzystne byłoby zawężenie analizowanego zbioru np. poprzez zastosowanie metod heurystycznych. Dla każdej wyznaczonej trajektorii przejścia $q^{(s)}_{p,\theta,i}$ obliczana jest miara dokładności odwzorowania względem jednej z typowych ścieżek użytkownika. Typowe ścieżki użytkowników mogą być określane za pomocą metod symulacyjnych lub na podstawie danych historycznych zebranych podczas działania systemu. Przyjętą miarą dokładności jest odległość Frécheta δ_{dF}, która pozwala na ocenę podobieństwa dwóch trajektorii przy uwzględnieniu kolejności ich punktów:

$$\mathcal{E}^{(s)}_{p,\theta} = \{\epsilon^{(s)}_{p,\theta,i} \mid i = 1 \dots N_{\text{det}}\} \tag{8}$$

$$\epsilon^{(s)}_{p,\theta,i} = \min\{\delta_{dF}(q^{(s)}_{p,\theta,i}, r_{p,\theta,j}) \mid j = 1 \dots N_{\text{ref}}\} \tag{9}$$

Następnie, otrzymane miary dokładności są agregowane dla poszczególnych stref:

$$\mathcal{Z}^{(s)}_{z,\theta} = \bigcup \{\mathcal{E}^{(s)}_{p,\theta} \mid \text{dla każdego przejścia } p \text{ w strefie } z\} \tag{10}$$

gdzie z jest identyfikatorem strefy. Dla każdego z analizowanych zestawów wyników obliczana jest wartość funkcji kosztu:

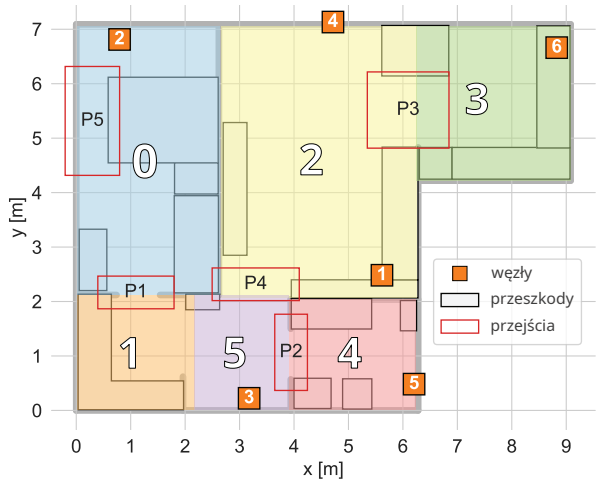
$$J(\mathcal{Z}^{(s)}_{z,\theta}) = \bar{\mathcal{Z}}^{(s)}_{z,\theta} + \min \mathcal{Z}^{(s)}_{z,\theta} \tag{11}$$

gdzie $\bar{\mathcal{Z}}^{(s)}_{z,\theta}$ i $\min \mathcal{Z}^{(s)}_{z,\theta}$ są średnią i minimalną wartością miary błędu dla każdej strefy z i zwrotu użytkownika θ. Ilość analizowanych zwrotów użytkownika jest zależna od umiejscowienia i kierunków przejść znajdujących się w danej strefie. Na przykład w strefie 1 na Rys. 1 analizowane będą cztery zwroty a w strefie 2 dwa.

Dla każdej kombinacji strefa-zwrot użytkownika dobierany jest zestaw węzłów o najmniejszej wartości funkcji kosztu. W celu maksymalizacji dostępności lokalizacji, wybór jest dokonywany dla wszystkich możliwych scenariuszy dostępności wyników pomiarów. Wyniki procedury są wykorzystywane podczas przyszłej rutynowej pracy systemu do adaptacyjnego wyboru wyników pomiarów.

4. BADANIA EKSPERYMENTALNE

Zaproponowana metoda została zweryfikowana eksperymentalnie. Badania zostały przeprowadzone w typowym umeblowanym mieszkaniu. którego rozkład przedstawiono na Rys. 4.

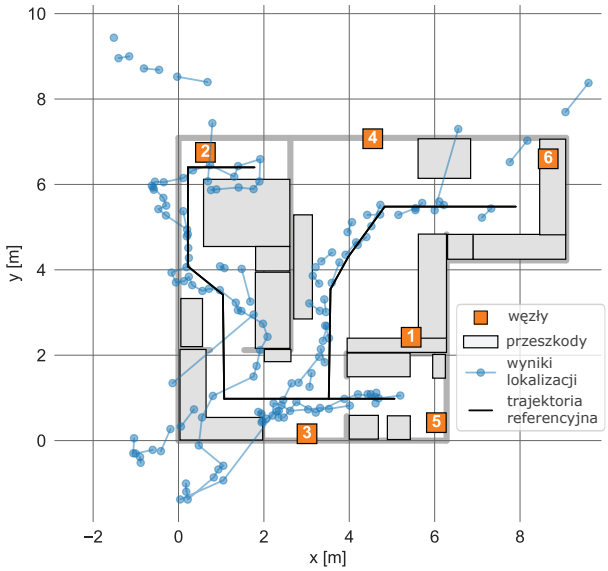


Rysunek 4: Schemat środowiska eksperymentalnego

Mieszkanie zostało podzielone na sześć stref obejmujących poszczególne pomieszczenia. Do badań wykorzystano ultraszerokopasmowy system lokalizacyjny zrealizowany na Politechnice Warszawskiej [5]. System składał się z sześciu węzłów rozmieszczonych tak by w każdym pomieszczeniu dostępny był co najmniej jeden. W trakcie eksperymentu użytkownik przemieszczał się po złożonej z prostych odcinków trajektorii obejmującej wszystkie pokoje. Lokalizacja użytkownika była wyznaczana na podstawie pomiarów TDOA z okresem 0,3 s. Do lokalizacji został wykorzystany algorytm oparty na rozszerzonym filtrze Kalmana.

Ocenę dokładności poprzedziła kalibracja metody. W tym celu zebrano zbiór danych odpowiadający swobodnym przemieszczaniu się po mieszkaniu przez 15 minut. Dane zostały następnie przetworzone zgodnie z procedurą opisaną w rozdziale 3.

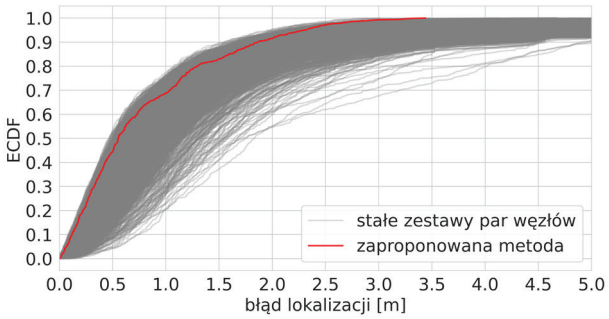
W celu oceny dokładności zaproponowanej metody, lokalizacja użytkownika została obliczona za jej pomocą oraz dla stałych zestawów par węzłów (2811 liniowo niezależnych zestawów par węzłów po 3,4 i 5 par). Przykładowe wyniki lokalizacji przedstawiono na Rys. 5.



Rysunek 5: Wybrane wyniki lokalizacji otrzymane za pomocą zaproponowanej metody

Zaproponowana metoda pozwala na odtworzenie trajektorii ruchu użytkownika. Najmniejszą dokładność lokalizacji otrzymano w skrajnych punktach pomieszczeń, dla których wyznaczone lokalizacje znajdowały się poza granicami mieszkania. Powyższy problem można zniwelować poprzez wprowadzenie dodatkowych ograniczeń przy obliczaniu lokalizacji.

Dokładność została oceniona na podstawie błędu lokalizacji. Użytkownik, podczas badań był wyposażony w stoper, którym rejestrował czasy przebycia kolejnych fragmentów trajektorii. Zakładając stałą prędkość ruchu użytkownika wyznaczono jego referencyjne lokalizacje, względem których obliczono błędy. Empiryczne dystrybuanty błędu zostały przedstawione na Rys. 6. Wybrane statystyki błędu zebrano w Tab. 1.



Rysunek 6: Empiryczna dystrybuanta błędu

Tabela 1: Parametry statystyczne błędu lokalizacji

parametr	*wartość [m]*	*kwantyl*
średnia	0,79	0,001
$Q_{0.25}$	0,27	0,03
$Q_{0.5}$ (mediana)	0,55	0,04
$Q_{0.75}$	1,14	0,12
$Q_{0.9}$	1,86	0,08

Zaproponowana metoda, w większości przypadków pozwala na lokalizację z większą dokładnością niż w przypadku zastosowania stałych par węzłów. Średnia wartość błędu i jego mediana wyniosły 0,79 m i 0,55 m co było wartościami lepszymi niż dla odpowiednio 99,9 $ i 96,0 % stałych zestawów par. W eksperymencie dla części stałych zestawów par węzłów otrzymano dokładniejsze wyniki. Niestety, manualny wybór tych zestawów z 2811 dostępnych jest na etapie instalacji systemu zadaniem praktycznie niemożliwym.

5. PODSUMOWANIE

W referacie przedstawiono metodę adaptacyjnej selekcji wyników pomiarów przeznaczoną do radiowych systemów lokalizacyjnych pracujących wewnątrz pomieszczeń. Metoda zakłada podział obszaru na strefy i dobór najkorzystniejszych węzłów dla obecnej lokalizacji i zwrotu użytkownika. Korzystne zestawy węzłów są określane poprzez analizę dokładności lokalizacji w wąskich przejściach między strefami.

Wyniki badań eksperymentalnych pokazały, że metoda pozwala na osiągnięcie wyższej dokładności niż w przypadku zastosowania większości z możliwych zestawów par węzłów. Metoda ma potencjał do dalszego rozwoju. Przykładowymi kierunkami mogą być zaproponowanie innej niż odległość Frécheta miary dokładności lub zastosowanie algorytmów heurystycznych w celu przyspieszenia doboru korzystnych zestawów węzłów dla stref.

W referacie przedstawiono wyniki prac realizowanych w ramach projektu finansowanego przez Narodowe Centrum Badań i Rozwoju w ramach programu AAL Call 2021 nr umowy AAL/AALCall2021/52/CAREUP/2022. Publikacja powstała przy wsparciu Fundacji Rozwoju Radiokomunikacji i Technik Multimedialnych.

LITERATURA

[1] Chen Yingying i in., 2019, „Impact of Anchor Placement and Anchor Selection on Localization Accuracy". „Handbook of Position Location", 435-465, doi:10.1002/9781118104750.ch13

[2] Courtay Antoine i in., 2019, „Anchor Selection Algorithm for Mobile Indoor Positioning using WSN with UWB Radio". 2019 IEEE Sensors Applications Symposium (SAS), 1-5, IEEE, doi: 10.1109/SAS.2019.8706113.

[3] Eiter Thomas, Mannila Heikki, 1994, „Computing Discrete Fréchet Distance". CD-TR 94/64, Technische Universität Wien

[4] Kaune Regina i in., 2011, „Accuracy analysis for TDOA localization in sensor networks". 14th International Conference on Information Fusion: 1-8

[5] Kołakowski Jerzy i in., 2020, „UWB/BLE Tracking System for Elderly People Monitoring". Sensors, 20: 1574, doi: 10.3390/s20061574.

[6] Kołakowski Marcin, 2021, „Adaptive Anchor Pairs Selection in a TDOA-based System Through Robot Localization Error Minimization". 2021 Signal Processing Symposium (SPSympo), 128–132, doi: 10.1109/SPSympo51155.2020.9593477.

[7] Kołakowski Marcin, Modelski Józef, 2022, „Anchor Pair Selection in TDOA Positioning Systems by Door Transition Error Minimization". 2022 24th International Microwave and Radar Conference (MIKON), 1-4, doi: 10.23919/MIKON54314.2022.9924777.

[8] Zhang Wei i in., 2020, „A Self-Adaptive AP Selection Algorithm Based on Multiobjective Optimization for Indoor WiFi Positioning". IEEE Internet of Things Journal, 8, 1406-1416, doi: 10.1109/JIOT.2020.3011402.

RADIOLOKACJA ENDOSKOPÓW KAPSUŁKOWYCH WYKORZYSTUJĄCA ADAPTACYJNY ALGORYTM OPRATY NA DETEKCJI FAZY

WIRELESS LOCALIZATION OF CAPSULE ENDOSCOPES USING AN ADAPTIVE ALGORITHM BASED ON PHASE DETECTION

Paweł Oleksy[1]; Łukasz Januszkiewicz[2];

[1] Politechnika Łódzka, Instytut Elektroniki, Łódź, pawel.oleksy@p.lodz.pl
[2] Politechnika Łódzka, Instytut Elektroniki, Łódź, lukasz.januszkiewicz@p.lodz.pl
…

DOI: 10.15199/59.2024.4.18

Streszczenie**: Diagnostykę przewodu pokarmowego można obecnie przeprowadzić za pomocą endoskopów kapsułkowych, które pozwalają wykryć zmiany chorobowe, nie powodując przy tym dyskomfortu dla pacjenta. W przypadku wykrycia jakichkolwiek nieprawidłowości nie mają jednak informacji o ich dokładnym położeniu. W artykule zaprezentowano metodę lokalizacji endoskopu kapsułkowego opartą na analizie fazy sygnału odebranego wykorzystującą adaptacyjny model ciała człowieka. Artykuł opracowano na podstawie rozprawy doktorskiej P.Oleksego.**

Abstract**: Diagnosis of diseases of the digestive system can now be performed using capsule endoscopes, which allow the detection of pathological changes without causing discomfort to the patient. Although doctors can obtain a clear image of the digestive tract, if any abnormalities are detected, they have no information about their exact location. The article presents a method of wireless endoscope localization based on the analysis of received signal phase using also an adaptive model of the human body. The article was prepared on the basis of P. Oleksy's doctoral dissertation.**

Słowa kluczowe**: : Endoskopy bezprzewodowe, FDTD, Modele ciała człowieka, Radiolokaliacja.**

Keywords**: Wireless endoscopes, FDTD, Human body models, Radiolocalization**

1. WSTĘP

W diagnostyce obrazowej chorób układu pokarmowego obecnie najczęściej wykorzystywaną techniką jest endoskopia. Pozwala ona zidentyfikować zmiany chorobowe powstałe we wnętrzu układu pokarmowego przy użyciu miniaturowych kamer. Niestety w endoskopii kapsułkowej dostępne są jedynie sekwencje obrazów zarejestrowanych w kolejnych fragmentach przewodu pokarmowego. Informacja o położeniu kapsułki podczas rejestracji konkretnego obrazu nie jest dostępna. W przypadku wykrycia zmiany chorobowej nie jest możliwa jej lokalizacja, która może być użyteczna w przypadku konieczności wprowadzenia precyzyjnego dozowania środków farmakologicznych czy też przeprowadzenia operacji.

W literaturze opisano wiele technik lokalizacji endoskopów kapsułkowych bazujących zarówno na analizie obrazu zarejestrowanego przez kapsułkę jak i parametrów sygnału radiowego wykorzystywanego do transmisji danych z kapsułki. Niektóre rozwiązania wprowadzają dodatkowe elementy do systemu, takie jak np. czujniki inercyjne umożliwiające śledzenie ruchu i orientacji kapsułki w przestrzeni trójwymiarowej[9]. Ciało człowieka zbudowane jest z wielu tkanek o różnych właściwościach elektrycznych co ma bezpośredni wpływ na parametry łączą pomiędzy endoskopem a odbiornikami umiejscowionymi na ciele. W zależności od częstotliwości sygnału ciało człowieka w różnym stopniu powoduje spowolnienie fali elektromagnetycznej co z kolei tak samo jak zjawiska ugięcia czy odbicia fali na granicy różnych tkanek wpływa na fazę i amplitudę sygnału w odbiorniku oraz czas jego propagacji. Złożone zjawiska propagacyjne we wnętrzu ludzkiego ciała sprawiają, że wiele technik lokalizacyjnych bazujących na analizie sygnału radiowego, które z powodzeniem są wykorzystywane w otwartej przestrzeni, nie mogą być w prosty sposób zaimplementowane w rozważanym przypadku.

Dotychczas opracowano wiele metod lokalizacji endoskopów kapsułkowych. Można wyróżnić pośród nich metody oparte na analizie pola elektromagnetycznego [1,2,5] lub przetwarzaniu obrazu, które mogą być uzupełnione o informację o orientacji kapsułki na podstawie informacji z czujników inercyjnych [9]. Istnieją metody pozwalające określić położenie endoskopu wzdłuż przewodu pokarmowego oraz takie, które pozwalają wyznaczyć położenie w trójwymiarowym układzie współrzędnych. Pierwsza grupa metod opiera się na śledzeniu przemieszczeń i orientacji kapsułki głównie z wykorzystaniem technik wizyjnych. W przypadku drugiej grupy najczęściej stosuje się metody oparte na analizie pola elektromagnetycznego. Rozwiązania wykorzystujące informacje o rozkładzie natężenia pola magnetycznego dają wyniki dla, których błąd nie przekracza 20 mm [1,5], co jest wielkością porównywalną z rozmiarem endoskopu. Inne metody wykorzystują algorytmy oparte na analizie parametrów sygnału odebranego, takich jak poziom mocy odbieranego sygnału, czas propagacji sygnału czy różnica faz odbieranego sygnału[2].

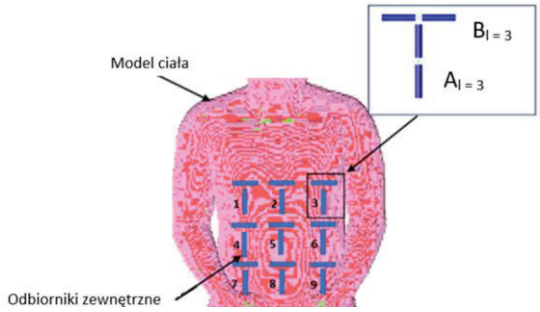
Opisane w literaturze metody mają pewne ograniczenia, które znacząco wpływają na dokładność pozycjonowania. W przypadku algorytmów opartych na analizie pola elektromagnetycznego, na parametry odbieranego

sygnału wpływa wielodrogowa propagacja wewnątrz ciała człowieka. Wykorzystanie dokładnych modeli ciała wykonanych np. za pomocą skanera NMR pozwala znacznie poprawić dokładność lokalizacji, jest to jednak rozwiązanie kosztowne i czasochłonne. W przypadku metod opartych na analizie obrazu lub śledzeniu ruchu z wykorzystaniem czujników inercyjnych, uzyskane wyniki mogą być obarczone błędem wynikającym z ruchu pacjenta lub zmiennej prędkości przemieszczania się kapsułki w poszczególnych odcinkach przewodu pokarmowego.

2. Algorytm lokalziacji

Celem badań było wykorzystanie algorytmu opartego na detekcji fazy do lokalizacji nadajników znajdujących się w ciele człowieka. Przedstawiona technika lokalizacji wykorzystuje pomiar różnicy fazy sygnałów odebranych od endoskopu kapsułkowego. Metodę tę podobnie jak w przypadku innych opracowanych dotychczas rozwiązań zaimplementowano z wykorzystaniem algorytmu Gauss-Newtona tak, aby możliwe było wyznaczenie współrzędnych geometryczne kapsułki w trójwymiarowym układzie współrzędnych [2]. Ponadto ze względu na złożony charakter środowiska propagacyjnego jakim jest ciało człowieka, zaimplementowano algorytm automatycznego doboru współczynnika skrócenia fali, który ma na celu poprawę dokładności lokalizacji.

W zaproponowanym podejściu współrzędne geometryczne kapsułki w przestrzeni trójwymiarowej wyznaczane są z wykorzystaniem metody PDoA (ang. Phase Difference of Arrival). Wykorzystywana jest w tym celu różnica fazy sygnałów o częstotliwościach f_1 i f_2 odebranych w poszczególnych odbiornikach zewnętrznych. Zaproponowane ułożenie odbiorników zewnętrznych przedstawiono na rysunku 1.



Rys. 1. Ułożenie anten zewnętrznych względem ciała człowieka[7].

Algorytm wyznacza położenie kapsułki na podstawie informacji o fazie sygnału z N=5 odbiorników. Wybierane są zatem te odbiorniki, dla których poziom odbieranej mocy jest najwyższy[7]. Ponadto w jednym kroku algorytmu nie mogą być wykorzystana jednocześnie sygnał odebrany dla anten A i B o tym samym indeksie. W kolejnym kroku z równania 1 na podstawie zmierzonej różnicy faz wyznaczane są odległości od kapsułki endoskopowej. Dla algorytmu wykorzystującego stałą wartość współczynnika skrócenia fali przyjęto względną przenikalność elektryczną ε_r = 52 oraz przewodność elektryczną właściwą σ = 1,8 S/m [4]. Jest to również wartość początkowa dal algorytmu z adaptacyjnym doborem współczynnika skrócenia fali.

$$d_n = \frac{v_d}{2\pi}\frac{\Delta\varphi_n}{\Delta f} = \frac{v_d}{2\pi}\frac{(\varphi_2-\varphi_1)}{(f_2-f_1)} \quad (1)$$

d_n [m]- odległość kapsułki od N-tej anteny umieszczonej na ciele człowieka,

v_d [m/s]- prędkość fali w modelu lub ciele,

ε_r - przenikalność dielektryczna modelu,

$\Delta\varphi_n$ [rad]- różnica fazy sygnału *f1* i *f2* w odbiorniku,

Δf [Hz]- różnica pomiędzy częstotliwością *f1* i *f2*.

Zastosowany model ciała charakteryzuje się taką samą wartość przenikalności elektrycznej w kierunku do każdej anteny odbiorczej, jednak wartość ta może być różna dla kolejnych pozycji kapsułki. W środowisku rzeczywistym uśrednione wartości przenikalności będą się różnić, dlatego wyznaczone w tym kroku odległości są jedynie przybliżone.

Następnie do wyznaczenia współrzędnych geometrycznych kapsułki wykorzystany jest algorytm Gaussa-Newtona. Celem algorytmu jest znalezienie takich współrzędnych, które pozwolą zminimalizować funkcję *SE* (2).

$$SE(p) = \frac{1}{2}\sum_{n=1}^{N} r_n^2 \quad (2)$$

N - liczba anten odbiorczych umieszczonych na ciele,

$$r_n = d_n - \sqrt{(x_k-u_n)^2+(y_k-v_n)^2+(z_k-w_n)^2}$$

p – współrzędne kapsułki x_k,y_k,z_k,

x_k, y_k, z_k - współrzędne kapsułki w kolejnych k-tych iteracjach algorytmu GN określające jej położenie p,

u_n, v_n, w_n współrzędne kolejnych anten umieszczonych na ciele.

Dla każdej iteracji algorytmu Gaussa-Newtona wyznaczana jest wartość funkcji err_k, opisana równaniem 3, która określa różnicę pomiędzy pozycjami kapsułki wyznaczonymi w dwóch ostatnich iteracjach algorytmu. Funkcja ta definiuje warunek zbieżności metody *GN*.

$$err_k = \sqrt{(x_k-x_{k-1})^2+(y_k-y_{k-1})^2+(z_k-z_{k-1})^2} \quad (3)$$

err_k – błąd k-tej iteracji algorytmu,

x_k, y_k, z_k – współrzędne kapsułki wyznaczone w k-tej iteracji, gdzie $0 < k < 50$.

Jeżeli wartość err_k znajdzie się w określonym przedziale tolerancji wówczas algorytm kończy swoje działanie. W przypadku niniejszej pracy przyjęto, że wartość ta będzie ułamkiem wielkości anteny kapsułkowej i została opisana zależnością 4.

$$err_k(\varepsilon_r, \Delta\varphi) < 0.04 * H \quad (4)$$

H - wysokość anteny kapsułkowej użytej w symulacji, w przypadku anteny meandrycznej H = 23 mm[6].

Warunek zbieżności wynikający z zależności 4 nie zawsze jest spełniony w przypadku rzeczywistego ciała człowieka jeżeli w metodzie przyjęta zostanie stała wartość przenikalności elektrycznej tkanek. Z tego względu

w algorytmie zastosowano metodę adaptacyjnego doboru przenikalności elektrycznej dla każdej pozycji kapsułki w celu osiągnięcia kryterium zbieżności i tym samym minimalizacji błędu lokalizacji. Metoda polega na znalezieniu w każdym punkcie, w którym wyznaczone jest położenie, takich wartości przenikalności modelu, dla których funkcja err_k osiąga minimum. Ponadto w jednorodnym środowisku propagacyjnym wartość funkcji *SE* opisanej równaniem 3 będzie bliska zeru ze względu na dokładniejszą estymację odległości. W przypadku, gdy środowisko propagacyjne jest bardziej złożone to faza odbieranego sygnału może zmieniać się nieliniowo wraz z odległością. W konsekwencji, wyznaczone przez algorytm odległości do kolejnych anten odbiorczych dadzą niejednoznaczną estymatę położenia. W związku z tym poszukiwana jest taka wartość ε_r która minimalizuje funkcję *SE* dla zakresu wartości przenikalności elektrycznej modelu, dla którego algorytm GN osiągnął zbieżność w poprzednim kroku co opisano zależnością 5[8].

$$SE_{min} = \min_{\varepsilon_r \epsilon <1,60>} |S(\varepsilon_r, \Delta\varphi)| \quad (5)$$

Minimalizacja funkcji *SE* nie będzie możliwa gdy zmierzona wartość różnicy fazy w jednym lub w kilku odbiornikach obarczona jest znacznym błędem. Estymata bazująca na takim pomiarze nie powinna być uznana za wiarygodną. Z tego powodu oszacowane położenie kapsułki endoskopowej poddawane jest klasyfikacji na podstawie wartości funkcji SE. Przyjęto, że jej wartość musi spełniać warunek (6).

$$SE\left(\varepsilon_{r,}, \Delta\varphi\right) \leq N * o_n^2 \quad (6)$$

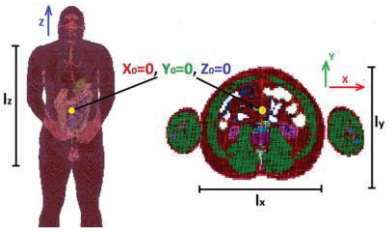
N — liczba anten odbiorczych;

o_n - stała wartość: 30 mm dla symulacji i 150 mm w przypadku pomiarów z wykorzystaniem modelu rzeczywistego.

3. WERYFIKACJA METODY

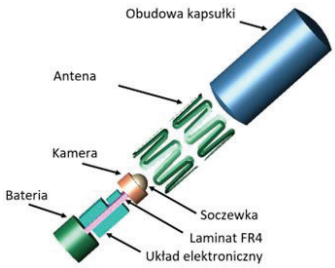
Pierwszy etap weryfikacji proponowanego algorytmu bazuje na komputerowych symulacjach przesunięcia fazowego sygnału pomiędzy anteną nadawczą endoskopu, a antenami odbiorczymi umieszczonymi na ciele. W tym celu wykorzystano model heterogeniczny NMR Hershey dostępny w środowisku symulacyjnym XFdtd Remcom, który zachowuje zarówno kształt ciała, jak i parametry elektryczne poszczególnych tkanek organizmu.

Na rysunku 6 przedstawiono model ciała człowieka, jego przekrój poprzeczny oraz kolorem żółtym zaznaczono punkt będący początkiem przyjętego układu współrzędnych. Wymiary modelu zaznaczone na rysunku wynoszą odpowiednio: *lx* = 330 mm, *ly* = 260 mm, *lz* = 650 mm. Uwzględniając wymiary modelu i układ współrzędnych przedstawiony na rysunku 6, podczas symulacji zmiano zarówno orientację kapsułki oraz jej położenie w zakresie: -100 mm < x < 100 mm, -95mm < y < 95 mm, -190 mm < z < 190 mm. Symulacje przeprowadzono dla częstotliwości 401 MHz i 406 MHz zgodnie z pasmem MICS dedykowanym dla endoskopów kapsułkowych [3].



Rys. 2. Model ciała człowieka wykorzystany w symulacjach oraz jego przekrój poprzeczny wykonany w płaszczyźnie XY dla z=z0[7]

Podczas prowadzonych badań algorytm lokalizacji zweryfikowano min. z wykorzystaniem przedstawionej na rysunku 3 autorskiej anteny meandrycznej, która zachowuje dobre dopasowanie impedancyjne w otoczeniu różnych tkanek w pasmie MICS 400 MHz[6]. Poprzez zastosowanie struktury meandrycznej uzyskano poprawę bilansu mocy łącza bezprzewodowego typu in-body. Jest to szczególnie istotne w przypadku endoskopów wykorzystujących miniaturowe źródło zasilania. W przypadku odbiorników zewnętrznych zastosowano anteny dipolowe.



Rys. 3. model kapsułki endoskopowej wykorzystany w symulacjach wraz z anteną [6]

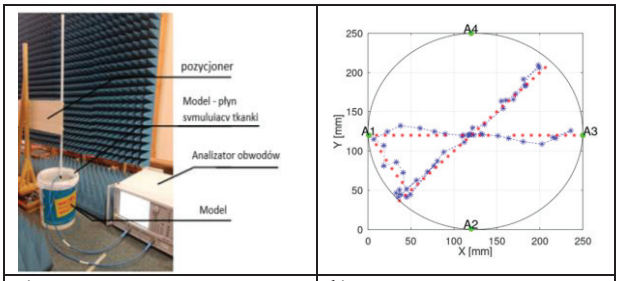
W tabeli 1 przedstawiono wartość błędu lokalizacji dla różnych wersji algorytmu. Porównano algorytm, który wykorzystuje stałą względną przenikalność modelu jak oraz różne warianty ułożenia i doboru odbiorników zewnętrznych. Wyniki opracowano dla około 160 położeń kapsułki w różnych częściach przewodu pokarmowego heterogenicznego modelu ciała człowieka. Dla ostatecznej wersji algorytmu średni błąd lokalizacji wyniósł 40 mm. Jest to wartość mniejsza o 31 mm w stosunku do algorytmu wykorzystującego stałą przenikalność modelu.

Tab. 1. Zestawieni błędu lokalizacji dla kolejnych wersji algorytmu lokalizacji.

Metoda	błąd [mm]
Stała względna przenikalność modelu ε_r = 52, odbiorniki 4,5,6 i 8.	71
Model adaptacyjny z minimalizacją funkcji err_k, odbiorniki 2,4,5,6 i 8.	56
Model adaptacyjny z minimalizacją funkcji *SE i* err_k odbiorniki 2,4,5,6 i 8.	50
Minimalizacja funkcji *SE i* err_k wraz z dynamicznym doborem odbiorników A i B	40

Eksperymenty z wykorzystaniem prototypowych nadajników mogą być niebezpieczne dla ludzi, a ponadto

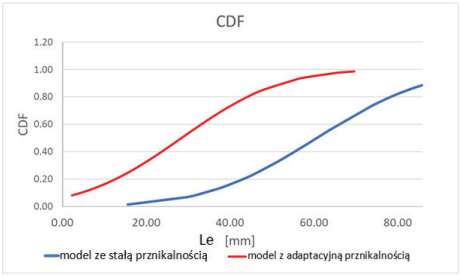
wymagają zgody komisji bioetycznej, co nie jest możliwe na początkowym etapie badań. Z tego względu algorytm zweryfikowano na podstawie pomiarów z wykorzystaniem modelu ciała wykonanego z cienkościennego naczynia z tworzywa sztucznego wypełnionego płynem imitującym tkanki. Zastosowano płyn o parametrach elektrycznych zbliżonych do średnich parametrów tkanek organizmu człowieka. W tym celu przygotowano roztwór wody, cukru i soli w następujących proporcjach: 1 l wody, 350 g cukru i 15 g soli. Proporcje dobrano eksperymentalnie otrzymując płyn przenikalności $\varepsilon_r = 57$ i przewodność $\sigma = 0{,}76$ S/m. Naczynie miało zbliżone wymiary do walca o średnicy 250 mm i wysokości 250 mm. Rysunek 4a przedstawia wykorzystany układ pomiarowy, w którego skład wchodzi pozycjoner liniowy, analizator obwodów i model ciała.



a) b)

Rys. 4. a)Układ pomiarowy wykorzystany do eksperymentalnej weryfikacji algorytmu [8], b) scenariusz eksperymentalnej weryfikacji algorytmu lokalizacji [8].

Na rysunku 4b przedstawiono scenariusz weryfikacji algorytmu w przypadku dwuwymiarowym. Ze względu na zastosowanie pozycjonera liniowego kapsułka poruszała się po 2 liniach prostych. Kolorem czerwonym zaznaczono rzeczywistą ścieżkę po jakiej poruszała się kapsułka, natomiast kolorem niebieskim jej estymowane położenie. Średni błąd lokalizacji dla przyjętego scenariusza wyniósł około 30 mm co jest wartością zbliżoną do rozmiaru kapsułki. W stosunku do algorytmu ze stałym współczynnikiem skrócenia fali uzyskano poprawę o ponad 30 mm co pokazano na rysunku 5.



Rys. 5. Dystrybuanta rozkładu prawdopodobieństwa błędu lokalizacji [8].

1. WNIOSKI

Wykazano, że zastosowanie iteracyjnego algorytmu doboru parametrów przenikalności elektrycznej ośrodka umożliwia poprawę dokładności metody lokalizacji endoskopu kapsułkowego. W porównaniu ze znaną z literatury metodą zakładająca stałą wartość przenikalności elektrycznej ośrodka uzyskano w symulacjach komputerowych poprawę dokładności o około 20 mm. Ponadto wyznaczenie położenia endoskopu było możliwe dla 95 % przypadków. Dla porównania dla algorytmu ze stałą przenikalnością było to około 70%.

Dokładność proponowanej techniki na poziomie około 4 cm wydaje się być wystarczająca, aby przyspieszyć analizę zarejestrowanego materiału obrazowego. Umożliwi to również skorelowanie wyników diagnostyki obrazowej uzyskane różnymi metodami w celu postawienia pełnej diagnozy. Ponadto zaletą zaprezentowanej metody jest to, że w jej przypadku budowa wewnętrzna ciała pacjenta i parametry elektryczne tkanek nie musza być znane przed przeprowadzeniem badania. Pozwala to skrócić czas potrzeby na przygotowanie badania ponieważ nie ma potrzeby opracowywania dedykowanych modeli ciała.

LITERATURA

[1] Cavlu V. and Brennan P., "Determining the Position and Orientation of In-body Medical Instruments Using Near-Field Magnetic Field Mapping," in IEEE Journal of Electromagnetics, RF and Microwaves in Medicine and Biology.doi: 10.1109/JERM.2019.2914402.

[2] Chandra R., Johansson A. J., and Tufvesson F., 'Localization of an RF source inside the Human body for Wireless Capsule Endoscopy', in *Proceedings of the 8th International Conference on Body Area Networks*, Boston, United States: ACM, 2013. doi: 10.4108/icst.bodynets.2013.253713.

[3] ITU-R RS. 1346, Sharing Between the Meteorological Aids Service and Medical Implant Communication Systems (MICS) Operating in the Mobile Service in the Frequency Band 401-406 MHz, 1998.

[4] Januszkiewicz Łukasz, Hausman Sławomir „Analiza wrażliwości jednorodnego modelu ciała człowieka na parametry płynu symulującego tkanki.", Przegląd Elektroniczny R.90 Nr 12/2014 Str. 214-215.

[5] Kim M.-C., Kim E.-S., Park J.-O., Choi E., and Kim C.-S., 'Robotic Localization Based on Planar Cable Robot and Hall Sensor Array Applied to Magnetic Capsule Endoscope', Sensors, vol. 20, no. 20, p. 5728, Oct. 2020, doi: 10.3390/s20205728.

[6] Oleksy Paweł., Januszkiewicz Łukasz., „Antena endoskopu kapsułkowego przeznaczona do pracy w pasmie MICS 400 MHz", Przegląd Telekomunikacyjny - Wiadomości Telekomunikacyjne, nr 4, 2023, str. 305-308

[7] Oleksy P., Januszkiewicz Ł., „Lokalizacja bezprzewodowych endoskopów kapsułkowych z wykorzystaniem algorytmu wyboru odbiorników", Przegląd Telekomunikacyjny - Wiadomości Telekomunikacyjne, nr 4, 2022, str. 624-627.

[8] Oleksy P. and Januszkiewicz Ł. , 'Wireless Capsule Endoscope Localization with Phase Detection Algorithm and Adaptive Body Model', Sensors, vol. 22, no. 6, p. 2200, Mar. 2022.

[9] Vedaei S. S. and Wahid K. A., 'A localization method for wireless capsule endoscopy using side wall cameras and IMU sensor', Sci Rep, vol. 11, no. 1, p. 11204, May 2021, doi: 10.1038/s41598-021-90523-w.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

ANALIZA KOSZTOWA PASYWNEJ OPTYCZNEJ SIECI XHAUL Z AGREGACJĄ RUCHU W WARSTWIE OPTYCZNEJ

COST ANALYSIS OF PASSIVE OPTICAL XHAUL NETWORK WITH OPTICAL-LAYER TRAFFIC AGGREGATION

Mirosław Klinkowski[1];

[1] Instytut Łączności - Państwowy Instytut Badawczy, ul. Swojczycka 38, 51-501 Wrocław, M.Klinkowski@il-pib.pl

DOI: 10.15199/59.2024.4.19

Streszczenie: Artykuł prezentuje wyniki analizy kosztowej optycznej sieci Xhaul wykorzystującej pasywną technologię zwielokrotniania długości fali (WDM) i agregującej ruch w warstwie optycznej przy pomocy optycznych multiplekserów add/drop (OADM). Badanie dotyczy sieci miejskiej i uwzględnia koszt elementów sieciowych oraz dzierżawy światłowodu jak i ograniczenia w zasięgu transmisji. Wyniki wskazują na opłacalność rozwiązania przy niskim i średnim obciążeniu ruchem, szczególnie dla systemów WDM obsługujących większą liczbę kanałów.

Abstract: The article presents the results of a cost analysis of an optical Xhaul network using passive wavelength division multiplexing (WDM) technology and enabling traffic aggregation in the optical layer with optical add/drop multiplexers (OADM). The study focuses on an urban network and considers the cost of network elements, the leasing of optical fibers, and transmission range limitations. The results indicate the solution's cost-effectiveness in low and medium-traffic load scenarios, especially for WDM systems supporting a larger number of channels.

Słowa kluczowe: pasywne sieci WDM, radiowe sieci dostępowe, sieci optyczne, Xhaul.

Keywords: optical networks, passive WDM networks, radio access networks, Xhaul.

1. WSTĘP

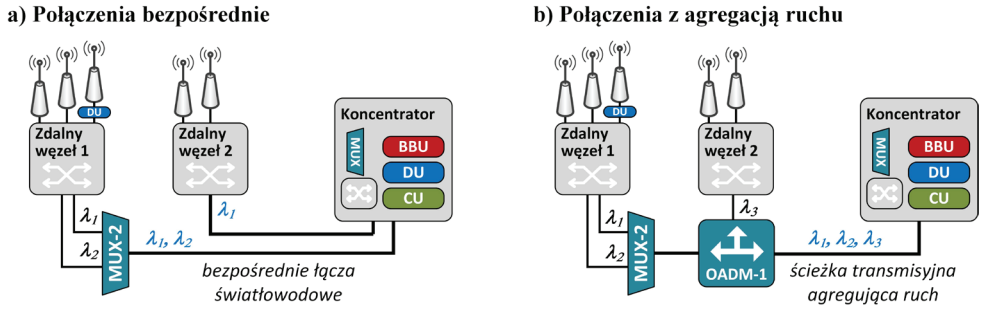
Sieci Xhaul stanowią infrastrukturę transportową łączącą elementy radiowych sieci dostępowych (RAN), takich jak jednostki radiowe (RU) rozproszone na pewnym obszarze, z zasobami obliczeniowymi "chmury radiowej" (jednostki rozproszone i centralne, DU/CU), zazwyczaj zlokalizowanymi w koncentratorze (hubie), oraz z siecią rdzeniową (CN). Xhaul obejmuje segmenty sieci fronthaul, midhaul i backhaul i jest obecny w sieciach mobilnych piątej generacji (5G) [9], a także przewidziany w sieciach 6G.

Sieci światłowodowe oparte na technologii WDM w połączeniu z transmisją Ethernet wykorzystującą komutację pakietów stanowią skalowalne i elastyczne rozwiązanie dla implementacji sieci Xhaul [6, 10]. WDM umożliwia agregację wielu połączeń o dużej przepustowości, realizowanych w różnych kanałach optycznych, w ramach pojedynczego łącza światłowodowego. W szczególności, pasywne systemy WDM są efektywne kosztowo dzięki uproszczonej eksploatacji nie wymagającej wzmacniania sygnału i kompensacji dyspersji, a jednocześnie dobrze przystosowane do sieci Xhaul, które charakteryzują się stosunkowo niewielkim zasięgiem transmisji [10]. Technologia Ethernet wzbogacona o mechanizmy sieciowe wrażliwe na czas (TSN), służące obsłudze przepływów danych wrażliwych na opóźnienia [6], zwiększa efektywność łączy optycznych poprzez multipleksację statystyczną pakietów przenoszących dane radiowe. Przykładowym rozwiązaniem komercyjnym jest system flexiHaul obejmujący przełącznik TSN oraz pasywne komponenty WDM [5].

Technologie WDM i Ethernet TSN zmniejszają zapotrzebowanie na bezpośrednie połączenia światłowodowe między zdalnymi lokalizacjami a koncentratorem, obniżając koszt wdrożenia sieci Xhaul. Dalsze ograniczenie zapotrzebowania na zasoby światłowodowe można osiągnąć poprzez zastosowanie pasywnych urządzeń OADM. OADM pozwala na agregację kanałów optycznych z wielu zdalnych lokalizacji w jednej optycznej ścieżce transmisyjnej (OST), zamiast używania kilku bezpośrednich połączeń światłowodowych między każdą zdalną lokalizacją a koncentratorem [7, 8]. Zastosowanie modułów OADM wprowadza jednak dodatkowe koszty i w przypadku niewielkich sieci może nie kompensować oszczędności wynikających z redukcji zapotrzebowania na łącza światłowodowe. Także, moduły OADM wprowadzają dodatkowe tłumienie zmniejszające zasięg transmisji, co ogranicza możliwość ich zastosowania na dłuższych ścieżkach transmisyjnych, które dają większe szanse na agregację ruchu ze zdalnych lokalizacji. Stąd też konieczna jest ocena efektywności i korzyści wynikających z zastosowania OADM do agregacji ruchu w warstwie optycznej sieci Xhaul.

Głównym celem pracy jest analiza korzyści kosztowych scenariusza pasywnej optycznej sieci Xhaul z agregacją ruchu w warstwie optycznej. W tym celu w badaniach wykorzystano realistyczną topologię sieci miejskiej z uwzględnieniem rzeczywistych kosztów elementów sieciowych (multiplekserów optycznych i OADM) i kosztu dzierżawy włókna światłowodowego oraz wartości parametrów transmisyjnych odpowiadające rzeczywistym systemom. Wyniki wskazują między innymi, które scenariusze są korzystne oraz jaki jest wpływ poszczególnych elementów kosztowych na zysk wynikający z agregacji ruchu.

Artykuł ma nastepującą strukturę. Sekcja 2 przedstawia model analizowanej sieci wraz z założeniami. Sekcja 3 prezentuje wyniki analizy kosztowej wraz z powiązaną dyskusją. Sekcja 4 stanowi podsumowanie pracy.



Rysunek 1: Pasywna optyczna sieć Xhaul łącząca zdalne węzły radiowe z koncentratorem: (a) z bezpośrednimi połączeniami światłowodowymi, (b) agregująca ruch ze zdalnych lokalizacji z użyciem OADM.

2. MODEL SIECI

Optyczna sieć Xhaul oparta na pasywnej transmisji WDM jest modelowana jako graf $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, gdzie $\mathcal{V}$ to zbiór węzłów, a $\mathcal{E}$ to zbiór łączy. Zdalne węzły za pomocą przełączników Ethernet agregują ruch radiowy, który przesyłany jest w kanałach optycznych WDM w kierunku koncentratora, gdzie znajdują się funkcje przetwarzania radiowego (DU/CU), jak pokazano na Rys.1. Zbiór $\mathcal{D}$ obejmuje zapotrzebowania ruchowe, gdzie każdy element jest powiązany z określonym zdalnym węzłem, a $W(d)$ to liczba żądanych kanałów WDM dla zapotrzebowania $d \in \mathcal{D}$. Łącza sieciowe posiadają wystarczającą ilość zasobów światłowodowych do realizacji połączeń, gdzie pojedynczy światłowód obsługuje W kanałów. Każde połączenie wykorzystuje dwa włókna, w kierunku od i do koncentratora. Zbiory $\mathcal{M}$ i $\mathcal{O}$ reprezentują dostępne moduły MUX i OADM.

Zbiór $\mathcal{P}(d)$ zawiera kandydujące ścieżki routingu dla zapotrzebowania $d \in \mathcal{D}$, gdzie ścieżka $p \in \mathcal{P}(d)$ jest trasą światłowodową po łączach sieciowych ($p \subseteq \mathcal{E}$) łącząca zdalny węzeł zapotrzebowania d z koncentratorem. Połączenia są realizowane za pomocą optycznych ścieżek transmisyjnych (OST) prowadzonych wzdłuż ścieżek kandydujących, z których każda obsługuje kanały WDM z jednego lub wielu zdalnych węzłów. Wszystkie kanały danego zapotrzebowania są obsługiwane przez tą samą OST. OST może być albo *bezpośrednia*, realizowana jako dedykowane połączenie światłowodowe między zdalnym węzłem a koncentratorem, jak pokazano na Rys. 1(a), albo *agregująca*, gdzie OST przechodzi przez kilka węzłów pośrednich agregując kanały z tych węzłów za pomocą modułów OADM (jak dla zdalnego węzła 2 na Rys. 1(b)). W pierwszym przypadku OST obsługuje tylko jeden zdalny węzeł, podczas gdy w drugim przypadku ruch z wielu lokalizacji jest obsługiwany przez pojedynczą OST zmniejszając zapotrzebowanie na zasoby światłowodowe.

Zgodnie z [4], zasięg transmisji dla pasywnego systemu transmisji optycznej (L) jest ograniczony budżetem mocy optycznej (P) i zależy od tłumienia elementów optycznych na ścieżce:

$$L = \Big(P - \big(\sum_{m\in\mathcal{M}} N_{\mathrm{M}}(m)A_{\mathrm{M}}(m) + \sum_{o\in\mathcal{O}} N_{\mathrm{O}}(o)A_{\mathrm{O}}(o) + 2A_{\mathrm{C}} + A_{\mathrm{MM}}\big)\Big)\alpha^{-1}. \quad (1)$$

gdzie $N_{\mathrm{M}}(m)$ i $A_{\mathrm{M}}(m)$ to odpowiednio liczba MUX i tłumienie MUX typu $m \in \mathcal{M}$, $N_{\mathrm{O}}(o)$ i $A_{\mathrm{O}}(o)$ to odpowiednio liczba OADM i tłumienie OADM typu $o \in \mathcal{O}$, $2A_{\mathrm{C}}$ reprezentuje tłumienie złączy końcowych, A_{MM} to margines konserwacji, a α to współczynnik strat światłowodu.

Model (1) zakłada, że tłumienie modułów MUX i OADM reprezentowane przez $A_{\mathrm{M}}(m)$ i $A_{\mathrm{O}}(o)$ obejmuje tłumienie ich złączy. Ze względu na krótkie założone odległości transmisji i transmisję w oparciu o system flexiHaul xWave 400G (pracujący w drugim oknie transmisyjnym), inne czynniki wpływające na transmisję (takie jak dyspersja) nie są brane pod uwagę. Czynniki te powinny być analizowane w przypadku systemów WDM pracujących w trzecim oknie transmisyjnym (jak CWDM i DWDM).

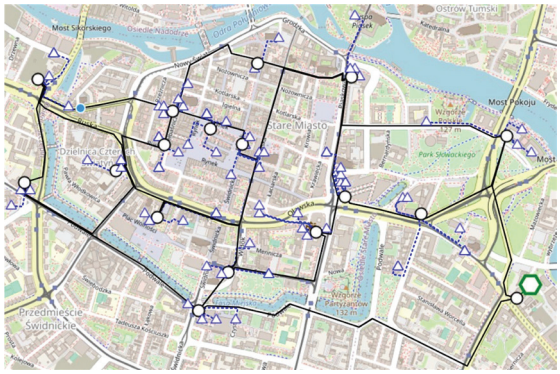
Rozważany w badaniu model kosztowy pasywnej optycznej sieci Xhaul obejmuje trzy elementy: (a) koszt dzierżawy włókien światłowodowych do realizacji ścieżek transmisyjnych; (b) koszt MUX-ów używanych do agregacji kanałów i filtrowania sygnału na końcach ścieżek (jak pokazano na Rys. 1); (c) koszt OADM-ów stosowanych do agregacji kanałów w węzłach pośrednich ścieżek transmisyjnych (Rys. 1(b)). Roczny koszt dzierżawy kilometra ciemnego włókna oznaczony jest przez κ^{F}. Przyjęto rosnący liniowo z liczbą obsługiwanych kanałów koszt urządzeń MUX i OADM, gdzie koszt pojedynczego kanału to κ^{M} i $2\kappa^{\mathrm{M}}$ odpowiednio dla MUX i OADM [7]. Całkowity roczny koszt dzierżawy łączy wraz z kosztem wdrożonych modułów MUX/OADM (z) jest wyrażony jako:

$$z = \kappa^{\mathrm{F}} D^{\mathrm{F}} + \kappa^{\mathrm{M}} \sum_{m\in\mathcal{M}} N_{\mathrm{M}}(m)W_{\mathrm{M}}(m) + 2\kappa^{\mathrm{M}} \sum_{o\in\mathcal{O}} N_{\mathrm{O}}(o)W_{\mathrm{O}}(o), \quad (2)$$

gdzie D^{F} to całkowita długość zestawionych ścieżek optycznych, $W_{\mathrm{M}}(m)$ to liczba kanałów multipleksowanych przez MUX typu m i $W_{\mathrm{O}}(o)$ to liczba kanałów wydzielanych/wstawianych przez OADM typu o.

3. WYNIKI ANALIZY

Analizę kosztową przeprowadzono dla syntetycznej sieci WRO-17, przedstawionej na Rys. 2. W sieci WRO-17 zdalne lokalizacje (okręgi) są umieszczone w pobliżu rzeczywistych lokalizacji anten (trójkąty) w centrum Wrocławia, a łącza pomiędzy węzłami i koncentratorem (sześciokąt) prowadzone są wzdłuż ulic, odzwierciedlając rzeczywiste odległości. Ścieżki kandydujące zostały wygenerowane za pomocą algorytmu k-najkrótszych ścieżek, przy założeniu limitu 10 km w celu ograniczenia opóźnień dla usług 5G. Rozważano transmisję optyczną z wykorzystaniem czterech i ośmiu kanałów WDM, $W \in \{4, 8\}$, jak w systemie flexiHaul xWave 400G [5]. Przyjęte wartości parametrów systemu transmisyjnego są podobne jak w [7] i przedstawione w Tabeli 1.



Rysunek 2: Topologia sieci WRO-17.

Tabela 1: Parametry systemu transmisyjnego.

Parametr	*Wartość*
Kanały WDM (W)	$\{4, 8\}$
Współ. strat świątł. (α) [dB/km]	0.5
Tłumienność złącza (A_{C}) [dB]	0.3
Margines konserwacji (A_{MM}) [dB]	2.0
Budżet mocy (P) [dBm]	17.0

Tabela 2 prezentuje rozważane moduły MUX i OADM wraz z liczbą obsługiwanych kanałów (multipleksowanych w przypadku MUX i wstawianych/wydzielanych w przypadku OADM) oraz ich tłumiennością "pass-through" (wpływającą na straty ścieżki), bez uwzględnienia tłumienności złączy. Przedstawione wartości są typowe dla produktów komercyjnych [1, 2].

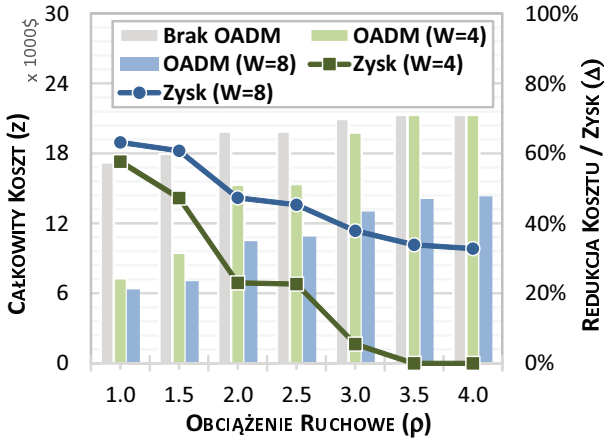
Tabela 2: Tłumienie OADM i MUX (bez złączy).

	Liczba kanałów				
	1	2	3	4	8
OADM [dB]	0.8	1.4	2.1	2.8	-
MUX [dB]	-	0.9	–	1.2	1.6

Liczba żądanych kanałów WDM przez każdą zdalną lokalizację była losowa, pomiędzy $W^{\min}$ a $W^{\max}$, gdzie $1 \leq W^{\min} \leq W^{\max} \leq 4$, a obciążenie ruchem wyrażono jako średnią $\rho = (W^{\min} + W^{\max})/2$. Cena dzierżawy kilometra ciemnego włókna wynosiła $\kappa^{\mathrm{F}} = 250$ USD rocznie [3], a koszt pojedynczego kanału MUX to $\kappa^{\mathrm{M}} = 15$ USD [7].

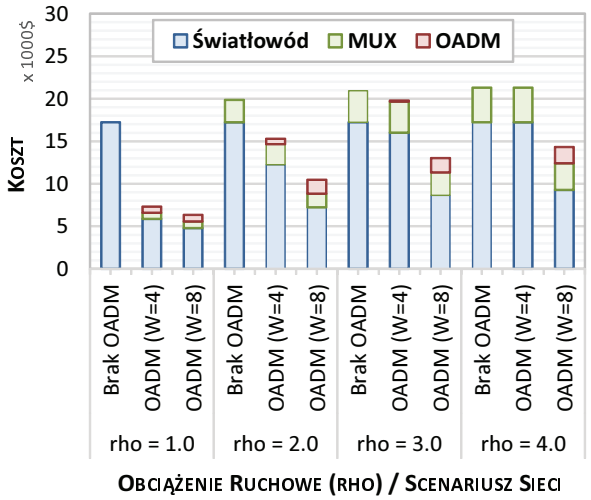
Wyniki analizy otrzymano z wykorzystaniem modelu programowania całkowitoliczbowego (MILP) przedstawionego w [8], służącego do optymalizacji wyboru ścieżek transmisyjnych wraz z doborem i rozmieszczeniem OADM. Badania przeprowadzono na stacji roboczej klasy Ryzen Threadripper 3.7 GHz (64 logicznych rdzeni, 128 GB RAM). Wszystkie wyniki były optymalne i uzyskane w czasie nie przekraczającym 16 sekund dla najbardziej wymagającego scenariusza.

Rys. 3 przedstawia koszt wdrożenia i rocznego użytkowania sieci optycznej (z, słupki) z połączeniami bezpośrednimi ("Brak OADM") i z agregacją ruchu ("OADM") oraz względną redukcję kosztu (zysk) uzyskaną dzięki zastosowaniu OADM (Δ, linie) w funkcji obciążenia ruchem (ρ), dla systemów WDM o różnej pojemności (W).



Rysunek 3: Koszt sieci bez i z agregacją ruchu oraz zysk wynikający z agregacji w funkcji obciążenia ruchowego i liczby kanałów WDM (po roku czasu).

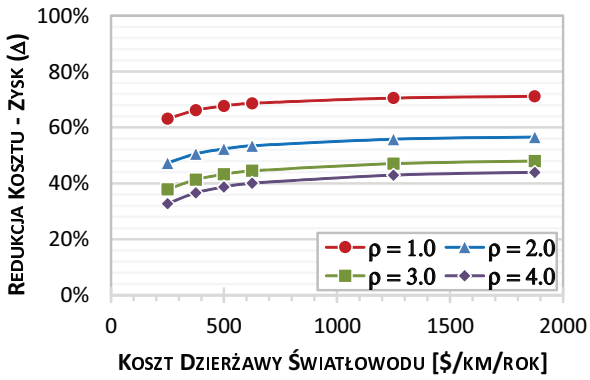
Rys. 3 pokazuje, że agregacja ruchu w warstwie optycznej pozwala na zmniejszenie kosztu sieci w porównaniu do sieci z połączeniami bezpośrednimi w scenariuszach niskiego i średniego obciążenia ruchem ($\rho \leq 3.0$) dla $W = 4$, oraz we wszystkich przypadkach dla $W = 8$. Zysk sięga około 60% dla $\rho = 1.0$ i ponad 30% dla $\rho = 4.0$ (przy zastosowaniu 8-kanałowego systemu WDM). W przypadku dużego obciążania ($\rho \geq 3.5$) obsługiwanego przez 4-kanałowy system WDM, zapotrzebowanie na kanały optyczne w każdym węźle jest zbliżone do pojemności WDM, co wymusza stosowanie bezpośrednich połączeń bez agregacji ruchu i wynikający z tego brak zysku.



Rysunek 4: Porównanie składowych kosztu dla różnych wymagań ruchowych i scenariuszy sieci.

Rys. 4 prezentuje szczegółowy udział poszczególnych składowych (dzierżawa światłowodu, moduły MUX i OADM) w całkowitym koszcie sieci, dla różnych scenariuszy jej realizacji i różnych wymaganiach ruchowych (ρ). Agregacja ruchu w warstwie optycznej wprowadza pewne koszty związane z wykorzystaniem OADM i MUX. Są one jednak niewielkie w porównaniu do znaczącego kosztu dzierżawy włókien światłowodowych, który jest dominujący w całkowitym koszcie wdrożenia i użytkowania sieci. Sieć bez OADM nie wymaga urządzeń MUX gdy $\rho = 1.0$.

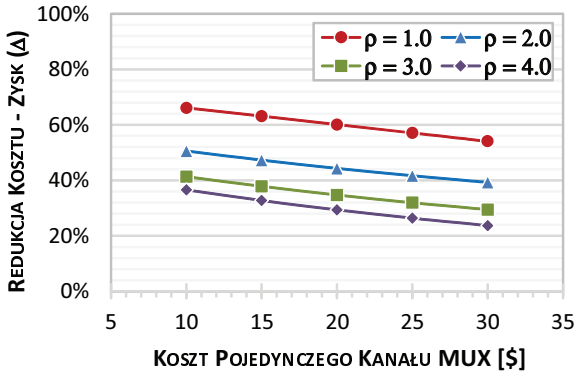
W takim przypadku zdalne lokalizacje mają zapotrzebowania na pojedyncze kanały, a zatem nie ma potrzeby ich multipleksacji na końcach połączeń światłowodowych.



Rysunek 5: Zysk (po roku użytkowania) wynikający z agregacji ruchu w funkcji kosztu dzierżawy światłowodu (κ^{F}) dla różnych wymagań ruchowych (ρ) dla systemu WDM obsługującego $W = 8$ kanałów, przyjmując koszt kanału MUX $\kappa^{\mathrm{M}} = 15$ USD.

Rys. 5 przedstawia wpływ kosztu dzierżawy włókna światłowodowego (κ^{F}) na zysk wynikający z agregacji ruchu ze zdalnych lokalizacji w warstwie optycznej (Δ) zakładając zastosowanie 8-kanałowego systemu WDM i koszt kanału MUX $\kappa^{\mathrm{M}} = 15$ USD. Zysk ten rośnie wraz z κ^{F} dla rozważanych wartości ρ, przy czym jest on szczególnie widoczny dla κ^{F} pomiędzy 250 i 700 USD. W tym zakresie wzrost Δ wynosi do około 8% procent, po czym stabilizuje się dla wyższych wartości κ^{F}.

Wyniki na Rys. 5 można też interpretować w kontekście zysku (Δ) po N latach eksploatacji sieci, gdzie dla $\kappa^{\mathrm{F}} = 250$ USD odpowiada wartość Δ po pierwszym roku eksploatacji, a w kolejnych latach zysk odczytywany jest dla $\kappa^{\mathrm{F}} = N \times 250$ USD.



Rysunek 6: Zysk (po roku użytkowania) wynikający z agregacji ruchu w funkcji kosztu MUX dla różnych wymagań ruchowych (ρ) dla systemu WDM obsługującego $W = 8$ kanałów, przyjmując koszt dzierżawy światłowodu $\kappa^{\mathrm{F}} = 250$ USD/km/rok.

Rys. 6 prezentuje wpływ kosztu urządzeń MUX/OADM, wyrażony przez koszt pojedynczego kanału κ^{M}, na zysk wynikający z agregacji ruchu w warstwie optycznej (Δ) zakładając 8-kanałowy system WDM i koszt dzierżawy włókna światłowodowego $\kappa^{\mathrm{F}} = 250$ USD/km/rok. Obserwowane wartości Δ maleją wraz z κ^{M} w sposób bliski liniowemu, z podobnym postępem bez względu na wartość ρ. W szczególności, nawet przy dwukrotnym zwiększeniu kosztu κ^{M}, z 15 USD do 30 USD, zysk Δ jest wysoki i wynosi od około 24% do 55% w zależności od wymagań ruchowych.

4. PODSUMOWANIE

Przeprowadzona analiza kosztowa wykazała, że zastosowanie pasywnych multiplekserów OADM w optycznej sieci Xhaul przynosi korzyści w scenariuszach o niskim i średnim obciążeniu ruchem, zwłaszcza dla systemów WDM obsługujących większą liczbę kanałów. W szczególności, możliwe jest znaczne obniżenie kosztów związanych z dzierżawą światłowodów poprzez efektywną agregację ruchu z wielu zdalnych lokalizacji. Dodatkowe koszty związane z OADM są kompensowane przez oszczędności wynikające z redukcji liczby potrzebnych bezpośrednich połączeń światłowodowych. Nasze przyszłe badania skupią się na opracowaniu skutecznych metod protekcji w optycznych sieciach Xhaul. Rozważane będą też technologie oparte na pasywnych sieciach optycznych (PON) takie jak NG-PON2 i "High Speed" PON.

PODZIĘKOWANIA

Praca powstała w wyniku realizacji projektu badawczego OPUS o nr 2018/31/B/ST7/03456 finansowanego ze środków Narodowego Centrum Nauki.

LITERATURA

[1] Auxora. „Simplex Bi-directional CWDM OADM Module data sheet". http://www.auxora.com/. (dostęp dnia 23.05.2024).

[2] Auxora. „Auxora 4Ch/8Ch Compact CWDM data sheet". http://www.auxora.com/. (dostęp dnia 23.05.2024).

[3] CTC Technology & Energy. 2012. „Dark Fiber Lease Considerations". https://www.ctcnet.us/DarkFiberLease.pdf. (dostęp dnia 23.05.2024).

[4] DeCusatis C. 2013. „Handbook of Fiber Optic Data Communication – A Practical Guide to Optical Networking". 4th edition. *Academic Press.*

[5] HFR Networks. 2021. „flexiHaul Packet xHaul Solution". https://hfrnet.com/. (dostęp dnia 23.05.2024).

[6] IEEE. 2018. „802.1CM-2018 – IEEE Standard for Local and Metropolitan Area Networks – Time-Sensitive Networking for Fronthaul".

[7] Klinkowski Mirosław, Jaworski Marek. 2023. „Dedicated Path Protection with Wavelength Aggregation in 5G Packet-Optical xHaul Access Networks". *IEEE J. Lightw. Technol.*, 41 (6) : 1591-1602.

[8] Klinkowski Mirosław, Jaworski Marek. 2023. „Cost-Aware Optimization of Optical Add-Drop Multiplexers Placement in Packet-Optical xHaul Access Networks". *Applied Sciences*, 13 (4862) : 1-16.

[9] O-RAN Alliance. 2021. „Xhaul Transport Requirements Tech. Spec. v1.0". https://www.o-ran.org/. (dostęp dnia 23.05.2024).

[10] O-RAN Alliance. 2024. „WDM-based Fronthaul Transport Tech. Spec. v4.0". https://www.o-ran.org/. (dostęp dnia 23.05.2024).

ATAKI NA PROTOKÓŁ PFCP SIECI SZKIELETOWEJ 5G ORAZ METODY ICH IDENTYFIKACJI

ATTACKS ON PFCP PROTOCOL IN 5G CORE NETWORK AND METHODS OF THEIR IDENTIFICATION."

Krzysztof Kosmowski[1]; Rafał Bryś[2]; Adam Dudko[3]

[1] Wojskowy Instytut Łączności – Państwowy Instytut Badawczy, Zegrze Południowe, Krzysztof.Kosmowski@wil.waw.pl
[2] Wojskowy Instytut Łączności – Państwowy Instytut Badawczy, Zegrze Południowe, Rafal.Brys@wil.waw.pl
[3] Wojskowy Instytut Łączności – Państwowy Instytut Badawczy, Zegrze Południowe, Adam.Dudko@wil.waw.pl

DOI: 10.15199/59.2024.4.20

Streszczenie**: Referat dotyczy możliwości przeprowadzenia ataków na sieć szkieletową 5G. Przedstawiono koncepcję ataków polegających na modyfikacji pakietów kontrolnych protokołu PFCP oraz ich realizację. Omówiono również możliwości wykrycia tego typu zagrożeń.**
Abstract**: The paper concerns the possibility of conducting attacks on the 5G Core Network. The concept of attacks involving modification of control packets of the PFCP protocol was presented, along with their implementation. In the subsequent part, the possibilities of detecting such threats were discussed.**

Słowa kluczowe: **Sieć Szkieletowa 5G; protokół PFCP; manipulacja pakietami;**
Keywords: **5G Core Network, PFCP protocol, packet manipulation.**

1. WSTĘP

Technologia 5G oferując nowe możliwości w zakresie realizacji usług wzbudza uzasadnione zainteresowanie na świecie. Jak zawsze w takiej sytuacji oprócz szans związanych z wdrażaniem nowych technologii wiąże się to również z nowymi zagrożeniami. Tym ostatnim zagadnieniem zajmuje się szereg zespołów badawczych w tym autorzy referatu, raportując potencjalne słabości wprowadzanych rozwiązań.

W literaturze przedmiotu raportowane są różnej klasy zagrożenia, które mogą wystąpić w bardziej lub mniej realnych scenariuszach. Generalnie rozpatrując scenariusz wielopoziomowego ataku cybernetycznego można wyróżnić kilka poziomów dostępu jakim dysponować powinien atakujący, żeby przeprowadzić poszczególne ataki. Najbardziej podstawowym warunkiem powodzenia jest posiadanie dostępu do sieci, w której następuje komunikacja między gNB a siecią szkieletową (CN). Na tym etapie można przyjąć, że atakujący uzyskał dostęp do kluczy TLS/SSL. Posiadając dostęp do sieci możliwe jest przeprowadzenie ataków z wykorzystaniem przechwyconych wiadomości przenoszonych np. za pomocą protokołu NGAP. Kolejny poziom dostępu zakłada, że w sieci jest „wystawiony" interfejs umożliwiający uruchomienie aplikacji zarządzającej siecią szkieletową. W takiej sytuacji możliwe jest przeprowadzenie ataku na aplikację chociażby typu *brute force* np. w celu zdobycia informacji o abonentach dla późniejszego wykorzystania w dalszych atakach. Zakładając, że atakujący uzyskał dostęp do serwerów, na których uruchomiona jest sieć szkieletowa możliwy jest jej *fingerprinting* za pomocą ustandaryzowanych API 3GPP. Atakujący może uzyskać dostęp do jednej lub kilku funkcji sieciowych, a po ich przejęciu wysyłać komunikaty, które spowodują brak możliwości realizacji usług poprzez atak DoS (*Denial of Service*) lub wyłączając niektóre funkcje sieciowe.

W dalszej części referatu przedstawiono wektory ataków na sieć szkieletową 5G zidentyfikowane w literaturze przedmiotu, wpisujące się w przedstawioną wyżej filozofię ataku. Następnie przedstawiono funkcje jakie w architekturze sieci szkieletowej 5G pełni protokół PFCP, który z kolei stanowi wektor ataku dla zaprezentowanych w dalszej części ataków na sieć szkieletową 5G wykonanych przez autorów referatu. W końcowej części wskazano na możliwości wykrywania symptomów tego typu ataków.

Zasadniczy wkład stanowi przedstawienie koncepcji ataków, ich realizacji oraz dyskusja nad możliwościami wykrycia tego typu zagrożeń[1].

2. PRZEGLĄD LITERATURY

W [1] przedstawiono badanie podatności protokołu TLS w implementacjach open-source sieci 5G. Badanie skupia się na analizie podatności TLS i zgodności z wymaganiami 3GPP w trzech różnych otwartych sieciach szkieletowych 5G: free5GC, Open5GS oraz OAI 5G CN. Analiza została przeprowadzona za pomocą narzędzi automatycznego skanowania i wykazała istnienie słabości w badanych implementacjach. Często wykorzystanie podatności na poziomie TLS jest koniecznym krokiem wstępnym do przeprowadzenia dalszych, wieloetapo-

[1] Projekt realizowany w konkursie CYBERSECIDENT CYBERSECIDENT/487845/IV/NCBR/2021. Zaawansowane metody i techniki wykrywania i przeciwdziałania atakom na infrastrukturę dostępową i aplikacje sieci 5G.

wych działań i w prezentowanych dalej pozycjach literatury przyjmuje się *a priori*, że to zabezpieczenie zostało przełamane.

W [5] przedstawiono atak na funkcję sieciową UPF z wykorzystaniem protokołu PFCP polegający na manipulowaniu sesjami użytkownika. Założono przy tym dostęp do interfejsu N4, co umożliwiło nieautoryzowane działania takie jak żądania usunięcia lub modyfikacji sesji PFCP, czy też zalewanie żądaniami ustanawiania sesji PFCP.

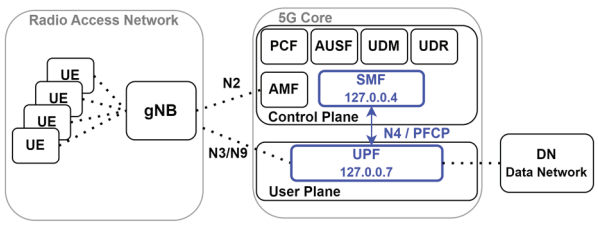
Kolejnym z prezentowanych w literaturze jest atak DDoS (*Distributed Denial of Service*) na sygnalizację [3]. Bazuje on na fakcie, że procedura rejestracji terminala użytkownika (UE, *User Equipement*) zakończona zestawieniem tunelu wymaga wymiany szeregu komunikatów między różnymi instancjami sieci 5G. W tej sytuacji wywołanie odpowiednio dużego ruchu sygnalizacyjnego za pomocą odpowiednio spreparowanych UE może doprowadzić do zablokowania części funkcji sieciowych w sieci szkieletowej.

Z kolei w [6] zaprezentowano atak DoS na funkcję AMF za pośrednictwem protokołu NGAP a także realizację jednego z testów bezpieczeństwa zaproponowanego w specyfikacji 3GPP. Wykorzystano narzędzie *open-source* o nazwie 5Grelpay, które pozwala na manipulację pakietami.

Inne podejście przedstawiono w [2]. Zaprezentowano wyniki ataku zrealizowanego w ramach *PwC & Aalto 5G Cybersecurity challenge*, gdzie atakujący usiłowali osiągnąć swój cel posiadając jedynie dostęp do sieci szkieletowej poprzez wystawiony na zewnątrz interfejs IP. Po rozpoznaniu poszczególnych funkcji sieciowych przerwano działanie sieci poprzez usunięcie części funkcji sieciowych (co było celem przedsięwzięcia).

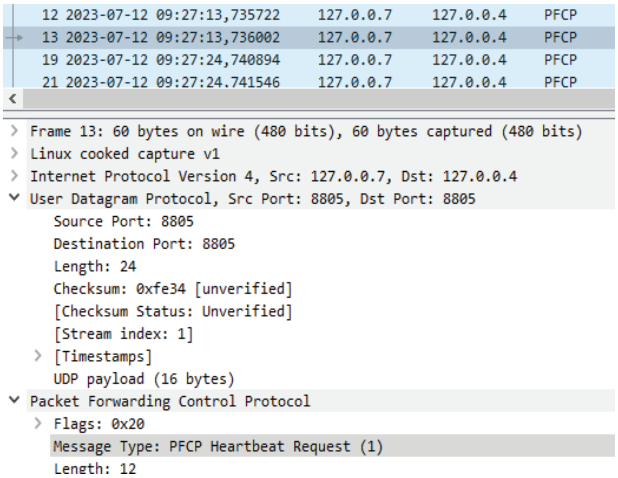
3. KOMUNIKACJA SIECIOWA POMIĘDZY FUNKCJAMI SIECIOWYMI SMF I UPF – PROTOKÓŁ PFCP

Wzmiankowany w przeglądzie wyżej protokół PFCP jest protokołem bezpołączeniowym, transportowanym przez protokół UDP. W sieciach 3GPP stosowany jest w komunikacji pomiędzy płaszczyzną sterowania (CP, *Control Plane*) a płaszczyzną użytkownika (UP, *User Plane*). Protokół ten nie jest szyfrowany i z tego względu podatny jest na modyfikacje danych zawartych w pakiecie. Uproszczoną architekturę sieci szkieletowej 5G wraz z zaznaczoną komunikacją PFCP oraz adresacją IP jak w badanym układzie testowym przedstawiono na Rys. 1.



Rys. 1 Komunikacja funkcji SMF i UPF za pomocą protokołu PFCP

Za pośrednictwem tego protokołu funkcja sieciowa UPF otrzymuje informacje od funkcji SMF na temat reguł wykrywania pakietów (PDR), ich przekazywania/routowania (FAR), QoS (QER) oraz raportowania użycia (URR) celem zestawienia/ usunięcia lub modyfikacji tunelu GPT-U pomiędzy UE – UPF a siecią danych (*Data Network*, DN) zestawianego za pomocą protokołu *GPRS Tunnelling Protocol*. W obrębie tego typu pakietów nie są załączane dodatkowe informacje o sesjach związanych z UE i służą one wyłącznie weryfikacji dostępności sieciowej lub nawiązaniu połączenia/asocjacji pomiędzy modułami. Przechwyconą za pomocą Wireshark komunikację kontrolną ilustruje Rys. 2.



Rys. 2 Protokół PFCP - komunikacja kontrolna

Inne typy komunikatów, identyfikowane za pomocą wartości od 50 do 57 (Tab. 1) związane są z zestawianiem, modyfikacją i usuwaniem sesji kontekstu/tunelu dla danego UE. W przypadku komunikatów związanych z daną sesją oprócz nagłówka protokołu PFCP może być dołączana lista IE (Information Element – Vendor-specific Information Element). Pełna lista typów wiadomości protokołu PFCP znajduje się w [4], poniżej przedstawiono fragment.

Tab. 1 Wybrane identyfikatory wiadomości PFCP [4]

Message Type value (Decimal)	Message
0	Reserved
…	
PFCP Session related messages	
50	PFCP Session Establishment Request
51	PFCP Session Establishment Response
52	PFCP Session Modification Request
53	PFCP Session Modification Response

4. KONCEPCJA ATAKÓW NA PFCP

Analizując przebieg sesji protokołu PFCP zaplanowano przeprowadzenie dwóch ataków polegających na 1). usuwaniu sesji pomiędzy UPF a UE oraz 2). odrzucaniu żądań zestawienia sesji przez UPF dla UE.

4.1. Usuwanie sesji pomiędzy UPF a UE

W momencie, gdy UE nawiązuje łączność następuje najpierw połączenie poprzez RAN (*Radio Access Network*) do AMF gdzie jest przeprowadzana rejestracja urządzenia a następnie (po zaakceptowaniu) zestawiane

jest połączenie do UPF. Funkcja sieciowa UPF jest sterowana przez SMF a komunikacja pomiędzy nimi jest realizowana przez protokół PFCP.

Jednym z typów wiadomości, przydatnym w planowanym ataku, jest polecenie usunięcia sesji i związanego z nią tunelu GTP-U co standardowo ma miejsce w czasie kończenia połączenia przez UE. Co istotne mimo usunięcia dla wskazanego UE tunelu GTP-U jest on nadal podłączony do RAN (skutkuje to brakiem dostępu przez UE do DN, np. Internetu).

Na każde żądanie usunięcia sesji *Session Deletion Request* wysłane przez SMF, UPF zwraca informację dotyczącą statusu sesji wraz z odpowiednim kodem. W komunikacie o usunięciu danej sesji musi być zawarty numer SEID (*sequence ID*), dla którego ma być usunięta sesja. Numeracja SEID w przypadku uruchomienia lub restartu modułu UPF zaczyna się od 1 i jest zwiększana o jeden dla każdej kolejnej sesji UPF.

Fakt ten może być wykorzystany do przeprowadzenia ataku, w którym wysyłane są do UPF żądania *Session Deletion Request* ze stopniowo zwiększanym numerem SEID co docelowo spowoduje usunięcie wszystkich sesji.

4.2. Odrzucanie żądań zestawienia sesji

Gdy UE próbuje zestawić połączenie do DN funkcja SMF wysyła do UPF żądanie zestawienia sesji *Session Establishment Request*. W przypadku prawidłowego zestawienia sesji następuje odpowiedź do SMF wraz z kodem zwrotnym pola *cause* ustawionym na wartość 1. Po prawidłowym zestawieniu sesji zostaje wysłany komunikat z SMF do UPF *Session Modification Request* w wyniku czego UPF zestawia tunel GTP-U, a po prawidłowym zestawieniu tunelu wysyła do SMF potwierdzenie w wiadomości *Session Modification Response*.

Z kolei w momencie, gdy nie jest możliwe zestawienie sesji przez UPF przekazywana jest w odpowiedzi odpowiednia wartość kodu pola *cause* wraz z pozostałymi informacjami uzupełniającymi [9].

Atak polega na modyfikacji danych zawartych w pakiecie PFCP typu *Session Establishment Response* kierowanym z UPF do SMF. W ten sposób każda próba zestawienia sesji a następnie tunelu pomiędzy UE a UPF będzie odrzucana z kodem błędu *Request Rejected* (*reason not specified*).

4.3. Realizacja ataków

Wykorzystano do tego celu środowisko Open5GS [6]. Ze względu na sposób konfiguracji środowiska, w którym poszczególne funkcje sieciowe uruchomiono na tym samym serwerze na adresacji IPv4 z zakresu 127.0.0.0/8 ataki wykonano również na tym serwerze. W tej sytuacji, aby możliwa była realizacja ataku typu *Man-in-The-Middle* i manipulacja pakietami konieczne było ich przekierowanie z kolejki systemowej do osobnej kolejki, co uzyskano za pomocą odpowiedniej konfiguracji zapory sieciowej *iptables*. Przekierowywano do osobnej kolejki (numer 1) ruch z adresu 127.0.0.4 (SMF) do adresu 127.0.0.7 (UPF), lub odwrotnie, na porcie UDP/8805:

iptables -I OUTPUT -s 127.0.0.7/32 -d 127.0.0.4/32 -p udp --dport 8805 -j NFQUEUE --queue-num 1

Kolejka numer 1 obsługiwana jest przez skrypt, w którym następuje przechwycenie danego pakietu, jego modyfikacja, modyfikowanie sum kontrolnych protokołu IP oraz UDP i ponowne skierowanie do kolejki systemowej.

Jako narzędzie użyto skryptów Python3, biblioteki *NetfilterQueue* [7] umożliwiającej dostęp do bufora sieciowego systemu operacyjnego oraz biblioteki *scapy* [8] zapewniającej m.in. dostęp do poszczególnych warstw protokołu PFCP.

Atak na PFCP – usuwanie sesji pomiędzy UPF a UE

Atak polegał na tym, że co czwarta ramka z adresu 127.0.0.4 do adresu 127.0.0.7 typu *Heartbeat Request* była modyfikowana. Pozostałych pakietów tego typu nie modyfikowano, aby nie dopuścić do całkowitego zaburzenia komunikacji między funkcjami sieciowymi. Atak przebiegał w następujący sposób:

1. Przechwycenie co czwartej ramki PFCP, która w polu *message_type* ma ustawioną wartość 1 *(heartbeat request*);
2. Modyfikacja pola *message_type* – ustawiana jest wartość 54 (*session deletion request*);
3. Ustawiana jest wartość dla pola SEID (zwiększana o 1 dla kolejnej, co czwartej ramki typu *heartbeat request*);
4. Usuwane są dodatkowe informacje przesyłane w pakiecie (*payload*);
5. Modyfikacja sum kontrolnych IP oraz UDP i ponownie skierowanie do kolejki systemowej.

Atak PFCP – odrzucanie żądań zestawienia sesji

Metodologia tego ataku jest identyczna z poprzednim a specyficzne dla ataku kroki są następujące:

1. Pakiety przekierowane do kolejki 1 są weryfikowane pod kątem wystąpienia w polu *message_type* w nagłówku protokołu PFCP wartości 51 (*Session Establishment Response*);
2. Przeszukiwana jest lista IE (*payload* protokołu PFCP) celem odnalezienia pola *IE_Type* wskazującego na wartość 19 (pole typu *cause*) i następnie zmieniana jest jego wartość z 1 (*Request Accepted*) na wartość 64 (*Request Rejected*).

4.4. Rezultaty przeprowadzonych ataków

W przypadku uruchomienia ataków z 4.1 potwierdzeniem skuteczności wykonanego ataku była informacja w logach UPF świadcząca o braku aktywnych sesji:

07/12 09:29:06.284: [upf] INFO: [Removed] Number of UPF-sessions is now 0 (../src/upf/context.c:212)

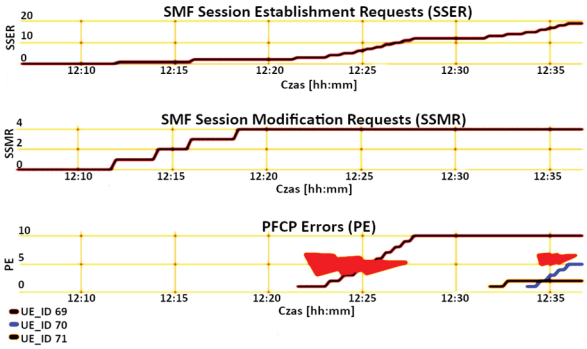
W podobny sposób zweryfikowano poprawność wykonania ataku opisanego w 4.2. W czasie, gdy był prowadzony atak odnotowano w logach SMF wpisy potwierdzające odrzucanie żądań zestawienia sesji:

07/10 09:22:31.963: [smf] ERROR: PFCP Cause [64]: Not Accepted (../src/smf/n4-handler.c:179)

5. WYKRYWANIE SYMPTOMÓW ATAKÓW

Ślady przedstawionych wyżej ataków są trudne do wykrycia i należy ich poszukiwać w plikach logów poszczególnych funkcji sieciowych. Analizując przebieg ataków oraz możliwości monitorowania można stwierdzić, iż atak na odrzucenie zestawienia sesji PDU podobnie, jak usuwający sesje, nie jest łatwy do wykrycia. Zasadnym wydaje się tu monitorowanie wskaźnika, jakim jest liczba błędów protokołu PFCP, zgłaszanych zarówno przez SMF jak i UPF.

Przeprowadzono weryfikację możliwości identyfikacji ataków z użyciem narzędzi Prometheus oraz Graphana. W tym celu opracowane zostały odpowiednie skrypty. Zaś przykładowe rezultaty są przedstawione na Rys. 3. Zaznaczono czerwonymi strzałkami wzrost liczby monitorowanych zdarzeń świadczący o atakach przeprowadzanych na protokół PFCP.



Rys. 3 Symptomy ataku odrzucania żądań zestawienia sesji raportowane przez funkcję SMF

Lawinowo przyrastająca liczba żądań zestawienia sesji PDU oraz jednocześnie liczba błędów mogą wskazywać na występowanie nieprawidłowości w sieci szkieletowej 5G. Pojedyncze, a zarazem trudno wykrywalne ataki usuwania sesji mogą być sygnalizowane jako wystąpienie dwóch czynników, tj. pojawienie się wiadomości *PDU session deletion request* oraz informacji w logach modułu UPF mówiącym o błędnym kontekście takiej wiadomości *No context errors*.

6. PODSUMOWANIE

W referacie przedstawiono raportowane w literaturze przedmiotu podatności sieci szkieletowej 5G mogące stanowić wektor ataku cybernetycznego. Następnie po przeanalizowaniu sekwencji wymiany komunikatów protokołu PFCP przeprowadzono własne dwa ataki skutkujące tym, że UE nie posiada dostępu do sieci danych, mimo że jest skojarzony z gNB.

Kolejnym elementem prac badawczych była analiza możliwości rozpoznawania symptomów tego typu ataków cybernetycznych. Wykorzystano w tym celu narzędzia Prometheus/ Graphana. Odpowiedni dobór monitorowanych wskaźników pobieranych, w tym wypadku, z logów systemowych umożliwił identyfikację zdarzeń niepożądanych. Przedstawione w opracowaniu wskaźniki wydają się być przydatne a ich wizualizacja może być pomocna osobom nadzorującym pracę sieci w ocenie sytuacji i identyfikacji zdarzenia oraz przeciwdziałaniu. Wykorzystany w badaniach mechanizm alarmów oprogramowania Graphana daje szereg możliwości detekcji zmian pojedynczych wskaźników, jak i powiązania zmian w kilku wskaźnikach jednocześnie. Podejście takie było wystarczające w pracy badawczej i docelowo zostało zautomatyzowane.

Natomiast w przypadku bardziej wyrafinowanych ataków na sieci 5G może być nie wystarczające do automatyzacji tego procesu, bez posiłkowania się wiedzą i doświadczeniem operatora. Wydaje się, iż celowym było by zastosowanie bardziej zaawansowanych mechanizmów detekcji zaburzeń takich, jak np. uczenie maszynowe lub algorytmy sztucznej inteligencji, które spośród zidentyfikowanych zdarzeń byłyby w stanie wykryć te o charakterze celowym.

LITERATURA

[1] Bui Nhat Linh, A. "Analysing Open-Source 5G Core Networks for TLS Vulnerabilities and 3GPP Compliance." RADBOUD UNIVERSITY, 2023. https://www.cs.ru.nl/bachelors-theses/2023/Alex_Bui_Nhat_Linh_1040308_Analysing_open-source_5G_core_networks_for_TLS_vulnerabilities.pdf.

[2] Dudek, S. "Intruding 5G SA Core Networks from Outside and Inside." *Penthertz* (blog), n.d. https://penthertz.com/blog/Intruding-5G-core-networks-from-outside-and_inside.html.

[3] Park, S., Byungsun Ch., Dowon K., and Ilsun Y. "Machine Learning Based Signaling DDoS Detection System for 5G Stand Alone Core Network." *Applied Sciences* 12, no. 23 (December 5, 2022): 12456. https://doi.org/10.3390/app122312456.

[4] Amponis, G., et all. "Threatening the 5G Core via PFCP DoS Attacks: The Case of Blocking UAV Communications." *EURASIP Journal on Wireless Communications and Networking* 2022, no. 1, Dec. 15, 2022. https://doi.org/10.1186/s13638-022-02204-5.

[5] Salazar, Z., Huu Nghia N., Wissam M., Ana R. Cavalli, and De Oca, E.M. "5Greplay: A 5G Network Traffic Fuzzer - Application to Attack Injection." In *Proceedings of the 16th International Conference on Availability, Reliability and Security*, 1–8. Vienna Austria: ACM, 2021. https://doi.org/10.1145/3465481.3470079.

[6] "Https://Open5gs.Org/." Accessed May 10, 2024.

[7] Fox, M., "NetfilterQueue: Python Bindings for Libnetfilter_queue." POSIX :: Linux, Cython, Python. Accessed May 10, 2024.

[8] "Scapy." https://scapy.net/. Acc. 05.10. 2024.

[9] "LTE; 5G; Interface between the Control Plane and the User Plane Nodes (3GPP TS 29.244 Version 15.8.0 Release 15)," January 2020.

KRiT 2024
KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

METODA SKALOWANIA WYDAJNOŚCI FUNKCJI UPF W SIECIACH 5G BAZUJĄCA NA SI

AI-BASED METHOD OF UPF'S PERFORMANCE SCALING IN THE 5G NETWORKS

Krzysztof Sadura[1]

[1] Politechnika Warszawska,
krzysztof.sadura.stud@pw.edu.pl

DOI: 10.15199/59.2024.4.21

Streszczenie: W niniejszej pracy przedstawiono metodę horyzontalnego skalowania wydajności funkcji UPF na potrzeby sieci 5G. Proponowany algorytm dopasowuje liczbę uruchomionych instancji UPF do zmieniającego się okresowo obciążenia sieci na podstawie analitycznego modelu z teorii kolejek i prognozy ruchu sieciowego dokonanej przez rekurencyjną sieć neuronową LSTM. Wyniki przeprowadzonych eksperymentów symulacyjnych wykazały skuteczność rozwiązania oraz jego przewagę nad metodą proponowaną przez innych autorów.

Abstract: In this paper, I present a method for horizontal scaling of the UPF's performance in the 5G network. The proposed algorithm adjusts the number of launched UPF instances to periodically evolving network load using an analytical model from queueing theory and network traffic prediction made by a recursive LSTM neural network. Simulation results showed the effectiveness of the solution and its advantage over the method from the literature.

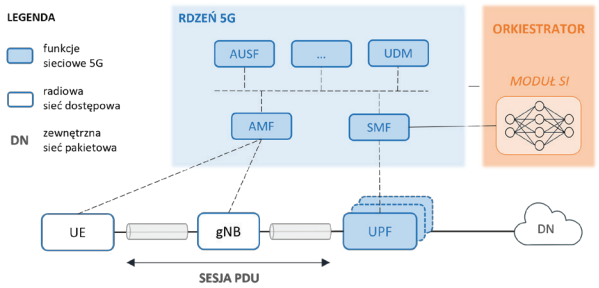
Słowa kluczowe: LSTM, sieć 5G, skalowanie, UPF

Keywords: 5G network, LSTM, scaling, UPF

1. WPROWADZENIE

Jedną z podstawowych różnic pomiędzy obecnie wdrażaną siecią 5G a starszymi generacjami jest to, iż odchodzi ona od rozwiązań czysto sprzętowych na rzecz rozwiązań bazujących na wirtualizacji funkcji sieciowych NFV (ang. *Network Function Virtualization*). Elementy tworzące rdzeń sieci 5G mają bowiem postać wirtualnych maszyn bądź kontenerów uruchamianych w chmurze obliczeniowej. Koncepcja ta pozwala operatorom na szybkie wdrażanie nowych funkcji, skalowanie ich wydajności czy automatyczne aktualizowanie oprogramowania zachowując ciągłość działania sieci.

W architekturze sieci 5G (zob. Rys. 1) wydzielono dwie podstawowe płaszczyzny: płaszczyznę sterowania odpowiedzialną m.in. za obsługę sygnalizacji oraz płaszczyznę przekazu danych odpowiedzialną za transfer pakietów pomiędzy terminalami użytkowników i zewnętrzną siecią. Przekaz ten jest realizowany w ramach sesji PDU (ang. *Protocol Data Unit*), zestawianej między terminalem a instancją brzegową funkcji UPF (ang. *User Plane Function*) nazywaną PSA-UPF (ang. ***P**DU **S**ession **A**nchor UPF*). Liczba aktywnych sesji ściśle zależy od aktywności użytkowników, która typowo zmienia się w cyklu dobowym, powodując zmiany obciążenia funkcji UPF. Zwymiarowanie wydajności UPF na wartość szczytową ruchu przyczynia się do nieefektywnego wykorzystania zasobów i w konsekwencji ich marnowania w okresach mniejszego obciążenia sieci 5G. Z drugiej strony, uruchomienie zbyt małej liczby instancji UPF względem aktualnego zainteresowania prowadzi do degradacji jakości oferowanych usług QoS (ang. *Quality of Service*). Zdefiniowany problem ukazuje konieczność opracowania algorytmu horyzontalnego skalowania wydajności funkcji UPF, który dostosowuje liczbę uruchomionych instancji funkcji UPF do aktualnego obciążenia sieci 5G. Potencjalne rozwiązanie pozwoliłoby na kompromis pomiędzy ilością zajętych zasobów, a utrzymaniem wymaganego poziomu jakości usług.



Rys. 1: Architektura sieci 5G

W literaturze można znaleźć publikacje dotyczące horyzontalnego skalowania wydajności funkcji sieciowych 5G. Autorzy pracy [1] proponują reakcyjny algorytm, w którym wprowadzają dwie wartości progowe. Jeśli liczba aktywnych sesji w systemie przekroczy jedną z nich, spowoduje to odpowiednio dodanie nowej lub usunięcie nadmiarowej instancji UPF z systemu. W artykułach [2, 3] przedstawiono podejście proaktywne. Polega ono na predykcji wolumenu ruchu sieciowego $\lambda(t+1)$ w następnym przedziale czasowym z wykorzystaniem metod sztucznej inteligencji (SI). Na podstawie tej prognozy, jest wyznaczana wymagana liczba instancji UPF zgodnie z zależnością (1), gdzie symbol γ oznacza maksymalną zdolność obsługową pojedynczej instancji UPF wyrażoną w Mbit/s.

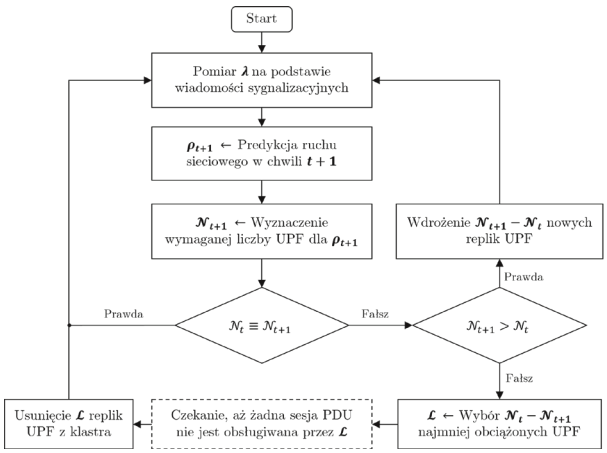
$$N = \frac{\lambda(t+1)}{\gamma} \tag{1}$$

W tym artykule zaproponowałem prewencyjną metodę skalowania wydajności funkcji UPF z wykorzystaniem rekurencyjnych sieci neuronowych (RNN). Przedstawiona metoda rozszerza obecne rozwiązania o nowy algorytm wyznaczania liczby wymaganych instancji UPF, który bierze pod uwagę zarówno predykcję ruchu, jak również założone *a'priori* prawdopodobieństwo przekroczenia zdolności obsługowej.

Organizacja pozostałej części artykułu jest następująca. W rozdziale 2 omówiono proponowany algorytm skalowania. Wyniki eksperymentów wraz z wnioskami zamieszczono w rozdziale 3, a ostatni rozdział 4 został poświęcony na podsumowanie całości.

2. PROPONOWANE ROZWIĄZANIE

Na Rys. 1 przedstawiono architekturę sieci 5G wraz z orkiestratorem oraz modułem SI. W zwirtualizowanej infrastrukturze sieciowo-obliczeniowej orkiestrator zarządza pełnym cyklem życia poszczególnych funkcji, w tym inicjuje proces zwiększenia bądź zmniejszenia ich wydajności. W zaproponowanej metodzie moduł SI jest odpowiedzialny za predykcję ruchu sieciowego. Na podstawie pomiarów ruchu w sieci dokonuje on prognozy, której wynik służy do wyznaczenia wymaganej liczby instancji funkcji UPF.



Rys. 2: Proponowany algorytm skalowania

Proponowany algorytm skalowania przedstawiono na Rys. 2 w postaci schematu blokowego. Jeśli prognozowana liczba instancji UPF nie różni się względem aktualnego stanu, to algorytm powraca to punktu wyjścia. W przeciwnym wypadku orkiestrator zwiększa lub redukuje wydajność funkcji poprzez dodanie nowych bądź usunięcie nadmiarowych instancji. To drugie następuje dopiero po zakończeniu wszystkich dotychczasowych sesji PDU obsługiwanych przez wytypowane do usunięcia instancje.

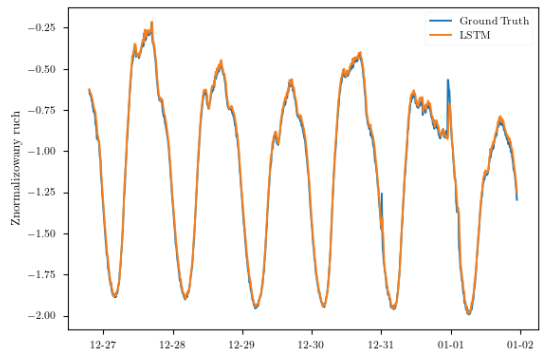
2.1 Pomiar intensywności ruchu

Za zestawienie sesji danych w 5G odpowiada funkcja sieciowa SMF (ang. *Session Management Function*). Jej działanie jest inicjowane w chwili otrzymania żądania na zestawienie sesji o treści `PDUSessionEstablishmentRequest`. Zliczając te wiadomości i dzieląc ich liczbę przez jednostkę czasu, w której dokonano pomiaru, otrzymujemy intensywność λ z jaką są zestawiane nowe sesje PDU. Na tej podstawie jest wyznaczana wartość ruchu oferowanego uwzględniając również średni czas trwania sesji $1/\mu$.

2.2 Predykcja ruchu sieciowego

Do przewidywania ruchu w sieciach, ze względu na jego cykliczną charakterystykę, są wykorzystywane techniki predykcji szeregów czasowych. Jedną z metod, która charakteryzuje się wysoką dokładnością na tle innych rozwiązań jest metoda bazująca na sieci **LSTM** (ang. *Long Short-Term Memory*) [4], która to z kolei jest rodzajem rekurencyjnej sieci neuronowej **RNN** (ang. *Recurrent Neural Network*). Podstawą działania tego typu sieci jest sprzężenie zwrotne. W trakcie przetwarzania sekwencyjnego sieć korzysta ze swoich wcześniejszych obliczeń (pamięci), dzięki czemu uwzględniony zostaje kontekst historyczny.

Do trenowania i testowania sieci LSTM posłużyły gromadzone przez dwa kolejne miesiące dane włoskiego operatora komórkowego [5]. Mają one postać szeregu czasowego $\{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i)\}$, gdzie x jest znacznikiem czasowym, a y oznacza liczbę aktywnych w tym czasie połączeń z Internetem. Czas pomiędzy kolejnymi próbkami wynosił 10 minut. W celu dostosowania danych do omawianego problemu, liczbę wygenerowanych połączeń podzielono przez okres pomiarowy, tak aby otrzymać średnią intensywność. Zbiór danych został podzielony na dane trenujące i testowe w stosunku 9:1. W opracowanym modelu sekwencja wejściowa ma postać wektora zawierającego sześć ostatnich pomiarów (1 godzina), a sieć LSTM na ich podstawie dokonuje prognozy na najbliższy 10-minutowy okres. Na Rys. 3 przedstawiono jak wytrenowana sieć radzi sobie z danymi testowymi. Na osi rzędnych pojawiły się wartości ujemne, gdyż dane zostały wcześniej znormalizowane. Błąd średniokwadratowy MSE (ang. *Mean Squared Error*) predykcji wyniósł kolejno 0.003 dla danych trenujących oraz 0.001 dla mniejszego zbioru testowego.



Rys. 3: Predykcja mobilnego ruchu sieciowego

2.3 Wyznaczanie liczby UPF

Elementem, który umożliwia określenie i kontrolę parametrów jakościowych podczas transmisji danych jest tzw. przepływ QoS (ang. *QoS Flow*). Każdy taki przepływ zawiera swój unikalny identyfikator oraz informację m.in. na temat gwarantowanej szybkości bitowej. Ze względu na fakt, że oferowane w ramach sieci 5G aplikacje i usługi mają odmienne wymagania co do przepływności, to zestawione sesje w różny sposób obciążają UPF.

Do obliczenia wymaganej liczby instancji UPF wykorzystano **uogólniony wielousługowy model Erlanga** (ang. *Generalised Multirate Erlang Loss Model*) [6], który rozszerza tradycyjny model Erlanga na wiele klas ruchu. Rozkład stacjonarny stanów $P[n]$ modelu można określić za pomocą rekurencyjnej formuły Kaufmana-Robertsa (2):

$$nP[n] = \sum_{k=1}^{M} d_k \rho_k P[n-d_k], \quad n = d_1 n_1 + \ldots + d_k n_k \quad (2)$$

gdzie M to liczba zdefiniowanych rodzajów sesji PDU, d_k jest wymaganą szybkością bitową dla sesji typu k, n_k oznacza aktualną liczbę sesji klasy k w systemie, a $\rho_k = {}^{\lambda_k}/_{\mu_k}$ to ruch oferowany. Prawdopodobieństwo blokady sesji należącej do danej klasy można obliczyć zgodnie z zależnością (3):

$$b_k = \sum_{n=m}^{C} P[n] \quad (3)$$

gdzie C to sumaryczna zdolność obsługowa wdrożonych funkcji UPF, a m jest wynikiem wyrażenia $C - d_k + 1$. Korzystając z przedstawionych wzorów, zaproponowany algorytm skalowania (zmieniając wartość C) wyznacza odpowiednią liczbę replik UPF dla zadanego *a'priori* prawdopodobieństwa odrzucenia napływającego zgłoszenia z powodu przekroczenia zdolności obsługowej.

3. OCENA EFEKTYWNOŚCI

Celem eksperymentów jest porównanie efektywności działania proponowanej metody skalowania względem rozwiązań proponowanych w literaturze. Badania uwzględniają również wpływ zastosowania predykcji bazującej na algorytmach sztucznej inteligencji w porównaniu do prostej predykcji, w której przewidywana wartość λ_{t+1} jest zawsze równa wartości ostatniego pomiaru λ_t.

3.1 Środowisko eksperymentalne

Do przeprowadzenia eksperymentów posłużył samodzielnie zaimplementowany symulator zdarzeniowy, w którym działanie systemu 5G zamodelowano na poziomie wywołań. Występują w nim trzy główne rodzaje zdarzeń.

- `PDUSessionRequest` - symuluje zestawienie sesji PDU przez funkcję SMF. Czas pomiędzy wystąpieniem zdarzeń tego typu jest opisany rozkładem wykładniczym, który modeluje napływ żądań od dużej liczby niezależnych użytkowników.
- `PDUSessionRelease` - czas trwania sesji. Ponieważ uogólniony model Erlanga jest systemem stratnym bez kolejki, zatem na stan systemu ma wpływ jedynie średni czas trwania sesji, więc może być zastosowany dowolny jego rozkład.
- `ArrivalRateMod` - umożliwia zmianę intensywności napływu żądań λ w cyklu dobowym zgodnie z charakterystyką danych zmierzoną w sieci operatora.

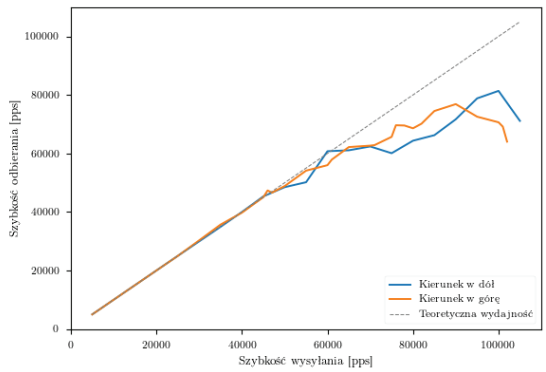
W tabeli 1 zamieszczono parametry symulacji. Ze względu na ograniczenia dostępnego zbioru danych testowych, symulacja została przeprowadzona jedynie dla pojedynczej usługi (jeden rodzaj sesji PDU). Warto jednak zaznaczyć, że model pozostaje spójny, a implementacja symulatora jest dostosowana do scenariusza z wieloma klasami ruchu sieciowego, które będą przedmiotem dalszych badań.

Tab. 1: Parametry symulacji

Parametr	*Wartość*	*Jednostka*
M	1	$[\cdot]$
d_1	128	[Mbit/s]
μ_1	0.1	$[1/min]$
b_1	0.01, 0.05, 0.1	$[\cdot]$

3.2 Badanie wydajności instancji UPF

Metoda skalowania wymaga określenia maksymalnej wydajności pojedynczej instancji UPF. Jako wartość odniesienia przyjęto wydajność funkcji UPF zaimplementowanej w ramach projektu *free5GC* [7], który dostarcza podstawowe funkcje sieciowe rdzenia 5G. W celu oceny jej maksymalnej zdolności obsługowej przeprowadzono test zgodny z zaleceniem RFC2544 [8]. Na Rys. 4 przedstawiono wyniki testu instancji UPF uruchomionej na serwerze z procesorem Intel® Xeon E5-2650. Maksymalne obciążenie przy jakim zachowany jest bezstratny przekaz danych wynosi około 500 pps (ang. *packets per second*). Ponieważ wysyłane podczas testu pakiety miały rozmiar 1280 B, to przyjęta na potrzeby symulacji przepustowość jednej instancji UPF wynosi 512 Mbit/s.



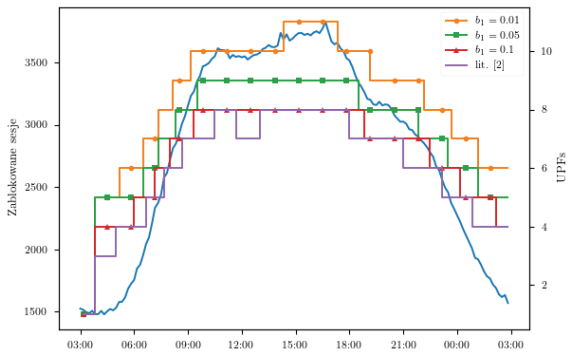
Rys. 4: Wyniki badania wydajności funkcji UPF

3.3 Porównanie efektywności metod skalowania

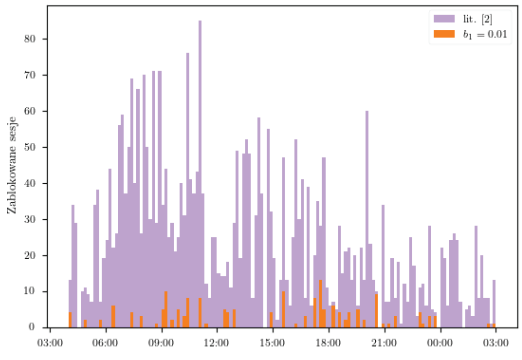
Symulacja została przeprowadzona trzykrotnie, zmieniając za każdym razem zakładane prawdopodobieństwo przekroczenia zdolności obsługowej b_1. Na Rys. 5 zaprezentowano otrzymane wyniki. Niebieska krzywa przedstawia dobową zmianę obciążenia sieci, natomiast wykresy schodkowe pokazują jak w ciągu doby zmieniała się liczba uruchomionych instancji funkcji UPF, w zależności od zastosowanego algorytmu skalowania. Zgodnie z przewidywaniami, przy mniejszej tolerancji na dopuszczalny spadek jakości, liczba uruchomionych instancji jest wyższa. Z tabeli 2 wynika, że założony poziom jakości został zachowany, gdyż prawdopodobieństwo blokady P_b w czasie trwania symulacji nie przekracza dopuszczalnego prawdopodobieństwa b_1.

Kolorem fioletowym zaznaczono wyniki uzyskane dla metody skalowania proponowanej przez autorów pracy [2], w której system jest wymiarowany na wartość średnią a prawdopodobieństwo przekroczenia zdolności obsługowej nie jest kontrolowane. Takie podejście spowodowało

straty na poziomie 10% przy średniej liczbie wdrożonych funkcji N_{avg} równej nieco ponad 6. Dzięki wykorzystaniu w proponowanej metodzie modeli bazujących na teorii kolejek, oprócz kontroli nad poziomem dostępności usługi osiągnięto również zysk multipleksacji. Wdrożenie kilku instancji UPF więcej zmniejszyło szansę przekroczenia zdolności obsługowej z 10% na 0.5%. Potwierdza to także histogram z Rys. 6, porównujący obie metody pod względem liczby zablokowanych (odrzuconych) żądań o zestawienie sesji PDU. Widać na nim znaczącą różnicę w liczbie odrzuconych sesji. Stworzony algorytm przez całą dobę zapewnił wysoki poziom dostępności, a zastosowanie predykcji pozwoliło ponadto uniknąć gwałtownych wzrostów liczby nieobsłużonych żądań.



Rys. 5: Porównanie metod skalowania przy zmianach obciążenia w cyklu dobowym



Rys. 6: Histogram zablokowanych sesji

Nieco zaskakującym wynikiem jest to, że zastosowanie rekurencyjnej sieci LSTM do predykcji ruchu, pomimo bardzo dobrej dokładności, nie wpłynęło istotnie na poprawę efektywności metody skalowania względem zastosowania uproszczonej predykcji. Przyczyną tego może być wysoka częstotliwość z jaką dokonano pomiarów ruchu w sieci względem jego dynamiki. Okres pomiędzy kolejnymi próbkami jest na tyle mały, że różnica między dwiema kolejnymi zmierzonymi wartościami jest nieznaczna i nie wpływa na rezultat dokonywanych przez algorytm obliczeń. Należy oczekiwać, że prosta predykcja nie sprawdzi się jednak w sytuacji, gdy prognoza ma sięgać dalej niż tylko 10 minut. Często skalowanie zwirtualizowanych zasobów jest inicjowane rzadziej, np. w 1-godzinnych interwałach, przez co konieczna tam będzie predykcja z użyciem bardziej wyrafinowanych technik.

Tab. 2: Wyniki symulacji

	LSTM		Predykcja prosta	
	P_b	N_{avg}	P_b	N_{avg}
●—	0.005	8.70	0.005	8.66
■—	0.027	7.55	0.027	7.52
▲—	0.066	6.71	0.066	6.72
—	0.098	6.26	0.109	6.18

4. PODSUMOWANIE

W artykule przedstawiono proaktywną metodę horyzontalnego skalowania funkcji UPF na potrzeby sieci 5G. Opisano jak dzięki założeniu, że rozwiązanie opiera swoje działanie na sesjach PDU, wykorzystać uogólniony model Erlanga w procesie prewencyjnego skalowania. Przeprowadzone eksperymenty symulacyjne pokazały efektywność zaproponowanego algorytmu względem metod proponowanych w literaturze. Relatywnie krótki okres predykcji spowodował, że prosta technika przewidywania ruchu dała zbliżony wynik co sieć LSTM. Kolejnym etapem pracy nad badanym tematem będzie wykorzystanie i weryfikacja zaprezentowanego rozwiązania w eksperymentalnej sieci 5G.

LITERATURA

[1] C. Rotter and T. Van Do, "A Queueing Model for Threshold-Based Scaling of UPF Instances in 5G Core," *IEEE Access*, vol. 9, pp. 81443–81453, 2021.

[2] T. Subramanya and R. Riggio, "Centralized and Federated Learning for Predictive VNF Autoscaling in Multi-Domain 5G Networks and Beyond," *IEEE Transactions on Network and Service Management*, vol. 18, no. 1, pp. 63–78, 2021.

[3] T. Subramanya, D. Harutyunyan, and R. Riggio, "Machine Learning-Driven Service Function Chain Placement and Scaling in MEC-enabled 5G Networks," *Computer Networks*, vol. 166, p. 106980, 2020.

[4] H. D. Trinh, L. Giupponi, and P. Dini, "Mobile Traffic Prediction from Raw Data Using LSTM Networks," in *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 1827–1832, 2018.

[5] Telecom Italia, "Telecommunications - SMS, Call, Internet - MI," 2015.

[6] K. W. Ross, "Chapter 2 - The Stochastic Knapsack," in *Multiservice Loss Networks for Broadband Telecommunications Networks*, pp. 18–23, Springer-Verlag, 1995.

[7] "free5GC." https://www.free5gc.org. [Online] Accessed: 2024-05-11.

[8] "Benchmarking Methodology for Network Interconnect Devices." RFC 2544, Mar. 1999.

KRiT 2024
KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

ORKIESTRACJA APLIKACJI I FUNKCJI UPF W SIECIACH B5G

ON APPLICATIONS AND UPF FUNCION ORCHESTRATION IN B5G NETWORKS

Witold Jóźwiak[1]

[1] Politechnika Warszawska, Warszawa, witold.jozwiak@pw.edu.pl

DOI: 10.15199/59.2024.4.22

Streszczenie: Artykuł dotyczy rozwoju sieci 5G ewoluujących w kierunku zintegrowanych systemów sieciowo-chmurowych. Celem badań jest opracowanie nowych strategii orkiestracji aplikacji i usług sieci 5G, które wykorzystują zasoby różnych dostawców, tj. operatorów sieci, dostawców MEC i chmurowych. Proponowane strategie wykorzystują model orkiestracji bazujący na metodach programowania liniowego. Przeprowadzone eksperymenty potwierdziły, że wspólna orkiestracja aplikacjami oraz funkcjami UPF pozwala uzyskać największą efektywność.

Abstract: The article deals with the evolution of 5G networks into edge-cloud continuum systems. The aim of the research is to develop new orchestration strategies for 5G network applications and services that utilize resources from different providers, i.e., network operators, MEC, and cloud providers. The proposed strategies use an orchestration model based on linear programming methods. The experiments confirmed that the joint orchestration of applications and UPF functions achieves the highest efficiency.

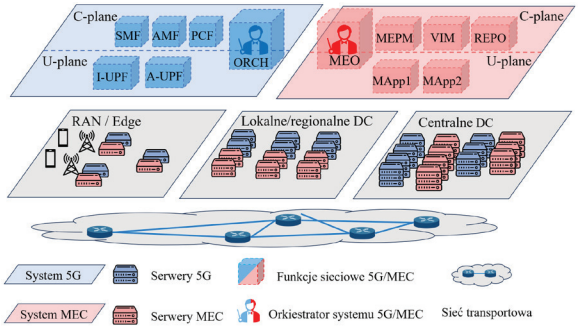
Słowa kluczowe: 5G/6G, orkiestracja, programowanie matematyczne, przetwarzanie na brzegu sieci, UPF

Keywords: 5G/6G, edge computing, orchestration, integer linear programming (ILP), UPF

1. WPROWADZENIE

Obecnie obserwujemy tendencję do wykorzystania technik chmurowych, wirtualizacji oraz softwaryzacji zarówno do budowy sieci nowej generacji, jak również do budowy infrastruktury umożliwiającej świadczenie usług. Biorąc pod uwagę, że systemy te wykorzystują podobne rozwiązania techniczne, należy oczekiwać ewolucji infrastruktury w kierunku zintegrowanego środowiska sieciowo-chmurowego, w literaturze określanego jako *edge-cloud continuum (ECC)*, którego zasoby będą wykorzystywane zarówno do wdrożenia funkcji sieciowych, jak również do świadczenia usług, w tym usług na brzegu sieci. Główne wyzwania stojące przed projektantami systemów ECC dotyczą: heterogeniczności stosowanych rozwiązań, rozdrobnienia podmiotów świadczących usługi, braku standardów i dobrych praktyk określających zasady współpracy podmiotów tego rynku, a także zróżnicowania wymagań wśród jego interesariuszy. Zagadnienia te są tematem prowadzonych badań, których celem jest opracowanie scenariuszy integracji takich systemów.

W systemach komórkowych 3G/4G możliwość integracji z systemami obliczeń na brzegu jest ograniczona ze względu na lokalizację bramy do sieci danych (*packet gateway*). Jej typowa lokalizacja w centrum danych operatora (DC) niweluje korzyści z umieszczenia serwerów obliczeniowych na brzegu sieci, ponieważ pakiety muszą



Rysunek 1: Przykładowy system ECC

pokonać podwójną drogę (od użytkownika do DC i od DC do serwera brzegowego). Sieci 5G z założenia wspierają przetwarzanie na brzegu sieci. W tym celu w pobliżu stacji bazowych gNB są umieszczane instancje funkcji I-UPF (*Intermediate UPF*) z funkcjonalnością ULCL *(Uplink Classifier*), które umożliwiają przekierowywanie sesji użytkownika do różnych sieci danych. Bramę między siecią 5G a siecią danych stanowią funkcje A-UPF (*PDU Session Anchor UPF*). Operator świadomy istnienia usług brzegowych umieszcza w swojej sieci wiele instancji A-UPF, w szczególności na brzegu sieci, co pozwala na świadczenie usług w lokalnych serwerach bez konieczności angażowania odległych serwerów w DC.

Rysunek 1 przedstawia architekturę przykładowego zintegrowanego systemu sieciowo-chmurowego, złożonego z elementów sieci 5G i systemu MEC (*Multi-access edge computing*), należącego do rodziny systemów ECC. Założono, że wszystkie komponenty systemu są dołączone do przewymiarowanej sieci transportowej operatora. Warstwa infrastruktury zawiera urządzenia fizyczne (stacje bazowe gNB, serwery obliczeniowe), które można podzielić na 3 grupy: główne, regionalne i brzegowe centra danych. Im bliżej brzegu sieci tym mniej zasobów jest dostępnych w DC. Zasoby warstwy infrastruktury są podzielone na zasoby sieci 5G zarządzane przez operatora (niebieskie) oraz zasoby MEC (czerwone) zarządzane przez zewnętrzny podmiot, np. dostawcę usług chmurowych. Na fizycznej infrastrukturze uruchomione są komponenty funkcjonalne sieci 5G i systemu MEC. Ze względu na oferowane funkcje, komponenty zostały podzielone na funkcje płaszczyzny sterowania (C-plane) i funkcje płaszczyzny użytkownika (U-plane). Każdy z systemów posiada własny orkiestrator. Współpraca między orkiestratorami i jej zakres zależy od stopnia integracji między systemami, co jest głównym przedmiotem rozważań. Zagadnienie orkiestracji w systemach ECC jest rozważane w literaturze. Jednakże, większość prac skupia się jedynie na wybranym problemie. Przykładowo modele optymalizujące rozmieszczanie serwerów/aplikacji brzegowych nie uwzględniają lokalizacji funkcji UPF, a modele dotyczące

wymiarowania UPF nie biorą pod uwagę usług świadczonych przez sieć 5G. Autorzy [1] zauważają, że rozmieszczając serwery brzegowe w sieci i zapewniając, że zakładane opóźnienie nie zostanie przekroczone, należy również wziąć pod uwagę lokalizację funkcji UPF. W proponowanym modelu nie uwzględniono jednak wszystkich ograniczeń, co czyni go niekompletnym. W [2] autorzy rozważają łańcuchy funkcji sieciowych UPF, z rozróżnieniem ich typów (A-, I-, MI-UPF). Napływ żądań jest modelowany jako zbiór aktywnych sesji PDU w każdym gNB. Wynik zawiera informację o tym, gdzie należy uruchomić instancje funkcji sieciowych. Zwracana jest również informacja o rozpływie ruchu. Zauważalnym mankamentem modelu jest założenie, że aplikacje działają na tych samych węzłach co A-UPF, co jest zbyt dużym uproszczeniem. Sama alokacja instancji aplikacji została pominięta. Ci sami autorzy w [3] idą o krok dalej zakładając, że alokacja odbywa się dynamicznie, w systemie zmieniającym się w czasie. W literaturze zauważalny jest również wzrost zainteresowania technikami sztucznej inteligencji i uczenia maszynowego. W [4] są proponowane algorytmy orkiestracji łańcuchami funkcji usługowych w heterogenicznym środowisku obliczeń na brzegu sieci złożonym z urządzeń IoT. Opracowano rozwiązanie optymalnie oraz heurezę wspartą technikami głębokiego uczenia maszynowego, przewidującą zdarzenia mobilności urządzeń oraz prognozującą zużycie energii.

Celem tego artykułu jest opracowanie strategii orkiestracji dla scenariuszy uwzględniających zróżnicowany stopień zintegrowania infrastruktury przyszłych sieci B5G (*Beyond 5G*), począwszy od niezależnych systemów projektowanych osobno, aż do w pełni zintegrowanych systemów, w których zasoby brzegowe i operatorskie są traktowane wspólnie. W artykule zaproponowano model orkiestracji wykorzystujący technikę programowania liniowego (MILP), który stanowi podstawę dla opracowania proponowanych strategii orkiestracji. Strategie te wyznaczają liczbę i rozmieszczenie instancji aplikacji MEC wraz z instancjami funkcji A-UPF, które przenoszą ruch generowany przez aplikacje. Wyznaczając rozpływ ruchu pomiędzy wybraną lokalizacją instancji aplikacji MEC, a lokalizacją terminala użytkownika algorytmy zapewniają wymaganą jakość przekazu danych. W rozdziale 2 przedstawiono proponowaną metodę orkiestracji. Wyniki eksperymentów wraz z uzyskanymi wnioskami zamieszczono w rozdziale 3. Podsumowanie artykułu zawiera rozdział 4.

2. PROPONOWANA METODA ORKIESTRACJI

W rozważanym systemie założono, że ścieżka przekazu danych zawsze składa się z dwóch UPF: (1) przewymiarowanego I-UPF z funkcjonalnością ULCL - rozmieszczanego przez operatora w stacjach bazowych, (2) A-UPF kończącego sesje PDU i zapewniającego dołączenie do sieci danych. Zasoby obliczeniowe w rozważanych systemach mogą być własnością operatora lub dostawcy usług MEC.

2.1 Strategie orkiestracji

Scenariusze współpracy między systemami MEC i 5G możemy podzielić ze względu na zakres współdzielenia zasobów obliczeniowych oraz sposób współpracy orkiestratorów. Zdefiniowano następujące strategie orkiestracji: S#1 niezależna orkiestracja, niezależne zasoby obliczeniowe – instancje A-UPF są wymiarowane i rozmieszczane przez operatora, bez wiedzy o aktualnym zapotrzebowaniu aplikacji MEC, orkiestratory nie współpracują; S#2 wspólna orkiestracja, niezależne zasoby obliczeniowe – wspólny orkiestrator podejmuje decyzje zarówno o alokacji aplikacji MEC, jak i o liczbie instancji A-UPF i ich rozmieszczeniu, z zastrzeżeniem, że instancje A-UPF mogą być umieszczane wyłącznie na zasobach operatorskich, a aplikacje wyłącznie na zasobach MEC; S#3 wspólna orkiestracja, współdzielone zasoby obliczeniowe – CNF mogą być alokowane na wszystkich dostępnych zasobach, orkiestrator posiada pełną wiedzę o stanie całego systemu.

2.2 Model systemu dla orkiestratora

Opisywane metody bazują na modelach wykorzystujących technikę mieszanego, liniowego programowania całkowitoliczbowego (*Mixed-Integer Linear Programming*).

Topologia systemu
Infrastrukturę sieciowo-chmurową zamodelowano jako skierowany graf $G(A, V, E)$, gdzie A, V reprezentują odpowiednio zbiór węzłów dostępowych oraz obliczeniowych, natomiast E reprezentuje zbiór połączeń pomiędzy węzłami obliczeniowymi. Każdy węzeł dostępowy $a_l \in A,\ l = 1, \ldots, |A|$ jest dołączony do węzła obliczeniowego $v_m \in V,\ m = 1, \ldots, |V|$. Przypisanie węzła dostępowego do węzła obliczeniowego jest opisane przez macierz $P_{A,Q}$, gdzie każdy element macierzy $p_{a,q} = [v_m, dist]$ jest wektorem opisującym połączenie danego węzła dostępowego do węzła obliczeniowego v_m, które scharakteryzowano odległością *dist*. Charakterystykę połączeń między węzłami obliczeniowymi $v1, v2 \in V$, określa macierz $D_{V,V}$. W rozważanym problemie orkiestracji, rezerwowane zasoby należą do dwuelementowego zbioru $R = \{CPU,\ BW\}$, który opisuje liczbę rdzeni obliczeniowych oraz przepustowość wirtualnego łącza wychodzącego, która uwzględnia wydajność implementacji warstwy przekazu danych w węźle. Zasoby węzłów obliczeniowych opisuje macierz $H_{V,R}$, natomiast jednostkowe koszty ich wykorzystania opisuje macierz $U_{V,R}$.

Deskryptor funkcji sieciowych i aplikacji
W rozważanym systemie są uruchamiane instancje aplikacji MEC i funkcji sieci 5G, nazywane dalej CNF (*Cloud Native Functions*). W zbiorze CNF, oznaczonym jako K, wyróżniono dwie kategorie: CNF reprezentujące aplikacje MEC oraz funkcję A-UPF. Ze względu na ich różne role w systemie, zdefiniowano podzbiór $K^- = K \setminus \{\text{A-UPF}\}$ zbioru K, który zawiera wyłącznie aplikacje MEC. Każda z funkcji CNF jest opisana przez macierze $\Pi_{K,R}$ oraz M_K określające odpowiednio ilość zasobów wymaganych do uruchomienia instancji danej CNF oraz opóźnienie propagacji tolerowane przez aplikację. Ponadto, zdolność obsługowa oferowana przez pojedynczą jej instancję opisana jest w macierzy M_K. Wartość $\mu_k \in M_K$ określa maksymalny wolumen ruchu, jaki może zostać obsłużony przez jedną instancję z zachowaniem wymaganej jakości obsługi. Dodatkowo, w modelu zdefiniowano macierz $S_{V,K}$ o wartościach binarnych, która określa czy w węźle $v_m \in V,\ m = 1, \ldots, |V|$ istnieje możliwość uruchomienia CNF typu $k_n \in K,\ n = 1, \ldots, |K|$. Macierz ta określa, które CNF mogą być uruchomione w ramach zasobów dostawcy MEC, a które w ramach zasobów operatora.

Napływ żądań użytkowników
W modelowanym systemie, w rozważanej chwili, do węzłów dostępowych $a_l \in A,\ l = 1, \ldots, |A|$ napływają żądania użytkowników dotyczące użycia danej aplikacji MEC. Liczbę strumieni danych związanych z aplikacją $k_n \in K^-,\ n = 1, \ldots, |K^-|$ i obsługiwanych przez węzeł dostępowy a_l opisuje macierz $L_{a,k}$.

Zmienne decyzyjne
Wynikiem działania algorytmu orkiestracji jest określenie: (a) wymaganej liczby instancji danego typu CNF, które należy uruchomić w poszczególnych węzłach oraz (b) rozpływu ruchu pomiędzy uruchomionymi instancjami aplikacji MEC, obsługującymi je instancjami funkcji A-UPF oraz węzłem dostępowym. Wartości te zawierają dwie zmienne decyzyjne opisujące stan systemu. Zmienna x_v^k opisuje liczbę instancji CNF typu $k_n \in K$, $n = 1, \ldots, |K|$, jakie należy uruchomić w węźle obliczeniowym $v_m \in V,\ m = 1, \ldots, |V|$. Zmienna $y_{v,u}^{a,k}$ opisuje rozpływ ruchu, tzn. określa jaka cześć ruchu napływającego

do węzła dostępowego $a_l \in A,\ l = 1, \ldots, |A|$ dotyczącego aplikacji k, jest obsługiwana przez instancje aplikacji MEC uruchomione w węźle obliczeniowym v_m oraz instancje A-UPF uruchomione w węźle $u_m \in V,\ m = 1, \ldots, |V|$.

Ograniczenia
Ograniczenie (1) zapewnia, że dla każdej aplikacji MEC, wolumen ruchu obsługiwany w danym węźle nie może przekroczyć sumarycznej zdolności obsługowej uruchomionych w nim instancji tej aplikacji.

$$\sum_{a \in A} \sum_{u \in V} y_{v,u}^{a,\ k} \ \leq\ x_v^k \ *\ \mu_k;\ \forall\, k\ \in K^-,\ \ \forall\, v\ \in V \qquad (1)$$

Ograniczenie (2) zapewnia, że wolumen ruchu z aplikacji MEC obsługiwanych przez funkcje A-UPF uruchomione w danym węźle nie może być większy niż zdolność obsługowa zaalokowanych w tym węźle funkcji A-UPF.

$$\sum_{a\ \epsilon\ A}\ \sum_{k\ \epsilon\ K^-}\ \sum_{v\ \epsilon\ V} y_{v,u}^{a,k} \ \leq\ x_u^{A-UPF} * \ \mu_{A-UPF};\ \ \forall\ u\ \epsilon\ V \qquad (2)$$

Ograniczenie (3) zapewnia, że nie mogą być przekroczone zasoby obliczeniowe węzłów.

$$\sum_{k\ \in K} x_v^k\ * \pi_{k,r}\ \leq\ H_{v,r};\ \ \forall\ v\ \in V,\ \ \forall\ r\ \in R \qquad (3)$$

Ograniczenie (4) wymusza uruchomienie instancji CNF jedynie w przeznaczonych dla nich węzłach.

$$x_v^k > 0\ \Leftrightarrow\ S_{v,k} = 1;\ \forall\ k\ \in K,\ \ \forall\ v\ \in V \qquad (4)$$

Ograniczenie (5) zapewnia, że nie zostanie przekroczona wartość opóźnienia tolerowana przez daną aplikację MEC. Wartość opóźnienia jest obliczana jako iloraz długości ścieżki, wyrażonej w km, oraz współczynnika szybkości propagacji $C_v \left[\frac{km}{ms}\right]$. W wyliczeniach uwzględniono wyłącznie czas propagacji, natomiast pomięto opóźnienia związane z przetwarzaniem pakietów.

$$C_v\ *\ (P_{a,Dist} + D_{P_{a,Node},u}\ + D_{v,u}\ ) \leq \tau_k$$
$$\forall\ a\ \epsilon\ A,\ \ \forall\ k\ \epsilon\ K^-,\ \ \forall\ v\ \epsilon\ V,\ \ \forall\ u\ \epsilon\ V;\ y_{v,u}^{a,\ k} > 0 \qquad (5)$$

Funkcja celu
Koszt uruchomienia instancji CNF typu k w węźle v jest wyznaczany jako iloczyn ilości zasobów r wykorzystywanych przez instancję k i kosztu jednostkowego specyficznego dla danego węzła v. Mnożąc ten iloczyn przez liczbę instancji (zmienna x) i sumując po wszystkich typach zasobów uzyskujemy koszt związany z alokacją wszystkich instancji k w węźle v. Sumaryczny koszt alokacji zasobów opisany jest funkcją f_1.

$$f_1(\ ) = \sum_{v \in V} \sum_{k\ \in K} \sum_{r \in R} x_v^k\ \ * \ \rho_{k,r}\ *\ U_{v,r} \qquad (6)$$

Obliczając koszty transmisji algorytm uwzględnia sumę odległości na ścieżce pomiędzy węzłem dostępowym, I-UPF, A-UPF i węzłem obliczeniowym, w którym działa instancja aplikacji. Założono, że funkcja I-UPF jest umieszczona w tym węźle obliczeniowym, do którego jest dołączony węzeł dostępowy. Koszty związane z transmisją na interfejsie radiowym pominięto. W celu zapewnienia spójności jednostek wprowadzono parametr $C_t \left[\frac{1}{Mbit\ *\ km}\right]$, rozumiany jako koszt przesłania 1 Mb danych na odległość 1 km. Sumaryczny koszt transmisji definiuje funkcja f_2.

$$f_2() = C_t * \sum_{a \in A} \sum_{k \in K^-} \sum_{v \in V} \sum_{u \in V} y_{v,u}^{a,k} \qquad (7)$$
$$* \pi_{k,BW} * (P_{a,Dist} + D_{P_{a,Node},u} + D_{u,v})$$

W nowoczesnych systemach dużą uwagę kładzie się na oszczędzanie energii. Minimalizując liczbę węzłów z uruchomionymi instancjami funkcji CNF minimalizujemy koszty działania systemu. Serwer, na którym nie są umieszczone jakiekolwiek CNF może pozostać w trybie niskiego poboru energii i oczekiwać na alokację. W funkcji celu uwzględniono sumaryczny koszt związany z uruchomieniem węzłów obliczeniowych opisany funkcją f_3 Wprowadzany jest parametr C_o będący jednostkowym kosztem uruchomienia węzła obliczeniowego.

$$f_3(\ ) = C_o * \sum_{v\ \epsilon\ V}\ \ F\left(\sum_{k\ \epsilon\ K} x_v^k\right)$$
$$gdzie\ F(z) = \begin{cases} 0\ dla\ z \leq 0 \\ 1\ dla\ z > 0 \end{cases} \qquad (8)$$

W funkcji celu uwzględniono również składnik związany z niepełnym zaspokojeniem żądań opisany funkcją f_4. Stopień niepełnego zaspokojenia żądań wyrażony jest jako różnica między wolumenem zaalokowanego ruchu a żądaniami. Uwzględnienie tego kryterium wymaga relaksacji ograniczenia (1), gdyż ograniczenie to wymusza zaspokojenie wszystkich żądań. Wolumen niezaspokojonych żądań jest mnożony przez współczynnik C_r będący karą za odrzucenie. Współczynnik musi być odpowiednio duży, by algorytm nie preferował odrzucania.

$$f_4(\ ) = C_r\ *\ \sum_{\substack{a \epsilon A \\ k \epsilon K^-}} \left[max\left(0, L_{a,k} - \sum_{\substack{v\ \epsilon\ V \\ u\ \epsilon\ V}} y_{v,u}^{a,k}\right)\right] \qquad (9)$$

Formalizacja algorytmu prowadzi do problemu wielokryterialnego ze sprzecznymi celami (minimalizacja kosztów transmisji oznacza umieszczanie zasobów bliżej brzegu sieci, co podnosi koszty alokacji). W celu przekształcenia modelu w jednokryterialny zastosowano sumę ważoną kryteriów. Przy założeniu, że w modelu uwzględniono Ω kryteriów, wprowadzono wagi γ takie, że:

$$\sum_i \gamma_i\ \ = 1;\ i\ =\ 1,\ 2,\ \ldots, \Omega \qquad (10)$$

Znormalizowanie sumy wag oraz odpowiedni dobór ich wartości pozwala wpływać na proces optymalizacji i zapewniać kompromis między kryteriami. Funkcję celu dla dowolnej liczby kryteriów można więc zdefiniować jako liniową kombinację kryteriów i zapisać następująco:

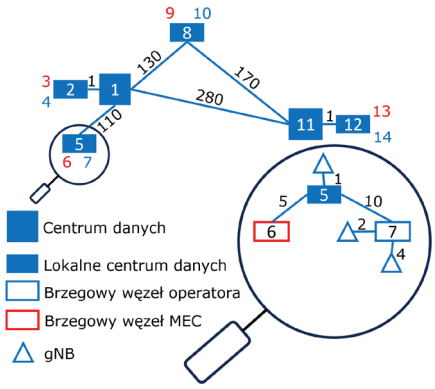
$$minimize\ f(\ )\ = \sum_i \gamma_i * f_i();\ i\ =\ 1,\ 2, \ldots, \Omega \qquad (11)$$

$s.t:\ (1) - (5)$

Model pozwala na realizację opisanych wcześniej strategii orkiestracji poprzez odpowiednie dane wejściowe, tj. wskazanie jakie CNF mogą być alokowane na jakich węzłach oraz, w przypadku S#1, wskazanie rozmieszczenia A-UPF. Model został zaimplementowany w języku MiniZinc oraz rozwiązany z wykorzystaniem solvera HiGHS (*High-performance parallel linear optimization software*).

3. EKSPERYMENTY

Celem przedstawionego eksperymentu jest porównanie efektywności proponowanych strategii orkiestracji. Badania przeprowadzono w warunkach dużego obciążenia systemu, na granicy występowania strat. Do strat, czyli niezaspokojenia części żądań, może dochodzić z powodu: niespełnienia wymagań na opóźnienie, braku dostępnych zasobów sieciowych lub obliczeniowych, lub przeciążenia funkcji A-UPF. Przyjęto równe wagi $\gamma_i = 0.25,\ i = 1, 2, 3, 4$. Testy przeprowadzono w systemie przedstawionym na Rysunku 2, złożonym z 14 węzłów obliczeniowych $(v_1 - v_{14})$ i 12 węzłów dostępowych $(a_1 - a_{12})$, reprezentującym średniego krajowego operatora. W systemie wyróżniono 4 rodzaje węzłów obliczeniowych: (1) centrum danych, (2) lokalne centrum danych, (3) brzegowy węzeł operatora, (4) brzegowy węzeł MEC. Do każdego węzła typu (2) jest dołączony jeden węzeł dostępowy (gNB), a do każdego węzła (3) są dołączone po 2 gNB. Topologia



Rysunek 2: Topologia testowanego systemu

zakłada istnienie operatorskich centrów danych w Warszawie i Poznaniu. Założono, że w węzłach (1) znajduje się po 10000 jednostek zasobów i kosztują one 1 jednostkę kosztu, w węzłach (2) znajduje się po 1000 jednostek zasobów i kosztują one 5 jednostek, a węzłach (3) i (4) znajduje się po 100 jednostek zasobów i kosztują one 10 jednostek. W typowych wdrożeniach koszty te są najmniejsze w dużych centrach danych ze względu na duży stopień agregacji zasobów oraz tym wyższe, im serwery obliczeniowe są umieszczone bliżej brzegu sieci. W systemie występują 3

Tabela 1: Deskryptory rozważanych CNF

CNF	Wymagane zasoby $\Pi_{K,R}$		Tolerowane opóźnienie	Zdolność obsługowa	Napływ żądań $\forall\ a \in A$
	CPU	BW	τ_k $[ms]$	μ_k $[req]$	$L_{a,k}[req]$
k_1	10	10	-	100	-
k_2	6	6	1	1	5
k_3	2	4	10	100	100

rodzaje CNF, różniące się znacząco wymaganiami: funkcja A-UPF jest reprezentowana jako aplikacja k_1, aplikacja MEC App1 (k_2) wymaga dużo zasobów i toleruje jedynie małe opóźnienia rzędu 1 ms oraz aplikacja MEC App2 (k_3) o umiarkowanych wymaganiach. Deskryptory aplikacji przedstawiono w Tabeli 1. Założono, że do wszystkich węzłów dostępowych napływa po 5 żądań dot. k_2 i 100 żądań dot. k_3, co łącznie daje 1260 żądań. Wartość ta pozwoliła na wprowadzenie systemu w stan dużego obciążenia. Dane te wykorzystano również do wyznaczenia wymaganej liczby A-UPF (13 instancji) w strategii S#1.

Tabela 2: Porównanie efektywności strategii

Wartość	*S#1*	*S#2*	*S#3*	*max S#3*
f_1	8 580	9 120	5 080	90 880
f_2	229 076	184 812	16 123	45 758 328
f_3	10 000	10 000	8 000	14 000
f_4	200 000	200 000	0	0
ψ [%]	0.16	0.16	0.0	0.0
F. celu	111 914	100 983	7 301	11 465 802

Wyniki eksperymentu przedstawiono w Tabelach 2 i 3. Analizując je stwierdzono, że w przypadku strategii S#1 i S#2 żądania dot. aplikacji k_3 zostały odrzucone ze względu na brak zasobów. Ponadto, niezależna orkiestracja założona w strategii S#1, prowadzi do wyższych kosztów transmisji ze względu na niedopasowaną lokalizację A-UPF do zaalokowanych aplikacji. Umożliwienie orkiestratorowi dopasowanie liczby i lokalizacji instancji A-UPF w ramach S#2, pozwoliło na zredukowanie tego zjawiska, kosztem uruchomienia większej liczby A-UPF. Zauważmy, że S#2 przy tym samym ruchu zaalokowała o 3 instancje A-UPF więcej niż S#1 (16 vs. 13). Dodatkowe instancje A-UPF pozwoliły na redukcję kosztów transmisji. O niedopasowaniu świadczy też występowanie strat. Wdrożenie strategii wspólnej orkiestracji S#3 pozwala na

Tabela 3: Rozmieszczenie instancji CNF w węzłach

W	*S#1*			*S#2*			*S#3*			*max load S#3*		
#	k_1	k_2	k_3	k_1	k_2	k_3	k_1	k_2	k_3	k_1	k_2	k_3
v_1	3	0	0	1	0	0	1	14	1	637	187	627
v_2	2	0	0	4	0	0	0	0	0	69	0	77
v_3	0	14	4	0	14	4	0	0	0	6	2	7
v_4	0	0	0	0	0	0	3	1	2	7	0	7
v_5	2	0	0	3	0	0	2	15	1	1	165	0
v_6	0	15	2	0	15	2	0	0	0	0	16	1
v_7	0	0	0	0	0	0	2	0	2	4	8	3
v_8	2	0	0	3	0	0	32	15	1	1	165	0
v_9	0	15	2	0	15	2	0	0	0	0	16	1
v_{10}	0	0	0	0	0	0	2	0	2	4	8	3
v_{11}	2	0	0	1	0	0	12	15	1	657	173	598
v_{12}	2	0	0	4	0	0	0	0	0	55	0	112
v_{13}	0	14	4	0	14	4	0	0	0	0	16	0
v_{14}	0	0	0	0	0	0	2	0	2	7	0	7

zmniejszenie zużycia zasobów i wyeliminowanie strat, co pozwala na zwiększenie obsługiwanego ruchu. Granicę obciążenia dla S#3 stanowią 63 żądania aplikacji k_2 per węzeł i 12000 żądań aplikacji k_3. Wyniki dla tego przypadku zamieszczono w ostatniej kolumnie Tabel 2 i 3. Strategie S#1-3 mogą obsługiwać aplikacje wymagające bardzo małego opóźnienia, przy czym: ograniczeniem dla S#1 jest poprawne zwymiarowanie przez operatora wydajności A-UPF na brzegu sieci, dla S#2 ograniczeniem jest ilość zasobów MEC dostępnych na brzegu sieci. Ze względu na swój charakter S#1 nie jest odporna na przeciążenia A-UPF – brak skalowania tych funkcji prowadzi do odrzucania żądań. Strategie S#2-3 w przypadku przeciążenia uruchamiają dodatkowe instancje A-UPF. Czas obliczeń zawierał się w przedziale od 2s (S#1) do 30s (S#3).

4. PODSUMOWANIE

Przeprowadzone eksperymenty pozwoliły ocenić efektywność proponowanych strategii orkiestracji. Pokazano, że aby uruchamiać w aplikacje wymagające bardzo małego opóźnienia niezbędne jest uruchamianie instancji funkcji A-UPF na brzegu sieci. W przypadku takich aplikacji znane z poprzednich generacji sieci podejście z bramą sieci pakietowej umieszczoną w centrum danych razem z siecią rdzeniową jest niepraktyczne. Samo rozmieszczenie A-UPF na brzegu nie będzie jednak wystarczające w sytuacji dużego obciążenia systemu. Ze względu na ograniczoną zdolność obsługową pojedynczej instancji A-UPF jest niezbędne horyzontalne skalowanie wydajności tej funkcji sieciowej, tak by możliwe było obsłużenie większej liczby jednoczesnych połączeń. Uzyskane wyniki wskazują, że orkiestracja jest efektywniejsza w sytuacji, w której wszystkie zasoby są traktowane wspólnie, a orkiestrator posiada pełną wiedzę o całym systemie. Oferowane zasoby są wtedy wykorzystywane efektywnie, przy jednoczesnym zachowaniu niskich kosztów. Uzyskane wnioski wskazują na potencjał zintegrowanych systemów sieciowo-chmurowych i wskazują jeden z kierunków dalszego rozwoju techniki 5G, a także przyszłych sieci 6G.

LITERATURA

[1] Y. Li, A. Zhou, X. Ma, S. Wang. 2022. "Profit-aware edge server placement". IEEE Internet of Things Journal 9 (1) : 55–67.

[2] I. Leyva-Pupo, C. Cervelló-Pastor. 2022. "Efficient solutions to the placement and chaining problem of userplane functions in 5g networks" Journal of Network and Computer Applications 197 : 103269.

[3] I. Leyva-Pupo, C. Cervelló-Pastor, C. Anagnostopoulos, D. P. Pezaros. 2022. "Dynamic upf placement and chaining reconfiguration in 5g networks" Computer Networks 215 : 109200.

[4] M. Laroui, H. I. Khedher, H. Moungla, H. Afifi. 2023. "Service function chains multi-resource orchestration in virtual mobile edge computing". Computer Networks, 224 : 109582.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

ALGORYTMY SZTUCZNEJ INTELIGENCJI W PRZETWARZANIU DANYCH ROZPOZNANIA RADIOELEKTRONICZNEGO ELINT

ARTIFICIAL INTELLIGENCE ALGORITHMS IN ELINT DATA PROCESSING

dr inż. Łukasz Rybak[1], dr hab. inż. Janusz Dudczyk, prof. WAT[2], inż. Piotr Mazur[3]

[1] Wojskowa Akademia Techniczna im. Jarosława Dąbrowskiego, Warszawa, lukasz.rybak@wat.edu.pl
[2] Wojskowa Akademia Techniczna im. Jarosława Dąbrowskiego, Warszawa, janusz.dudczyk@wat.edu.pl
[3] Wojskowa Akademia Techniczna im. Jarosława Dąbrowskiego, Warszawa, piotr.mazur03@student.wat.edu.pl

DOI: 10.15199/59.2024.4.23

Streszczenie: **Artykuł ukazuje znaczenie sztucznej inteligencji AI dla rozpoznania radioelektronicznego ELINT. Celem było przedstawienie procesu przetwarzania danych ELINT ze wskazaniem potencjału jego automatyzacji przez zastosowanie nadzorowanych i nienadzorowanych algorytmów uczenia maszynowego ML. Scharakteryzowano etapy procesu przetwarzania danych ELINT i wybrane algorytmy klasyfikacji i grupowania. Wyniki wykazały potencjał aplikacyjny poszczególnych algorytmów AI w różnych problemach z zakresu przetwarzania danych ELINT.**

Abstract: **The article shows the significance of artificial intelligence (AI) for electronic intelligence (ELINT). The aim was to present the ELINT data processing and indicate the potential for its automation through the use of supervised and unsupervised machine learning algorithms. The stages of the ELINT data processing and selected classification and clustering algorithms were characterized. The results showed the application potential of particular AI algorithms in various problems in the field of ELINT data processing.**

Słowa kluczowe: **algorytmy, ELINT, rozpoznanie radioelektroniczne, sztuczna inteligencja, uczenie maszynowe**

Keywords: **algorithms, artificial intelligence, electronic intelligence, ELINT, machine learning**

1. WPROWADZENIE

Od kilkunastu lat rozwój sztucznej inteligencji AI (ang. *Artificial Intelligence*) dostarcza praktycznych rozwiązań stosowanych w otaczającej człowieka rzeczywistości. Kontekstem rozważań, opisanych w artykule, dotyczących użyteczności algorytmów AI, jest sektor militarny.

AI należy postrzegać jako umiejętność systemu, dzięki której potrafi on poprawnie interpretować zgromadzone dane, uczyć się na ich podstawie, adaptować do aktualnych warunków, a przetworzone informacje wykorzystać w procesie realizacji założonych celów. Istnieją dwa obszary AI - uczenie maszynowe ML (ang. *Machine Learning*) i uczenie głębokie DL (ang. *Deep Learning*). Wspomniane dziedziny koncentrują się na aspekcie zastosowania zdobytego przez system doświadczenia, do poprawy jego efektywności w ponownym wykorzystaniu do rozwiązania tego samego lub podobnego problemu [1]. W zbiorze algorytmów ML istnieją następujące podzbiory metod: uczenie nadzorowane, uczenie nienadzorowane, uczenie częściowo-nadzorowane i uczenie ze wzmocnieniem. W artykule poddano analizie potencjał zastosowania wybranych metod uczenia nadzorowanego i nienadzorowanego w procesie automatyzacji przetwarzania danych ELINT (ang. *Electronic Intelligence*). Cechą dystynktywną algorytmów nadzorowanych jest fakt, że proces rozpoznawania oparty jest na wiedzy *a priori*. Oznacza to, że uczenie się bazuje na pewnym zbiorze treningowym, w którym każdy znajdujący się obiekt jest opisany cechami i ma przypisaną etykietę, na podstawie której system w przyszłości będzie w stanie klasyfikować nieznane obiekty. Podejścia w ramach uczenia nienadzorowanego przetwarzają zbiory danych, w których obiekty są opisane cechami, ale nie są opisane etykietami. Celem nienadzorowanych algorytmów uczenia maszynowego jest rozpoznawanie wzorców i korelacji występujących w przetwarzanych danych, najczęściej prowadzące do tworzenia maksymalnie homogenicznych grup obiektów [2].

W mariażu AI z dziedziną walki radioelektronicznej EW (ang. *Electronic Warfare*) na uwagę zasługują urządzenia oraz systemy rozpoznawania sygnałów radarowych. Istotnym ich komponentem są algorytmy klasyfikacji i grupowania danych, stosowane w ramach prowadzenia rozpoznania radioelektronicznego. Wymienione instrumenty oferują szerokie spektrum wyspecjalizowanych funkcjonalności z zakresu C2I (ang. *Command, Control and Intelligence*) [3].

Celem artykułu było przedstawienie procesu przetwarzania danych rozpoznania radioelektronicznego ze wskazaniem potencjału automatyzacji jego czynności przez zastosowanie nadzorowanych i nienadzorowanych algorytmów ML

2. PROCES PRZETWARZANIA SYGNAŁÓW RADAROWYCH

Przetwarzanie sygnałów radarowych jest stosowane w budowaniu przewagi informacyjnej na współczesnym polu walki [4]. Obecnie skomplikowanie przetwarzania danych ELINT rośnie w postępie geometrycznym. Jest to pokłosiem występującego na współczesnym polu walki bogactwa tła elektromagnetycznego, nasycenia obszaru działań mnogością źródeł emisji elektromagnetycznej

w bardzo szerokim zakresie widmowym oraz zaawansowaną i dynamiczną charakterystyką pracy radarów [5].

Dostępne na rynku urządzenia do akwizycji sygnałów radarowych i ekstrakcji ich cech bazują na analizie widmowej i czasowej odbieranej przestrzeni elektromagnetycznej. W podstawowym ujęciu produktem ich działania jest widmo amplitudowe oraz zestaw wektorów cech PDW (ang. *Pulse Description Word*), z których każdy stanowi zbiór wartości wyrażających wyekstrahowane parametry pojedynczego impulsu elektromagnetycznego [6]. W kontekście wiodącej problematyki artykułu, do relewantnych składowych PDW można zaliczyć: amplitudę (AMP), czas trwania (PD), częstotliwość nośną (RF) i czas przybycia (TOA).

Manualne przetwarzanie danych ELINT, które prowadzi do rozplatania sygnałów radarowych nie jest procesem trywialnym, który w łatwy sposób może zostać poddany automatyzacji. Przedmiotowa problematyka obejmuje analizę serii danych opisujących zarejestrowane widmo amplitudowe oraz wielowymiarowe przetwarzanie danych opisujących parametry impulsów w dziedzinie czasu ich nadejścia. Ze względu na często bardzo dużą liczebność takich zbiorów danych prowadzone są również analizy histogramów relewantnych cech impulsów. W zaawansowanym przetwarzaniu korelowane są ze sobą poszczególne parametry techniczne impulsów, tworząc wielowymiarowe przestrzenie cech [7].

Automatyzacja procesu przetwarzania sygnałów radarowych, w szczególności w celu zgrupowania impulsów, sklasyfikowania lub zidentyfikowania wykrytych radarów wymaga zastosowania standaryzacji wartości cech. W przetworzonej przestrzeni cech, celem grupowania impulsów zarejestrowanych sygnałów jest skonstruowanie maksymalnie homogenicznych grup. Taka strategia maksymalizuje prawdopodobieństwo, że impulsy, należące do grupy posiadającej tę cechę, zostały wyemitowane przez ten sam radar. Immanentną cechą tworzenia homogenicznych grup impulsów jest minimalizacja liczby wartości odstających, co w kontekście analizy sygnałów jest tożsame z usuwaniem szumu. W przypadku radarów stosowanych we współczesnej walce radioelektronicznej, które do przeciwdziałania zakłócaniu wykorzystują rozpraszanie widma poprzez zmiany częstotliwości FHSS (ang. *Frequency-hopping Spread Spectrum*), modulację czasu trwania impulsu PDM (ang. *Pulse-duration Modulation*) i zmianę okresu powtarzania impulsów PRIM (ang. *Pulse-repetition Interval Modulation*) wydajny algorytm grupowania może pozwolić na ekstrakcję trybów pracy takich źródeł emisji [8]. Finalnym celem procesu automatycznego przetwarzania danych ELINT jest podjęcie decyzji o typie źródła emisji lub identyfikacja, czyli wskazanie egzemplarza radaru [7], na podstawie referencyjnej bazy danych w czasie quasi-rzeczywistym. W zagadnieniu klasyfikacji i identyfikacji źródeł emisji istotne jest również określenie stopnia pewności decyzji.

3. ALGORYTMY AI W PRZETWARZANIU DANYCH ELINT

Pierwszą z analizowanych, w ramach prac badawczych opisanych w tym artykule, grup algorytmów sztucznej inteligencji są klasyfikatory. Należą one do nadzorowanych metod uczenia maszynowego. Każdy z algorytmów klasyfikacji cechuje się unikalnymi właściwościami do wyodrębnienia informacji z przetwarzanych danych oraz może być stosowany w jednym lub wielu rodzajach klasyfikacji: binarnej, wieloklasowej czy wieloetykietowej. Powszechnie są one znane z zastosowania w rozpoznawaniu obrazów, rozpoznawaniu mowy, wykrywaniu spamu, diagnostyce medycznej, klasyfikacji tekstu czy wykrywaniu oszustw internetowych. W kontekście prowadzenia walki radioelektronicznej klasyfikatory mogą zostać zastosowane w procesie rozpoznawania sygnałów radarowych – na różnych poziomach analizy wektorów cech: opisujących pojedyncze impulsy lub pojedyncze typy źródeł emisji, a w szczególnych przypadkach egzemplarze radarów.

Pierwszym z wybranych klasyfikatorów był *k*-NN (ang. *k–Nearest Neighbors*). Ten klasyfikator jest jednym z najstarszych oraz najprostszych podejść do rozpoznawania wzorców. Założeniem jest, że dane są reprezentowane w wielowymiarowej przestrzeni, której każdy z wymiarów odpowiada poszczególnym atrybutom danych. Ideą tego algorytmu jest rozpoznanie obiektu na podstawie porównania wektora cech opisującego nową próbkę ze wszystkimi referencyjnymi wektorami cech próbek treningowych. Rezultat jego działania powstaje w drodze obliczenia dystansu, zgodnie z zastosowaną przestrzenią metryczną, pomiędzy rozpoznawanym obiektem a wszystkimi znanymi obiektami zgromadzonymi w referencyjnej bazie danych. W procesie określenia etykiety klasyfikowanej próbki używane są etykiety *k* najbliżej niej położnych w przestrzeni elementów. W stosowanej metryce możliwe jest również zróżnicowanie priorytetu poszczególnych atrybutów poprzez zastosowanie wag cech. Dobór wartości parametru *k* zależy od charakterystyki przetwarzanego zbioru danych. Zwiększenie wartości parametru *k* zmniejsza wpływ szumu na wynik klasyfikacji. Natomiast zmniejszenie wartości parametru *k* prowadzi do większego dopasowania granicy decyzyjnej do danych treningowych [9].

Kolejnym z wybranych nadzorowanych algorytmów uczenia maszynowego był naiwny klasyfikator Bayesa. Działa on w oparciu o model warunkowy prawdopodobieństwa wyprowadzony z twierdzenia Bayesa. Założeniem opisywanego podejścia jest fakt, że analizowane cechy są od siebie niezależne, dlatego nazywane jest ono „naiwnym". Dla atrybutów dyskretnych na podstawie zbioru danych treningowych obliczane jest warunkowe prawdopodobieństwo wystąpienia każdej z dostępnych dyskretnych wartości dla każdej z istniejących etykiet (klas). Natomiast w przypadku cech ciągłych przyjmowane jest założenie, że dla poszczególnych klas, rozkład ich wartości jest zbliżony do rozkładu normalnego. Wówczas wartość prawdopodobieństwa warunkowego jest wyznaczana w oparciu o funkcję gęstości prawdopodobieństwa. Na podstawie tak zbudowanego modelu predykcyjnego algorytm, klasyfikując nowy obiekt, wybiera klasę, dla której wektor cech opisujących przedmiotowy element uzyskał najwyższe prawdopodobieństwo. Klasyfikator ten jest znany ze swojej prostoty, szybkości i wysokiej skuteczności predykcji [10].

Drugą z grup metod sztucznej inteligencji poddanych analizie, której rezultaty przedstawiono w tym arty-

kule, są algorytmy grupowania. Aplikacyjność tych nienadzorowanych metod uczenia maszynowego w procesie prowadzenia walki radioelektronicznej może dotyczyć wyszukiwania grup podobnych do siebie impulsów, co może być użyteczne w procesie rozplatania sygnałów pochodzących z różnych radarów. Dysponując danymi opisującymi nie pojedyncze impulsy, a źródła emisji, metody grupowania mogą posłużyć wyszukaniu typów radarów lub ich egzemplarzy, których charakterystyka pracy jest do siebie podobna.

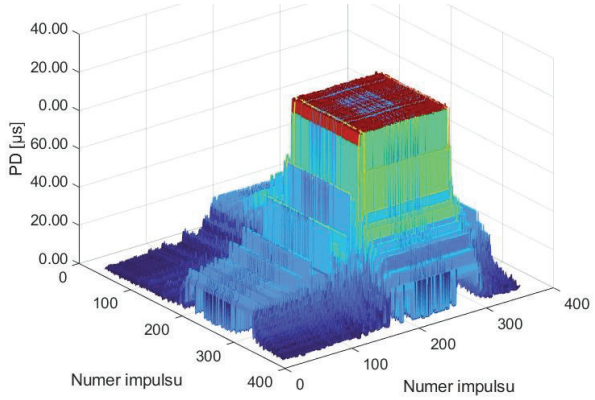
Pierwszym wybranym z tego zbioru był algorytm DBSCAN (ang. *Density Based Spatial Clustering of Applications with Noise*). Został on zaproponowany w 1996 jako algorytm grupowania baz danych przestrzennych zawierających szum. Oparty jest on na ekstrakcji obszarów przestrzeni cech o skrajnych gęstościach. Ideą tej metody grupowania jest wyszukanie w przestrzeni cech regionów o wysokiej oraz niskiej gęstości, w taki sposób, że obszary o niskiej gęstości, które zawierają punkty szumu są odseparowane od regionów o wysokiej gęstości. Cechą DBSCAN jest jego zdolność do wykrywania grup, w których rozkład obiektów składowych tworzy różne kształty w przestrzeni cech. Kolejną z cech opisywanego podejścia jest jego wysoka efektywność w usuwaniu szumu. Ostatnią z istotnych zalet algorytmu DBSCAN jest brak konieczności definiowania liczby grup, które mają zostać utworzone – jej wartość wynika bezpośrednio ze struktury zbioru danych oraz parametryzacji metody. W kontekście ostatniej wspomnianej kwestii, do parametrów charakteryzowanego podejścia należy ε, oznaczający promień koła, w którym poszukiwany jest region gęsty. Natomiast drugim z parametrów jest *minPts*, który określa, wymaganą do utworzenia regionu gęstego, liczbę obiektów należących do koła. DBSCAN jest najczęściej używanym algorytmem grupowania opartym na analizie gęstości obszarów przestrzeni cech [11].

Ostatnim z wybranych nienadzorowanych algorytmów uczenia maszynowego był opracowany w 1999 roku OPTICS (ang. *Ordering Points To Identify the Clustering Structure*). OPTICS, podobnie jak podejście DBSCAN, należy do zbioru metod grupowania opierających się na analizie gęstości regionów przetwarzanej przestrzeni cech. W związku z tym zestaw parametrów niniejszego algorytmu jest tożsamy z ich zbiorem stosowanym w metodzie DBSCAN. Pierwszy z parametrów - ε określa promień koła, w którym poszukiwany jest region gęsty. Natomiast drugim parametrem jest *minPts*, który definiuje minimalną liczbę obiektów należących do koła, wymaganą do utworzenia regionu gęstego. Algorytm OPTICS charakteryzuje się dobrą wydajnością w przetwarzaniu przestrzeni cech, w których istnieją grupy obiektów o różnych gęstościach. Idea przedmiotowego podejścia opiera się na zdefiniowaniu struktury klastrowej opartej na analizie gęstości obiektów znajdujących się w przestrzeni cech, utworzonej na podstawie bazy danych. Powstały w ten sposób opis „osiągalności" każdego z obiektów jest stosowany do utworzenia grup elementów, w taki sposób, że byty o zbliżonej „osiągalności" są przypisywane do tej samej grupy. W rezultacie działania algorytmu powstają maksymalnie homogeniczne grupy o zróżnicowanej gęstości [12].

4. WYNIKI

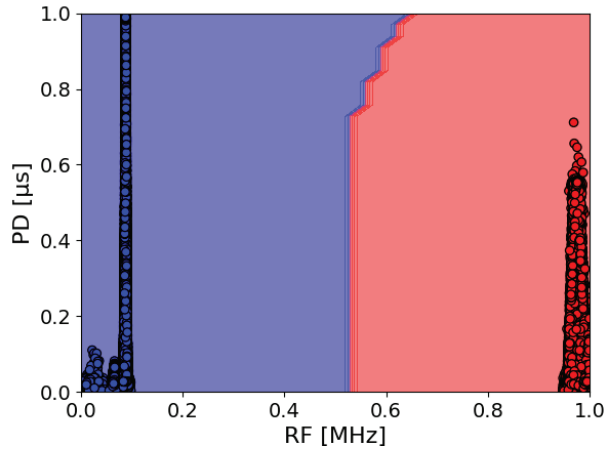
W badaniach zastosowano zbiór danych, który zawierał 641 942 impulsów radaru #1 oraz 1 592 400 impulsów radaru #2. Radar #1 cechował się pracą na jednej częstotliwości z jednym czasem trwania impulsów. Natomiast radar #2 był opisany jedną częstotliwością i dwoma czasami trwania impulsów.

Na Rys. 1. zobrazowano fragment rejestracji przedstawiający wartość PD kolejnych impulsów, pochodzących ze źródła emisji #1. Widoczne maksimum zawiera informację o PD relewantnych impulsów pochodzących z tego radaru.



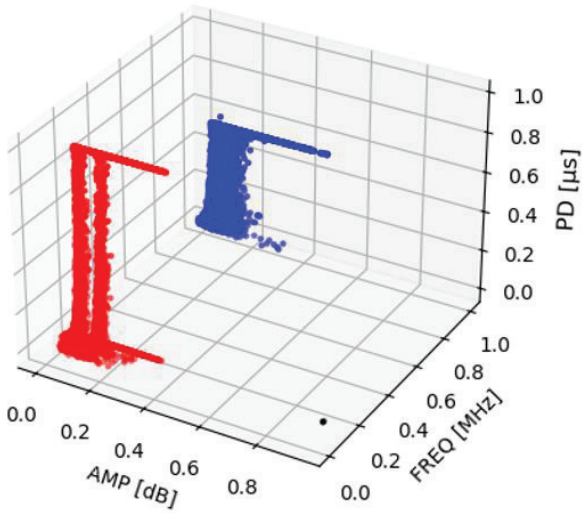
Rys. 1. Wartość PD kolejnych impulsów radaru #1
Źródło: opracowanie własne

Rys. 2 przedstawia rezultat działania algorytmu *k*-NN, w którym wartość parametru $k = 1495$. Dla użytego zbioru danych, metoda *k*-NN utworzyła granicę decyzyjną umożliwiającą skuteczne rozpoznanie źródła emisji impulsów. Na zastosowanym zbiorze danych naiwny klasyfikator Bayesa uzyskał podobną skuteczność.



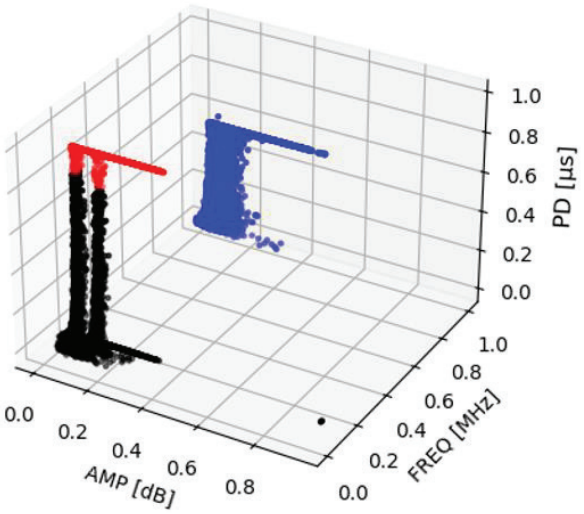
Rys. 2. Obszary decyzyjne algorytmu k-NN w znormalizowanej przestrzeni cech
Źródło: opracowanie własne

Na Rys. 3. przedstawiono wynik działania algorytmu grupowania DBSCAN, w którym *minPts* = 200, a $\varepsilon = 0.5$. Metoda wyekstrahowała trzy grupy impulsów - punkty niebieskie (radar #1) oraz punkty czerwone (radar #2). DBSCAN oznaczył również impulsy niepasujące do żadnej z grup (szum) - punkty czarne. Można również zaobserwować, że impulsy o wyraźnie różnym PD, opisujące radar #2, zostały sklasyfikowane do jednej grupy.



Rys. 3. Grupy utworzone przez algorytm DBSCAN w znormalizowanej przestrzeni cech
Źródło: opracowanie własne

Rys. 4. obrazuje wynik działania algorytmu OPTICS, który został sparametryzowany w identyczny sposób, jak DBSCAN. Podejście utworzyło trzy grupy – punkty niebieskie (radar #1), punkty czerwone (radar #2) oraz punkty czarne (szum). W porównaniu do podejścia DBSCAN, OPTICS część istotnych impulsów pochodzących z radaru #2 uznał za szum, co reprezentuje duża liczba czarnych punktów. Podział ostatniej z wymienionych grup punktów mógłby ostatecznie doprowadzić do ekstrakcji dwóch grup impulsów, pochodzących z radaru #2.



Rys. 4. Grupy utworzone przez algorytm OPTICS w znormalizowanej przestrzeni cech
Źródło: opracowanie własne

5. PODSUMOWANIE

W artykule przedstawiono proces przetwarzania danych ELINT oraz opisano wybrane nadzorowane i nienadzorowane algorytmy ML. Badania wykonano na rzeczywistym zbiorze danych impulsów pochodzących z dwóch różnych typów radarów. Eksperymenty wykazały, że istnieje potencjał automatyzacji procesu przetwarzania danych ELINT z zastosowaniem algorytmów AI. Nadzorowane algorytmy ML umożliwiają rozpoznanie typu radaru, z którego pochodzi impuls. Natomiast algorytmy nienadzorowane charakteryzują się efektywnością w grupowaniu sygnałów pochodzących z tego samego radaru.

Kierunkiem dalszych badań może być porównanie różnych algorytmów AI na rzeczywistych zbiorach danych zawierających impulsy, pochodzące z większej liczby radarów.

LITERATURA

[1] Kaplan Andreas, Haenlein Michael. 2019. “Siri, Siri, in my hand: Who’s the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence”. Business Horizons, 62 (1) : 15-25.

[2] Xin Yang i in. 2018. „Machine learning and deep learning methods for cybersecurity”. IEEE Access, 6 : 35365-35381.

[3] Wiley Richard. 2006. „ELINT: „The interception and analysis of radar signals”. Artech House Publishers.

[4] Rybak Łukasz, Dudczyk Janusz. 2019. „Increasing the information superiority on the modern battlefield through the use of virtual reality systems”. Security and Defence Quarterly, 25 (3) : 86-98.

[5] Matuszewski Jan. 2021. „Rozpoznawanie sygnałów radarowych z wykorzystaniem sieci neuronowych”. Przegląd Elektrotechnicznych, 10 : 125-128.

[6] Iglesias Victor i in. 2014. “Real-time radar pulse parameter extractor”. W: 2014 IEEE Radar Conference, Cincinnati, OH, USA, 371-375.

[7] Dudczyk Janusz, Rybak Łukasz. 2023. “Application of data particle geometrical divide algorithms in the process of radar signal recognition”. Sensors, 23 (19).

[8] Kang Shiqian, Wang Cong, Liu Zhangmeng. 2021. “Pulse Group Extraction of Multi-Function Radar”. W: 2021 CIE International Conference on Radar (Radar), Haikou, Hainan, Chiny, 2322-2325.

[9] Guo Gongde i in. 2003. „KNN Model-Based Approach in Classification”. Lecture Notes in Computer Science, 2888 : 986-996.

[10] Yang Zhiwei i in. 2023. "A New Three-Way Incremental Naive Bayes Classifier". Electronics, 12 (7).

[11] Ester Martin i in. 1996. “A density-based algorithm for discovering clusters in large spatial databases with noise”. W: Second International Conference on Knowledge Discovery and Data Mining, Portland, OG, USA, 226-231.

[12] Ankerst Mihael i in. 1999. “OPTICS: ordering points to identify the clustering structure”. W: 1999 ACM SIGMOD international conference on Management of data, 49-60.

GŁĘBOKIE UCZENIE DO KOREKCJI FAZY SYGNAŁÓW GMSK W RZECZYWISTYM ŚRODOWISKU WEWNĄTRZBUDYNKOWYM

DEEP LEARNING FOR PHASE CORRECTION OF GMSK SIGNALS IN REAL INDOOR ENVIRONMENT

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: 10.15199/59.2024.4.24

Streszczenie: **Niniejszy artykuł prezentuje zastosowanie modelu głębokiej sieci neuronowej do estymacji średniego odchylenia fazy sygnałów odebranych i jest elementem badań obejmujących szersze zagadnienie, jakim jest odbiór sygnałów GMSK wspomagany uczeniem maszynowym. Analiza pozwoliła potwierdzić wysoką skuteczność sieci neuronowej, a wyniki obejmowały kanały ETU i EPA oraz dane pomiarowe zebrane w rzeczywistym środowisku wewnątrzbudynkowym. Rozwiązanie porównano z klasycznymi metodami estymacji średniego odchylenia fazy.**

Abstract: **The given paper presents the application of a deep neural network for phase rotation estimation and is part of a study covering the wider issue of GMSK signal reception based on machine learning. The analysis confirmed the high efficiency of the neural network, including ETU and EPA channels and measurement data collected in a real indoor environment. The proposed solution was compared with classical methods for the estimation of mean phase rotation.**

Słowa kluczowe: **głębokie uczenie, GMSK, GSM, uczenie maszynowe.**

Keywords: **deep learning, GMSK, GSM, machine learning.**

1. WSTĘP

Zainteresowanie metodami uczenia maszynowego ML (ang. machine learning) zapoczątkowane w obszarach przetwarzania dźwięku, obrazu czy języka naturalnego coraz szerzej obejmuje również inne dziedziny naukowo-technologiczne, w tym szeroko rozumianą radiokomunikację. Jednym z zagadnień podejmowanych w ramach licznych publikacji naukowych jest odbiór sygnałów radiowych – wspomagany lub w pełni realizowany przy użyciu sztucznej inteligencji.

Problem odbioru radiowego podejmowany jest zazwyczaj na jeden z trzech sposobów. Pierwszym jest tzw. deep unfolding, w którym to iteracyjne algorytmy przetwarzania sygnałów w warstwie fizycznej są osobno optymalizowane w ramach struktur przypominających warstwy sieci neuronowej. W [5] autorzy zaprezentowali metodę deep unfolding wspomaganą RIS (ang. Reconfigurable Intelligent Surface) w celu detekcji sygnałów SM/SSK (ang. Spatial Modulation/Space Shift Keying). Analiza obejmowała kanały zmienne i niezmienne czasowo, a końcowa skuteczność porównana została z klasycznym modelem ML.

Drugim ze sposobów jest podejście „black box" łączące wszystkie funkcje toru odbiorczego lub nawet torów nadawczego i odbiorczego równocześnie, w ramach jednej (lub kilku połączonych) sieci neuronowej. W [10] zaproponowano głęboką sieć DNN (ang. Deep Neural Network) zastępującą w pełni odbiornik OFDM (ang. Orthogonal Frequency-Division Multiplexing). Analiza skuteczności obejmowała charakterystyki BER (ang. Bit Error Rate) w zależności od występowania w sygnale przedrostka cyklicznego oraz liczby symboli pilotowych rozmieszczonych w siatce zasobów OFDM. System realizujący zarówno część nadawczą, jak i odbiorczą przedstawiono w [11] – symetrycznie rozmieszczone warstwy pojedynczej sieci neuronowej odpowiadały kolejno funkcjom nadajnika i odbiornika.

Rozwiązaniem pośrednim, łączącym niejako oba wymienione podejścia jest natomiast sposób trzeci, wykorzystujący modele uczenia maszynowego do realizacji osobnych elementów odbiornika, jak np. filtr pasmowo-przepustowy [13], estymacja kanału [2], algorytm Viterbiego [8]. W [12] autorzy zaprezentowali system składający się z dwóch sieci osobno realizujących estymację kanału oraz detekcję sygnałów OFDM z HPA (ang. High Power Amplifier). Modele miały na celu optymalizację klasycznych metod LS (ang. Least Squares) oraz ZF (ang. Zero Forcing). Rozdzielenie części estymacji i detekcji zaprezentowali również autorzy [3], a badania objęły sygnały o strukturze zgodnej ze standardem LTE (ang. Long Term Evolution).

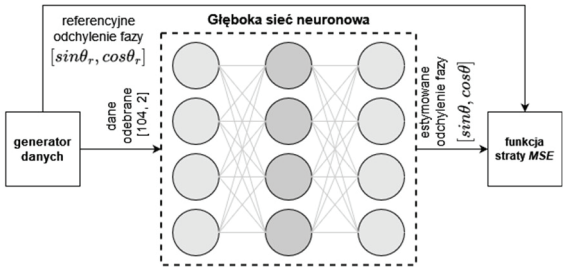
Niniejszy artykuł stanowi kontynuację badań podejmujących tematykę odbioru sygnałów GMSK prezentowanych w [6-7]. Dotychczasowe prace dowiodły, że w przypadku degradacji sygnałów przez propagację wielodrogową naiwne podejście „black box" wykorzystujące pojedynczy model ML do niejawnej estymacji kanału oraz detekcji sygnału GMSK jest niewystarczające, szczególnie w przypadku rozbudowanych pakietów, zgodnych ze strukturą ramkową definiowaną w standardach komunikacyjnych. W celu poprawy skuteczności odbioru GMSK autorzy zdecydowali się na dodanie bloku korekcji fazy wstępnie przetwarzającego dane przekazywane do detektora DL (ang. Deep Learning). Należy zwrócić uwagę, że dostępna literatura dot. odbioru sygnałów z użyciem ML, obejmuje wyłącznie transmisję sygnałów rozmieszczonych na wielu podnośnych, np. OFDM, gdzie proces estymacji kanału realizowany jest przez

predykcję pełnej macierzy współczynników kanału. Z uwagi na przyjętą strukturę sygnałów GMSK [9] w niniejszym artykule proponowane jest uproszczenie procesu estymacji do predykcji wyłącznie średniego odchylenia fazy transmitowanych sygnałów.

W poniższych rozdziałach zaprezentowany zostanie zaprojektowany model DL do estymacji rotacji fazy sygnałów GMSK w środowisku wewnątrzbudynkowym. W rozdziale 2 scharakteryzowane zostanie środowisko symulacyjne, w rozdziale 3 przedstawiony zostanie etap projektowania modelu DL, a w rozdziale 4 zaprezentowane zostaną wyniki ewaluacji sieci neuronowej z uwzględnieniem pomiarów w rzeczywistym środowisku propagacyjnym. Całość artykułu podsumuje rozdział 5.

2. ŚRODOWISKO SYMULACYJNE

Schemat prezentujący ogólną koncepcję projektowanego modelu DL realizującego estymację średniego odchylenia fazy przedstawiono na Rys. 1.



Rys. 1. Schemat modelu do estymacji średniego odchylenia fazy

Środowisko symulacyjne jest tożsame z tym prezentowanym w [6]. Wykorzystane zostały sygnały zgodne ze strukturą pakietową TCH/F (ang. Traffic Channel/Fullrate) z systemu GSM [9] oraz profile kanałów EPA (ang. Extended Pedestrian A) i ETU (ang. Extended Typical Urban) [1]. W ramach zbioru uczącego przygotowano 40 bloków kodowych (każdy blok składał się z 2004 pakietów) dla 5 różnych ciągów uczących, w zakresie E_b/N_0 = 1-20 dB. W tym miejscu należy podkreślić, że średnie odchylenie fazy θ określane jest na podstawie ciągu uczącego (104 próbki) umiejscowionego pośrodku pakietu i służy do korekcji wszystkich próbek w pakiecie. Zatem do modelu estymującego fazę dostarczane były jedynie zsynchronizowane czasowo odebrane ciągi uczące.

Dodatkowo, symulacyjny charakter realizowanych prac umożliwił wygenerowanie danych referencyjnych do uczenia modelu DL w postaci idealnego średniego odchylenia fazy wyrażonego w postaci trygonometrycznej $[\sin\theta_r, \cos\theta_r]$ i wyznaczonego dla próbek z wyjścia symulatora kanału przed dodaniem szumu na podstawie maksimum korelacji wzajemnej, określonej wzorem:

$$R_{pp'}[m] = \sum_{n=-\infty}^{\infty} p'[n] \cdot p[n-m] \qquad (1)$$

gdzie p i p' reprezentują kolejno oryginalny (zmodulowany) ciąg uczący oraz sygnał odebrany (bez szumu). Należy mieć na uwadze, że jest to wariant idealny, nierealizowalny w przypadku rzeczywistych danych pomiarowych. W celu uzyskania metody porównawczej dla modelu DL, możliwej do wykorzystania w praktyce, korelację wzajemną wyznaczono również dla danych odebranych, tj. z szumem AWGN.

3. EWALUACJA MODELI NEURONOWYCH

Zgodnie z ogólną koncepcją zaprezentowaną w poprzednim rozdziale (Rys. 1), punktem wyjścia etapu projektowania modelu DL był dobór struktury odpowiadającej charakterowi analizowanych danych. Z uwagi na fakt, iż analizowane ciągi uczące stanowiły wektor wejściowy w postaci zespolonych próbek sygnału modulowanego GMSK zdecydowano się na strukturę neuronowej sieci splotowej CNN (ang. Convolutional Neural Network) umożliwiającą przetwarzanie oraz ekstrakcję cech charakterystycznych dla danych wielowymiarowych [4]. Oprócz warstw splotowych, zastosowano również warstwy sieci FNN (ang. Feedforward Neural Network) wraz z wyjściową, liniową warstwą skalującą, pozbawioną funkcji aktywacji.

W procesie uczenia testowanych modeli DL za miarę błędu estymowanego średniego odchylenia fazy θ przyjęto funkcję straty MSE (ang. Mean Square Error) wyrażoną wzorem:

$$MSE = \frac{1}{d}\sum_{i=1}^{d}(s_i - \hat{s_i})^2 \qquad (2)$$

gdzie s oraz $\hat{s}$ odpowiadają wektorom referencyjnego $[\sin\theta_r, \cos\theta_r]$ i estymowanego $[\sin\theta, \cos\theta]$ średniego odchylenia fazy.

Zbiór podstawowych hiperparametrów przyjętych podczas ewaluacji modeli w ramach poszukiwania struktury o najwyższej skuteczności usystematyzowano w Tabeli 1.

Tab. 1. Zakresy hiperparametrów analizowanych modeli dla fazy treningowej.

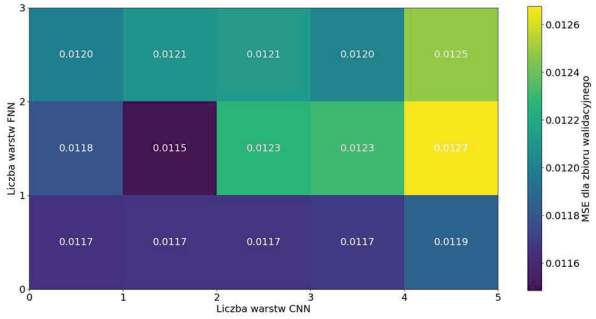
parametr	*zakres*	*rozkład*
batch_size	[100,1000]	równomierny (wartości całkowite)
wsp. uczenia	[0,00009; 0,001]	równomierny
rozmiar zbioru danych	[100 000; 1 700 000]	równomierny (wartości całkowite)
kernel	[2; 24]	równomierny (wartości całkowite)
l. iteracji	{200}	-
dropout	[0,1; 0,5]	równomierny

gdzie batch_size odnosi się do rozmiaru zestawu danych, po przetworzeniu których parametry wewnętrzne sieci są aktualizowane, współczynnik uczenia określa krok tejże aktualizacji, a kernel odnosi się do rozmiaru okna splotu w ramach przetwarzania wektorów danych wewnątrz warstw splotowych sieci.

Podczas gdy hiperparametry ściśle związane ze strukturą sieci (m.in. liczba warstw, funkcje aktywacji)

analizowano w sposób systematyczny (ang. grid search), parametry wymienione w Tab. 1 były każdorazowo dobierane losowo (ang. random search) wraz z inicjalizacją nowego modelu.

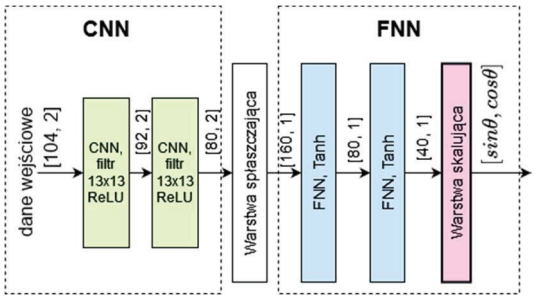
Na Rys. 2 przedstawiono mapę cieplną prezentującą zależność pomiędzy końcową skutecznością modelu DL oraz jego głębokością, zarówno w części CNN jak i FNN.



Rys. 2. Zależność wartości MSE w zbiorze walidacyjnym dla modeli o różnej głębokości

Należy zauważyć, że zwiększanie głębokości sieci powyżej 4 warstw (2 warstwy splotowe CNN oraz 2 warstwy FNN) nie wpływa na dalszą poprawę skuteczności modelu, a determinuje wręcz efekt odwrotny. Przyczyną takiej degradacji jest m.in. problem przeuczenia sieci – bardzo złożony model dobrze dopasowuje się do danych treningowych, tracąc zdolność generalizacji. Tego efektu nie udało się skompensować nawet po zastosowaniu dodatkowych warstw (ang. dropout) uzmieniających parametry sieci w trakcie fazy treningowej.

Strukturę końcową, dla jakiej uzyskano najmniejszą wartość MSE zaprezentowano na Rys. 3.

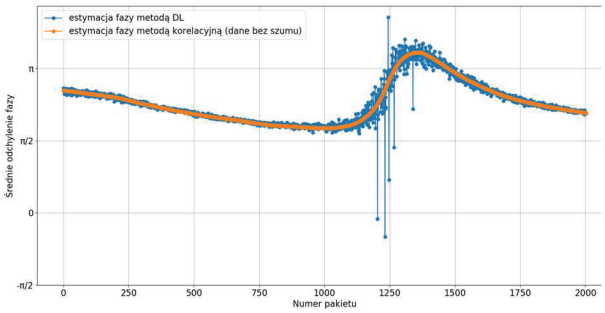


Rys. 3. Proponowana struktura głębokiej sieci neuronowej do korekcji fazy sygnałów GMSK

Wybrany model trenowany był ze współczynnikiem uczenia 0,000095, a rozmiar zestawu danych batch_size wynosił 344, przy czym pełen zestaw treningowy obejmował 1 682 807 pakietów. Operacja splotu obejmowała 13 próbek sygnału, a zastosowane funkcje aktywacji to: ReLU (ang. Rectified Linear Unit) w części CNN oraz tangens hiperboliczny w części FNN.

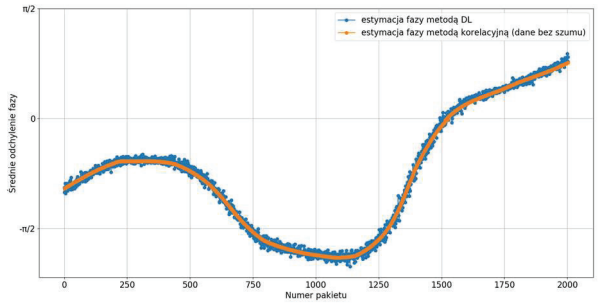
4. WYNIKI

Mając na uwadze, że sama wartość funkcji straty MSE, może różnić się zależnie od analizowanego bloku pakietów (a właściwie zaników występujących w tymże bloku), a tym samym utrudniać ogólną ocenę efektywności obu prezentowanych metod (korelacja i DL), analizę skuteczności estymacji średniego odchylenia fazy można przeprowadzić bazując na wykresach prezentujących kolejne wartości odchyleń fazy w pojedynczym bloku 2004 pakietów (Rys. 4-5).



Rys. 4. Predykcja średnich odchyleń fazy dla bloku danych testowych, Eb/No = 7 dB, profil EPA

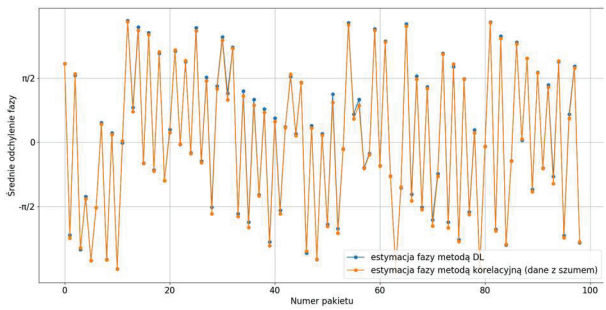
W przypadku kanału EPA można zauważyć, że estymacja modelu DL odbiega od rzeczywistej (idealnej) rotacji fazy w miejscu występowania zaników. Wierną estymację (bez rozbieżności) prezentuje natomiast Rys. 5, dla którego przykładowy blok kodowy (bez zaniku) był transmitowany w kanale ETU.



Rys. 5. Predykcja średnich odchyleń fazy dla bloku danych testowych, Eb/No = 7 dB, profil ETU

W celu weryfikacji skuteczności estymacji średniego odchylenia fazy prezentowanego modelu dla rzeczywistych sygnałów GMSK przeprowadzono dodatkową serię pomiarów stacjonarnych w środowisku wewnątrzbudynkowym, na terenie kampusu Politechniki Gdańskiej (budynek WETI B). Akwizycję danych zrealizowano przy pomocy urządzenia USRP x310 oraz anteny dookólnej OA2-0.3-10.0V firmy Cobham, na częstotliwości 939,6 MHz (ARFCN 22, E-GSM 900) z częstotliwością próbkowania 1,92 MHz.

Po przepróbkowaniu, metodą eksperymentalną wyodrębniono ciąg uczący występujący w analizowanej transmisji i określono wartości średnich odchyleń fazy dla sekwencji uczących metodą korelacyjną (analizowano dane z szumem) oraz z wykorzystaniem omawianego modelu DL. Wykres prezentujący uzyskane wyniki przedstawiono na Rys. 6. Wyniki potwierdziły skuteczność estymacji modelu DL, również w przypadku rzeczywistych sygnałów GMSK. Dużo większe zmiany fazy danych pomiarowych względem danych symulacyjnych wynikają z ciągłości analizy – rzeczywiste próbki analizowano w co ósmej szczelinie czasowej.



Rys. 6. Predykcja odchylenia fazy dla danych pomiarowych w rzeczywistym środowisku wewnątrzbudynkowym

W tym miejscu należy podkreślić, że z punktu widzenia bezwzględnej złożoności obliczeniowej, przy zbliżonej skuteczności estymacji średniego odchylenia fazy, klasyczna metoda korelacyjna jest rozwiązaniem lepszym od prezentowanego modelu DL. Analizując oba rozwiązania, złożoność obliczeniową korelacji można wyrazić jako: $O(3N \log N)$, gdzie N to długość ciągu uczącego.

W przypadku modelu sieci neuronowej, dla samej części FNN będzie to wielokrotnie więcej, tj. $O\left(n^{(1)}n^{(2)}n^{(3)} + n^{(1)}n^{(2)} + n^{(1)} + n^{(2)}\right)$, gdzie $n^{(1)}, n^{(2)}$ oraz $n^{(3)}$ wyznaczają kolejno liczbę neuronów w warstwie pierwszej, drugiej i warstwie skalującej.

Wybór metody tak nieoptymalnej obliczeniowo chcielibyśmy uzasadnić dwojako: jako weryfikację przystawalności sieci neuronowych do analizy fazy zespolonych próbek sygnału oraz podkreślając, iż nadrzędnym celem badań, których wycinek jest prezentowany w niniejszym artykule, jest odbiór sygnałów GMSK. W szerszym ujęciu, możliwe jest więc uczenie dwóch sieci jednocześnie, tj. trenowanie detektora i douczanie korektora fazy w oparciu o wspólną metrykę oceny skuteczności, jaką stanowi BER. Taki sposób przetwarzania może dawać możliwość uzyskania wyższej skuteczności detekcji niż zastosowanie detektora DL i korelacji.

5. PODSUMOWANIE

Niniejszy artykuł umożliwił potwierdzenie skuteczności modeli sieci neuronowych do estymacji średniego odchylenia fazy sygnałów GMSK, również w rzeczywistym środowisku propagacyjnym. Wyniki pokazały, że model DL wiernie odtwarza rotację fazy, jednak nie radzi sobie w obecności głębokich zaników. Należy pamiętać, że modele neuronowe nie zawsze stanowią uzasadnioną alternatywę dla metod klasycznych, szczególnie z punktu widzenia złożoności obliczeniowej, jednak w omawianym przypadku ostateczna ocena prezentowanego rozwiązania zostanie wydana na podstawie analizy pełnego systemu odbioru sygnałów GMSK, które to autorzy planują zaprezentować w przyszłości.

LITERATURA

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

[2] Bao, Yicheng, Zeyu Tan, Haifeng Sun, Zhikang Jiang. 2020. "SimNet: Simplified Deep Neural Networks for OFDM Channel Estimation." *2020 3rd IEEE International Conference on Information Communication and Signal Processing, ICICSP 2020*, 348–52.

[3] Gao, Xuanxuan, Shi Jin, Chao Kai Wen, Geoffrey Ye Li. 2018. "ComNet: Combination of Deep Learning and Expert Knowledge in OFDM Receivers." *IEEE Communications Letters*. 22 (12): 2627–30.

[4] Goodfellow, Ian, Yoshua Bengio, Aaron Courville. 2016. *Deep Learning*. Londyn: MIT Press.

[5] Liu, Jiang, Marco Di Renzo. 2021. "Data-Driven and Model-Driven Deep Learning Detection for RIS-Aided Spatial Modulation." *IEEE 4th 5G World Forum, 5GWF*, 88–92.

[6] Olejniczak, Alicja, Krzysztof K. Cwalina, Jaroslaw Sadowski, Jacek Stefanski. 2023. "Analiza Skuteczności Detekcji Sygnałów GMSK w Kanale z Propagacją Wielodrogową Przy Użyciu Metod Głębokiego Uczenia." *Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne*. 4 (2023): 234–37.

[7] Olejniczak, Alicja, Krzysztof K. Cwalina, Jarosław Sadowski, Jacek Stefański. 2022. "Analiza Efektywności Metod Głębokiego Uczenia w Odbiorze Sygnałów GMSK." *Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne*. 4 (2022): 351–54.

[8] Shlezinger, Nir, Nariman Farsad, Yonina C. Eldar, Andrea J. Goldsmith. 2020. "ViterbiNet: A Deep Learning Based Viterbi Algorithm for Symbol Detection." *IEEE Transactions on Wireless Communications*. 19 (5): 3319–31.

[9] Stefanski, Jacek. 2000. "Efektywność Odbioru Adaptacyjnego w Systemie Łączności Komórkowej GSM." Politechnika Gdanska.

[10] Yao, Rugui, Shengyao Wang, Xiaoya Zuo, Juan Xu, Nan Qi. 2019. "Deep Learning Aided Signal Detection in OFDM Systems with Time-Varying Channels." *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, 1–5.

[11] Zhang, Haotian, Lin Zhang, Yuan Jiang, Zhiqiang Wu. 2021. "LSTM and ResNets Deep Learning Aided End-to-End Intelligent Communication Systems." *2021 2nd Information Communication Technologies Conference, ICTC*, 156–60.

[12] Zheng, Shuntian, Sheng Wu, Haihan Li, Chunxiao Jiang, Xiaojun Jing. 2021. "Deep Learning-Aided Receiver Against Nonlinear Distortion of HPA in OFDM Systems." *13th International Conference on Wireless Communications and Signal Processing, WCSP*, 1–6.

[13] Zhou, Xue, Zhuo Sun, Hengmiao Wu, Qianqian Wu. 2019. "Embeddable Convolutional Layer-Based Filter for Wireless Signal Detection." *2019 IEEE Wireless Communications and Networking Conference Workshop, WCNCW*, 1–6.

METODY DETEKCJI ORAZ ROZPOZNAWANIA SYGNAŁÓW DRONÓW

Wojciech Flakowski[1]; Maciej Krasicki[2]; Rafał Krenz[3]

[1] Politechnika Poznańska, Poznań, wojciech.flakowski@doctorate.put.poznan.pl
[2] Politechnika Poznańska, Poznań, maciej.krasicki@put.poznan.pl
[3] Politechnika Poznańska, Poznań, rafal.krenz@put.poznan.pl

DOI: 10.15199/59.2024.4.25

Streszczenie: **Niniejsza publikacja prezentuje przegląd metod detekcji oraz klasyfikacji sygnałów radiowych wykorzystywanych do komunikacji z bezzałogowymi statkami powietrznymi. Artykuł przedstawia zarówno techniki wstępnego przetwarzania sygnału, algorytmy detekcyjne, jak również wybrane sposoby klasyfikacji sygnałów, bazujące na technikach uczenia maszynowego oraz sieciach neuronowych. Dodatkowo prezentowane są architektury wybranych algorytmów detekcyjnych w ujęciu złożoności obliczeniowej oraz skuteczności detekcji pożądanych sygnałów.**

Abstract: **This publication presents an overview of detection and classification methods of radio signals used to communicate with unmanned aerial vehicle. The article presents signal pre-processing techniques, detection algorithms, as well as selected signal classification problems based on machine learning techniques and neural networks. Additionally, the architectures of selected detection algorithms are presented in terms of computational complexity and effectiveness of detecting the desired signals.**

Słowa kluczowe: **bezzałogowe statki powietrzne, detekcja sygnałów radiowych, radio definiowane programowo, sieci neuronowe, uczenie maszynowe**

Keywords: **Machine Learning (ML), Neural Networks, Radio signals detection, Software Defined Radio (SDR), Unmanned Aerial Vehicles (UAV)**

1. WSTĘP

Bezzałogowe Statki Powietrzne - BSP, zwane również dronami, były na przestrzeni ostatnich lat przedmiotem znacznego rozwoju technologicznego. Umożliwiło to wykorzystanie różnych typów dronów do wyspecyfikowanych zadań, zarówno w odniesieniu do zastosowań amatorskich, jak również komercyjnych. Istotnym problemem związanym z wykorzystaniem dronów w przestrzeni powietrznej jest konieczność monitorowania ich obecności z uwagi na zapobieganie kolizjom z innymi statkami powietrznymi czy monitorowanie stref niedozwolonych dla dronów. Istnieje wiele metod detekcji obecności BSP bazujących na metodach detekcji dźwięku wirników napędowych [4] czy metodach wizyjnych, wykorzystujących kamery o wysokiej rozdzielczości [7]. Niestety, metody te cechują się znacznymi ograniczeniami, wynikającymi z warunków fizycznych i środowiskowych, uniemożliwiającymi efektywną detekcję w pewnych sytuacjach. W przypadku metod wizyjnych pogorszenie warunków atmosferycznych (np. wystąpienie gęstej mgły) całkowicie lub znacznie ogranicza możliwości detekcyjne. Natomiast środowisko o znacznym poziomie szumów akustycznych (np. duża aglomeracja miejska) uniemożliwia odfiltrowanie pożądanego sygnału przy użyciu metod akustycznych. Oprócz uprzednio wymienionych metod detekcji należy zwrócić uwagę na możliwość wykrycia drona z wykorzystaniem technik radarowych , wykorzystujących fale radiowe o częstotliwościach mikrofalowych. Metody radarowe, pomimo znacznie większej uniwersalności w porównaniu z metodami wizyjnymi oraz akustycznymi, nadal nie rozwiązują problemu wykrycia typu oraz charakterystyki drona – poprzez zjawisko odbicia fal generowanych przez urządzenie radarowe od statku powietrznego, możliwe jest ustalenie położenia drona w przestrzeni, jednak bez szczegółowej wiedzy na temat jego typu i zastosowania.

Ponieważ jednak drony jako BSP wyposażone są w systemy łączności radiowej, służące do sterowania za pomocą stacji naziemnej oraz przekazywania zbieranych danych (np. obrazów), możliwe jest stwierdzenie ich obecności na podstawie analizy odebranych sygnałów radiowych. Dzięki temu, z wykorzystaniem anten detekcyjnych o dużym zysku energetycznym możliwe jest wykrycie zbliżającego się statku powietrznego z odległości wielu kilometrów, znacznie wyprzedzając możliwość optycznej identyfikacji drona. Wszystkie te zalety przemawiają za wykorzystaniem metody detekcji sygnałów radiowych jako optymalnej metody wykrywania obecności drona w zdefiniowanej przestrzeni. Dodatkowo, poddając analizie odebrany sygnał, można ustalić konkretny typ drona i jego przeznaczenie, co jest istotne w przypadku monitorowania krytycznych obszarów przestrzeni powietrznej. Należy jednak zauważyć, iż w przypadku wykonywania misji z odgórnie zaprogramowaną trasą lotu, odbywaną bez połączenia z centrum kontroli naziemnej nie jest możliwe stwierdzenie obecności drona za pomocą technik detekcji radiowej. Także rozpoznanie dronów innych niż komercyjnych – np. wojskowych czy wytworzonych samodzielnie wiąże się z koniecznością zapewnienia szerszego zakresu analizy specyficznych pasm innych, oprócz powszechnie stosowanego w tego

typu aplikacjach (ISM). W niniejszej publikacji zaprezentowane zostały metody detekcji sygnałów dronów wykorzystujące różnorodne techniki wykrywania oraz przetwarzania sygnałów radiowych.

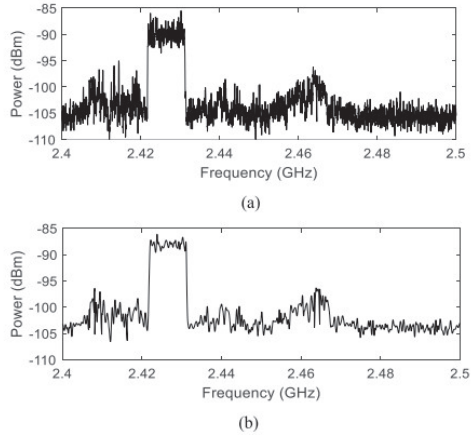
2. KLASYFIKACJA METOD DETEKCJI SYGNAŁÓW RADIOWYCH

Najważniejszym z mechanizmów, rozpoczynających proces podjęcia decyzji, czy odebrany sygnał jest sygnałem związanym z obecnością drona, jest detekcja widma sygnału. Istnieje wiele metod detekcji sygnału: od pierwotnych metod detekcji energii, polegających na porównaniu wybranej metryki z wartością progową, poprzez metody porównawcze, wymagające znajomości charakterystyki nadawanego sygnału, kończąc na metodach analizy widmowej z wykorzystaniem technik wizyjnych [5]. Należy zwrócić uwagę, że metody niewymagające technik przetwarzania wizyjnego cechują się znacznie mniejszą złożonością obliczeniową, zapewniając jednak proporcjonalnie mniejszą skuteczność. Wzrost liczbydronówwprzestrzenipowietrznejwymuszajednakstosowanie technik o wysokim poziomie skuteczności oraz elastyczności w wykrywaniu nowego rodzaju sygnałów, co spowodowało wyparcie mniej efektywnych metod bazujących na stałych progach detekcyjnych.

2.1. Metoda analizy statystycznej

Pierwszą z metod detekcji widmowej, stosowanej w przypadku wykrywania sygnałów dronów jest metoda bazująca na analizie statystycznej sygnału [8]. Odebrany sygnał w dziedzinie czasu jest przepuszczany przez filtr pasmowy i sprowadzany do pasma podstawowego. Po przekształceniu sygnału do dziedziny częstotliwości z wykorzystaniem szybkiej transformaty Fouriera (FFT – Fast Fourier Transform), wykonywana jest na nim operacja mająca na celu wyeliminowanie szumów występujących w kanale radiowym – EMD (ang. Empirical Mode Decomposition). Proces ten odbywa się na zasadzie dekompozycji sygnału na wiele sygnałów składowych, zwanych wewnętrznymi funkcjami modowymi – IMF (Intristic Mode Functions) oraz na sygnał rezydualny, który pozbawiony jest szumów. Dzięki temu można zrealizowaćwłaściwąanalizęstatystycznąsygnału,bazującąnaocenie trzech podstawowych kryteriów. W celach poglądowych na rys. 1 przedstawiony został sygnał w postaci zaszumionej oraz po przeprowadzeniu dekompozycji EMD. Dzięki niej odfiltrowane zostały szumy oraz interferencje występujące w badanym sygnale.

Pierwszym z kryteriów analizowanych w omawianym algorytmie jest obecność zboczy sygnału (ang. Slope detection). Detekcja zbocza sygnału (rosnącego i opadającego) realizowana jest poprzez badanie kąta nachylenia zbocza widma sygnału względem pionowej linii odniesienia i obliczenie kąta pochyłu zbocza. W celu wykrycia szerokości widma sygnału wykorzystywane jest ruchome okno detekcyjne, skanujące zadane pasmo w celu wykrycia początku oraz końca widma sygnału. W przypadku detekcji dronów pracujących w ogólnodostępnym paśmie ISM (ang. Industrial, Scientific and Medical) w zakresie częstotliwości 2,4 – 2,4835 GHz, wykorzystuje się okno detekcyjne o szerokości ok. 500 kHz.



Rys. 1. Rezultat dekompozycji sygnału metodą EMD [8]
a. Widmo gęstości mocy sygnału zaszumionego przed dekompozycją
b. Widmo gęstości mocy sygnału odfiltrowanego po dekompozycji

Następnym parametrem definiującym charakterystykę wykrywanego sygnału jest współczynnik skośności (ang. Skewness). Współczynnik ten opisuje asymetrię widma sygnału. W przypadku całkowitej symetrii, współczynnik skośności powinien przyjąć wartość 0. Przy detekcji sygnału z wykorzystaniem analizy współczynnika skośności wykorzystuje się ruchome okno detekcyjne o szerokości wynoszącej ok. 10 MHz, umożliwiające analizę symetrii większych fragmentów widma sygnału. Ostatnim z parametrów określających cechy statystyczne sygnałujestwspółczynnikkurtozy(ang.Kurtosis).Parametrtenpozwala scharakteryzować miarę podobieństwa rozkładu statystycznego próbek analizowanego sygnału względem rozkładu normalnego i zarazem przedział, w którym rozkład przyjmuje wartości zerowe (ang. Tail). Trzy ww. parametry tworzą wektor klasyfikujący, którego składowe, dla podjęcia decyzji, są porównywane ze zdefiniowanymi wartościami progowymi. Należy zauważyć, iż metoda ta nie umożliwia dokładnego rozpoznania typu oraz aktywności drona, a jedynie jego obecność na podstawie analizy parametrów odnoszących się do kształtu obserwowanego widma sygnału.

2.2. Metoda automatycznego rozpoznawania modulacji – AMR

Następną z metod wykrywania sygnałów dronów w paśmie radiowym jest metoda oparta na analizie typów modulacji, jakie wykorzystywane są przez BSP i odpowiedniej ich klasyfikacji z wykorzystaniem technologii głębokiego uczenia maszynowego. Analiza typów modulacji – AMR (ang. Automatic Modulation Recognition) cechuje się automatyzacją procesu detekcji sygnałów, jednak wymaga dokładnie sprecyzowanych i odfiltrowanych z szumów próbek sygnałów w celu zapewnienia efektywnego procesu uczenia algorytmu klasyfikującego [3].

Główną zaletą wykorzystania automatycznego rozpoznawania modulacji jest możliwość równoczesnej detekcji wielu sygnałów za pomocą rozproszonych ośrodków decyzyjnych.

Algorytm AMR jest w stanie rozpoznać typy sygnałów zmodulowanych analogowo (AM, SSB), jak i cyfrowo (M-ASK, FSK, QAM). Możliwe jest wykrycie powyższych sygnałów w szerokim zakresie odstępu sygnału od szumu – SNR (ang. Signal to Noise Ratio), przez co algorytm może znaleźć zastosowanie w rzeczywistych warunkach radiowych.

Należy wspomnieć, że każdy z sygnałów wejściowych posiada swój zestaw parametrów (tzw. hiperparametrów) pozwalających wstępnie określić typ sygnału w kontekście modulacji oraz jego charakterystykę. Biorąc pod uwagę specyfikę konkretnych typów modulacji (analogowe oraz cyfrowe), poszczególne wektory hiper-parametrów różnią się między sobą liczbą i rodzajem. Dla przykładu, modulacja ASK posiada zestaw trzech zasadniczych parametrów: częstotliwość środkową sygnału, czas trwania symbolu oraz parametr M, określający wartościowość modulacji.

Biorąc pod uwagę fakt, iż algorytm przystosowany został do detekcji sygnałów w dziedzinie częstotliwości, na odebranym sygnale wykonywana jest transformata FFT. Następnie za pomocą detekcji kształtów widm sygnałów (z wykorzystaniem technik rozpoznawania obrazów), tworzony jest dla każdego z wykrytych sygnałów zestaw uprzednio opisanych hiper-parametrów. Tak przygotowany sygnał przesyłany jest następnie do sieci neuronowej w celu jego klasyfikacji. W kontekście technik rozpoznawania kształtów widma trzeba podkreślić, że algorytm detekcyjny pracować może w dwóch trybach: jednostopniowym oraz dwustopniowym. Tryb jednostopniowy polega na podziale widma na określoną liczbę komórek – w przypadku, gdy częstotliwość środkowa sygnału znajduje się w konkretnej komórce, komórka ta odpowiada za jego detekcję. W trybie dwustopniowym do detekcji kształtów widm sygnałów wykorzystywane jest ruchome okno detekcyjne. Tryb ten, pomimo większej dokładności, wiąże się ze znacznym spadkiem szybkości wykrywania sygnałów, co uniemożliwia jego użycie do detekcji sygnałów w czasie rzeczywistym.

2.3. Metoda dwuwymiarowej detekcji 2D-CNN

Następną z metod wykrywania sygnałów dronów paśmie radiowym jest metoda wykorzystująca dwuwymiarową sieć neuronową o charakterze splotowym – 2D-CNN (ang. 2-Dimensional Convolutional Neural Network) [9]. Jednym z podstawowych założeń algorytmu jest umożliwienie wykrycia, analizy oraz klasyfikacji wielu sygnałów w analizowanym paśmie radiowym oraz ich równoczesnego przetwarzania. Pierwszym etapem przetwarzania odbieranego sygnału jest wyznaczenie spektrogramu przy pomocy transformacji STFT (ang. Short Time Fourier Transform). Uzyskana reprezentacja czasowo-częstotliwościowa sygnału umożliwia kompleksową analizę w sieci neuronowej w dalszej fazie działania algorytmu. Wykorzystanie reprezentacji sygnału w dwóch dziedzinach jest istotne z punktu widzenia kompatybilności danych wprowadzanych do dwuwymiarowej sieci neuronowej, gdzie każdy z wymiarów sieci odpowiada za przetwarzanie sygnału w danej dziedzinie, co zwiększa ogólną wydajność algorytmu detekcyjnego i jego zdolności klasyfikacyjne.

Główną cechą omawianego algorytmu jest klasyfikacja sygnałów transmitowanych przez drony z wykorzystaniem dostępnych baz danych, zawierających próbki sygnałów dronów określonych producentów. Takie podejście umożliwia bezpośrednie określenie typu drona, protokołu komunikacyjnego, jaki wykorzystuje do transmisji sygnału, a także jego aktywności. Poprzez pojęcie aktywności drona rozumie się następujące tryby jego pracy:stanaktywności–załączeniazasilania(ang.PowerOn-Mode), podłączenia do stacji kontrolnej (ang. Connected Mode), tryb lotu (ang. Flying Mode), tryb transmisji strumienia wizji z wbudowanych kamer (ang. Video Mode) i tryb żeglowania (ang. Hovering Mode). Każdy z trybów różni się rodzajem transmitowanego sygnału, również dla konkretnego modelu drona. Możliwe jest zatem rozpoznanie stanu aktywności w odniesieniu do konkretnego typu drona, jednak wymaga to wprowadzenia klasyfikatorów bazy danych, umożliwiających jednoznaczne określenie typu sygnału na podstawie ww. kryteriów. W celu jednoznacznej identyfikacji próbek pochodzących od różnych typów dronów, do każdego wektora próbek dołączany jest wektor BUI (ang. Binary Unique Identifier), reprezentujący cechy poszczególnych próbek.

2.4. Metoda detekcji YOLO

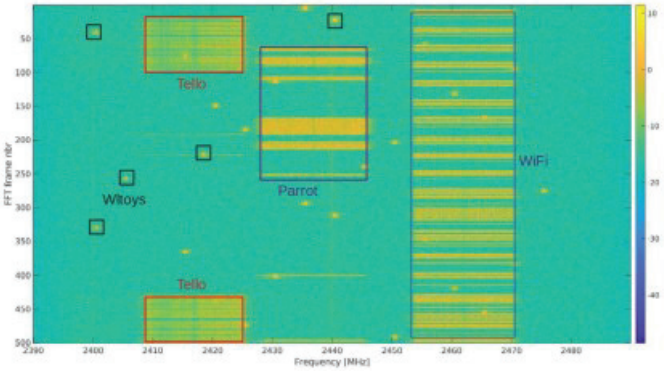
Kolejną z metod detekcji sygnałów dronów jest metoda YOLO (ang. You Only Look Once) [6]. Metoda ta opiera się na analizie sygnatury odebranego sygnału w paśmie radiowym oraz jego klasyfikacji. Godnym podkreślenia jest fakt, że omawiany algorytm detekcji posiada zdolności wykrycia zarówno pojedynczego sygnału, jak i wielu równolegle transmitowanych sygnałów w analizowanym paśmie radiowym. Pierwszym krokiem jest tu przekonwertowanie próbek sygnału do postaci spektrogramu. W tym celu sygnał w dziedzinie czasu poddawany jest transformacie STFT. W wyniku tej operacji sygnał dzielony jest na kolejne bloki o zadanej szerokości pasma. Dzięki temu możliwe jest umieszczenie konkretnych bloków w macierzy spektrogramu i reprezentowanie sygnału w dziedzinie czasu oraz częstotliwości. Dodatkowo uwzględnienie amplitudy sygnału w trakcie konwersji pozwala na pełną reprezentację sygnału za pomocą spektrogramu.

Zastosowanie macierzy spektrogramowej w analizie i klasyfikacji sygnałów generowanych przez BSP pozwala określić więcej cech wykrytych sygnałów w porównaniu z klasyczną konwersją do dziedziny częstotliwości dla uzyskania widmowej gęstości mocy PSD (ang. Power Spectral Density). Do podstawowych cech zaliczyć można częstotliwość środkową sygnału, szerokość kanału transmisji (pasmo zajmowane przez sygnał w dziedzinie częstotliwości), czas trwania sygnału czy wzór przeskoków częstotliwości oraz ich częstość (w przypadku metod rozpraszania widma opartych na przeskokach

częstotliwości). W nawiązaniu do omawianych powyżej cech przetwarzanego sygnału, należy podkreślić, iż znaczna grupa dronów dostępnych komercyjnie wykorzystuje techniki rozpraszania widma DSSS (ang. Direct Sequence Spread Spectrum) lub FHSS (ang. Frequency Hopping Spread Spectrum). Dzięki zastosowaniu przeskoków częstotliwości nośnej według określonych sekwencji, możliwa jest praca wielu systemów (w tym dronów) z wykorzystaniem tych samych zasobów częstotliwościowych. Mając na uwadze specyfikę działania powyższego mechanizmu rozpraszania widma, reprezentacja analizowanego sygnału w postaci spektrogramu pozwala uzyskać jeszcze lepszą jakość detekcji sygnału wykorzystującego techniki rozpraszające. Istotnie, możliwa jest szczegółowa analiza wzoru i sposobu przeskoków sygnału, uwzględniająca zarówno dziedzinę czasu, jak i częstotliwości [2].

Spektrogram daje informację o różnicach w reprezentacji sygnału w odniesieniu do konkretnych typów dronów, co pozwala tak zmodyfikować algorytm detekcyjny, aby uwzględniał dodatkowe cechy sygnału, takie jak nieliniowości nadajników konkretnych modeli dronów czy fakt zastosowania konkretnych filtrów pasmowych po stronie nadawczej. Mechanizm detekcji oraz klasyfikacji sygnałów generowanych przez BSP wykorzystuje rozpoznawanie reprezentacji sygnałów metodami analizy wizualnej. Analiza ta polega na zastosowaniu algorytmu rozpoznawania kształtów oraz elementów obecnych w obrazie wejściowym. W przypadku algorytmu detekcji sygnałów radiowych, obrazem tym jest spektrogram. Wówczas na wygenerowanej macierzy następuje obrazowa detekcja bloków częstotliwości o konkretnym czasie trwania, świadczących o obecności pożądanego sygnału. Równocześnie analizowane są cechy takie jak częstotliwość środkowa czy szerokość pasma zajmowanego przez sygnał.

Kolejną istotną cechą metody YOLO jest zdolność do wykrycia sygnałów nakładających się w dziedzinie częstotliwości. Zdolność detekcji sygnałów nakładających się ma istotne znaczenie w przypadku pracy algorytmu w rzeczywistym środowisku radiowym, gdzie obecnych jest wiele sygnałów pochodzących od różnych systemów radiowych. Dotyczy to w szczególności ogólnodostępnego, nielicencjonowanego pasma ISM, które używane jest przez zdecydowaną większość dronów dostępnych komercyjnie. Dodatkowo należy zaznaczyć, iż skuteczność detekcji sygnałów przy użyciu algorytmu YOLO zależna jest od liczby sygnałów obecnych w analizowanym paśmie – regułą jest proporcjonalny wzrost skuteczności detekcji sygnałów wraz ze zwiększeniem ich liczby w analizowanym paśmie. Poprzez skuteczność rozumieć należy uśrednione prawdopodobieństwo detekcji wszystkich sygnałów w odniesieniu do zdefiniowanego zakresu wartości SNR (ang. Signal to Noise Ratio), dla których obliczone zostało prawdopodobieństwo detekcji. Detekcja sygnału uznawana jest za prawdziwą jeśli współczynnik prawdopodobieństwa detekcji obrazu widmowego konkretnego sygnału CS (ang. Confidence Score) przekroczy statycznie zdefiniowany próg. Co istotne, niezależnie od wartości SNR, przy wzroście liczby współistniejących sygnałów w analizowanym pasmie wzrasta wydajność algorytmu. Dzięki temu algorytm ten może służyć do wykrywania wielu sygnałów o małej mocy lub w środowiskach o znacznych poziomach szumów oraz interferencji. Na rysunku 2 przedstawione zostały przykładowe sygnały dronów rozpoznane przez algorytm.



Rys.2. Reprezentacja wykrytych sygnałów w spektrogramie [2]

2.5. Metoda detekcji GoF

Ostatnią z prezentowanych metod detekcji jest metoda GoF (ang. Goodness of Fit), polegająca na analizie statystycznej sygnału oraz obliczeniu miary dopasowania wykrytego sygnału względem uprzednio wyuczonych wzorców [2]. Miara dopasowania statystycznego określa, w jakim stopniu wykryty sygnał pasuje do uśrednionej, statystycznej funkcji wzorcowej.

Algorytm wykorzystujący metodę GoF posiada dwa zasadnicze bloki funkcjonalne, współpracujące ze sobą. Pierwszy z nich pełni rolę analizatora dopasowania statystycznego wykrytego sygnału, odpowiada zatem za właściwą detekcję sygnału. Drugim blokiem, znacznie rozszerzającym możliwości algorytmu GoF, jest blok klasyfikacyjny, wykorzystujący sieć neuronową. Metoda GoF cechuje się ślepą detekcją sygnału w uprzednio zdefiniowanym paśmie częstotliwości. Oznacza to, że w odróżnieniu od metody detekcyjnej YOLO, nie jest możliwe wyodrębnienie bezpośrednio na etapie detekcji cech sygnału takich jak: szerokość zajmowanego pasma transmisyjnego, czas trwania sygnału czy częstotliwość środkowa. Metoda ta umożliwia jednak uzyskanie lepszych rezultatów w porównaniu z klasycznymi metodami detekcji widma, takimi jak metody analizy energii sygnału. Dzięki temu możliwa jest detekcji sygnałów przy znacznie zredukowanej liczbie próbek wejściowych – a zatem krótszym czasie trwania, a także w rzeczywistych warunkach radiowych o dużych poziomach interferencji, w których sygnały cechują się niskimi wartościami SNR.

Wysoką czułość detekcji algorytmu GoF w przypadku analizy sygnałów o małej mocy można uzyskać poprzez podział analizowanego pasma na wąskie przedziały. Wówczas w każdym z przedziałów przeprowadzana jest wąskopasmowa detekcja GoF, co umożliwia w pewnym stopniu zredukowanie wąskopasmowych zakłóceń występujących tylko w niektórych przedziałach i tym samym uzyskanie większej dokładności po uśrednieniu wyniku ze wszystkich analizowanych przedziałów. Do-

kładność detekcji rozumieć należy jako stosunek liczby poprawnie wykrytych przedziałów częstotliwości przez algorytm do wszystkich przedziałów częstotliwości, w których znajduje się pożądany sygnał. Proces podziału pasma na poszczególne przedziały realizowany jest z wykorzystaniem szybkiej transformaty Fouriera – FFT. W ramach poszczególnych przedziałów wykonywana jest operacja detekcji GoF bazująca na modelu testu statystycznego AD (ang. Anderson Darling). Dane wejściowe procesów detekcji oraz klasyfikacji dostarczane są w początkowej fazie w dziedzinie czasu. Następnie wykonywana jest na nich transformata FFT, dokonująca podziału analizowanego pasma na wspomniane uprzednio wąskopasmowe przedziały. Detekcja bazująca na modelu AD polega na stwierdzeniu, czy w analizowanym przedziale obecny jest pożądany sygnał, czy też występuje w nim tylko szum. Analiza obecności sygnału odbywa się na zasadzie porównania wartości funkcji dopasowania (testu AD) z uśrednioną wartością tzw. widmowego współczynnika mocy w analizowanym przedziale. W przypadku braku sygnału (obecny tylko szum) widmowy współczynnik mocy będzie podobny do wartości funkcji dopasowującej. W przypadku wykrycia sygnału wartość testu AD wykaże znaczne rozbieżności w stosunku do widmowego współczynnika mocy. Należy również pamiętać, iż moc szumu jest uśredniana z wykorzystaniem histogramu obrazującego odebrane sygnały w dziedzinie czasu oraz częstotliwości.

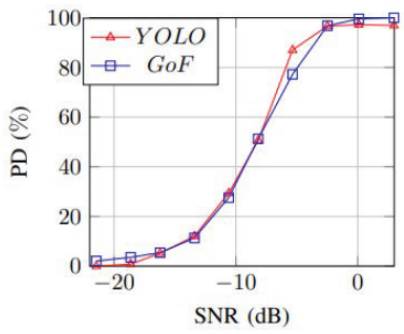
3. OCENA WYDAJNOŚCI ORAZ WNIOSKI KOŃCOWE

Należy zaznaczyć, iż każdy z omawianych algorytmów cechuje się pewnymi fluktuacjami wydajności w odniesieniu do konkretnych warunków radiowych (np. SNR), typów modulacji, czy liczby przetwarzanych równolegle sygnałów w czasie rzeczywistym. Powoduje to, że pewne algorytmy sprawdzają się lepiej w ściśle określonych przypadkach, w porównaniu z algorytmami zapewniającymi lepszą wydajność w przeciętnych warunkach. Poniżej zostaną omówione wyniki zaczerpnięte z pozycji literaturowych, odnoszące się do algorytmów omawianych w rozdziałach poprzednich.

Na podstawie [1] należy stwierdzić, że największą skutecznością cechuje się rozwiązanie oparte na metodzie realizującej detekcję z wykorzystaniem algorytmu YOLO. Ogólna skuteczność algorytmu w zakresie zdolności poprawnej klasyfikacji wynosiła 70,3 % przy 1 % oznaczonych próbek w całym zbiorze treningowym. Zwiększenie liczby oznaczonych próbek (a zatem znanych sygnałów) do poziomu 10 % skutkuje wzrostem odsetka poprawnych klasyfikacji do 97,4 %. Dalszy wzrost liczby oznaczonych próbek do poziomu 30 % umożliwia uzyskanie współczynnika poprawnych detekcji o wartości 99,6 %. Dowodzi to faktu, że algorytm cechuje się znaczną sprawnością detekcyjną, a przez to nie ma potrzeby zapewnienia pełnego zestawu oznaczonych danych; algorytm uczy się w sposób ciągły już w trakcie swego działania. Wystąpienie fałszywych detekcji sygnału objawia się minimalnymi wahaniami przesunięcia pasma sygnału w dziedzinie częstotliwości podczas procesu detekcyjnego. Sytuacja ta występuje jednak w niewielkim odsetku wykrywanych sygnałów, wynoszącym 0,4 %. Niezaprzeczalnym atutem metody bazującej na detekcji YOLO jest możliwość klasyfikacji sygnałów na podstawie wektora cech sygnału, dzięki czemu można uzyskać wiedzę o podstawowych właściwościach sygnału [1].

Biorąc pod uwagę wyłącznie skuteczność detekcji bez uwzględnienia mechanizmu klasyfikacji wykrytych sygnałów, algorytm YOLO sprawdza się lepiej dla wielu sygnałów analizowanych równolegle, w porównaniu z algorytmem ślepej detekcji GoF. W przypadku algorytmu GoF ogólna skuteczność zwiększa się wraz ze wzrostem SNR. Jest to zjawisko obserwowane również w przypadku metody YOLO, jednak tylko dla algorytmu GoF wydajność detekcji może osiągnąć poziom 99,9 %. Przy niewielkiej wartości SNR (od -20 dB do -3 dB) algorytmy wykazują bardzo zbliżoną skuteczność detekcji [2]. Skuteczność detekcyjna algorytmu GoF określona została na zasadzie weryfikacji wyniku predykcji obecności sygnału w analizowanych przedziałach pasma za pomocą testu AD z użyciem danych testowych. Na tej podstawie obliczony został współczynnik PD (ang. Probability of Detection) obrazujący skuteczność algorytmu w funkcji SNR. Obserwacje te dotyczą scenariusza, w którym w analizowanym paśmie częstotliwości obecny jest tylko jeden sygnał.

Skuteczność algorytmu GoF w scenariuszu detekcji wielu sygnałów odbieranych jednocześnie zależy jedynie od wartości współczynnika SNR, natomiast nieistotna jest liczba współistniejących sygnałów. Omawiana charakterystyka algorytmu GoF wykazuje zatem cechy odmienne w stosunku do zdolności detekcyjnych algorytmu YOLO, gdzie liczba równocześnie występujących sygnałów ma znaczący wpływ na wydajność detekcji. Powyższa zależność wynika z faktu, iż metoda GoF jest metodą ślepą, traktującą określony przedział częstotliwości za pomocą stałych, ściśle określonych kryteriów (progu detekcji λ). Dodatkowo, jeśli w tym samym paśmie obecnych jest jednocześnie wiele sygnałów o różnych wartościach współczynnika SNR, algorytm GoF dokonuje estymacji poziomu szumu (podłogi szumu), co może mieć wpływ na ogólną wydajność detekcji. Wówczas ogólny poziom szumów (będący poziomem odniesienia) ulega podwyższeniu, co może być powodem błędnych detekcji w wyniku przekroczenia progu detekcji λ. Powyższy problem nie występuje w przypadku algorytmu YOLO. Na rysunku 3 przedstawiono porównanie skuteczności detekcji algorytmów YOLO oraz GoF dla różnych wartości współczynnika SNR analizowanego sygnału. W omawianym przypadku wykrywany był tylko jeden sygnał za pomocą każdego z algorytmów. Do celów treningowych użytych zostało 4900 próbek, natomiast testy skuteczności przeprowadzone zostały przy użyciu 2100 próbek sygnału testowego [2].



Rys. 3. Porównanie wydajności metod YOLO oraz GoF [2]

Odwrotna zależność, wiążąca liczbę jednocześnie analizowanych sygnałów oraz wydajność, ma miejsce w przypadku algorytmu automatycznego rozpoznawania modulacji sygnału. W tym przypadku zwiększenie liczby sygnałów w paśmie powoduje proporcjonalny spadek ogólnej wydajności algorytmu [3]. Metoda ta nie jest zatem efektywna w przypadku skanowania szerokiego przedziału pasma, w którym jednocześnie może wystąpić wiele sygnałów.

Analogicznie jak w przypadku metody AMR, spadek wydajności podczas równoległej detekcji wielu sygnałów zauważalny jest również w metodzie przetwarzania parametrów statystycznych sygnałów [8]. Aby możliwe było zwiększenie skuteczności detekcji wymagana jest operacja dekompozycji szumowej EMD, co wiąże się ze wzrostem zapotrzebowania pod względem zasobów obliczeniowych. Algorytm ten podejmuje również decyzję względem arbitralnie ustawionych progów, nie ma on zatem możliwości dynamicznego aktualizowania zbioru danych podczas pracy algorytmu. Najbardziej efektywną metodą spośród uprzednio omawianych pozostaje zatem metoda bazująca na detekcji YOLO.

PODSUMOWANIE

Metody detekcji radiowej bazujące na sieciach neuronowych cechują się największą wydajnością oraz elastycznością zapewniającą szybkie przystosowanie się do nowych standardów transmisji wykorzystywanych przez drony. Autorzy planują ulepszyć mechanizm detekcji bazujący na metodzie YOLO, aby w sposób autonomiczny był on w stanie klasyfikować wykrywane sygnały poprzez modyfikację mechanizmu uczenia sieci neuronowej. Zapewni to jeszcze większą elastyczność działania algorytmu, a także ograniczy procedurę wstępnego treningu algorytmu. Dzięki temu możliwa będzie szybka, automatyczna klasyfikacja wykrytych sygnałów.

INNE UWAGI

Niniejsza publikacja przygotowana została w roku 2024 w ramach projektu pt. „Metody zwiększania efektywności widmowej, energetycznej oraz zwiększanie bezpieczeństwa w sieciach bezprzewodowych” o numerze 0312/SBAD/8166 .

BIBLIOGRAFIA

[1] S. Basak, S. Rajendran, S. Pollin, B. Scheers, „Autoencoder based framework for drone RF signal classification and novelty detection”, 2023 25th International Conference on Advanced Communication Technology (ICACT), DOI:10.23919/ICACT56868.2023.10079363

[2] S. Basak, S. Rajendran, S. Pollin, B. Scheers, „Combined RF-Based Drone Detection and Classification”, October 2023 International Journal for Research in Applied Science and Engineering Technology 11(10):110-115, DOI:10.22214/ijraset.2023.55928

[3] Ch. Hou, D. Fu, Z. Zhou, X. Wu, „A Deep Learning-Based Multi-Signal Radio Spectrum Monitoring Method for UAV Communication”, Drones 2023, 7(8), 511; https://doi.org/10.3390/drones7080511

[4] M. Nijim, N. Mantrawadi, “Drone classification and identification system by phenome analysis using data mining techniques,” 2016 IEEE Symposium on Technologies for Homeland Security (HST), pp. 1–5

[5] K. N. R. Surya Vara Prasad, V. K. Bhargava, “A Classification Algorithm for Blind UAV Detection in Wideband RF Systems” 2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall), IEEE, DOI: 10.1109/VTC2020-Fall49728.2020.9348513

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788

[7] Z. Zhang, Y. Cao, M. Ding, L. Zhuang, W. Yao, “An intruder detection algorithm for vision based sense and avoid system” International Conference on Unmanned Aircraft Systems (ICUAS), 2016, pp. 550–556

[8] H. Zhang, C. Cao, L. Xu, T. A. Gulliver, „A UAV Detection Algorithm Based on an Artificial Neural Network”, May 2018, IEEE Access PP(99):1-1, DOI:10.1109/ACCESS.2018.2831911

[9] Z. Zhao, Q. Du, X. Yao, „A Two-Dimensional Deep Network for RF-based Drone Detection and Identification Towards Secure Coverage Extension”, August 2023, eprint arXiv:2308.13906, DOI: 10.48550/arXiv.2308.13906

OPTYMALIZACJA PRZEPUSTOWOŚCI ŁĄCZA SATELITARNEGO ZA POMOCĄ UCZENIA MASZYNOWEGO

OPTIMIZING SATELLITE LINK THROUGHPUT USING MACHINE LEARNING

Ilia Zainutdinov[1]; Rafał Krenz[2]

[1] Politechnika Poznańska, Poznań, ilia.zainutdinov@doctorate.put.poznan.pl
[2] Politechnika Poznańska, Poznań, rafal.krenz@put.poznan.pl

DOI: 10.15199/59.2024.4.26

Streszczenie: **Artykuł ten bada wykorzystanie uczenia maszynowego do optymalizacji komunikacji satelitarnej, gdzie tradycyjne informacje zwrotne są ograniczone. Integracja wcześniej obliczonych planów sesji z modelami Random Forest i sekwencyjnymi, szkolonymi na danych symulacyjnych, ma na celu skuteczne radzenie sobie z dynamicznymi wyzwaniami komunikacyjnymi. Wstępne wyniki wskazują na obiecujący potencjał w zakresie zwiększenia efektywności i niezawodności komunikacji satelitarnej.**

Abstract: **This paper explores using machine learning to optimize satellite communications where traditional feedback is limited. Integrating pre-calculated session plans with Random Forest and Sequential models, trained on simulated data, these models aim to handle dynamic communication challenges effectively. Preliminary results suggest promising potential for enhancing satellite communication efficiency and reliability.**

Słowa kluczowe: **Algorytmy adaptacyjne, komunikacja satelitarna, modelowanie predykcyjne, optymalizacja transmisji, uczenie maszynowe**

Keywords: **Adaptive Algorithms, Machine Learning, Predictive Modeling, Satellite Communications, Transmission Optimization**

1. INTRODUCTION

In traditional satellite communication systems, feedback channels are very important for adjusting transmission settings dynamically to make signal quality and data flow better. These adjustments are made in response to real-time feedback from ground stations, which allows continuous fine-tuning of transmission settings like bandwidth, code rate, and others.

However, not all satellite communication setups have the capability for such interactive adjustments. In our specific scenario [3], we face a challenge because of the separation of functionalities between different types of ground stations. While the control ground station is allowed to change the satellite's transmitter settings, other ground stations can only receive data and cannot send feedback information to the satellite. This separation requires a different approach to manage and optimize satellite transmissions.

To tackle this limitation, we have developed a new protocol where a session plan is pre-calculated and uploaded to the satellite by the control ground station before the satellite passes. This session plan sets important transmission parameters: it chooses a fixed bandwidth for the duration of the pass and plans a schedule for changing the code rates at specific times. These pre-planned settings aim to make data transmission more efficiently by adapting to expected changes in link quality because of the satellite's changing position relative to the ground station and other phenomena.

The session plan is generated using a custom program that uses Two-Line Element (TLE) data to predict satellite pass characteristics like latitude, longitude, altitude, and velocity along with expected environmental conditions. This program calculates the best bandwidth and code rate changes based on a theoretical model to maximize the satellite's data throughput. Although this method is theoretically efficient, it is limited by its inability to adapt to unexpected real-time changes such as actual weather conditions, antenna misalignments, and other imperfections.

Recognizing these limitations, this paper looks at the use of machine learning techniques to make the adaptability and accuracy of satellite transmission settings better. By using historical data and predictive modeling, we plan to develop dynamic algorithms that can adjust transmission parameters in ways that traditional pre-calculated session plans cannot. This approach marks a major shift towards more resilient and efficient satellite communication strategies in scenarios where real-time interactive feedback is not available.

2. MACHINE LEARNING INTEGRATION

As we get closer to the launch of our satellite, it's very important to make sure that machine learning predictions match up with the outcomes from traditional, formula-based methods. These formulas are seen as the 'correct' standard during the pre-launch phase and set a reliable baseline for training our machine learning models. In this preparation stage, our models are trained with simulated data that imitates expected satellite

conditions, including various environmental and operational factors. This type of training is essential for developing a relationship with the predictions made by the formulas, ensuring our models are theoretically prepared when we start getting actual satellite data.

The main goal is to reach a high level of accuracy with our machine learning predictions in relation to the traditional methods, laying a solid groundwork for making adjustments after the launch. When the models are updated with real-world data after the launch, we plan to make the needed changes to handle the complexities that simulations can only approximate.

To manage the dynamic and complex nature of satellite communications, we use two advanced machine learning models: the Random Forest and the Sequential Model using TensorFlow. The Sequential model is a type of recurrent neural network (RNN) implemented using Keras, specifically designed to handle sequential data. It is well-suited for time-series predictions inherent in satellite communication, allowing for predictions of transmission parameters during the satellite pass.

The Random Forest Model is selected for its strength in controlling overfitting [2]. It uses a group method that brings together predictions from many decision trees, each trained on different sets of data. This method provides stable and reliable results, making it very good for predicting the best transmission settings. For the Sequential Model, hyperparameters such as the number of layers, neurons per layer, learning rate, and batch size were initially selected as a starting point. In future work, we plan to optimize these hyperparameters using a genetic algorithm.

The Sequential Model is great at making separate predictions for each moment of the transmission [1]. Its flexibility lets it effectively work with and learn from a large number of distinct data points. This ability is key for adjusting to fast changes in satellite communication conditions and for figuring out the best code rate at every second.

3. DATA PREPARATION

The success of any machine learning model heavily relies on the quality and setup of the data collected and provided to the algorithm in the training phase.

3.1 Input Features

Both models use the same inputs but are preprocessed differently to meet the specific needs of each model.

Bandwidth Prediction Inputs: Total Satellite Time [s], Minimum & Maximum Distance [m], Minimum & Maximum Elevation [m], Minimum & Maximum Antenna Misalignment [°], Average Weather (related) Attenuation [dB].

Code Rate Prediction Inputs: Distance [m], Elevation [°], Antenna Misalignment [°], Bandwidth [MHz], Weather (related) Attenuation [dB].

3.2 Preprocessing

No normalization is applied to the Random Forest Model. The strong ability of decision trees to manage different data scales and the combined method used in Random Forests means that normalization is not necessary for optimal predictions.

All inputs are normalized in the Sequential Model to help the network learn more effectively, making sure that the adjustments during training are handled well.

3.3 Outputs

For the Random Forest Model, the outputs for the code rate and bandwidth predictions are seen as separate classes that match specific operational settings (e.g., code rates from 0.91, 0.83, 0.76, 0.57, 0.38, 0.285, and 0.19; and bandwidth options from 1, 1.25, 5, 10, and 20 MHz - system-specific values [4]).

Outputs for the Sequential Model are formatted in hot-one encoding across multiple nodes for each prediction task, which fits the model's requirement for processing categorical outputs in neural networks.

4. MODEL LEARNING AND RESULTS

The development and fine-tuning of machine learning models in satellite communications needs careful monitoring of how these models progress in learning from data over time and how their predictions compare to theoretical baselines.

Both the Random Forest and the Sequential models go through training phases where they learn from a simulated dataset made to imitate real satellite transmission situations. The learning curve for each model is shown in the performance graph which demonstrates improvement and adaptation as the training moves forward (Fig. 1).

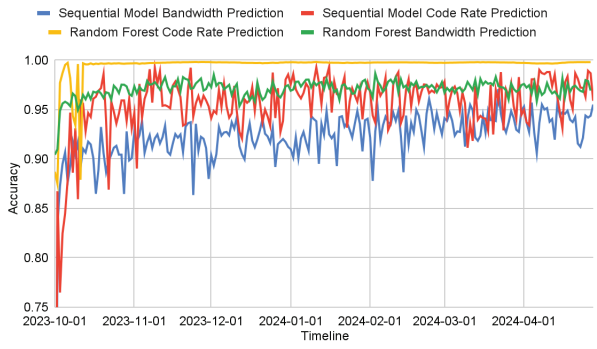


Fig 1. The learning curve of the Sequential Model and the Random Forest Model

The Random Forest model has been showing an impressive average accuracy of 97% in predicting the right bandwidth and an even more remarkable 99.5% accuracy for code rate prediction. The Sequential model has also shown strong results, with an average of 92% accuracy in bandwidth prediction and 95.3% accuracy in code rate prediction. Throughout a seven-month evaluation, the Random Forest's performance was

consistently stable. On the other hand, the Sequential model's predictions were a bit more uneven.

Decision trees' size inside Random Forest models is directly reflected in their ability to achieve high prediction accuracy. For bandwidth prediction, as visualized in the provided graphs (Fig. 2 and Fig. 3), the decision tree structure used within our Random Forest model shows about ten layers of depth and around one hundred nodes. This complexity allows the model to capture the nuances of satellite link quality effectively, contributing to the impressive 97% accuracy in bandwidth prediction.

Fig 2. Decision Tree for Bandwidth Prediction for Aerocube 6A on 01.04.2024

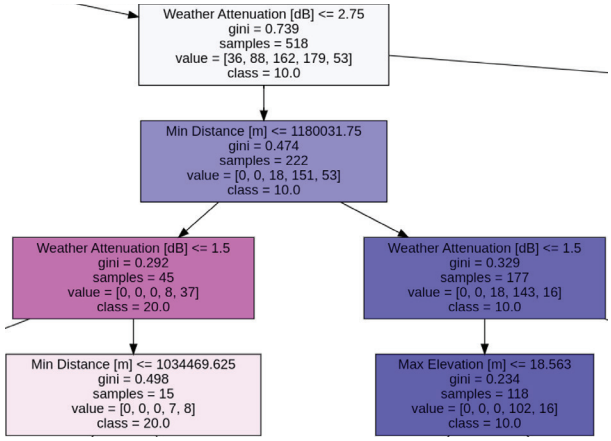


Fig 3. Zoomed in Decision Tree for bandwidth prediction for Aerocube 6A on 01.04.2024

The decision trees used for code rate prediction are significantly larger, with about fifteen layers of depth and approximately 8000 nodes for each day's prediction. Due to the sheer size and complexity of these trees, incorporating them directly into the paper as figures is impractical. The massive scale of these trees helps explain the extremely high accuracy of 99.5% in code rate predictions, illustrating a sophisticated understanding and handling of the different conditions that affect satellite communication.

To further analyze the performance of the models, a single satellite pass was chosen arbitrarily on the 3rd of May. The current machine learning models can't be compared by bits transmitted due to the lack of real-life experiments and are only pre-trained on simulated data. Therefore, we focus on comparing the accuracy of the predictions made by our models. These visualizations (Fig. 4 and Fig. 5) compare the baseline (formula-based) predictions with those from our machine learning models. The first graph compares the baseline code prediction and the Random Forest model across various bandwidths, where it closely aligns with only minor deviations of approximately 2-3 seconds in non-optimal selections of the code rate by the Random Forest.

The second graph shows the baseline compared to the Sequential Model across different bandwidths. Although it slightly struggles with timing the exact moment to change the code rate, it remains close to the baseline, exhibiting about 30 seconds of suboptimal predictions for every bandwidth.

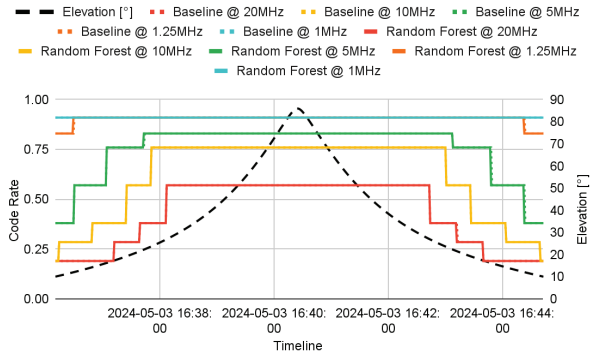


Fig 4. Baseline vs Random Forest Model code rate prediction for various bandwidths

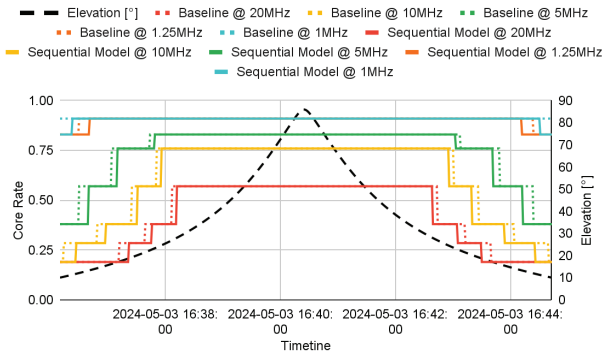


Fig 5. Baseline vs Sequential Model code rate prediction for various bandwidths

To conclude this section, a graph that shows the performance of the models in terms of goodput across different bandwidth options (Fig. 6) is shown. This graph offers a direct comparison of the average goodput achieved by the Random Forest and Sequential models against the baseline during the same satellite pass.

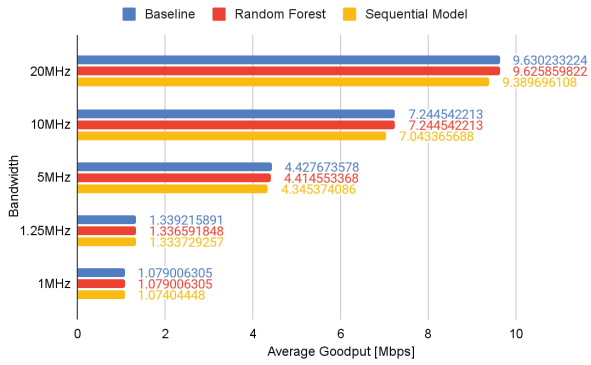


Fig 6. Comparison of average goodput by different models across different bandwidths

As illustrated in the graph (Fig 6), both machine learning models have a close result in goodput compared to the baseline, with the Random Forest model often outperforming the Sequential model. These results are quite encouraging as they show the models' high accuracy in predicting code rate even before advanced tuning. The consistent performance across different bandwidths suggests that the models can adapt effectively to varying conditions in satellite communications.

Following the evaluation of code rate prediction accuracy, we can also look at the bandwidth prediction results shown in Fig. 7 and a detailed breakdown made in Tab. 1. The satellite passes were sorted in ascending order by baseline bandwidth values to better visualize the details.

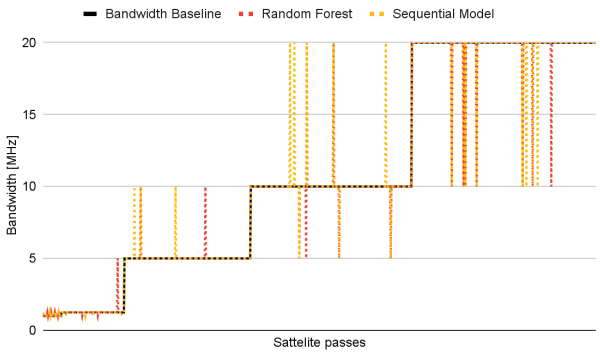


Fig 7. Comparison of bandwidth predictions for Baseline vs Random Forest Model vs Sequential Model

Tab. 1. Bandwidth predictions summary

Bandwidth [MHz]	*Total Sattelite Passes*	*Random Forest Misses*	*Sequential Model Misses*
20	296	7	8
10	259	6	8
5	203	2	3
1.25	101	3	5
1	30	8	6
Total	889	26	30

Both machine learning models, Random Forest, and Sequential, show good accuracy in predicting bandwidth values. Most mistakes happen when predicting the 1 MHz bandwidth. This is most likely due to the simulated conditions of the satellite passes in the training data being too extreme.

5. CONCLUSION

The application of machine learning to predict bandwidth and code rate settings for satellite communications has shown promising results. Comparing these models to traditional formula-based methods, we found that the Random Forest model achieved 97% accuracy in bandwidth prediction and 99.5% accuracy in code rate prediction. The Sequential Model also performed well, with 92% accuracy in bandwidth prediction and 95.3% accuracy for code rate predictions.

While these results are impressive, they do not yet match the theoretical formula-based approach, which is considered correct in the current phase. This indicates that while machine learning models provide significant advantages in adaptability and potentially reduce the need for fixed formula-based settings, further tuning is necessary to reach optimal performance. Looking forward, more data should also be collected, especially for bandwidth predictions where current data inputs may be insufficient.

5.1 Tuning Strategies and Future Directions

To enhance the performance of these models, several tuning strategies are planned. For Supervised Learning Models there is a plan to implement a genetic algorithm to fine-tune the supervised learning models. This method will help in optimizing the model parameters more effectively by simulating a process of natural selection, therefore improving accuracy. For the Random Forest model, adding weights to the trees is considered. This adjustment will prioritize more accurate predictions and minimize errors, particularly in situations where the model struggles with data complexity or variability. Furthermore, exploring hybrid solutions that integrate elements from both the Random Forest and Sequential models along with traditional methods could lead to a more robust system. This combined approach aims to leverage the strengths of each method to improve overall accuracy and reliability.

ACKNOWLEDGMENTS

This article is supported by grant number 0312/PRJG/8130. The authors would like to express their gratitude for the financial support provided, which has been essential in conducting this study.

REFERENCES

[1] Bishop, Christopher M. 2006. "Pattern Recognition and Machine Learning". *Information Science and Statistics* : 605 - 640

[2] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome. 2009. "The Elements of Statistical Learning Data Mining, Inference, and Prediction". *Springer Series in Statistics* : 587–602.

[3] Krenz, Rafał; Michalski, Marek; Zainutdinov, Ilia. 2023. "System Łączności Dla Satelitów Cubesat – Architektura i Protokoły". *Przegląd Telekomunikacyjny - Wiadomości Telekomunikacyjne,* 4/2023 : 224–227.

[4] Krenz, Rafał; Sybis, Michał; Zainutdinov, Ilia. 2022. "Wykorzystanie pasma C do łączności z nano-satelitami CubeSat". *Przegląd Telekomunikacyjny - Wiadomości Telekomunikacyjne*, 4/2022 : 371–375.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

KRYPTOGRAFICZNA OCHRONA UŻYTKOWNIKÓW SIECI 5G PRZED ATAKIEM TYPU MITM PRZY UŻYCIU WĘZŁÓW PRZEKAŹNIKOWYCH

ENSURING CRYPTOGRAPHIC PROTECTION OF 5G NETWORK USERS AGAINST MAN-IN-THE-MIDDLE ATTACK USING RELAY NODES

Mariusz Borowski[1]; Mateusz Buczkowski[2]

[1] Wojskowy Instytut Łączności - PIB, Zegrze Płd., m.borowski@wil.waw.pl
[2] Grandmetric sp. z o.o., Poznań, mateusz.buczkowski@grandmetric.com

DOI: 10.15199/59.2024.4.27

Streszczenie: **W artykule zaprezentowano metodę kryptograficznej ochrony użytkowników sieci 5G przed atakiem typu MitM przy użyciu węzłów przekaźnikowych. Metoda opiera się na wyznaczeniu kryptograficznej sumy kontrolnej poprzez mechanizm kryptograficzny HMAC, którego danymi wejściowymi są: klucz symetryczny K_PHYint oraz konkatenacja standardowej sumy kontrolnej wyliczonej w warstwie MAC i znacznika bloku, skonstruowanego poprzez połączenie zmiennych identyfikatorów użytych zasobów podczas komunikacji z prawdziwą stacją bazową.**

Abstract: **The article presents a method for cryptographic protection of 5G network users against Man-in-the-Middle (MitM) attacks using relay nodes. The method relies on determining cryptographic checksum through the HMAC cryptographic mechanism, with the input data being the symmetric key K_PHYint and the concatenation of the standard checksum calculated at the MAC layer and the block marker, constructed by combining the variable identifiers of the resources used during communication with the actual base station.**

Słowa kluczowe: **atak typu MitM, HMAC, integralność, kod uwierzytelniający, węzeł przekaźnikowy.**

Keywords: **authentication code, HMAC, integrity, MitM attack, relay node.**

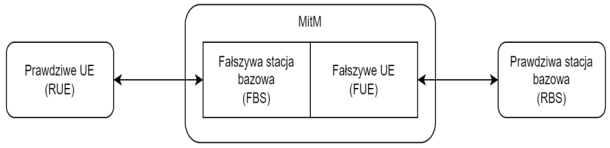
1. WSTĘP

Fałszywe stacje bazowe, znane również jako złośliwe stacje bazowe, to nieautoryzowane urządzenia, które naśladują legalne stacje bazowe sieci komórkowej. Te fałszywe stacje bazowe wykorzystują luki w infrastrukturze sieci 5G, umożliwiając atakującym przechwytywanie i manipulowanie wrażliwymi danymi użytkowników. Podszywając się pod legalną sieć, fałszywe stacje bazowe oszukują pobliskie urządzenia, aby się z nimi połączyły, umożliwiając atakującym podsłuchiwanie komunikacji, przeprowadzanie ataków typu MitM (ang. *Man-in-the-Middle*) i zbieranie danych o użytkownikach sieci i naruszenie ich prywatności. Standard 5G wprowadził znaczne ulepszenia w zakresie zwalczania fałszywych stacji bazowych, takie jak ukrywanie SUPI, gwarantowane odświeżanie GUTI, chronione przekierowania i ogólne ramy wykrywania informacyjnego. Niemniej ataki tego typu nadal są możliwe, zwłaszcza gdy zostaną wsparte poprzez węzły przekaźnikowe.

Praca powstała w projekcie „Zaawansowane metody i techniki wykrywania i przeciwdziałania atakom na infrastrukturę dostępową i aplikacje sieci 5G", w ramach programu Cyberbezpieczeństwo i e-Tożsamość, przy wsparciu Narodowego Centrum Badań i Rozwoju.

2. WĘZEŁ PRZEKAŹNIKOWY W ATAKU TYPU MITM

Węzeł przekaźnikowy w ataku typu MitM zazwyczaj zawiera fałszywą stację bazową (FBS) i fałszywe urządzenie abonenckie (FUE). Atakujący warstwy drugiej może umieścić węzeł przekaźnikowy MitM między prawdziwym UE (RUE), a prawdziwą stacją bazową (RBS). Jak pokazano na Rys.1, atakujący poprzez atak typu MitM może przechwytywać, dokładać i przesyłać dalej wiadomości z niezmienioną lub zmienioną treścią. Potencjalne manipulacje przesyłanymi komunikatami mogą prowadzić do wielu ataków, w tym między innymi: aLTEr [6], IMP4GT [7], błędnej konfiguracji sieci lub 5Greasoner.



Rys.1 Atak typu MitM przy użyciu węzłów przekaźnikowych

Istnieje zatem potrzeba zapobieżenia niewykrytej manipulacji zawartością wymienianych komunikatów i uzyskiwania tą droga możliwości prowadzenia kolejnych ataków [3].

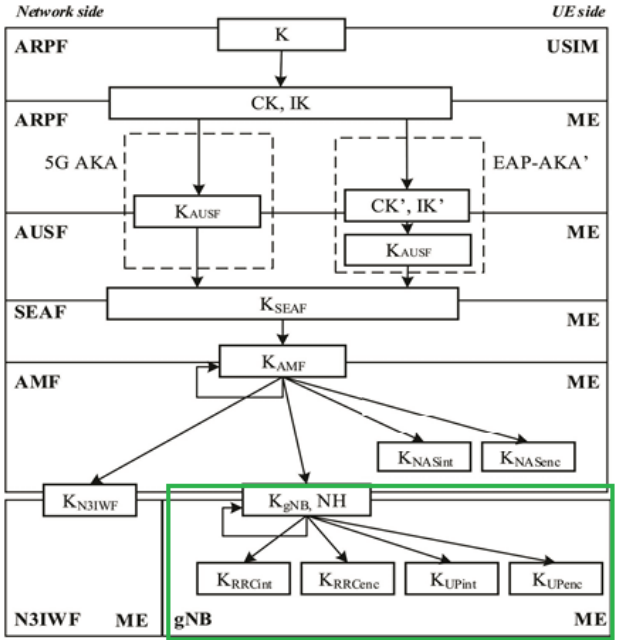
3. UŻYCIE KRYPTOGRAFICZNEJ SUMY KONTOROLNEJ DO OCHRONY PRZED ATAKAMI TYPU MITM

W celu zabezpieczenia przed atakiem typu MitM przy użyciu węzłów przekaźnikowych opracowano metodę (MF6), która dzięki nowej, kryptograficznej sumie kontrolnej 5gSTAR_MAC, wykrywa i blokuje powtarzające się bloki zasobów fizycznych w warstwie MAC/PHY bez konieczności sprawdzania czasu, które wymaga ścisłej synchronizacji czasowej [2], [4]. Metoda wykorzy-

stuje dodatkowy klucz symetryczny K_PHYint, wyprowadzony z materiału kluczowego dostępnego już w urządzaniu UE oraz stacji gNB. Każdy blok transportowy (ang. *transport block*, TB) występujący w warstwie MAC jest alokowany na blok zasobów fizycznych (ang. *physical resource block*, RB) użytych do jego transmisji. W warstwie fizycznej następuje podmiana standardowej 24-bitowej sumy kontrolnej na kryptograficzną sumę kontrolną o tej samej długości, wyznaczoną według schematu HMAC wykorzystującego zmienne identyfikatory zasobów radiowych użytych podczas komunikacji z rzeczywistą stacją bazową gNB.

3.1. Współdzielony klucz symetryczny K_PHYint

Metoda MF6 wyprowadza klucz K_PHYint z klucza K_gNB oraz dodatkowych, zmiennych identyfikatorów. Oznacza to, że metoda zaczyna działać dopiero po ustanowieniu zabezpieczeń AS (ang. *Access Stratum*). Klucz symetryczny K_PHYint jest wyprowadzany z klucza K_gNB w podobny sposób jak K_RRCenc, K_RRCint, K_UPenc oraz K_UPint, gdzie wartość 0x07 służy jako wyróżnik typu algorytmu w klauzuli A.8 przedstawionej w dokumencie TS 33.501 [1], co pokazano na Rys.2.



Rys.2 Schemat hierarchii generacji kluczy sieci 5G [1]

3.2. Działanie prawdziwego UE (RUE)

Prawdziwe urządzenie UE wysyłając blok transportowy:

- RUE oblicza 24-bitową wartość standardowej CRC.
- RUE uzyskuje informacje o przydzielonych blokach zasobów fizycznych RB do wysłania bloku transportowego i wyprowadza znacznik bloku blockID.
- RUE oblicza nową sumę 5gSTAR_MAC jako HMAC-SHA-256(K_PHYint, CRC || blockID).
- RUE umieszcza nową kryptograficzną sumę kontrolną 5gSTAR_MAC w polu sumy CRC.

3.3. Działanie prawdziwej stacji bazowej (RBS)

Prawdziwa RBS odbierając blok transportowy:

- RBS odbiera wiadomość w pewnych blokach RB i wyprowadza znacznik bloku blockID.
- RBS oblicza 24-bitową wartość standardowej sumy CRC' dla odebranego bloku transportowego.
- RBS oblicza kryptograficzną sumę 5gSTAR_MAC' na podstawie CRC' i blockID.
- RBS sprawdza, czy wyliczona suma 5gSTAR_MAC' i odebrana suma 5gSTAR_MAC są identyczne. Jeśli wartości nie zgadzają się to blok transportowy jest odrzucany. Jeśli wartości są identyczne, blok jest akceptowany, co oznacza, że nie zawiera błędów i został przesłany przy wykorzystaniu tych samych zasobów fizycznych tj. nie został odtworzony.
- RBS określa obecność atakującego typu MitM sprawdzając, czy liczba nieudanych kontroli sumy przekracza zadany próg.

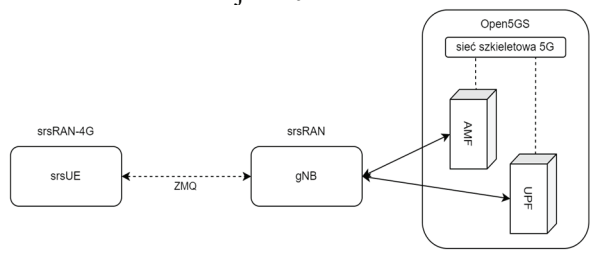
4. IMPLEMENTACJA METODY MF6

Metoda zabezpieczająca poprzez kryptograficzną sumę kontrolną przed atakami typu MitM przy wykorzystaniu węzłów przekaźnikowych (MF6) została oprogramowana przez specjalistów z firmy Grandmetric sp. z o.o. pod nadzorem merytorycznym naukowców z WIŁ-PIB [5]. Implementacja metody MF6 została wykonana w ramach środowiska opartego na srsRAN oraz srsUE, które pozwala na stosunkowo łatwą konfigurację w trybie programowym, omijając tym samym konieczność posiadania końcówki radiowej (np. USRP) do transmisji próbek. Zatem całość komunikacji między UE a gNB może ograniczyć się do pojedynczej maszyny wirtualnej.

Projekt srsRAN to projekt typu open source 5G CU/DU firmy SRS. Jest to rozwiązanie RAN zgodne ze specyfikacjami organizacji 3GPP i O-RAN Alliance. Projekt srsRAN obejmuje pełny stos L1/2/3 z minimalnymi zależnościami zewnętrznymi. Oprogramowanie jest przenośne między różnymi architekturami procesorów i jest skalowalne. W skład projektu wchodzi srsRAN gNB oraz srsUE. Urządzenia gNB i UE wymieniają miedzy sobą próbki IQ, które w rzeczywistości transmitowane są kanałem radiowym. W tym przypadku kanał radiowy został zamieniony na kolejkę ZMQ. ZMQ to wysokowydajna biblioteka do asynchronicznego przesyłania wiadomości, przeznaczona do stosowania w aplikacjach rozproszonych lub współbieżnych. Zapewnia kolejkę komunikatów i może działać bez dedykowanego brokera komunikatów. Można zatem warstwę komunikacji w środowisku radiowym zastąpić przesyłaniem próbek przez ZMQ.

Konfiguracja ta została uruchomiona na maszynie wirtualnej o następujących parametrach [5]:

- 6vCPU (rdzeni wirtualnych),
- 16GB RAM,
- 80GB przestrzeni dyskowej,
- z zainstalowanym systemem operacyjnym Ubuntu w wersji 22.04.2 LTS.



Rys. 3. Schemat środowiska wykorzystanego do implementacji i weryfikacji metody ochrony RAN

4.1. Wykonane modyfikacje programowe

Metoda opiera się na wyznaczeniu kryptograficznej sumy kontrolnej w oparciu o mechanizm kryptograficzny HMAC-SHA-256, którego danymi wejściowymi są: klucz symetryczny K_PHYint oraz konkatenacja standardowej sumy kontrolnej CRC wyznaczonej w warstwie MAC i znacznika bloku blockID, który skonstruowany jest poprzez połączenie zmiennych identyfikatorów wykorzystywanych zasobów podczas komunikacji z daną stacją bazową gNB.

W celu implementacji opisanej metody konieczne było wprowadzenie następujących zmian w projekcie srsRAN:

- Wyprowadzanie klucza K_PHYint w UE i gNB;
- Implementacja wyznaczania znacznika blockID;
- Implementacja wyznaczania nowej sumy kryptograficznej 5gSTAR_MAC;
- Implementacja podmiany standardowej sumy CRC na sumę 5gSTAR_MAC podczas wyliczania i weryfikacji sumy kontrolnej.

4.2. Wyprowadzenie znacznika bloku blockID

Znacznik bloku blockID został skonstruowany poprzez połączenie zmiennych identyfikatorów wykorzystywanych zasobów podczas komunikacji z prawdziwą stacją bazową gNB w następujący sposób:

blockID = SFN || SubF || Slot || ZasobyR.

Do wyznaczenia znacznika blockID użyto:

1. Numer ramki (SFN – ang. *System Frame Number*);
2. Numer ramki pomocniczej SubF (ang. *subframe index*);
3. Gniazdo Slot (ang. slot subframe index);
4. Identyfikator przydzielonych zasobów radiowych *ZasobyR*.

Pierwsze trzy parametry wchodzą w skład struktury zwanej *slotpoint*, natomiast parametr *ZasobyR* wynika wprost z działania mechanizmu przydzielana zasobów radiowych. Blok ten będzie dołączany do standardowej sumy kontrolnej CRC w celu wyznaczenia nowej, kryptograficznej sumy kontrolnej 5gSTAR_MAC.

4.3. Wyznaczanie klucza współdzielonego chroniącego integralność

Klucz symetryczny K_PHYint, współdzielony w urządzeniu UE oraz na stacji gNB, jest wyprowadzany z klucza K_gNB zgodnie z klauzulą A.8 dokumentu TS 33.501 [1], oraz Aneksem B dokumentu TS 33.220 w sposób podany poniżej:

K_PHYint = HMAC-SHA-256(K_gNB, S).

Klucz K_PHYint wyznaczany jest za pomocą zaimplementowanej metody *security_generate_k_phy*. Metoda ta przyjmuje na wejściu argumenty:

- Klucz główny K_gNB;
- ID algorytmu wyznaczania klucza zapewniającego integralność danych oraz wyróżnik algorytmu (string S).

Zwracana jest tablica 32 bajtów zawierająca klucz K_PHYint. Implementacja metody znajduje się w części oprogramowania, która jest wspólna dla terminala UE oraz stacji bazowej, więc implementacja wyprowadzania klucza jest identyczna dla obu komunikujących się stron.

Dodatkowo zdefiniowano nowy wyróżnik dla algorytmu:

```
#define ALGO_EPS_DISTINGUISHER_PHY_INT_ALG 0x07
```

Klucz K_gNB ma długość 256 bitów. Wyprowadzony klucz K_PHYint służy do wyznaczenia nowej kryptograficznej sumy kontrolnej 5gSTAR_MAC. Do poprawnego działania metody MF6 wymagana jest między innymi zmienność klucza K_gNB.

4.4. Wyznaczanie nowej sumy kryptograficznej

Metoda MF6 wykorzystuje tylko 24 najbardziej znaczące bity wyjściowe uzyskane na wyjściu algorytmu wyznaczającego kod uwierzytelniający

HMAC-SHA-256(K_PHYint, CRC || blockID).

W tym celu algorytm HMAC-SHA-256 został zmodyfikowany w taki sposób, że z wyliczonych 32 bajtów kodu pobierane są trzy najstarsze bajty, stanowiące nową kryptograficzną sumę kontrolną 5gSTAR_MAC.

5. WERYFIKACJA METODY MF6

W celu weryfikacji implementacji metody MF6 zabezpieczającej przed atakami typu MitM przy wykorzystaniu węzłów przekaźnikowych uruchomiony został minimalny zestaw testowy składający się z:

- sieci rdzeniowej,
- stacji bazowej,
- terminala abonenckiego UE.

5.1. Weryfikacja zmienności nowej kryptograficznej sumy kontrolnej 5gSTAR_MAC

Do sprawdzenia poprawności działania metody MF6 posłużono się narzędziem *ping*, które wysyła cyklicznie pakiet o tej samej zawartości. W ten sposób wyznaczono kolejne wartości znacznika bloku blockID, które były zmienne w czasie, gdyż zawierały takie parametry jak: numer ramki, numer podramki, nr slotu, identyfikator zasobu radiowego, co pokazano na Rys.4. Poniższy fragment logów ukazuje parametry wykorzystane do wyznaczenia znacznika blockID w przypadku cyklicznej wymiany danych między stacją bazową, a terminalem UE.

```
[5Gstar] handle_crc: sl_rx.sfn 13
[5Gstar] handle_crc: sl_rx.subframe_index 5
[5Gstar] handle_crc: sl_rx.slot_index 5
[5Gstar] handle_crc: sl_rx.subframe_slot_index 0
[5Gstar] handle_crc: Handling CRC
[5Gstar] handle_crc: sl_rx.sfn 108
[5Gstar] handle_crc: sl_rx.subframe_index 7
[5Gstar] handle_crc: sl_rx.slot_index 7
[5Gstar] handle_crc: sl_rx.subframe_slot_index 0
[5Gstar] handle_crc: Handling CRC
[5Gstar] handle_crc: sl_rx.sfn 109
[5Gstar] handle_crc: sl_rx.subframe_index 5
[5Gstar] handle_crc: sl_rx.slot_index 5
[5Gstar] handle_crc: sl_rx.subframe_slot_index 0
```

Rys.4 Rzeczywiste wartości parametrów wykorzystywanych do wyznaczenia znacznika blockID

Dzięki użyciu w teście pakietów o tej samej zawartości, ich standardowe sumy CRC są identyczne. Wtedy wyprowadzany znacznik bloku blockID będzie zależał od dodatkowo kontrolowanych w metodzie zmiennych parametrów czyli numeru ramki, numeru podramki, numeru slotu, identyfikatora zasobu radiowego (Rys.4). Uzyskujemy zatem zmienny wskaźnik bloku blockID, który podany na wejście algorytmu wyznaczającego kod uwierzytelniający HMAC-SHA-256 prowadzi do uzyskania różnych wartości nowej, kryptograficznej sumy kontrolnej

5gSTAR_MAC (8ddf93, b89c1a, 6aa2d6 odpowiednio) dla pakietów o takiej samej zawartości.

5.2. Weryfikacja poprawności działania

Na Rys.5, Rys.6 oraz Rys.7 przedstawiono zrzuty ekranu ze środowiska testowego [5]. Pokazują one, że połączenie pomiędzy stacją bazową gNB a terminalem UE zostało ustanowione poprawnie [5].

Stacja bazowa początkowo sygnalizuje logiem „*gNodeB started*" gotowość do przyjęcia połączeń od terminali UE (Rys.5).

```
root@dc-poz-p-5g-srsran:/home/service/srsRAN_Project/build/apps/gnb# ./gnb -c /home/service/srsRAN_Project/confi
gs/gnb_zmq.yaml
Lower PHY in executor blocking mode.

--== srsRAN gNB (commit 549d12131) ==--

Connecting to AMF on 10.53.1.2:38412
Available radio types: zmq.
Cell pci=1, bw=10 MHz, dl_arfcn=368500 (n3), dl_freq=1842.5 MHz, dl_ssb_arfcn=368410, ul_freq=1747.5 MHz

==== gNodeB started ===
Type <t> to view trace
[5Gstar] handle_crc: Handling CRC
[5Gstar] handle_crc: sl_rx.sfn 34
[5Gstar] handle_crc: sl_rx.subframe_index 4
[5Gstar] handle_crc: sl_rx.slot_index 4
[5Gstar] handle_crc: sl_rx.subframe_slot_index 0
[5Gstar] payload:
[5Gstar] handle_crc: Handling CRC
[5Gstar] handle_crc: sl_rx.sfn 40
[5Gstar] handle_crc: sl_rx.subframe_index 7
[5Gstar] handle_crc: sl_rx.slot_index 7
[5Gstar] handle_crc: sl_rx.subframe_slot_index 0
```

Rys.5 Logi z uruchomionej stacji gNB

Następnie, uruchamiany jest terminal UE (widoczny na Rys.6), który po krótkiej chwili dołącza się do stacji bazowej i wymienia komunikaty zgodnie z procedurą *RRC Connect*. Terminal oraz stacja bazowa wymieniają początkowe wiadomości, po których następuje przerwa w logach ze względu na brak transmisji.

```
Active RF plugins: libsrsran_rf_zmq.so
Inactive RF plugins:
Reading configuration file ../../srsRAN_Project/configs/ue_zmq.conf...

Built in Release mode using commit fa56836b1 on branch master.

Opening 1 channels in RF device=zmq with args=tx_port=tcp://127.0.0.1:2001,rx_port=tcp://127.0.0.1:2000,base_sra
te=11.52e6
Supported RF device list: zmq file
CHx base_srate=11.52e6
Current sample rate is 1.92 MHz with a base rate of 11.52 MHz (x6 decimation)
CH0 rx_port=tcp://127.0.0.1:2000
CH0 tx_port=tcp://127.0.0.1:2001
Current sample rate is 11.52 MHz with a base rate of 11.52 MHz (x1 decimation)
Current sample rate is 11.52 MHz with a base rate of 11.52 MHz (x1 decimation)
Waiting PHY to initialize ... done!
Attaching UE...
Random Access Transmission: prach_occasion=0, preamble_index=0, ra-rnti=0x39, tti=334
Random Access Complete.     c-rnti=0x4601, ta=0
RRC Connected
RRC Connected3
PDU Session Establishment successful. IP: 10.45.1.2
RRC NR reconfiguration successful.
```

Rys.6 Logi poprawnie podłączonego urządzenia UE

Dopiero wyzwolenie wysłania pakietów za pomocą narzędzia *ping* (Rys.7) generuje kolejne logi diagnostyczne wskazujące wartości *sfn*, *subframe index, slot index* itd. użyte także do weryfikacji zmienności nowej sumy kryptograficznej, co omówiono w podrozdziale 5.1.

```
PING 10.45.1.1 (10.45.1.1) 56(84) bytes of data.
64 bytes from 10.45.1.1: icmp_seq=1 ttl=64 time=0.056 ms
64 bytes from 10.45.1.1: icmp_seq=2 ttl=64 time=0.041 ms
64 bytes from 10.45.1.1: icmp_seq=3 ttl=64 time=0.055 ms
64 bytes from 10.45.1.1: icmp_seq=4 ttl=64 time=0.046 ms
64 bytes from 10.45.1.1: icmp_seq=5 ttl=64 time=0.053 ms
^C
--- 10.45.1.1 ping statistics ---
5 packets transmitted, 5 received, 0% packet loss, time 4091ms
rtt min/avg/max/mdev = 0.041/0.050/0.056/0.005 ms
```

Rys.7 Logi z narzędzia diagnostycznego ping

5.3. Weryfikacja zmienności kluczy K_gNB oraz K_PHYint

Każde uruchomienie środowiska testowego wyznacza nowy klucz K_gNB, co prowadzi do uzyskania unikalnego klucza chroniącego integralność K_PHYint. Uzyskane przykłady kluczy, w notacji ang. *little endian*, szesnastkowo, pokazano poniżej:

K_PHYint (1) : `e2e122bb09455bdb437d8cff002541f0 47aa4c35e49c5fa8e00cce167793384a`

K_PHYint (2) : `3c2f005bc4b83a1e7d2b8f695e654dd6 ddcf285aa605abaf18ce4a87f273ec59`

Unikalność klucza K_PHYint w schemacie wyznaczającym kod uwierzytelniający HMAC z funkcją skrótu SHA-2 pozwala na uzyskanie unikalnej kryptograficznej sumy zabezpieczającej 5gSTAR_MAC, zależnej od zasobów radiowych wykorzystanych do transmisji bloku transportowego. Nie ma potrzeby częstego odświeżania klucza K_PHYint, bo suma 5gSTAR_MAC jest obliczana i sprawdzana tylko do momentu, gdy numer ramki SFN w blockID ponownie przyjmie wartość początkową. To wystarczy, aby rzeczywista stacja bazowa wykryła obecność ataku typu MitM i podjęła stosowne działania informując abonentów UE oraz sieć o ataku.

6. PODSUMOWANIE

Metoda wprowadzająca ochronę przed atakami typu MitM przy użyciu węzłów przekaźnikowych, została z sukcesem zaimplementowana i przebadana w wydzielonej sieci 5G wykorzystywanej na potrzeby projektu 5gSTAR. Zabezpiecza ona, za pomocą mechanizmów kryptograficznych, przed możliwością niezauważonej manipulacji przesyłanymi komunikatami prowadzącej do materializacji kroków ataków na warstwę 5G.

Zweryfikowane metody przeciwdziałające atakom kryptograficznym na RAN sieci 5G umożliwiają „utwardzenie" tego standardu, co jest szczególnie pożądane w wydzielonych instalacjach specjalnych np. wojskowych.

LITERATURA

[1] 3GPP TS 33.501 v.17 (2022-01) Security architecture and procedures for 5G system.

[2] 3GPP 3GPP TR 33.809 v.18 (2023-09): Study on 5G security enhancements against False Base Station (FBS).

[3] Borowski Mariusz. „Ocena bezpieczeństwa algorytmów zapewniających poufność oraz uwierzytelnienie danych w sieci 5G". KSTiT, 2022.

[4] Borowski Mariusz. „Analiza zagrożeń oraz problemów z prywatnością w scenariuszach zakładających występowanie fałszywych stacji bazowych w sieci 5G, 5gSTAR, 2023.

[5] Buczkowski Mateusz, Borowski Mariusz. „Implementacja wybranych mechanizmów chroniących prywatność w atakach wykorzystujących fałszywe stacje bazowe w sieci 5G". 5gSTAR, 2023.

[6] Rupprech David, Kohls Katharina, Holz Thorsten. „Breaking LTE on Layer Two". IEEE Symposium on Security and Privacy, 2019.

[7] Rupprech David, Kohls Katharina, „IMP4GT: IMPersonation attacks in 4G neTworks". Network and Distributed System Security Symposium, 2020.

WYZWANIA I PERSPEKTYWY WDROŻENIA EFEKTYWNEJ KRYPTOGRAFII POSTKWANTOWEJ

CHALLENGES AND PERSPECTIVES OF IMPLEMENTING EFFECTIVE POST-QUANTUM CRYPTOGRAPHY

Michał Bajor, Piotr Chołda, Bartłomiej Gdowski, Daniel Karcz, Marcin Niemiec, Filip Opiłka, Miłosz Stypiński, Wojciech Szczepanik

AGH Akademia Górniczo-Hutnicza, Instytut Telekomunikacji, al. Mickiewicza 30, 30-059 Kraków

DOI: 10.15199/59.2024.4.28

Streszczenie: Kryptografia postkwantowa oferuje rozwiązania mające na celu ochronę danych przed potencjalnymi zagrożeniami wynikającymi z rozwoju komputerów kwantowych. Niniejszy artykuł analizuje globalne wysiłki i strategie wdrożenia tego rozwiązania, skupiając się na działaniach Unii Europejskiej oraz postępie prac w Polsce. Szczególną uwagę poświęcono projektowi PQ-REACT, który ma na celu rozwój i implementację zaawansowanych algorytmów postkwantowych w systemach teleinformatycznych i infrastrukturze krytycznej.

Abstract: Post-quantum cryptography offers solutions aimed at protecting data from potential threats arising from the development of quantum computers. This article analyzes global efforts and strategies for implementing this solution, focusing on the actions of the European Union and the progress of work in Poland. Special attention is given to the PQ-REACT project, which aims to develop and implement advanced post-quantum algorithms in telecommunication systems and critical infrastructure.

Słowa kluczowe: bezpieczeństwo, algorytmy postkwantowe, kryptografia asymetryczna, ochrona danych.

Keywords: security, post-quantum algorithms, asymmetric cryptography, data protection.

1. WSTĘP

W obliczu szybkiego rozwoju technologii kwantowych, bezpieczeństwo danych i infrastruktury cyfrowej staje się coraz bardziej zagrożone. Komputery kwantowe, dzięki nowemu podejściu do wykonywania obliczeń, mogą zrewolucjonizować wiele dziedzin, ale jednocześnie stanowią poważne wyzwanie dla współczesnej kryptografii. Kryptografia postkwantowa (Post-Quantum Cryptography — PQC) staje się zatem nieodzownym narzędziem w budowaniu systemów teleinformatycznych odpornych na ataki z użyciem komputerów kwantowych.

W ramach globalnych wysiłków na rzecz bezpieczeństwa w cyberprzestrzeni, Unia Europejska podejmuje konkretne działania mające na celu stworzenie harmonijnego planu wdrażania kryptografii postkwantowej. Rekomendacja Komisji Europejskiej [1] zachęca państwa członkowskie do opracowania strategii adaptacji PQC, które zapewnią spójność i synchronizację działań w całej Unii. Podejmowane są także inicjatywy badawcze i rozwojowe w tym zakresie. Jednym z kluczowych projektów jest PQ-REACT (*Post-Quantum Cryptography Framework for Energy Aware Contexts*), finansowany przez Unię Europejską w ramach programu Horyzont Europa. Projekt ten koncentruje się na projektowaniu i implementacji koncepcji płynnego przejścia od obecnie stosowanych szyfrów asymetrycznych do algorytmów postkwantowych.

W niniejszym artykule przedstawiono zarówno globalne wysiłki na rzecz implementacji PQC, jak i postęp prac w Polsce, ze szczególnym uwzględnieniem projektu PQ-REACT. Przedstawione zostały również wyniki badań dotyczących wydajności algorytmów postkwantowych w różnych scenariuszach użycia, co ma kluczowe znaczenie dla dalszego rozwoju tej dziedziny.

2. KRYPTOGRAFIA POSTKWANTOWA

W 1997 roku Peter Shor opublikował badania dotyczące algorytmu, który rozkłada na czynniki liczby złożone i znajduje logarytmy dyskretne w czasie wielomianowym [2]. W efekcie algorytm Shora zagraża dzisiejszej kryptografii asymetrycznej, której bezpieczeństwo warunkowe opiera się na wspomnianych problemach algebraicznych. Do wdrożenia omawianego algorytmu niezbędny jest komputer kwantowy z odpowiednio dużą liczbą kubitów, czyli bitów kwantowych.

Kryptografia postkwantowa (PQC) ma na celu opracowanie i implementację algorytmów odpornych na próby łamania zarówno na tradycyjnym komputerze, jak i na komputerze kwantowym. Dotychczas zdefiniowano szereg algorytmów spełniających ten warunek. Tak jak w przypadku standardowych algorytmów asymetrycznych, algorytmy typu PQC opierają swoje bezpieczeństwo na braku możliwości odwrócenia pewnych działań algebraicznych. Kody korekcyjne, operacje na kratach, funkcje jednokierunkowe oraz izogenie krzywych eliptycznych to niektóre mechanizmy umożliwiające implementację algorytmów PQC [3].

Znaczącym krokiem w badaniach algorytmów PQC był konkurs ogłoszony przez jednostkę standaryzacyjną NIST (*National Institute of Standards and Technology*). Celem konkursu było wyłonienie bezpiecznych algorytmów mogących zastąpić aktualnie wykorzystywaną kryptografię asymetryczną [4].

Zaznaczyć jednak trzeba, że kryptografia postkwantowa wiąże się z pewnymi wyzwaniami, które stoją na drodze jej powszechnego wykorzystania. Najbardziej znaczącym wydaje się brak szeroko rozumianych standardów związanych z kryptografią. Co za tym idzie, nie ma zbyt wiele gotowych rozwiązań (zarówno sprzętowych jak i programowych), które pozwalałyby na szybką integrację w już istniejących produktach, a także brak uniwersalnych bibliotek programistycznych, które pozwoliłyby na relatywnie łatwą implementację algorytmów postkwantowych. Należy jednak zauważyć, że aby skutecznie zagrozić obecnie wykorzystywanym standardom kryptografii klasycznej potrzeba komputerów kwantowych o znacznie większej liczbie kubitów niż obecnie dostępne.

3. POSTĘP PRAC NAD KRYPTOGRAFIĄ POSTKWANTOWĄ

W ostatnich latach, w miarę postępu prac nad komputerami kwantowymi, które zagrażają obecnym rozwiązaniom kryptograficznym, kryptografia postkwantowa coraz bardziej zyskuje na znaczeniu. Rozwój PQC jest kluczowy dla zapewnienia bezpieczeństwa danych w przyszłym ekosystemie cyfrowym zawierającym komputery kwantowe. Zatem potencjalne zastosowanie dla kryptografii postkwantowej to tak naprawdę zastąpienie we wszystkich sektorach gospodarki obecnie stosowanej kryptografii asymetrycznej, wykazującej słabości w kontekście algorytmów kwantowych. Jako główne przykłady można wymienić ochronę komunikacji w sieci przy użyciu certyfikatów cyfrowych, uwierzytelnianie przy użyciu kluczy prywatnych czy bezpieczne podpisy cyfrowe.

Niezwykle istotny jest aspekt związany ze skutecznością i efektywnością nowych rozwiązań PQC. Za skuteczną kryptografię uznamy taką, której wynikiem działania jest szyfrogram odporny na próby jego złamania — zarówno za pomocą klasycznych komputerów jak i komputerów kwantowych. Natomiast efektywna kryptografia musi również brać pod uwagę wydajność poszczególnych algorytmów — w tym czas i zasoby potrzebne na wykonanie obliczeń.

3.1 Perspektywa globalna

Obecnie amerykański NIST [4] pełni wiodącą rolę w dziedzinie standaryzacji algorytmów postkwantowych. Rozpoczęty w 2016 roku proces standaryzacji skupia się na wyłonieniu najbezpieczniejszych i najbardziej efektywnych algorytmów kryptograficznych odpornych na ataki z użyciem komputerów kwantowych. W ramach tego procesu analizowane są różne rodzaje algorytmów, w tym algorytmy kratowe takie jak CRYSTALS-KYBER i NTRU, algorytmy oparte na kodach korekcyjnych, np. McEliece, oraz algorytmy bazujące na izogeniach krzywych eliptycznych, takie jak SIKE.

Równocześnie, Europa aktywnie uczestniczy w tych działaniach dzięki inicjatywom takim jak Quantum Flagship [5]. Program ten, powołany przez Komisję Europejską, obejmuje projekty badawcze i demonstracyjne mające na celu przyspieszenie komercjalizacji technologii kwantowych, w tym bezpieczeństwa kwantowego i PQC.

Giganci technologiczni tacy jak Google, IBM i Microsoft również intensywnie inwestują w badania nad kryptografią kwantową (opartą na ustalaniu kluczy kryptograficznych poprzez przesyłanie informacji kwantowej) i postkwantową. Te firmy opracowują własne laboratoria badawcze oraz współpracują ze środowiskiem akademickim i jednostkami państwowymi, badając wykorzystanie algorytmów takich jak Picnic [6] i FrodoKEM [7], w celu rozwoju bezpiecznych technologii szyfrowania.

Na poziomie akademickim, uniwersytety takie jak MIT, Stanford, oraz Uniwersytet w Toronto prowadzą zaawansowane badania w dziedzinie PQC. Równolegle, start-upy i przedsiębiorstwa technologiczne, takie jak ISARA i PQShield [8], specjalizują się w tworzeniu rozwiązań kryptograficznych, które są zarówno wydajne jak i bezpieczne w obliczu zagrożeń kwantowych. Zatem rozwiązania kryptograficzne już ewoluują w odpowiedzi na postępy w informatyce kwantowej, a kontynuacja globalnych wysiłków w badania, rozwój i implementację nowych rozwiązań jest niezbędna, aby zapewnić bezpieczeństwo danych cyfrowych w przyszłości.

3.2 Postęp prac w Polsce

Obecnie w sektorze cywilnym w Polsce, w kontekście oficjalnych rekomendacji wydawanych przez organy publiczne, nie istnieje oficjalny program ani zestaw zaleceń dotyczących kryptografii postkwantowej. Odpowiedzialność za wydawanie Narodowych Standardów Cyberbezpieczeństwa, będących ciągle aktualizowaną serią bieżących rekomendacji zgodnych przede wszystkim z NIST SP, spoczywa na Pełnomocniku Rządu ds. Cyberbezpieczeństwa. Obecnie te standardy nie obejmują wytycznych dotyczących PQC. Również Strategia Cyberbezpieczeństwa Rzeczypospolitej Polskiej na lata 2019-2024 nie zawiera zaleceń dotyczących kryptografii postkwantowej.

W sektorze wojskowym odpowiedzialność za bezpieczeństwo w cyberprzestrzeni przede wszystkim powierzona jest Wojskom Obrony Cyberprzestrzeni, które przekształcone zostały z Narodowego Centrum Bezpieczeństwa Cyberprzestrzeni. Działania specjalistów wojskowych pracujących nad protokołami kryptograficznymi uznawane są za poufne, dlatego nie można ujawnić oficjalnych informacji na temat ich pracy w polskim sektorze wojskowym. Niemniej jednak ogłoszono pewne inicjatywy w dziedzinie kryptografii postkwantowej.

Jednym z projektów z tego zakresu jest APQ (Opracowanie, implementacja i ocena bezpieczeństwa algorytmów postkwantowych) [9], realizowany w latach 2022-2024. Cel projektu to rozwój i ocena asymetrycznych algorytmów kryptograficznych odpornych na ataki komputerów kwantowych. Ogłoszono także nowy program badawczy Narodowego Centrum Badań i Rozwoju pod nazwą NCBR 1/PERUN/2023 [10]. Program ten składa się z serii projektów zleconych przez wojsko. Jeden z obszarów badawczych uwzględnionych w programie jest poświęcony technologiom kwantowym i rozwojowi technologii w dziedzinie metod automatycznej oceny algorytmów postkwantowych. Wybrane projekty badawcze mają rozpocząć się w drugiej połowie 2024 roku.

Warto wspomnieć o polskich start-upach poświęconych badaniom nad kryptografią postkwantową, które działają w ramach inicjatywy znanej jako Klaster Technologii Kwantowych, lub krócej „Klaster Q” [11]. Zrzesza on prywatne przedsiębiorstwa, uczelnie i instytuty naukowe afiliowane przy Polskiej Akademii Nauk. Wśród nich warto wspomnieć o dwóch opisanych poniżej.

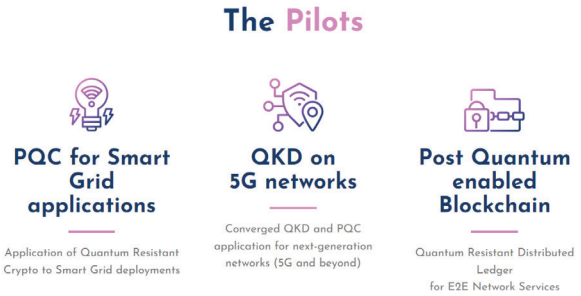
- Quantum Blockchains, skupiający się na tworzeniu rozwiązań opartych na technologii blockchain, które obejmują także algorytmy postkwantowe, w tym CRYSTALS-Dilithium [13].

- ResQuant, specjalizujący się w rozwoju rozwiązań sprzętowych opartych na algorytmach postkwantowych, szczególnie z wykorzystaniem technologii FPGA. Oferta obejmuje dostosowywane akceleratory kryptograficzne implementujące algorytmy CRYSTALS-Dilithium, XMSS i CRYSTALS-Kyber. ResQuant jest finansowany przez NCBR w ramach programu Bridge Alfa dla start-upów [12].

4. PROJEKT PQ-REACT

We wrześniu 2023 roku rozpoczęła się realizacja 3-letniego projektu badawczego PQ-REACT (*Post-Quantum Cryptography Framework for Energy Aware Contexts*), finansowanego w ramach programu ramowego Unii Europejskiej Horyzont Europa. Projekt PQ-REACT ma na celu zaprojektowanie, opracowanie i zweryfikowanie koncepcji płynnego przejścia od współcześnie używanych szyfrów asymetrycznych do kryptografii postkwantowej — w różnych sektorach, zastosowaniach i scenariuszach użycia [14]. W skład konsorcjum wchodzi 12 jednostek z 7 krajów UE (Grecja, Włochy, Bułgaria, Niemcy, Polska, Estonia i Hiszpania). Koordynatorem jest grecka jednostka badawcza — National Centre of Scientific Research „Demokritos".

Zadania projektu PQ-REACT koncentrują się wokół problemów bezpieczeństwa i wydajności algorytmów postkwantowych (głównie rekomendowanych przez NIST), a także aspektów prawnych migracji kryptografii asymetrycznej w kierunku szyfrów odpornych na ataki z użyciem komputerów kwantowych. Jednak kluczowym elementem projektu — przedstawionym schematycznie na rysunku 1 — są praktyczne implementacje kryptografii kwantowej lub postkwantowej w trzech różnych zastosowaniach: urządzeniach inteligentnej sieci energetycznej (tzw. *smart grid*), usługach oferowanych przez sieć mobilną piątej generacji (5G) oraz platformach obsługiwanych przez technologię blockchain.



Rysunek 1: Planowane implementacje algorytmów postkwantowych w ramach projektu PQ-REACT [14]

Instytut Telekomunikacji AGH aktywnie uczestniczy w pracach projektu PQ-REACT. Dotychczasowa aktywność skupiała się przede wszystkim na analizie bibliotek programistycznych, które zawierają programowe implementacje algorytmów postkwantowych, śledzenia działań związanych z oceną szyfrów postkwantowych przez międzynarodowe jednostki standaryzacyjne oraz analizą działań z zakresu kryptografii postkwantowej w Europie. Obecnie zespół z AGH rozpoczyna prace nad usługami bezpieczeństwa opartymi na praktycznych implementacjach kwantowej dystrybucji kluczy kryptograficznych w sieci 5G hiszpańskiego operatora telekomunikacyjnego. Kontynuowane są także badania efektywności szyfrów postkwantowych w różnych środowiskach programistycznych, umożliwiające praktyczną implementację usługi podpisu cyfrowego.

5. ANALIZA WYDAJNOŚCI

W ramach dotychczasowych badań przeprowadzone zostały testy wydajności algorytmów postkwantowych wykorzystywanych do podpisów cyfrowych. Analizowano implementację wybranych algorytmów, takich jak CRYSTALS-Dilithium, Falcon i SPHINCS+, przy użyciu biblioteki `liboqs` [15]. Testy obejmowały procesy generowania par kluczy, podpisywania plików oraz weryfikacji podpisów. Uzyskane wyniki posłużyły do testów porównawczych z powszechnie używanym algorytmem RSA. Szczegółowe wyniki zostały zaprezentowane w pracy [16]. Kluczowym elementem analizy był kompromis między bezpieczeństwem, a wydajnością stosowanych rozwiązań.

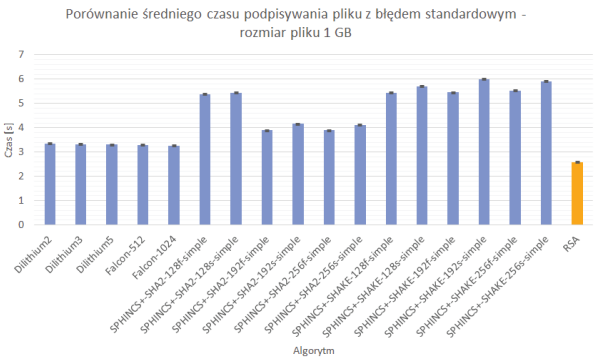
Środowisko testowe zostało skonfigurowane na maszynie wirtualnej z systemem operacyjnym Ubuntu 22.04.1 LTS. Platforma dysponowała 12 GB pamięci RAM oraz 8 rdzeniami procesora Intel Core i7-9750H o taktowaniu 2,60 GHz. Każda funkcjonalność została przetestowana w zestawieniu ze wszystkimi algorytmami, a w celu zapewnienia wiarygodności pomiarów, każdy test został powtórzony 101 razy. Warto wspomnieć, że czas pierwszego pomiaru dla danego algorytmu był zauważalnie dłuższy, dlatego był odrzucany. Dodatkowy narzut czasu wynikał z faktu, że podczas pierwszego powtórzenia dane były odczytywane z dysku, a następnie ładowane do pamięci podręcznej. Dzięki pominięciu pierwszego pomiaru w każdej serii, możliwe jest wykluczenie czynnika prędkości zapisu oraz odczytu danych. Sporządzone wykresy przedstawiają średni czas wykonania algorytmu wraz z słupkami błędów oznaczającymi błąd standardowy średniej (SEM).

W pierwszej kolejności przetestowana została wydajność czasowa generowania par kluczy. Algorytmy Dilithium osiągają najlepszą wydajność czasową wśród testowanych. Analizując wyniki algorytmów z grupy SPHINCS+ można zauważyć znaczącą dysproporcję pomiędzy tymi z przyrostkiem „s", a tymi z przyrostkiem „f". Jest to spowodowane różnymi strategiami przy ich implementacji. Przyrostek „s" wskazuje na algorytm zoptymalizowany pod kątem minimalizacji rozmiaru generowanego podpisu cyfrowego, natomiast „f" oznacza, że priorytetem była optymalizacja szybkości działania algorytmu. Uzyskane wyniki potwierdzają zgodność z założeniami różnych podejść implementacyjnych.

Innym badanym elementem była wydajność czasowa algorytmów postkwantowych podczas procesu weryfikacji podpisów cyfrowych. Testy przeprowadzono na plikach o trzech różnych rozmiarach: 10 MB, 100 MB oraz 1 GB. W przeciwieństwie do generowania par kluczy średnie czasy dla większości algorytmów były bardzo zbliżone. Jest to zgodne z założeniami teoretycznymi. Wyjątek stanowiła jedynie grupa czterech algorytmów z rodziny algorytmów opartych na funkcjach skrótu — SPHINCS+.

Głównym elementem poddanym analizie był proces podpisywania plików. Testy w tym zakresie również przeprowadzono na plikach o trzech różnych rozmiarach: 10 MB, 100 MB oraz 1 GB. Na rysunku 2 przedstawione zostały wyniki ostatniego z wymienionych przypadków. Zauważyć można, że algorytmy należące do grup Dilithium i Falcon wykazały najkrótsze czasy podpisywania.

Dodatkowo, wyniki te zostały zestawione z powszechnie stosowanym algorytmem RSA. Biorąc pod uwagę złożoność algorytmów postkwantowych, można oczekiwać pewnego narzutu czasowego w porównaniu z RSA. Jednak różnice te nie są drastyczne, a w zamian za dodatkowy czas, algorytmy postkwantowe oferują znaczącą poprawę bezpieczeństwa.



Rysunek 2: Porównanie średniego czasu podpisywania pliku o rozmiarze 1 GB

Podsumowując, algorytmy z grupy Dilithium, a w szczególności najlepszy z nich — oznaczony jako Dilithium5, wydają się odpowiednimi kandydatami do zastąpienia RSA w erze kwantowej transformacji systemów kryptograficznych. OpenQuantumSafe [17] deklaruje, że Dilithium5 osiąga najwyższy, piąty poziom bezpieczeństwa w pięciostopniowej skali NIST, podczas gdy testowana wersja RSA z kluczem o długości 2048 bitów szacowana jest jako poziom pierwszy [18]. W przeprowadzonych testach na pliku o rozmiarze 1 GB, Dilithium5 osiągnął średni czas podpisu wynoszący 3,31 s, w porównaniu do 2,56 s dla RSA. Stanowi to narzut 27,7%, co jest ceną za odporność na ataki z użyciem komputera kwantowego.

6. PODSUMOWANIE

Rozwój technologii kwantowych wymaga wdrożenia kryptografii postkwantowej, być może już w niedalekiej przyszłości. Standaryzacja PQC jest kluczowa dla bezpieczeństwa danych przesyłanych w sieciach teleinformatycznych. Unia Europejska już teraz dąży do harmonijnego wdrażania PQC, a projekty takie jak PQ-REACT pokazują zaangażowanie w rozwój tej technologii. W Polsce, mimo braku oficjalnych programów, prowadzone są ważne inicjatywy badawcze. Badania wykazują, że algorytmy postkwantowe są gotowe do praktycznego zastosowania, mimo wyzwań związanych z efektywnością tych rozwiązań. Również edukacja i świadomość społeczna na temat kryptografii postkwantowej będą w przyszłości niezwykle ważne. Szczególnie, że stosowanie słabych algorytmów kryptograficznych do zabezpieczania danych wrażliwych, może być poważnym problemem w szerokim horyzoncie czasowym — gdy komputery kwantowe umożliwią skuteczne łamanie dziś stosowanych zabezpieczeń.

LITERATURA

[1] Komisja Europejska. „Recommendation on a Coordinated Implementation Roadmap for the transition to Post-Quantum Cryptography". https://digital-strategy.ec.europa.eu/en/library/recommendation-coordinated-implementation-roadmap-transition-post-quantum-cryptography/ Data dostępu: 15.05.2024.

[2] Shor Peter. 1997. „Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer". *Journal of Computing* 26, pp. 1484-1509.

[3] Dam, Duc-Thuan, Thai-Ha Tran, Van-Phuc Hoang, Cong-Kha Pham, and Trong-Thuc Hoang. 2023. „A Survey of Post-Quantum Cryptography: Start of a New Race". *Cryptography* 7, no. 3: 40.

[4] National Institute of Standards and Technology (NIST). 2022. „Post-Quantum Cryptography". https://www.nist.gov/pqcrypto/ Data dostępu: 15.05.2024.

[5] Quantum Flagship. https://qt.eu/ Data dostępu: 15.05.2024.

[6] Picnic, A Family of Post-Quantum Secure Digital Signature Algorithms, Strona internetowa projektu. https://microsoft.github.io/Picnic/ Data dostępu: 15.05.2024.

[7] FrodoKEM, Practical quantum-secure key encapsulation from generic lattices, Strona internetowa projektu. https://frodokem.org/ Data dostępu: 15.05.2024.

[8] PQShield, Strona internetowa projektu. https://pqshield.com/ Data dostępu: 15.05.2024.

[9] Narodowe Centrum Badań i Rozwoju. Opracowanie, implementacja i ocena bezpieczeństwa algorytmów postkwantowych pk. APQ. https://www.gov.pl/web/ncbr/konkurs-nr-122022/ Data dostępu: 15.05.2024.

[10] Narodowe Centrum Badań i Rozwoju. Konkurs nr 1/PERUN/2023. https://www.gov.pl/web/ncbr/konkurs-nr-1perun2023/ Data dostępu: 15.05.2024.

[11] Klaster Q — Klaster Technologii Kwantowych, Strona internetowa projektu. https://klasterkwantowy.pl/ Data dostępu: 15.05.2024.

[12] ResQuant, Strona internetowa projektu. https://www.resquant.com/ Data dostępu: 15.05.2024.

[13] Quantum Blockchains, Strona internetowa projektu. https://www.quantumblockchains.io/ Data dostępu: 15.05.2024.

[14] PQ-REACT, Strona internetowa projektu. https://pqreact.eu/ Data dostępu: 15.05.2024.

[15] liboqs, Strona internetowa projektu. https://github.com/open-quantum-safe/liboqs/ Data dostępu: 15.05.2024.

[16] Opiłka Filip, Niemiec Marcin, Gagliardi Maria, Kourtis Michail Alexandros. 2024. „Performance Analysis of Post-Quantum Cryptography Algorithms for Digital Signature". *Applied Sciences* 14, no. 12: 4994.

[17] Douglas Stebila, Michele Mosca. 2017. „Post-quantum key exchange for the Internet and the Open Quantum Safe project". In Roberto Avanzi, Howard Heys, editors, *Selected Areas in Cryptography (SAC)* 2016, LNCS, vol. 10532, pp. 1–24. Springer. https://openquantumsafe.org Data dostępu: 15.05.2024.

[18] National Institute of Standards and Technology (NIST). 2020. „Status Report on the Second Round of the NIST Post-Quantum Cryptography Standardization Process". NISTIR 8309.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

METODA LOKALIZACJI UWZGLĘDNIAJĄCA CHARAKTERYSTYCZNE TRAJEKTORIE RUCHU

POSITIONING METHOD CONSIDERING CHARACTERISTIC MOVEMENT TRAJECTORIES

Marcin Kołakowski

Instytut Radioelektroniki i Technik Multimedialnych, Politechnika Warszawska, Warszawa
marcin.kolakowski@pw.edu.pl

DOI: 10.15199/59.2024.4.29

Streszczenie: W referacie zaprezentowano koncepcję metody lokalizacji przemieszczającej się osoby za pomocą smartfona. Metoda wykorzystuje dane z sensorów inercyjnych i pomiary poziomu sygnałów WiFi oraz fakt, że ze względu na ograniczenia przestrzenne część trajektorii ruchu może być zaobserwowana jedynie w wybranych miejscach.

Abstract: The paper presents a smartphone-based method for indoor positioning of a moving person. The method uses data from inertial sensors and WiFi signals power measurements. The method utilizes a fact that in indoor environment, the movement trajectories are constrained by obstacles and some patterns can be observed only in selected places.

Słowa kluczowe: lokalizacja wewnątrzbudynkowa, uczenie maszynowe, WiFi

Keywords: indoor positioning, machine learning, WiFi

1. WSTĘP

Szybko postępujące starzenie się populacji w krajach rozwiniętych poskutkowało w ostatnich latach dużą liczbą nowoczesnych platform sprzętowo-programowych mających na celu wsparcie osób starszych. Celem większości proponowanych rozwiązań jest zapewnienie wsparcia poprzez ciągłe monitorowanie stanu osoby starszej oraz reakcja w sytuacjach zagrażających jej życiu bądź zdrowiu [7]. Bardziej wysublimowane platformy pozwalają również na monitorowanie zachowania użytkownika poprzez zastosowanie systemów lokalizacyjnych.

Typowe systemy lokalizacyjne składające się z lokalizowanego urządzenia (etykiety), węzłów oraz komputera przetwarzającego dane często spotykają się z niechęcią osób starszych. Ma na to wpływ ich stosunkowo wysoka cena, konieczność instalacji infrastruktury w często już umeblowanym mieszkaniu oraz wymagane noszenie dodatkowego urządzenia. Ostatni problem można wyeliminować poprzez zastosowanie systemów typu "device-free" [2]. Niestety, ich ograniczona dokładność może być niewystarczająca do niektórych zastosowań.

Rozwiązaniem powyższych problemów może być implementacja systemu opartego na urządzeniach już posiadanych i używanych przez docelowego użytkownika np. smartfon czy smartwatch. Takie podejście znacznie zmniejszyłoby barierę wejścia oraz pozwoliło na umasowienie rozwiązania.

W referacie zaprezentowano koncepcję metody lokalizacji wykorzystującą dane inercyjne oraz poziomy sygnałów sieci WiFi zarejestrowane za pomocą smartfona. W zaproponowanej metodzie, do wyznaczenia lokalizacji, oprócz pomiarów wykonywanych na bieżąco, wykorzystywana jest trajektoria ruchu osoby z ostatnich kilku sekund. Takie podejście pozwala na zwiększenie dokładności lokalizacji poprzez uwzględnienie faktu, że niektóre trajektorie, ze względu na ograniczenia przestrzenne są charakterystyczne dla wybranych punktów środowiska. W zaproponowanej metodzie jest to uwzględniane poprzez zastosowanie modułu prognozy lokalizacji użytkownika wytrenowanego na trajektoriach typowych dla środowiska pracy systemu.

W literaturze można znaleźć inne rozwiązania wykorzystujące charakterystyczne cechy punktów przestrzeni. W [4] zaproponowano rozwiązanie wykrywające moment otwarcia drzwi co pozwala na określenie lokalizacji użytkownika w ich okolicy. Podobne rozwiązanie zastosowano w [1] gdzie oprócz przejścia przez drzwi wykrywane jest wchodzenie po schodach i zmiana kierunku.

Metody korzystające z faktu że możliwe trajektorie ruchu są ograniczone przez rozkład pomieszczenia i przeszkód są mniej popularne. W [6] wyznaczone lokalizacje są poddawane dodatkowym ograniczeniom by uniknąć trajektorii nie mających sensu. W [3] analizowane jest prawdopodobieństwo wystąpienia danej lokalizacji użytkownika na podstawie konfiguracji przestrzennej pomieszczenia.

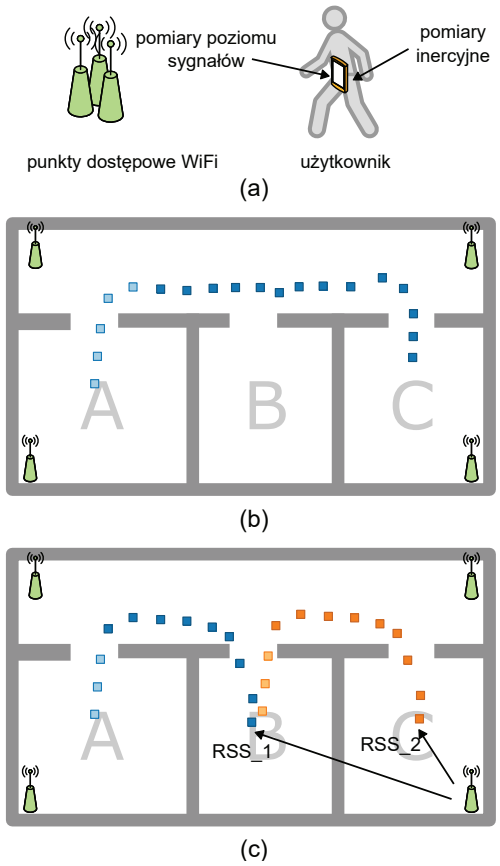
2. KONCEPCJA METODY

Zaproponowana metoda została opracowana z myślą o lokalizacji użytkownika za pomocą smartfona. Schemat ideowy koncepcji został przedstawiony na Rys. 1.

W metodzie użytkownik jest wyposażony w smartfon wykonujący pomiary inercyjne (przyspieszenia oraz prędkości kątowej) pozwalające na określenie jego trajektorii ruchu oraz pomiary mocy sygnałów WiFi nadawanych przez punkty dostępowe obecne w pobliżu Rys. 1.a.

Metoda, do lokalizacji użytkownika, korzysta z faktu, że w przestrzeniach wewnątrzbudynkowych, ze względu na ich rozkład i przeszkody w nich występujące, pewne określone trajektorie (np. długie proste ścieżki lub sekwencje skrętów), mogą być zarejestrowane w wybranych

punktach. Przykład takiego środowiska i trajektorii został przedstawiony na Rys. 1.b. Ścieżka składająca się z długiego prostego fragmentu zakończona skrętem w prawo może być zarejestrowana wyłącznie w jednym przypadku – przejścia z pokoju A do C. Informacja, że użytkownik pokonał taką drogę daje nam informację o jego lokalizacji.



Rysunek 1: Koncepcja metody: (a) architektura systemu; (b) jednoznaczna trajektoria; (c) niejezdnoznaczne trajektorie

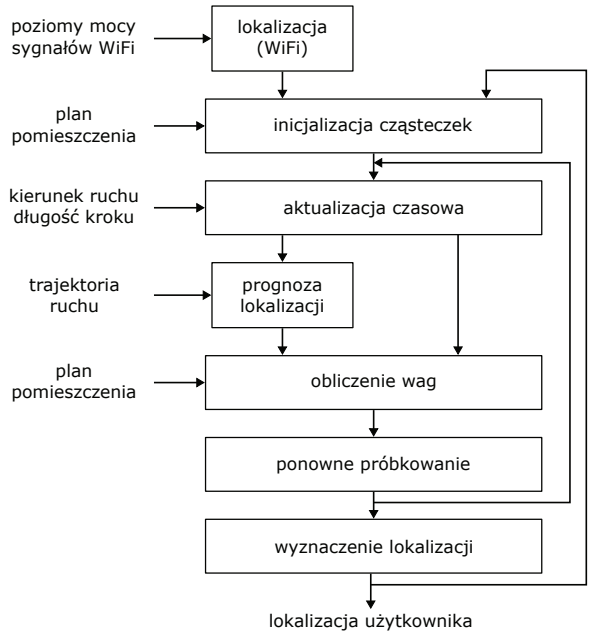
W środowiskach o powtarzalnej strukturze np. biurowych, może się zdarzyć, że podobne trajektorie będą możliwe do zaobserwowania w wielu punktach. Przykładowo przejście z pokoju A do B i B do C skutkuje zbliżonymi trajektoriami (Rys. 1.c). W takim przypadku, do ich rozróżnienia metoda wykorzystuje dane n.t. poziomów zarejestrowanych sygnałów WiFi – poziomy w pokojach B i C są różne. Pomiary mocy sygnałów WiFi wykorzystywane mogą być również do lokalizacji nieporuszających się od dłuższego czasu użytkowników lub ustalenia piętra budynku, na którym przebywa użytkownik.

Dzięki zastosowaniu jedynie sensorów inercyjnych smartfona oraz jego modułu WiFi metoda może być zaimplementowana w praktycznie każdym środowisku wewnątrzbudynkowym. Jedynym wymaganiem jest posiadanie co najmniej jednego punktu dostępowego WiFi.

3. ALGORYTM LOKALIZACYJNY

3.1 Filtr cząsteczkowy

Do wyznaczenia lokalizacji użytkownika wykorzystywany jest algorytm, w którym dane są przetwarzane za pomocą filtru cząsteczkowego i towarzyszących modułów. Schemat blokowy algorytmu został przedstawiony na Rys. 2.



Rysunek 2: Schemat blokowy algorytmu lokalizacyjnego

Działanie algortymu rozpoczyna się od inicjalizacji cząsteczek $\boldsymbol{p_k^{(i)}}$. Każda cząsteczka zawiera informacje na temat lokalizacji i zwrotu użytkownika w chwili k:

$$\boldsymbol{p_k^{(i)}} = \begin{bmatrix} x_k^{(i)} \\ y_k^{(i)} \\ \theta_k^{(i)} \end{bmatrix} \tag{1}$$

gdzie i to numer cząsteczki, $x_k^{(i)}$ i $y_k^{(i)}$ to współrzędne użytkownika a $\theta_k^{(i)}$ to jego zwrot.

W celu dokładniejszej inicjalizacji filtru, użytkownik jest wstępnie lokalizowany na podstawie poziomu sygnałów WiFi zarejestrowanych przez smartfon. Cząsteczki są generowane losowo przy założeniu rozkładu jednostajnego wokół wyznaczonej lokalizacji. Podczas generacji jest nałożone dodatkowe ograniczenie - cząsteczki muszą być zlokalizowane w środowisku gdzie działa system oraz nie mogą znajdować się w obrębie przeszkód. Po inicjalizacji algorytm wchodzi w swoją główną pętlę.

Pierwszym krokiem filtru jest aktualizacja czasowa. Jest ona wykonywana w momencie wykrycia kroku użytkownika i polega na obrocie cząsteczek o kąt $\Delta\theta$ obliczony na podstawie pomiarów inercyjnych i ich przesunięcie o długość kroku L_{step}.

$$\boldsymbol{\hat{p}_{k+1}^{(i)}} = \boldsymbol{p_k^{(i)}} + \begin{bmatrix} L_{step}\cos(\theta + \Delta\theta) \\ L_{step}\sin(\theta + \Delta\theta) \\ \Delta\theta \end{bmatrix} \tag{2}$$

Następnie obliczane są nowe wagi cząsteczek. W zaproponowanej implementacji na wagę cząsteczki mają wpływ trzy czynniki:

$$w_i = \eta \cdot (a \cdot w_{F,i} + b \cdot w_{T,i}) \cdot w_{V,i} \tag{3}$$

gdzie $w_{F,i}$ to czynnik związany z odległość cząsteczki od prognozowanej lokalizacji, $w_{T,i}$ - stopień w jakim dla danej cząsteczki trajektoria pokrywa się z przeszkodami,

$w_{V,i}$ - to czy cząsteczka znajduje się w dozwolonym miejscu. Wartości a i b to parametry pozwalające na uwzględnienie czynników z różną wagą a η jest stałą skalującą zapewniając sumowanie się wag do jedności.

Pierwszy czynnik $w_{F,i}$ związany jest z tym jak prawdopodobne jest wystąpienie danej cząsteczki $\hat{\boldsymbol{p}}_{\boldsymbol{k+1}}^{(\boldsymbol{i})}$. Współrzędne cząsteczki porównywane są z lokalizacją użytkownika $[x_F, y_F]$ prognozowaną na podstawie kawałka uprzednio zarejestrowanej trajektorii ruchu T.

$$w_{F,i} = \frac{1}{||[x_F, y_F] - [\hat{x}_{k+1}^{(i)}, \hat{y}_{k+1}^{(i)}]||} \tag{4}$$

Drugi czynnik określa w jakim stopniu, dla danej cząsteczki (jej lokalizacji i określonemu przez niej zwrotowi użytkownika) uprzednio zarejestrowana trajektoria T pokrywa się z przeszkodami występującymi w środowisku:

$$w_{T,i} = 1 - \frac{N_O}{N_T} \tag{5}$$

gdzie N_O to liczba punktów trajektorii T leżących poza monitorowanym środowiskiem lub w obrębie przeszkód a N_T to łączna liczba analizowanych punktów.

Ostatni czynnik związany jest z możliwością wystąpienia cząsteczki w danym punkcie. Podobnie jak w przypadku inicjalizacji, przyjmuje się że cząsteczki nie mogą znajdować się obrębie przeszkód i poza monitorowanym środowiskiem.

$$w_{V,i} = \begin{cases} 1, & \text{jeśli } \hat{\boldsymbol{p}}_{\boldsymbol{k+1}}^{(\boldsymbol{i})} \text{w dozwolonym miejscu} \\ 0, & \text{w przeciwnym przypadku} \end{cases} \tag{6}$$

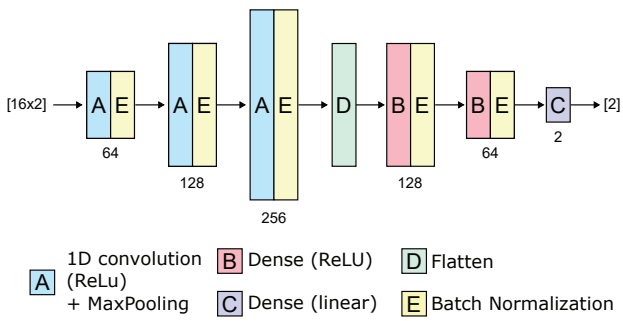
Lokalizacja użytkownika obliczana jest jako średnia ważona cząsteczek:

$$\boldsymbol{p_{k+1}} = \frac{\sum_i w_i \hat{\boldsymbol{p}}_{\boldsymbol{k+1}}^{(\boldsymbol{i})}}{\sum_i w_i} \tag{7}$$

Wynik lokalizacji użytkownika jest okresowo porównywany z wynikiem lokalizacji na podstawie sygnałów sieci WiFi. Jeśli różnice między dwoma lokalizacjami są znaczące, np. są w różnych pomieszczeniach, filtr jest inicjalizowany ponownie.

3.2 Prognoza lokalizacji użytkownika

Zaproponowana koncepcja nie wymusza konkretnej formy modułu predykcji lokalizacji użytkownika – do tego celu można wykorzystać różne metody uczenia maszynowego czy deterministyczne. W ramach poniższych badań zdecydowano się na wykorzystanie splotowej sieci neuronowej, której architekturę przedstawiono na Rys. 3.



Rysunek 3: Architektura modelu prognozy lokalizacji użytkownika

Sieć składa się z dwóch części: ekstraktora cech i modułu regresji. Na wejście modelu podawana jest trajektoria składająca się z 16 punktów (współrzędne x, y). Wyjściem modelu jest kolejna lokalizacja użytkownika. Model zakłada, że okres z jakim wyznaczana jest lokalizacja to 0.3 s – trajektoria obejmuje więc blisko 5 sekund ruchu. Gdy lokalizacja w systemie jest wyznaczana z inną częstotliwością należy dokonać interpolacji lub decymacji.

Ekstraktor cech składa się z trzech bloków, w których wykonywany jest jednowymiarowy splot, maksymalna agregacja (ang. max pooling) oraz normalizację wsadowa (ang. batch normalization). Za regresję odpowiadają trzy warstwy gęste.

W badaniach, model został wstępnie wytrenowany na podstawie symulowanych trajektorii użytkownika. Symulowane trajektorie zostały otrzymane za pomocą algorytmu RRT* (Rapidly Exploring Random Trees*), który pozwala na wyznaczenie ścieżki między dowolnymi punktami pomieszczenia. Zbiór treningowy obejmował trajektorie odpowiadające trzem godzinom ruchu. Średni błąd prognozy następnej lokalizacji użytkownika obliczony dla zbioru testowego to 0.02 m. W przypadku zebrania wystarczającej ilości danych przez użytkownika model może zostać ponownie wytrenowany tak by lepiej dopasować go do lokalizowanej osoby.

4. BADANIA EKSPERYMENTALNE

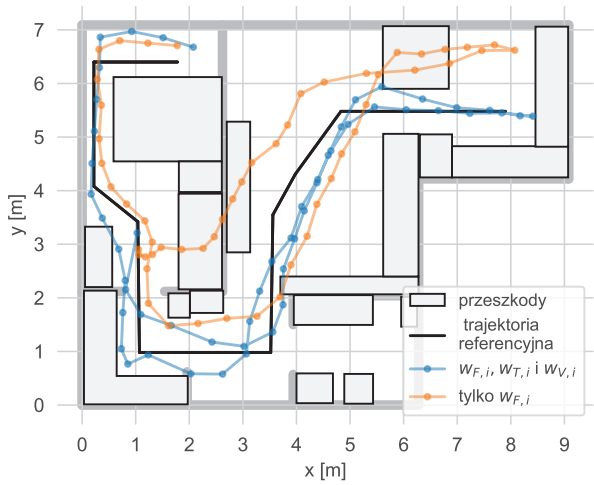
Zaproponowana koncepcja została zweryfikowana eksperymentalnie. Badania zostały przeprowadzone w mieszkaniu o powierzchni 48 m2 i obejmowały lokalizację osoby przemieszczającej się po referencyjnej trajektorii. Plan mieszkania oraz trajektoria referencyjna zostały przedstawione na Rys. 4.

Pomiary zostały wykonane za pomocą smartfona SAMSUNG Galaxy A51. Na smartfonie została zainstalowana prosta aplikacja zapisujaca dane z żyroskopu i sensora obrotu oraz wykonująca pomiary mocy sygnałów WiFi. Podczas eksperymentu telefon znajdował się w przedniej kieszeni spodni.

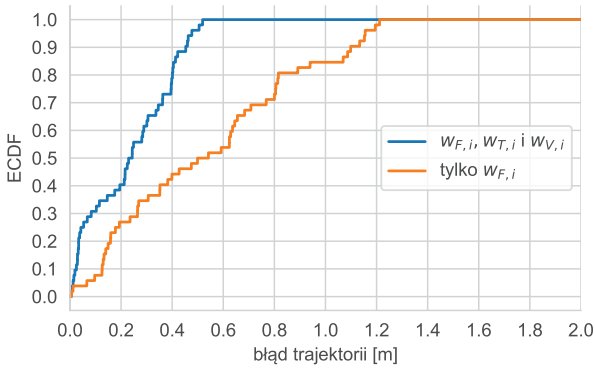
Pomiary inercyjne wykonywane były z częstotliwością około 100 Hz. Zmiany orientacji użytkownika zostały określone bezpośrednio za pomocą sensora obrotu (Game Rotation Vector sensor). Sensor obrotu przetwarza jedynie dane z akcelerometru i żyroskopu. Pominięcie danych z magnetometru skutkuje brakiem informacji na temat zwrotu użytkownika w stosunku do północy ale pozwala na dokładniejsze wyznaczenie względnych zmian kierunku. Dane z żyroskopu zostały również wykorzystane do detekcji kroku użytkownika za pomocą metody opisanej w [5]. Przyjęto stałą długość kroku 0.65 m.

Pomiary poziomów sygnałów WiFi, ze względu na ograniczenie narzucone przez system Android, były wykonywane co 30 sekund. Przed rozpoczęciem eksperymentu została wykonana zgrubna 10-punktowa mapa radiowa pomieszczenia. Wstępna lokalizacja na podstawie mocy sygnałów pozwoliła na efektywniejsze zainicjowanie cząsteczek.

Dane zebrane podczas eksperymentu zostały przetworzone z wykorzystaniem skryptów napisanych w języku Python. Zbadano algorytm w dwóch wersjach - pełnej w której przy wyliczaniu wag brane są pod uwagę wszystkie czynniki ($a = b = 1$) oraz ograniczonej gdzie uwzględniana jest tylko odległość od prognozowanej lokalizacji ($a = 1, b = 0$). Liczba cząsteczek w filtrze wynosiła 1000. Wyniki lokalizacji przedstawiono na Rys. 4.



Rysunek 4: Wyniki lokalizacji



Rysunek 5: Empiryczne dystrybuanty błędu trajektorii

Dokładność lokalizacji określono na podstawie błędu trajektorii zdefiniowanego jako najmniejsza odległość wyznaczonej lokalizacji użytkownika od rzeczywistej trajektorii ruchu. Empiryczne dystrybuanty błędu ECDF (ang. Empirical Cumulative Distribution Function) przedstawiono na Rys. 5.

Zaproponowana metoda pozwala na lokalizację użytkownika z wysoką dokładnością. W przypadku zastosowania wszystkich wag średni błąd trajektorii wyniósł 0,24 m. Gdy uwzględniono jedynie prognozowaną lokalizację średni błąd był większy i wynosił 0,55 m. Wynikało to głównie z faktu, że w tym przypadku, filtr wymagał zebrania większej ilości informacji zanim cząsteczki znalazły się w okolicy rzeczywistej lokalizacji użytkownika.

W scenariuszu pomiarowym użytkownik zaczynał i kończył ruch w tym samym punkcie. Ze względu na fakt, że pomiary mocy sygnałów WiFi są rzadko wykonywane oraz że metoda bazuje na analizie fragmentów trajektorii, pierwsza lokalizacja użytkownika dostępna jest po kilku sekundach ruchu. Stąd na wykresie pominięto fragment trajektorii. Problem ten można potencjalnie rozwiązać przez interpolację brakującej części na podstawie zapisanych wskazań czujników inercyjnych.

5. PODSUMOWANIE

W referacie zaprezentowano koncepcję metody lokalizacji przemieszczającej się osoby za pomocą smartfona. Metoda, do lokalizacji, wykorzystuje fakt, że w środowisku wewnątrzbudynkowym możliwe do zarejestrowania trajektorie są fizycznie ograniczone przez układ pomieszczeń oraz występujące w nich przeszkody.

W zaproponowanej implementacji, lokalizacja użytkownika jest wyznaczana na podstawie pomiarów inercyjnych oraz pomiarów sygnałów sieci WiFi zarejestrowanych za pomocą smartfona. Dane te są przetwarzane za pomocą filtru cząsteczkowego, który aktualizuje swoje wagi poprzez porównanie cząsteczek z prognozowanymi lokalizacjami użytkownika.

Zaproponowana koncepcja ma potencjał do dalszego rozwoju. Potencjalne kierunki to usprawnienie modułu prognozy lokalizacji użytkownika lub dodanie do filtru cząsteczkowego dodatkowych wag. Docelowo planowana jest implementacja koncepcji w postaci jednego modelu uczenia maszynowego przyjmującego jako wejście nieprzetworzone pomiary ze smartfona.

W referacie przedstawiono wyniki prac realizowanych w ramach projektu finansowanego przez Narodowe Centrum Badań i Rozwoju w ramach programu AAL Call 2021 nr umowy AAL/AALCall2021/52/CAREUP/2022.

LITERATURA

[1] Al Mamun Md Abdullah, Rasit Yuce Mehmet, 2021, „Map-Aided Fusion of IMU PDR and RSSI Fingerprinting for Improved Indoor Positioning". 2021 IEEE Sensors: pp. 1-4, doi: 10.1109/SENSORS47087.2021.9639778

[2] Chen Geng i in., 2023, „A Review of Device-Free Indoor Positioning for Home-Based Care of the Aged: Techniques and Technologies". Computer Modeling in Engineering & Sciences, 135 (3): 1901–1940, doi: 10.32604/cmes.2023.024901

[3] Ghaou Mohamed Anis i in., 2023, „Human Motion Likelihood Representation Map-Aided PDR Particle Filter". IEEE Sensors Journal, 23 (1): 484-494, doi: 10.1109/JSEN.2022.3222639

[4] Jackermeier Robert, Ludwig Bernd, 2019, „Door Transition Detection for Long-Term Stability in Pedestrian Indoor Positioning". 2019 International Conference on Indoor Positioning and Indoor Navigation (IPIN): 1–8, doi: 10.1109/IPIN.2019.8911810

[5] Kołakowski Marcin i in., 2024, „Wrist-to-Tibia/Shoe Inertial Measurement Results Translation Using Neural Networks". Sensors 24 (1): 293, doi: 10.3390/s24010293

[6] Ramadhan Hani i in., 2019, „Extracting valid indoor semantic trajectories using movement constraints". 2019 IEEE International Conference on Big Data: 6201-6202, doi: 10.1109/BigData47090.2019.9006105

[7] Stavropoulos Thanos G. i in., 2020, „IoT Wearable Sensors and Devices in Elderly Care: A Literature Review". Sensors, 20 (10) : 2826, doi: 10.3390/s20102826

POMIAROWA WERYFIKACJA EFEKTYWNOŚCI PRACY ASYNCHRONICZNEJ METODY JEDNOCZESNEJ ESTYMACJI POŁOŻENIA I ORIENTACJI OBIEKTU

MEASUREMENT PERFORMANCE VERIFICATION OF ASYNCHRONOUS METHOD FOR SIMULTANEOUS ESTIMATION OF OBJECT POSITION AND ORIENTATION

Jarosław Sadowski[1]; Olga Błaszkiewicz[2]; Krzysztof Cwalina[3];
Alicja Olejniczak[4]; Piotr Rajchowski[5]; Jacek Stefański[6]

Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk
[1]jarsad@eti.pg.edu.pl, [2]olga.blaszkiewicz@pg.edu.pl, [3]kkcwalina@eti.pg.edu.pl,
[4]aliolejn@pg.edu.pl, [5]piorajch@eti.pg.edu.pl, [6]jstef@eti.pg.edu.pl

DOI: 10.15199/59.2024.4.30

Streszczenie: **W referacie opisano wyniki pomiarowej weryfikacji efektywności pracy asynchronicznej metody lokalizowania obiektu na płaszczyźnie za pomocą sygnałów lokalizacyjnych wysyłanych jednocześnie z dwóch nadajników umieszczonych na tym obiekcie w znanej odległości od siebie.**

Abstract: **The paper describes the results of a measurement verification of the effectiveness of an asynchronous method of locating an object on a plane using localization signals sent simultaneously from two transmitters placed on that object at a known distance from each other.**

Słowa kluczowe: **radiolokalizacja, radionawigacja, systemy asynchroniczne.**

Keywords: **asynchronous systems, radiolocalization, radionavigation**

1. WSTĘP

Opisywane w literaturze systemy radiolokalizacyjne przeznaczone do estymacji położenia obiektów w większości opierają się na synchronizacji czasowej pomiędzy elementami systemu, tj. synchroniczna emisja sygnałów ze stacji referencyjnych lub synchroniczna detekcja sygnałów odbieranych z węzłów ruchomych. Niewielka liczba opisywanych systemów dotyczy metod asynchronicznych. W niniejszym referacie przedstawiono pomiarową weryfikację asynchronicznej metody, która polega na jednoczesnej transmisji sygnałów z dwóch nadajników umieszczonych na lokalizowanym obiekcie w znanej odległości od siebie (np. końce skrzydeł samolotu, dziób i rufa statku, itp.) [1]. Sygnały te są odbierane przez kilka odbiorników, umieszczonych na stacjach referencyjnych (stacjach bazowych) pracujących asynchronicznie. Proponowane rozwiązanie umożliwia jednoczesne określenie położenia i orientacji lokalizowanego obiektu.

Artykuł jest zorganizowany w następujący sposób: w rozdziale 2 w skrócie przedstawiono opis matematyczny zaproponowanej metody; w rozdziale 3 zostały opisane zastosowane algorytmy do rozwiązywania układu nieliniowych równań pozycyjnych dla opisanej metody; w kolejnych dwóch rozdziałach zaprezentowano środowisko pomiarowe oraz wyniki badań pomiarowych efektywności pracy asynchronicznej metody radiolokalizacyjnej; ostatnia sekcja stanowi podsumowanie zrealizowanych prac.

2. OPIS MATEMATYCZNY METODY

Dla przypadku dwuwymiarowego, dwa zsynchronizowane nadajniki są umieszczone na lokalizowanym obiekcie. Nadajniki znajdują się w znanej odległości d od siebie i jednocześnie (synchronicznie) emitują sygnały lokalizacyjne. Sygnały te są odbierane przez cztery stacje referencyjne (SR) – analogicznie do zastosowanej konfiguracji układu pomiarowego. Stosując unikatowe identyfikatory w każdym nadajniku, poszczególne stacje referencyjne mogą niezależnie odbierać te sygnały i określać różnicę odległości pomiędzy nadajnikami i antenami odbiorczymi [2]. Układ równań nieliniowych, który opisuje rozważaną metodę jest następujący [1]:

$$\begin{cases} d_{n2} - d_{n1} = \Delta t_n \cdot v = \Delta d_n \quad n = 1,...,4 \\ d^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2 \end{cases}, \tag{1}$$

gdzie d_{n1} i d_{n2} reprezentują odpowiednio odległości pierwszego nadajnika na obiekcie o współrzędnych (x_1,y_1) oraz drugiego nadajnika na obiekcie o współrzędnych (x_2,y_2) od n-tej stacji referencyjnej; Δt_n stanowią różnicę czasów odbioru w poszczególnych stacjach referencyjnych sygnałów emitowanych z obu nadajników, v jest prędkością propagacji sygnałów radiowych w ośrodku propagacji ($v \approx 3 \cdot 10^8$ m/s). Ponadto, d_{n1} i d_{n2} można opisać następującymi zależnościami:

$$\begin{cases} d_{n1} = \sqrt{(X_n - x_1)^2 + (Y_n - y_1)^2} \\ d_{n2} = \sqrt{(X_n - x_2)^2 + (Y_n - y_2)^2} \end{cases}, \tag{2}$$

gdzie (X_n,Y_n) są współrzędnymi SR. W kolejnym punkcie zostaną przedstawione przykładowe dwa algorytmy, które umożliwiają estymowanie współrzędnych obiektu na podstawie układu równań (1).

3. ALGORYTMY ESTYMACJI POŁOŻENIA OBIEKTU

Do rozwiązania układu równań nieliniowych (1), w którym niewiadomymi są współrzędne obu anten nadawczych (x_1,y_1) oraz (x_2,y_2), zastosowano uogólnioną wersję algorytmu Newtona (NA) [3] oraz algorytm genetyczny (GA) [4].

Zakładając przypadek dwuwymiarowy z czterema SR, problem estymacji położenia obiektu sprowadza się do rozwiązania układu równań nieliniowych postaci:

$$\mathbf{f}(x_1,y_1,x_2,y_2)=\begin{cases} d_{12}-d_{11}=\Delta d_1 \\ d_{22}-d_{21}=\Delta d_2 \\ d_{32}-d_{31}=\Delta d_3 \\ d_{42}-d_{41}=\Delta d_4 \\ d^2=(x_2-x_1)^2+(y_2-y_1)^2 \end{cases} . \quad (3)$$

Zlinearyzowaną wersję powyższego układu równań można zapisać

$$\mathbf{J}(x_1,y_1,x_2,y_2)\cdot\mathbf{h}=-\mathbf{f}(x_1,y_1,x_2,y_2), \quad (4)$$

gdzie $\mathbf{J}(x_1,y_1,x_2,y_2)$ jest macierzą Jakobiego, a wektor $\mathbf{h}$ reprezentuje poprawkę współrzędnych nadajników w kolejnych iteracjach. Algorytm NA jest wrażliwy na postać wektora startowego, który został przyjęty w pierwszej iteracji. Wiadomo, że im bliższy poprawnemu rozwiązaniu jest wektor rozwiązań początkowych, tym szybciej algorytm znajduje to rozwiązanie. W naszych rozważaniach wektor początkowy był zawsze taki sam, co oznacza stałą pozycję w punkcie początkowym i stałą orientację. W procesie estymacji położenia nadajników przyjęto kryterium zbieżności w postaci normy opisanej zależnością $\|(\mathbf{J}^T\cdot\mathbf{J})^{-1}\cdot\mathbf{J}^T\cdot\mathbf{f}\|$. Maksymalna liczba iteracji wynosiła 1 000. Jeżeli w trakcie działania algorytmu po 1 000 iteracji wartość normy nie spadła poniżej przyjętego progu, obliczenia przerywano i wybierano losową postać wektora startowego i cały proces znajdowania rozwiązania był powtarzany.

Dla celów porównawczych, do rozwiązania układu równań nieliniowych (1) wykorzystano również algorytm genetyczny, który pomimo stosunkowo dużej złożoności obliczeniowej zapewnia zbieżność do właściwego rozwiązania w niemal 100% rozpatrywanych przypadków. Jego praktyczna implementacja została zaproponowana w [5], przy czym $\mathbf{x}_0=[x_1^0,y_1^0,x_2^0,y_2^0]$ reprezentuje wektor początkowy współrzędnych nadajników na obiekcie. Pseudokod zaimplementowanego algorytmu jest następujący:

Step 0. Given W numerous $\mathbf{x}_0\in\Re^n$ (initial population of chromosomes);
Set the threshold calculations *thold*;
Set the number of iteration *n_iter*;
Compute objective function $\mathbf{f}(\mathbf{x}_0)$ and $\|\mathbf{f}(\mathbf{x}_0)\|$;
Sort $\mathbf{x}_0$ by values $\|\mathbf{f}(\mathbf{x}_0)\|$;
If the lowest $\|\mathbf{f}(\mathbf{x}_0)\|\le thold$, then STOP, otherwise, set i:=1.

Step 1. Creation of a new population $\mathbf{x}_i$ for ½ W less matched population using ½ W better matched population (process of selection and mutation of chromosomes).

Step 2. Compute $\mathbf{f}(\mathbf{x}_i)$ and $\|\mathbf{f}(\mathbf{x}_i)\|$ for W numerous population of chromosomes;
Sort $\mathbf{x}_0$ by values $\|\mathbf{f}(\mathbf{x}_i)\|$;
If the lowest $\|\mathbf{f}(\mathbf{x}_i)\|\le thold$, then STOP.

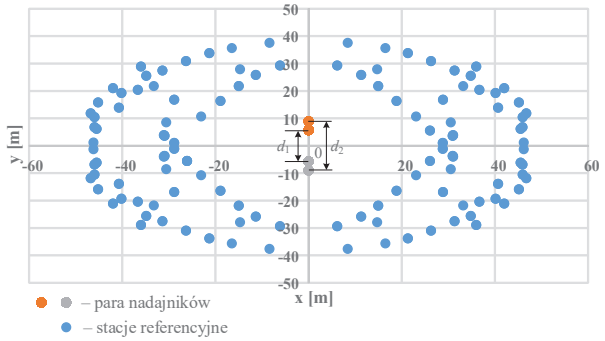
Step 3. i:=i+1;
If $i \ge n_iter$, then STOP, otherwise go to ***Step 1.***

Podczas badań symulacyjnych przyjęto: W = 1000; *thold* = 0,0001 oraz *n_iter* = 1 000. Wektory wartości początkowych $\mathbf{x}_0$ (współrzędne nadajników) były wyznaczane losowo z obszaru o promieniu 50m wokół przyjętego środka układu współrzędnych.

Powyższe algorytmy estymacji położenia obiektu został zaimplementowany w uniwersalnym środowisku obliczeń matematycznych MATLAB.

4. ŚRODOWISKO POMIAROWE

Efektywność pracy zaproponowanej metody została zweryfikowana eksperymentalnie. Pomiary dokładności estymacji położenia obiektu zostały przeprowadzone w okolicach byłego pasa startowego (teren płaski, pozbawiony przeszkód) na gdańskiej Zaspie w odległości ok. 150 m od najbliższej zabudowy. Nadajniki emitujące w sposób synchroniczny sygnały z widmem rozproszonym DSSS (*Direct Sequence Spread Spectrum*) i modulacją BPSK (*Binary Phase Shift Keying*) były nieruchome, a odległość między nimi wynosiła d_1 = 11,39 m lub d_2 = 17,91 m (rys. 1).

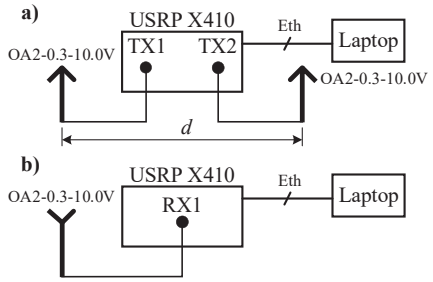


Rys. 1. Struktura testowanego systemu radiolokalizacyjnego

Wokół pary nadajników rozmieszczane były stacje referencyjne (odbiorniki), które mierzyły różnicę nadejścia sygnałów pochodzących od tych nadajników. W celu rozróżnienia nadawanych sygnałów do rozpraszania widma sygnałów wykorzystano różne ciągi Kasamiego o długości 2^{12}-1. Pomiary były wykonane na dwóch częstotliwościach nośnych 2,375 GHz oraz 5,750 GHz odpowiednio z szerokością kanału testowego 150 MHz i 200 MHz. Odbiór sygnałów lokalizacyjnych był realizowany z częstotliwością próbkowania 250 MHz. Zakresy częstotliwości pracy nadajników pokrywają się z pasmami amatorskimi i zostały dobrane zgodnie z Krajową Tablicą Przeznaczeń Częstotliwości [6].

Natomiast do wyznaczania różnic czasów propagacji sygnałów pochodzących od pary nadajników zastosowano metodę korelacyjną przebiegów nadpróbkowanych

120 razy. Układ pomiarowy został wykonany na bazie urządzeń USRP X410 firmy NI Ettus Redearc oraz laptopów z dedykowanym oprogramowaniem (rys. 2). W części nadawczej i odbiorczej zastosowano ten sam rodzaj anten dookólnych OA2-0.3-10.0V firmy Cobham.



Rys. 2. Schemat blokowy a) stacji nadawczej, b) stacji referencyjnej (odbiorczej).

Synchronizacja sygnałów nadawanych była zapewniona na poziomie sprzętowym w urządzeniach USRP. Z kolei, pomiary odległości pomiędzy stacjami referencyjnymi a poszczególnymi nadajnikami zostały wykonane przy użyciu taśmy mierniczej z dokładnością ±1 cm.

5. WYNIKI BADAŃ POMIAROWYCH

Podczas kampanii pomiarowej zarejestrowano 212 zestawów próbek sygnałów nadawanych na częstotliwości 2,375 GHz oraz 148 zestawów na częstotliwości 5,750 GHz. Pojedynczy przypadek lokalizacyjny (cztery stacje referencyjne odbierające sygnały emitowane z dwóch nadajników) wymagał każdorazowo analizy 4 zestawów próbek.

Na podstawie wyników uzyskanych z badań pomiarowych wykreślono dystrybuantę CDF (*Cumulative Distribution Functions*) błędu bezwzględnego δ, opisanego zależnością

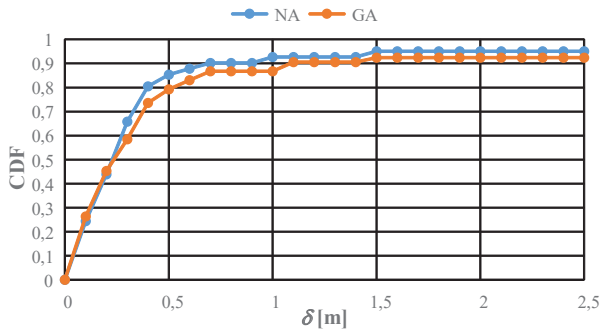
$$\delta = \sqrt{(\hat{x}-x)^2 + (\hat{y}-y)^2}\,, \tag{5}$$

gdzie $(\hat{x}, \hat{y})$ reprezentują estymaty współrzędnych obiektu, natomiast (x,y) to rzeczywiste wartości współrzędnych obiektu, przy czym założono, że

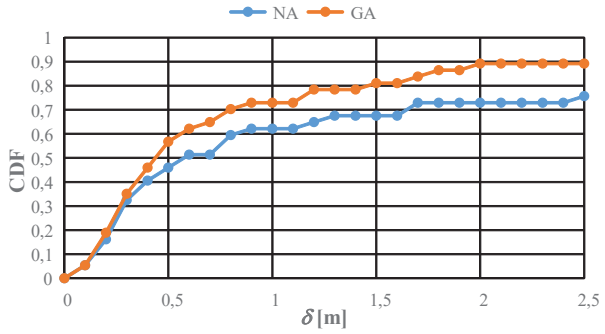
$$\hat{x} = \frac{\hat{x}_1 + \hat{x}_2}{2},\quad \hat{y} = \frac{\hat{y}_1 + \hat{y}_2}{2}, \tag{6}$$

$$x = \frac{x_1 + x_2}{2},\quad y = \frac{y_1 + y_2}{2}. \tag{7}$$

Wyniki pomiarów CDF = $f(\delta)$ przedstawiono na rys. 3 i rys. 4. Dla przyjętej topologii sieci radiolokalizacyjnej otrzymane wyniki są obiecujące i potwierdzają przyjęte założenia systemowe. Dla częstotliwości nośnej 2,375 GHz w 90% przypadków błąd estymacji położenia obiektu nie przekracza 1 m bez względu na zastosowany algorytm do rozwiązywania nieliniowego układu równań pozycyjnych. Na uwagę zasługuje komentarz do wyników otrzymanych na częstotliwości nośnej 5,750 GHz. Pomimo zwiększenia pasma nadawanych sygnałów lokalizacyjnych, a co za tym idzie, precyzji wykonywanych pomiarów, w 90% przypadków błąd estymacji położenia obiektu dla algorytmu GA jest mniejszy jedynie od 2,5 m. Pogorszenie wyników należy upatrywać w konstrukcji zestawów pomiarowych i braku informacji o sposobie rozwiązania synchronizacji w urządzeniach USRP lub też w zwiększonym na tej częstotliwości wpływem odbić sygnałów od nierównej powierzchni ziemi na rezultaty pomiarów. Przykładowo, niewielkie różnice w długościach kabli połączeniowych wewnątrz i na zewnątrz urządzeń przy tej częstotliwości nośnej mogą prowadzić do błędów, które przełożyły się na otrzymane wyniki. Za pomocą algorytmu NA zastosowanego do rozwiązywania układu równań pozycyjnych można uzyskać lepsze rezultaty dla stosunkowo małych wartości błędów pomiarów, z kolei przy większych wartościach błędów widać zdecydowaną przewagę algorytmu GA.



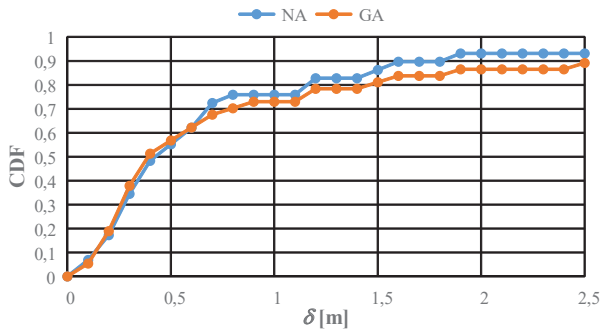
Rys. 3. Dystrybuanty (CDF) błędu bezwzględnego położenia obiektu δ z zastosowaniem algorytmu Newtona (NA) oraz algorytmu genetycznego (GA) dla częstotliwości nośnej 2,375 GHz.



Rys. 4. Dystrybuanty (CDF) błędu bezwzględnego położenia obiektu δ z zastosowaniem algorytmu Newtona (NA) oraz algorytmu genetycznego (GA) dla częstotliwości nośnej 5,750 GHz.

W dalszej kolejności podjęto próbę korekcji danych uzyskanych na częstotliwości nośnej 5,750 GHz w celu zwiększenia dokładności estymacji położenia obiektu. Zakładając, że w układzie pomiarowym występuje błąd systematyczny spowodowany brakiem symetryczności połączeń w poszczególnych torach nadawczych, każdy pomiar różnicy odległości pomiędzy parą nadajników a SR korygowano o stałą wartość odległości z krokiem 1 cm. Kryterium jakościowym był minimalny, wypadkowy błąd RMSE (*Root Mean Square Error*) położenia obiektu dla wszystkich zestawów danych na tej częstotliwości stosując algorytm NA. W efekcie tych analiz stwierdzono, że minimum błędu RMSE przypada dla róż-

nic odległości skorygowanych o 11 cm. Po korekcji danych o wskazaną wartość stałą wykreślono ponownie rodzinę krzywych CDF = $f(\delta)$, którą przedstawiono na rys. 5.



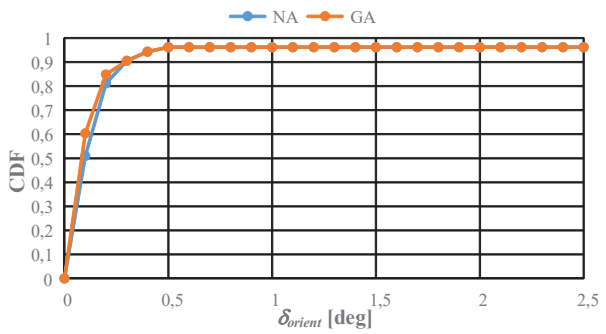
Rys. 5. Dystrybuanty (CDF) błędu bezwzględnego położenia obiektu δ z zastosowaniem algorytmu Newtona (NA) oraz algorytmu genetycznego (GA) dla częstotliwości nośnej 5,750 GHz po korekcie wyników pomiarowych.

Korekcja danych pomiarowych przyniosła zdecydowaną poprawę wyników, które teraz są zbieżne z wynikami uzyskanymi na częstotliwości nośnej 2,375 GHz. Warto w tym miejscu wspomnieć, że podobna próba korekcji danych pomiarowych na częstotliwości 2,375 GHz nie przyniosła istotnych zmian wyników w odniesieniu do estymacji położenia, natomiast nieznacznie poprawiła wyniki estymacji orientacji obiektu, o której będzie mowa w dalszej części artykułu.

Jak wiadomo, badana metoda umożliwia jednoczesne oszacowanie położenia obiektu i jego orientacji. Korzystając z danych pomiarowych wykreślono również rodzinę krzywych CDF błędu orientacji δ_{orient} określonego zależnością

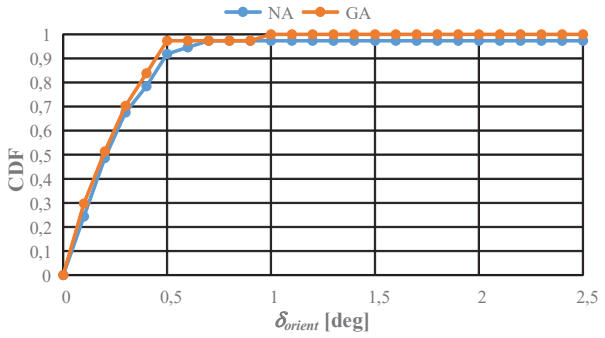
$$\delta_{orient} = \left| \arctan\left(\frac{y_2 - y_1}{x_2 - x_1}\right) - \arctan\left(\frac{\hat{y}_2 - \hat{y}_1}{\hat{x}_2 - \hat{x}_1}\right) \right|. \qquad (8)$$

Wyniki analizy CDF = $f(\delta_{orient})$ zostały przedstawione na rys. 6 i rys. 7. Na każdej częstotliwości nośnej, na której zostały przeprowadzone pomiary oraz bez względu na zastosowany algorytm rozwiązywania układu równań nieliniowych, w 90% przypadków błąd estymacji orientacji obiektu jest mniejszy od 0,5 stopnia.



Rys. 6. Dystrybuanty (CDF) błędu orientacji δ_{orient} z zastosowaniem algorytmu Newtona (NA) oraz algorytmu genetycznego (GA) dla częstotliwości nośnej 2,375 GHz.

W krótkim podsumowaniu wyników badań pomiarowych można stwierdzić, że zweryfikowana praktycznie metoda lokalizacji obiektów na podstawie dwóch synchronicznie emitowanych sygnałów ma duży potencjał aplikacyjny. Umożliwia jednoczesną estymację położenia i orientacji obiektów, w tym co istotne, będących w spoczynku.



Rys. 7. Dystrybuanty (CDF) błędu orientacji δ_{orient} z zastosowaniem algorytmu Newtona (NA) oraz algorytmu genetycznego (GA) dla częstotliwości nośnej 5,750 GHz.

6. PODSUMOWANIE

W referacie przedstawiono wyniki badań eksperymentalnych efektywności pracy asynchronicznej metody radiolokalizacyjnej, w której sygnały są nadawane w sposób synchroniczny z dwóch nadajników umieszczonych na lokalizowanym obiekcie w znanej odległości od siebie. Badania pomiarowe zostały wykonane na dwóch częstotliwościach nośnych na obszarze o średnicy ok. 50 m. Uzyskane dokładności estymacji położenia poniżej 1 m należy uznać za obiecujące. Na podstawie tych samych danych pomiarowych, obok szacowania położenia, możliwa jest estymacja orientacji tego obiektu. Otrzymane wyniki pomiarów orientacji również są obiecujące. Uzyskane dokładności poniżej 1 stopnia rozszerzają znacząco obszary potencjalnych zastosowań tej metody.

LITERATURA

[1] Stefański Jacek, Jarosław Sadowski. 2022. "Asynchroniczna metoda jednoczesnej estymacji położenia i orientacji obiektu za pomocą dwóch nadajników". Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne, 4/2022: 567-570

[2] Sadowski Jarosław, Stefański Jacek. 2021. „Sposób jednoczesnego wyznaczania położenia i orientacji obiektu, zwłaszcza będącego w ruchu". *Zgłoszenie patentowe w UPRP*, nr P.43919

[3] Foy Wade H. 1976. "Position-Location Solutions by Taylor-Series Estimation". *IEEE Transactions on Aerospace and Electronic Systems AES-12* (2): 187-194

[4] Karr Charles L., Weck Barry, Freeman L. Michael. 1998. „Solutions to Systems of Nonlinear Equations via a Genetic Algorithm". Engineering Applications of Artificial Intelligence 11 (3): 369-375

[5] Stefański Jacek, Sadowski Jarosław. 2016. „Analiza porównawcza metody TDOA i ATDOA w systemie WAM". Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne, 6/2016: 297-300

[6] Prezes Rady Ministrów. 2022. „Obwieszczenie w sprawie ogłoszenia jednolitego tekstu rozporządzenia Rady Ministrów w sprawie Krajowej Tablicy Przeznaczeń Częstotliwości". Dz. U. poz. 1988

SYSTEM DO LOKALIZOWANIA OSÓB W POMIESZCZENIACH Z WYKORZYSTANIEM CZUJNIKÓW PODCZERWIENI

INDOOR LOCALIZATION SYSTEM UTILIZING INFRARED SENSORS

Vitomir Djaja-Jośko[1]; Jan Tuchowski[2]

[1] Politechnika Warszawska, IRTM, Warszawa, vitomir.djaja-josko@pw.edu.pl
[2] Politechnika Warszawska, Warszawa, jan.m.tuchowski@gmail.com

DOI: 10.15199/59.2024.4.31

Streszczenie: **Rosnące zapotrzebowanie na usługi wykorzystujące informacje o lokalizacji osób wewnątrz pomieszczeń implikuje dynamiczny rozwój systemów udostępniających takie dane. Jednym z zastosowań jest wspomaganie osób starszych, dla których najwygodniejszym mógłby okazać się system nie wymagający noszenia dodatkowych urządzeń. Takie podejście jest możliwe dzięki wykorzystaniu czujników podczerwieni. W referacie przedstawiono koncepcję systemu lokalizacyjnego wykorzystującego dwa sensory, będące uproszczonymi kamerami termowizyjnymi, pozwalające na wyznaczanie lokalizacji w jednym pomieszczeniu.**

Abstract: **The growing demand for indoor location-based services (LBS) implies the dynamic development of the systems providing such data. Exemplary application of LBS is assisting elderly individuals, for whom a system that does not require additional devices to be worn could be most convenient. This approach is possible through the use of the infrared sensors. The paper introduces the concept of the localization system utilizing two sensors, which are simplified thermal cameras, allowing for location determination within a single room.**

Słowa kluczowe: **lokalizacja wewnątrz pomieszczeń, czujniki podczerwieni, przetwarzanie obrazu, termografia.**

Keywords: **image processing, indoor localization, infrared sensors, thermography.**

1. WSTĘP

Rosnące zapotrzebowanie na usługi wykorzystujące informacje o lokalizacji osób wewnątrz pomieszczeń implikuje dynamiczny rozwój systemów udostępniających takie dane. Systemy lokalizacyjne mają wiele zastosowań – przemysłowych, takich jak usprawnianie procesów w fabrykach czy magazynach, biznesowych, takich jak optymalizacja zysków w sklepach wielkopowierzchniowych czy lokalizowanie osób na lotniskach bądź medycznych, jak lokalizowanie personelu w szpitalu czy wspomaganie osób starszych. Ze względu na różnorodność zastosowań, wymagania stawiane różnym rozwiązaniom mogą się od siebie znacząco różnić. W części systemów kluczowym czynnikiem może być dokładność wyznaczania lokalizacji, w innych z kolei częstość jej wyznaczania bądź minimalizacja infrastruktury oraz liczby urządzeń z jakich złożony jest system.

Biorąc pod uwagę, że według aktualnych danych WHO żyjemy w szybko starzejącym się społeczeństwie, rozwój systemów wspomagających osoby starsze w ich życiu codziennym jest wyjątkowo istotny. Do ich zadań może należeć wykrywanie upadków, analizowanie schematów poruszania się i obecności w różnych obszarach mieszkania w celu wykrywania zmian w zachowaniu świadczących o pojawianiu się potencjalnych stanów chorobowych czy asystowanie w codziennych czynnościach.

Do analizy schematów poruszania się po domu niezbędna jest implementacja systemu lokalizacyjnego. Istnieje wiele sposobów klasyfikacji takich systemów. Jednym z nich jest podział ze względu na konieczność noszenia dodatkowego urządzenia przez lokalizowaną osobę. Do najpopularniejszych należą rozwiązania radiowe, w których do wyznaczania lokalizacji osób wykorzystywane są pomiary poziomu odbieranego sygnału (BLE, Wi-Fi [2]), kąta przybycia sygnału (BLE, UWB [8]), bądź czasu przybycia sygnału (UWB [9]). Systemy te, zależnie od implementacji, zapewniają dokładność wyznaczania położenia w zakresie od kilkunastu centymetrów do pojedynczych metrów. W kontekście lokalizowania osób starszych jako wadę można wskazać konieczność ciągłego noszenia tzw. etykiety nadającej sygnały, bądź wykonującej pomiary.

Druga grupa systemów, niewymagająca dodatkowych urządzeń noszonych przez lokalizowaną osobę obejmuje przede wszystkim systemy radarowe oraz wizyjne [4]. Klasyczne systemy wizyjne wykorzystujące kamery mogą rodzić obawy w kwestii prywatności użytkowników. Pozbawione tego problemu są kamery termowizyjne, które z kolei są drogie. W związku z tym, rozwiązaniem mogą być matrycowe czujniki podczerwieni, będące kompromisem w kwestii kosztów, prywatności i dokładności. W niniejszej pracy przedstawiono koncepcję systemu wykorzystującego dwa takie czujniki do wyznaczania położenia osoby w pomieszczeniu.

2. ANALIZA LITERATURY

2.1. Termografia

Czujniki podczerwieni wykorzystują termografię, będącą techniką obrazowania i estymacji temperatury za pomocą promieniowania podczerwonego. Do celów obrazowania najczęściej wykorzystywana jest podczerwień średnia o długości fali w zakresie 2,5 – 25 μm. Działanie czujników podczerwieni opiera się na prawie Stefana-Boltzmana, które w uproszczeniu mówi, że ciało doskonale czarne o temperaturze wyższej od zera absolutnego emituje promieniowanie. Podstawowym zadaniem czujników podczerwieni jest konwersja tego promieniowania na sygnał elektryczny.

2.2. Stan wiedzy

Koncepcja wyznaczania lokalizacji z wykorzystaniem czujników podczerwieni jest znana, ale często wiąże się z dużymi kosztami bądź niewystarczającą dokładnością. Z.-J. Feng, X.-L. Zhang, L.-Y. Yuan i J.-N. Wang w [6] zaproponowali metodę lokalizacji osób łączącą obrazy wykonane w widmie widzialnym, jak i w podczerwieni. Zaawansowane przetwarzanie danych umożliwia detekcję zarówno poruszających się osób, jak i obiektów statycznych.

Z. Tang i X. Wang w [1] udowodnili, że z wykorzystaniem podczerwieni można lokalizować nawet bardzo małe obiekty. Zaproponowali algorytm, który niezależnie od stopnia skomplikowania tła, umożliwia wyodrębnienie z niego pożądanego obiektu. Algorytm ten działa dwustopniowo – najpierw lokalizowany jest obszar, na którym występuje obiekt, a następnie opierając się na metodzie fraktalnej, obiekt lokalizowany jest z pożądaną dokładnością.

Z kolei N. Faulkner, F. Alam, M. Legg i S. Demedienko w [5] przedstawili sposób wyznaczania lokalizacji obiektów wykorzystujący kamery podczerwieni o niskiej rozdzielczości oraz algorytmy uczenia maszynowego. Zaprezentowane wyniki pokazują, że po odpowiednim wyuczeniu sieci neuronowej z wykorzystaniem jednej osoby w danym pomieszczeniu, możliwe jest, bez znaczącej degradacji dokładności lokalizacji, wyznaczanie położenia innych osób w innych pomieszczeniach. Co więcej, zaobserwowano, że zwiększenie liczby kamer ma pozytywny wpływ na dokładność lokalizacji, ale negatywnie wpływa na rozmiar obszaru, w którym działa system, ze względu na konieczność zapewnienia bezpośredniej widoczności między lokalizowaną osobą a każdym z czujników.

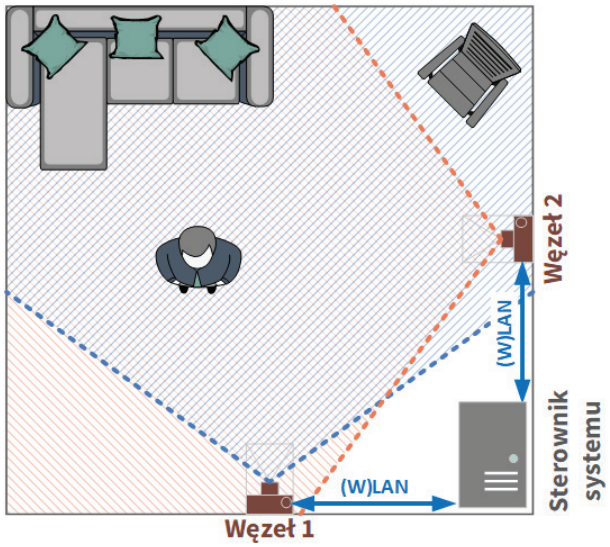
Inne podejście, wykorzystujące macierze czujników termoparowych bądź sieć piroelektrycznych czujników podczerwieni zaproponowali K. Ha, K. Lee i S. Lee w [3].

Większość rozwiązań wymaga dość skomplikowanych, często kosztownych urządzeń, co pokazuje, że istnieje pewna luka w kwestii lokalizacji z wykorzystaniem podczerwieni.

3. KONCEPCJA ROZWIĄZANIA

3.1. Architektura

Proponowane rozwiązanie polega na wykorzystaniu matrycowych czujników podczerwieni, które umożliwiają dostarczenie prostego obrazu termalnego. Fakt, że temperatura ludzkiego ciała jest wyższa od temperatury otoczenia pozwala na wyodrębnienie sylwetki a następnie, na podstawie otrzymanego obrazu, estymację kąta, pod jakim osoba znajduje się względem pionowej płaszczyzny prostopadłej do płaszczyzny czujnika. Znając położenie czujnika oraz kąt, można wyznaczyć równanie prostej, na której znajduje się osoba. Dysponując dwoma czujnikami możliwe jest określenie położenia osoby w punkcie przecięcia dwóch prostych. Na rysunku 1 przedstawiono architekturę proponowanego systemu. Zaznaczono na nim również obszary w postaci trójkątów (w trzech wymiarach byłyby to ostrosłupy czworokątne) reprezentujących pola widzenia poszczególnych czujników. Warto wspomnieć, że obszar poprawnego działania systemu, to obszar przecięcia się obu trójkątów.



Rys. 1. Architektura systemu

Do realizacji zadania przewidziano wykorzystanie dwóch węzłów wyposażonych w czujniki podczerwieni oraz sterownika systemu. Węzły powinny być umieszczone w dogodnej pozycji, najlepiej na dwóch prostopadłych ścianach, tak aby zapewnić jak największe pokrycie pomieszczenia. Zadaniem węzłów jest rejestracja kolejnych obrazów termalnych z częstotliwością wynoszącą 2 Hz, co biorąc pod uwagę potencjalne zastosowanie systemu jest wystarczającą wartością, przetwarzanie ich w celu wyznaczenia kątów pod jakimi osoba znajduje się względem czujnika oraz przekazywanie ich za pośrednictwem sieci lokalnej do sterownika systemu, który je agreguje i na ich podstawie wyznacza położenie osoby. Na aktualnym etapie prac przyjęto założenie, że lokalizacja będzie się odbywać w dwóch wymiarach i że jednocześnie lokalizowana może być tylko jedna osoba.

3.2. Wykorzystane moduły i urządzenia

Rolę sterownika węzła systemu pełniły minikomputery RaspberryPi 4B. Jako czujnik termalny wybrano moduł MLX90640 firmy Melexis [7], wyposażony w matrycę czujników podczerwieni, zdolny do generowania obrazu termalnego o rozdzielczości 32 na 24 piksele i polu widzenia 75° w pionie i 110° w poziomie. Komunikacja ze sterownikiem węzła odbywała się za pomocą interfejsu I2C. Rolę sterownika systemu pełnił komputer PC.

4. OPROGRAMOWANIE

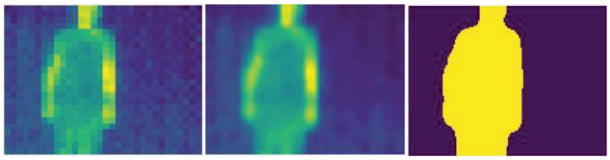
Zarówno oprogramowanie węzła, jak i sterownika systemu zostało zrealizowane w języku Python.

Proces komunikacji sterownika węzła z czujnikiem, pobieranie oraz wstępne przetwarzania obrazu zostały zrealizowane z wykorzystaniem gotowych bibliotek dostarczonych przed producenta modułu.

4.1. Algorytm detekcji kąta

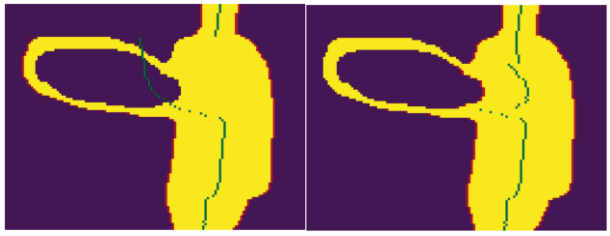
Algorytm detekcji kąta składa się z kilku następujących po sobie etapów. Pierwszym krokiem są interpolacja i filtracja obrazu. Po wstępnych badaniach i empirycznej analizie otrzymanych wyników zdecydowano się na zastosowanie czterokrotnej interpolacji oraz filtru medianowego o rozmiarze okna wynoszącym 5, jako że dawały one najlepsze wyniki.

Następnie, na podstawie danych dostarczanych przez czujnik, ustawiany jest próg detekcji (do wykrywania osoby wykorzystywany jest fakt, że jej temperatura jest wyższa niż temperatura otoczenia) a następnie tworzony jest obraz binarny. Na rysunku 2 przedstawiono obraz na kolejnych etapach przetwarzania.



Rys. 2. Obraz surowy, obraz po interpolacji i filtracji, obraz binarny z wykrytą sylwetką człowieka

Ostatnim etapem jest wykrywanie kąta pod jakim osoba znajduje się względem czujnika. Odbywa się to w dwóch krokach. Najpierw każdy wiersz w obrazie binarnym jest analizowany od lewej i od prawej strony, tak żeby znaleźć krawędzie sylwetki (oznaczone kolorem czerwonym na rysunku 3). W drugim kroku, operacja jest powtarzana, z tą różnicą, że analizowanie obrazu odbywa się od środka, zamiast od zewnętrznych krawędzi. Ma to zniwelować negatywny wpływ sytuacji, w której na obrazie występują pętle, które mogą zniekształcać sylwetkę, co pokazano na rysunku 3. Po określeniu skrajnych współrzędnych sylwetki, dla każdego wiersza wyznaczany jest jej środek (oznaczony kolorem zielonym). Następnie uzyskane wartości są uśredniane i uzyskiwana jest współrzędna środka sylwetki.



Rys. 3. Wykrywanie środka sylwetki; krok 1 i krok 2

W proponowanym rozwiązaniu wartość kąta można wyznaczyć ze wzoru 1:

$$\alpha = -55° + \frac{n}{32*4} * 110°, \quad (1)$$

gdzie n to indeks kolumny odpowiadającej współrzędnej środka sylwetki. Rozdzielczość wyznaczania kąta opisano wzorem 2:

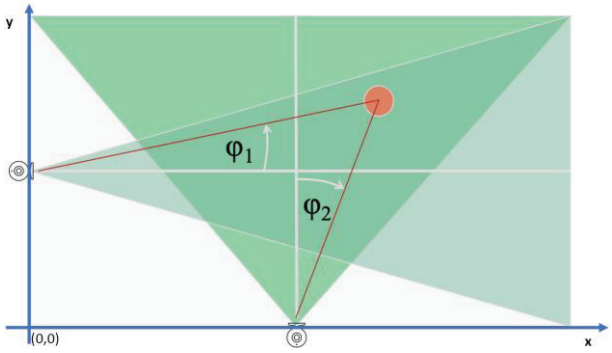
$$\Delta\alpha = \frac{1}{32*4} * 110° \cong 0{,}86°. \quad (2)$$

4.2. Algorytm lokalizacyjny

W aktualnej wersji systemu przyjęto analityczne podejście do określania położenia osoby. Zakładając, że w pomieszczeniu przyjęto układ współrzędnych, w którym wyrażono położenia węzłów, znając kąty (φ_1 i φ_2) pod jakimi osoba znajduje się względem nich, jej lokalizację (x,y) można wyznaczyć jako punkt przecięcia dwóch prostych (rysunek 4), rozwiązując układ równań 3:

$$\begin{cases} y = a_1 x + b_1 \\ y = a_2 x + b_2 \end{cases}, \quad (3)$$

gdzie x i y to współrzędne przecięcia się prostych, natomiast a_1 i a_2 to ich współczynniki kierunkowe, a b_1 i b_2 to ich wyrazy wolne. Wartości a_1, a_2, b_1, b_2 można wyznaczyć na podstawie współrzędnych węzłów i wartości kątów.



Rys. 4. Wyznaczanie lokalizacji osoby

5. BADANIA

Badania laboratoryjne opracowanego systemu, mające na celu zarówno weryfikację zasadności koncepcji wyznaczania lokalizacji za pomocą matrycowych czujników podczerwieni, jak i określenie błędu wyznaczania położenia przeprowadzono w dwóch etapach. Najpierw zweryfikowano dokładność wyznaczania kąta, następnie przeprowadzono testy dokładności wyznaczana lokalizacji.

5.1. Badanie błędu wyznaczania kąta

Badanie przeprowadzono z wykorzystaniem jednego węzła. Ustawiono go tak, aby polem widzenia obejmował prawie całe pomieszczenie a następnie przeprowadzono w 6 punktach pomiar kąta pod jakim znajduje się osoba w odległości 1 m od czujnika. W każdym punkcie testowym wykonywano 200 pomiarów. Wyniki przedstawiono w tabeli 1.

Tab. 1. Wyniki pomiarów dokładności wyznaczania kąta

Kąt	*0°*	*10°*	*20°*	*30°*	*40°*	*50°*
Średni błąd [°]	-1,2	0,5	0,2	0,6	-0,4	1,9
Odchylenie standardowe [°]	0,5	0,4	0,6	0,5	0,6	0,9

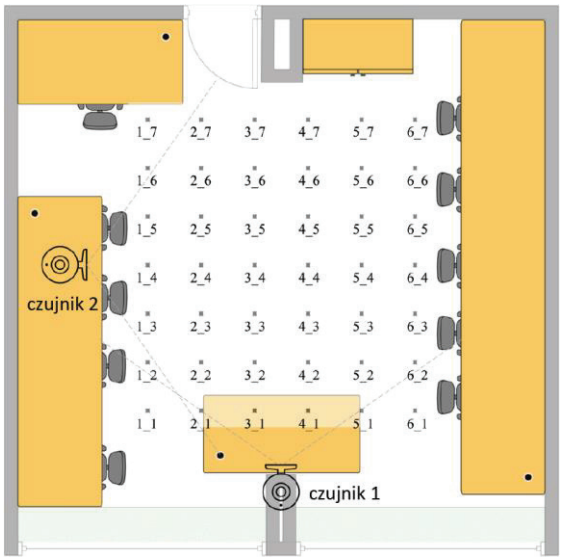
Błędy wyznaczania kąta mogą wynikać z faktu, że zbieranie wyników w każdym punkcie trwało ok. 100 sekund, w trakcie których osoba mogła się delikatnie odchylać, co z kolei w połączeniu z niepomijalną szerokością kątową człowieka mogło sprawić, że rzeczywisty środek osoby nie zawsze pokrywał się ze środkiem określonym przez algorytm. Niemniej, wyniki badań pokazują, że możliwe jest wykorzystanie matrycowych czujników w procesie wyznaczana lokalizacji.

5.2. Badanie błędu wyznaczania lokalizacji

Badanie przeprowadzono w pomieszczeniu laboratoryjnym o rozmiarach 6 na 6 m w 42 punktach pomiarowych. Dla każdego położenia osoby zebrano 200 wyników. Na rysunku 5 przedstawiono poglądowy schemat pomiarowy.

W tabeli 2 przedstawiono wyniki dla kilku punktów pomiarowych, natomiast na rysunku 6 zobrazowano wyniki graficznie. Za miarę dokładności przyjęto średni błąd wyznaczania lokalizacji, natomiast za miarę precyzji

przyjęto tzw. CEP75 (ang. Circular Error Probability), stanowiący promień okręgu, wewnątrz którego skupione jest 75% wyników pomiarów dla danego punktu testowego.



Rys. 5. Schemat pomiarowy badania błędu wyznaczania lokalizacji

Tab. 2. Przykładowe wyniki pomiarów wyznaczania lokalizacji

Identyfikator punktu	*Średni błąd wyznaczania lokalizacji [m]*	*CEP75 [m]*
1_3	0,28	0,12
2_3	0,28	0,04
3_5	0,18	0,05
4_5	0,24	0,00
5_4	0,11	0,04
6_3	0,09	0,07

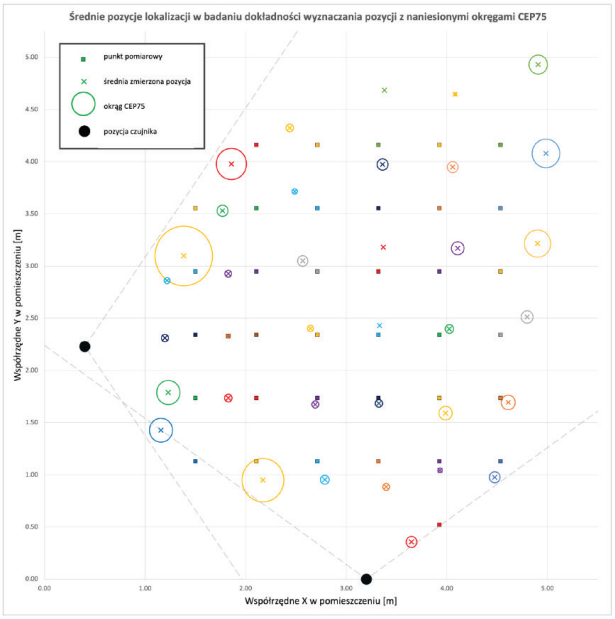
Analizując otrzymane wyniki, warto wspomnieć, że w dniu wykonywania pomiarów temperatura wewnątrz pomieszczenia sięgała 36°C, co miało negatywny wpływ na algorytm wykrywania sylwetki osoby. Widać również, że na rozrzut i błąd wyznaczania lokalizacji istotny wpływ ma położenie względem czujników – najlepsze wyniki uzyskiwane są w okolicy środka ich pola widzenia.

W odniesieniu do innych systemów, przedstawione rozwiązanie pod kątem dokładności plasuje się między systemami ultraszerokopasmowymi a systemami BLE/Wi-Fi. Jego zaletą jest brak konieczności noszenia urządzeń, a wadą istotny wpływ otoczenia (np. przysłanianie przez obiekty w pomieszczeniu).

6. PODSUMOWANIE

W pracy przedstawiono system do wyznaczania lokalizacji z wykorzystaniem matrycowych czujników podczerwieni. Przeprowadzone testy wskazują, że takie rozwiązanie można z sukcesem zastosować do wyznaczania położenia osoby w pomieszczeniu. Wyniki są na akceptowalnym poziomie. Dalsze prace mogą skupiać się na usprawnieniu algorytmu detekcji osoby i dodaniu możliwości lokalizowania kilku osób.

W referacie przedstawiono wyniki prac realizowanych w ramach projektu finansowanego przez Narodowe Centrum Badań i Rozwoju w ramach programu AAL Call 2021 nr umowy AAL/AALCall2021/52/CAREUP/2023.



Rys. 6. Wyniki pomiarów dokładności wyznaczania lokalizacji

LITERATURA

[1] Tang Z., Wang X. 2009. „An Efficient Algorithm for Infrared Small Target Detection". 2009 Second Int. Conf. on Information and Computing Science, Manchester, England, UK: IEEE, s. 51–54.

[2] Maung N.A., Lwi B. Y., Thida S. 2020. „An Enhanced RSS Fingerprinting-based Wireless Indoor Positioning using Random Forest Classifier", 2020 Int. Conf. on Advanced Information Technologies (ICAIT), Yangon, Myanmar: IEEE, s. 59–63.

[3] Ha K., Lee K., Lee S. 2016. „Development of PIR sensor based indoor location detection system for smart home", 2006 SICE-ICASE Int. Joint Conf., Busan Exhibition & Convention Center-BEXCO, Busan, Korea: IEEE, s. 2162–2167.

[4] Alam F., Faulkner N., Parr B. 2021. „Device-Free Localization: A Review of Non-RF Techniques for Unobtrusive Indoor Positioning". IEEE Internet Things J., t. 8, nr 6, s. 4228–4249.

[5] Faulkner N., Alam F., Legg M., Demidenko S. 2021. „Device-Free Localization Using Privacy-Preserving Infrared Signatures Acquired from Thermopiles and Machine Learning", IEEE Access, t. 9, s. 81786–81797.

[6] Feng Z.-J., Zhang X.-L., Yuan L.-Y., Wang J.-N. 2013. „Infrared Target Detection and Location for Visual Surveillance Using Fusion Scheme of Visible and Infrared Images". Math. Probl. Eng., s. 1–7.

[7] Melexis, „MLX90640 32x24 IR array Datasheet", Revision 12, 2019. Dostęp: 28.05.2024. [Online].

[8] He S., Long H., Zhang W. 2021. "Multi-antenna Array-based AoA Estimation Using Bluetooth Low Energy for Indoor Positioning,". 2021 7th Int. Conf. on Computer and Communications (ICCC), Chengdu, China, s. 2160-2164.

[9] Chantaweesomboon W. i inni, 2016. "On performance study of UWB real time locating system," 2016 7th Int. Conf of Inf. and Communication Tech. for Embedded Systems (IC-ICTES), Bangkok, Thailand, s. 19-24.

PRZEGLĄD METOD NAMIERZANIA EMITERÓW RADIOWYCH
OVERVIEW OF METHODS FOR TRACKING RADIO EMITTERS

Karolina BŁASZCZUK1, Zbigniew PIOTROWSKI2

[1] Wojskowa Akademia Techniczna, Warszawa, karolina.blaszczuk@student.wat.edu.pl
[2] Wojskowa Akademia Techniczna, Warszawa, zbigniew.piotrowski@wat.edu.pl

DOI: 10.15199/59.2024.4.32

Streszczenie: **Niniejszy przegląd omawia trzy główne metody namierzania emiterów radiowych: AOA, TDOA i DOA. Każda z metod ma swoje unikalne zalety i wady, a wybór odpowiedniej metody zależy od specyficznych wymagań, takich jak liczba dostępnych odbiorników, wymagania dotyczące dokładności oraz odporność na zakłócenia.**

Abstract: **This review discusses three main methods of radio emitter localization: AOA, TDOA, and DOA. Each method has its unique advantages and disadvantages, and the choice of the appropriate method depends on specific requirements such as the number of available receivers, accuracy requirements, and resistance to interference.**

Słowa kluczowe: **emitery radiowe, kąt nadejścia, kierunek nadejścia, różnica czasu nadejścia, walka radioelektroniczna**

Keywords: **AOA, DOA, electronic warfare, radio emitters, TDOA,**

1. WSTĘP

Celem niniejszego przeglądu jest analiza metod namierzania emiterów radiowych, skupiając się na trzech głównych technikach: AOA (Angle of Arrival), TDOA (Time Difference of Arrival) oraz DOA (Direction of Arrival). Przegląd ten ma na celu zidentyfikowanie zalet i wad każdej z metod oraz określenie ich najbardziej odpowiednich zastosowań w różnorodnych kontekstach zastosowań praktycznych.

1.1. Znaczenie emiterów radiowych

Precyzyjne namierzanie emiterów radiowych ma kluczowe znaczenie w kontekście współczesnych konfliktów oraz zarządzania kryzysowego. Techniki te pozwalają na wykrycie, identyfikację oraz lokalizację źródeł sygnałów radiowych, co jest niezbędne do skutecznego zarządzania operacjami wojskowymi i cywilnymi [11].

1.2. Znaczenie emiterów radiowych

Namierzanie emiterów radiowych jest istotnym elementem walki radioelektronicznej, umożliwiającym identyfikację oraz neutralizację zagrożeń. Dzięki zaawansowanym metodom lokalizacji możliwe jest wykrywanie i przeciwdziałanie działaniom przeciwnika w sposób bardziej precyzyjny i efektywny [11].

2. DEFINICJE I PODSTAWY TEORETYCZNE

Wprowadzenie do metod AOA, TDOA i DOA obejmuje definicje oraz podstawy teoretyczne działania tych technik, a także omówienie ich zastosowań praktycznych w różnych kontekstach zastosowań.

2.1. Angle of Arrival

AOA działa na zasadzie pomiaru różnicy w fazie lub czasie przybycia sygnału do różnych anten. Te różnice są następnie przekształcane na kąty przybycia sygnału stosując matematyczną analizę triangulacji lub zaawansowane algorytmy przetwarzania sygnałów [5]. Kąt nadejścia sygnału (θ) z dwóch anten może być wyliczony przy pomocy prostego wzoru opartego na różnicy faz sygnałów (Δϕ) i odległości między antenami (d):Proponujemy włączenie opcji dzielenia wyrazów (komendy: Narzędzia, Język, Dzielenie wyrazów, Dziel wyrazy automatycznie). Uniknie się w ten sposób pustych miejsc szczególnie widocznych w tekście dwukolumnowym. Przypominamy jednak o konieczności wyboru języka polskiego jako języka tworzenia dokumentu w edytorze MS-Word.

$$\theta = arcsin\frac{\lambda\triangle\varphi}{2\pi d} \quad (1)$$

gdzie λ to długość fali sygnału radiowego [18]. Dane potrzebne do obliczeń: długość fali (λ) sygnału, zależna od jego częstotliwości, różnica czasu przybycia sygnału lub różnica faz (Δϕ). odległość między antenami (d), która musi być znana i stała.

2.2. Zalety i wady AOA

Jako pierwszą z zalet metody AOA należy wymienić możliwość zastosowania jej w różnych warunkach środowiskowych - duża skuteczność w środowiskach miejskich. Metoda często jest stosowana w algorytmach do minimalizacji efektów wielodrogowości [16]. Wyróżnia ją również skuteczność w pasmach częstotliwości o wysokiej zajętości, gdzie inne metody mogą zawodzić. Metoda pozwala na oddzielenie sygnałów pochodzących od różnych nadajników [16].

Jej główną wadą jest zależność odliczby i jakości anten – im wyższa jakość i większa liczba anten, tym lepsze wyniki, ale również wyższe koszty. Jest także wrażliwa na błędy synchronizacji, a nieprecyzyjna synchronizacja czasowa między antenami może znacznie wpłynąć na dokładność pomiarów [16]. Dodatkowo potrzebuje wolnej przestrzeni wokół anten, celem uniknięcia zakłóceń i odbić sygnału [16].].

2.3. Time Difference of Arrival

TDOA (Time Difference of Arrival) jest metodą lokalizacji, która wykorzystuje różnice czasu przybycia sygnału do różnych odbiorników. Te różnice są następnie używane do triangulacji pozycji źródła sygnału [18]. Jeśli Δt oznacza różnicę w czasie przybycia między dwoma odbiornikami, a v to prędkość sygnału, to odległość d między źródłem a każdym z odbiorników można obliczyć jako:

$$d = v \triangle t \tag{2}$$

Metoda ta jest skuteczna w środowiskach, gdzie możliwe jest precyzyjne mierzenie czasów przybycia z różnych kierunków np. w środowiskach miejskich mimo potencjalnych zakłóceń spowodowanych odbiciami sygnału od budynków, TDOA jest użyteczna dzięki zdolności do rozróżniania sygnałów [8].

2.4. Zalety i wady TDOA

Do zalet metody TDOA należy wysoka dokładność lokalizacji źródła sygnału, szczególnie w środowiskach o wysokim poziomie szumów. Pomiar różnic czasowych jest mniej podatny na zakłócenia niż pomiar amplitudy. TDOA może być stosowana w systemach z dużą liczbą odbiorników, co pozwala na uzyskanie bardzo dokładnych wyników. Większa liczba odbiorników poprawia dokładność triangulacji. Dodatkowo metoda jest wysoce odporna na zakłócenia w zatłoczonych pasmach częstotliwości, co czyni ją idealną w środowiskach z dużą ilością zakłóceń radiowych.

Jako pierwszą z wad tej metody można uznać wymóg precyzyjnej synchronizacji czasowej pomiędzy odbiornikami, co może być trudne do osiągnięcia i utrzymania, zwłaszcza w dużych systemach [8]. Algorytmy TDOA mogą być skomplikowane i wymagać zaawansowanego przetwarzania sygnałów, co może wymagać dużej mocy obliczeniowej. Dodatkowo wielotorowość sygnału może prowadzić do błędów w pomiarach różnic czasowych, co negatywnie wpływa na dokładność lokalizacji [8].

2.5. Direction of Arrival

DOA (Direction of Arrival) DOA, podobnie jak AOA, określa kierunek, z którego przychodzi sygnał. Metoda ta jest często używana w zaawansowanych systemach radarowych i telekomunikacyjnych. Zasada działania: DOA używa różnic fazowych lub czasowych sygnału przychodzącego do pojedynczego, ale zaawansowanego odbiornika lub szyku anten, aby określić kierunek jego przybycia [20]. Definicja i zasada działania: DOA określa kierunek przybycia sygnału i jest stosowana, gdy ważne jest określenie kierunku nadajnika bez konieczności jego dokładnego lokalizowania. Algorytmy DOA są zazwyczaj prostsze i szybsze niż te wymagane do dokładnej lokalizacji. W sytuacjach, gdzie czas reakcjijest krytyczny (np. w radarach lotniczych czy systemach obrony), szybkie określenie kierunku może być bardziej wartościowe niż dokładna lokalizacja. Poniżej przedstawiono podstawowy wzór używany do obliczeń:

$$\theta = arcsin \frac{\triangle \varphi \lambda}{2\pi d} \tag{3}$$

2.6. Wady i zalety DOA

Metoda DOA jest stosunkowo prostą metodą do zaimplementowania w systemach z jednym zaawansowanym odbiornikiem, co redukuje koszty i złożoność systemu. Może być stosowana do szybkiego określania kierunku źródła sygnału, co jest szczególnie przydatne w systemach radarowych i obronnych. Jej wszechstronność: zastosowań pozwala na użycie jej w różnych aplikacjach, w tym w astronomii radiowej, monitorowaniu środowiskowym i systemach satelitarnych, gdzie ważne jest określenie kierunku sygnału [20].

Jako pierwszą z wad należy wymienić fakt, że DOA określa jedynie kierunek przybycia sygnału, a nie jego dokładną pozycję, co może być niewystarczające w niektórych aplikacjach wymagających precyzyjnej lokalizacji. Dokładność metody DOA może być ograniczona w warunkach z dużą ilością zakłóceń i odbić, co negatywnie wpływa na pomiar kierunku sygnału [20]. Dodatkowo metoda ta często wymaga użycia wieloelementowych anten lub szyków antenowych, co może zwiększać koszty i złożoność instalacji systemu [20].

3. MINIMALNA LICZBA ODBIORNIKÓW WYMAGANA W AOA, TDOA I DOA

Metoda AOA wymaga co najmniej dwóch anten, ponieważ zasada działania tej techniki opiera się na pomiarze różnicy czasu lub fazy, z jakim sygnał radiowy dociera do każdej z anten. Dwie anteny umożliwiają stworzenie linii bazowej, która jest konieczna do zastosowania triangulacji i określenia kierunku przybycia sygnału. Użycie dwóch lub więcej anten pozwala na precyzyjne określenie kąta względem linii bazowej, z której nadchodzi sygnał.

Metoda TDOA wymaga co najmniej dwóch odbiorników, aby obliczyć różnicę czasową przybycia sygnału między nimi. Jednak do dokładnego określenia pozycji źródła sygnału w dwóch wymiarach (2D) potrzebne są trzy odbiorniki, a w trzech wymiarach (3D) cztery odbiorniki [8]. Dwa odbiorniki pozwalają na obliczenie jednej różnicy czasu przybycia, co definiuje hiperbolę możliwych pozycji źródła sygnału. Trzeci odbiornik natomiast wprowadza dodatkową hiperbolę, która przecina się z pierwszą, pozwalając na określenie jednego punktu w przestrzeni 2D. Większa liczba odbiorników niż trzy poprawia dokładność lokalizacji, ponieważ dostarcza więcej punktów przecięcia hiperbol, co zmniejsza błąd pomiaru.

Metoda DOA wymaga użycia szyków antenowych, które mogą składać się z wielu elementów antenowych ustawionych w określony sposób. Zasada działania tej techniki opiera się na pomiarze kierunku, z jakiego sygnał przybywa do szykuantenowego. DOA używa analizy fazowej lub amplitudowej sygnałów odbieranych przez różne elementy antenowe w szyku, aby określić kąt przybycia sygnału. Metoda ta jest skuteczna w określaniu kierunku źródła sygnału, ale nie dostarcza bezpośrednio informacji o jego dokładnej pozycji.

Minimalna liczba elementów potrzebna do określenia kierunku przybycia sygnału za pomocą podstawowych metod fazowych to dwa elementy antenowe. Trzy elementy antenowe umożliwiają

natomiast określenie kierunku sygnału w dwóch płaszczyznach (azymut i elewacja). Cztery lub więcej elementów antenowych: Pozwalają na jeszcze bardziej precyzyjną analizę sygnału, uwzględniając bardziej złożone algorytmy i metody przetwarzania sygnału takie jak: MUSIC lub ESPRIT [1].

4. WPŁYW ZAKŁÓCEŃ NA WYBRANE METODY ORAZ ICH DOKŁADNOŚCI

Tab. 1. Porównanie metod namierzania emiterów radiowych

Metoda	*Dokładność*	*Odporność na zakłócenia*	*Typowe zastosowania*
AOA	1-2°	średnia	radar, lotnictwo
TDOA	Kilka metrów	wysoka	telekomunikacja, ratownictwo
DOA	1-2°	zmienna	astronomia, badania naukowe

Metody AOA, TDOA i DOA odgrywają kluczową rolę w namierzaniu emiterów radiowych. Każda z tych technik ma swoje unikalne zastosowania i cechy, które czynią je odpowiednimi do różnych zastosowań praktycznych.

AOA (Angle of Arrival): wymaga co najmniej dwóch anten i jest precyzyjna w aplikacjach radarowych i lotniczych. Z tabeli nr 1 wynika, że oferuje ona dokładność 1-2°, co wynika z precyzyjnych pomairów kątów przybycia sygnału względem anten i jest średnio odporna na zakłócenia w środowiskach z dużą liczbą odbić i interferencji [3].

TDOA (Time Difference of Arrival): wymaga trzech lub więcej odbiorników. Jest bardzo skuteczna w środowiskach z wysokim poziomem szumów i zakłóceń, co czyni ją idealną do telekomunikacji i operacji ratunkowych [5]. Zapewnia dokładność rzędu kilku metrów, zależnie od warunków środowiskowych i jakości sprzętu pomiarowego [3].

DOA (Direction of Arrival): wykorzystuje szyki antenowe i jest kluczowa w astronomii oraz badaniach naukowych. Pomaga określić kierunek przybycia sygnału bez konieczności dokładnej lokalizacji źródła, oferując dokładność 1-2°. DOA wykazuje zmienną odporność na zakłócenia. Zależy to od użytej technologii i konfiguracji anten. Wykorzystanie zaawansowanych algorytmów może poprawić odporność na zakłócenia [3].

Wybór odpowiedniej metody zależy od specyficznych potrzeb operacyjnych, takich jak wymagania dotyczące dokładności, liczba dostępnych odbiorników i odporność na zakłócenia.

5. PODSUMOWANIE

Namierzanie emiterów radiowych jest krytycznym aspektem wielu dziedzin, od wojskowości po telekomunikację i badania naukowe. Znajomość i właściwy wybór metody (AOA, TDOA, DOA) umożliwiają efektywne i precyzyjne operacje w różnych warunkach i scenariuszach. Każda z metod ma swoje zalety i wady, a wybór odpowiedniej techniki zależy od specyficznych wymagań operacyjnych, takich jak potrzeba dokładności, liczba dostępnych odbiorników oraz odporność na zakłócenia. Wybór właściwej metody pozwala na skuteczne i precyzyjne namierzanie emiterów radiowych w różnych warunkach operacyjnych.

LITERATURA

[1] Choi Lee S. 2018. "Advanced DOA Algorithms for Environmental Monitoring". *Environmental Monitoring and Assessment 190(6): 355-362.*

[2] White, M. Green. 2016. "Analysis of AOA Methods in Air Traffic Control*". Journal of Air Traffic Management 8(4): 220-234.*

[3] AI-Sadoon M. A. G., Ali T. 2017. „A New Low Complexity Angle of Arrival Algorithm for 1D and 2D Direction Estimation in MIMO Smart Antenna Systems". *Sensors.*

[4] Florio A., Avitabile G. 2020 „An Experimental Validation of the Naïve Approach to Angle of Arrival Estimation for Green WSNs" *IEEE 20th Mediterranean Electrotechnical Conference (MELECON).*

[5] . Smith Johnson B. 2022. "Angle of Arrival Estimation Using Advanced Antenna Arrays." *Journal of Communication Engineering 15(3): 255-267*.

[6] Sharma Patel, R.. 2014. "Comparative Analysis of TDOA and DOA in Various Scenarios." *International Journal of Signal Processing 22(7): 481-490.*

[7] Park Kim, J. 2019. "Comparative Study of AOA and TDOA Techniques in Radar Systems.". *International Journal of Electronics and Communications 112: 10-22.*

[8] Li X., Deng Z.D. 2016," Contributed Review: Source-localization algorithms and applications using time of arrival and time difference of arrival measurements". *Review of Scientific Instruments.*

[9] Hernandez Garcia, P. 2013. "Effective AOA Estimation in Dense Signal Environments." *IEEE Transactions on Wireless Communications 12(8): 3420-3430.*

[10] Zhang, Y. Zhang, Z. He, J. Wei. 2023. "Efficient TDOA-Based Localization Method for Mobile Sensor Networks". IEEE Sensors Journal 23(1): 1-11.

[11] Allah M. Khhalaf 2021. „Emitter Location with Azimuth and Elevation Measurements Using a Single Aerial Platform for Electronic Support Missions', Sensors. 1

[12] KRECZMER B. 2019. „Estimation of the Azimuth Angle of the Arrival Direction for an Ultrasonic Signal by Using Indirect Determination of the Phase Shift", Archives of Acoustics.

[13] Tanaka. Nakamura, Y. 2017. "High-Precision TDOA Systems for Emergency Response." IEEE Access 5: 22789-22800.

[14] Baker. Turner, L 2012. "Innovations in DOA Techniques for Smart Grids." *IEEE Transactions on Smart Grid 3(1): 55-65.*
[15] Brown, T. Wilson. 2015. "Integration of DOA Techniques in Satellite Communication." Satellite Communications Journal 31(2): 67-75
[16] J. Li, S. Lv,2023, „Joint TDOA, FDOA and PDOA Localization Approaches and Performance Analysis". Proceedings of the IEEE.
[17] AI-Jazzar S. O, Muchkaev A. 2011. „Low complexity and high accuracy angle of arrival estimation using eigenvalue decomposition with extension to 2D AOA and power estimation". Proceedings of the IEEE.
[18] Wang, D., Hu Z.., 2024, „Optimal station placement based on grey wolf optimizer for regional target localization". Published in the EURASIP Journal on Wireless Communications and Networking.
[19] Chen H., Ballal T., 2019, „Realtime 2-D DOA Estimation using Phase-Difference Projection". *27th European Signal Processing Conference (EUSIPCO).*

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

LEVERAGING LORAWAN FOR INTELLIGENT GARBAGE COLLECTION IN CITIES USING MATERIAL COMPOSITION DATA FROM SMART BINS

Angela Ruiz Alvarez [1]; Andres Vejar [2]; Faysal Marzuk [3]

Departamento de Ingeniería de Sistemas Telemáticos, ETSI de Telecomunicación,
Universidad Politécnica de Madrid, Madrid, Spain
[1]angela.ruiz.alvarez@alumnos.upm.es

Institute of Telecommunications, AGH University of Krakow, Kraków, Poland,
[2]avejar@agh.edu.pl [3]marzuk@agh.edu.pl

DOI: 10.15199/59.2024.4.33

Abstract: **Smart cities are increasingly utilizing IoT sensing technologies to address urban challenges more effectively. Traditional 4G/5G technology-based measurement systems may not be suitable for applications requiring thousands of IoT nodes, making Low Power Wide Area Network (LPWAN) technologies like LoRaWAN more suitable. This article focuses on waste management, a critical aspect of smart cities. Despite societal pressure for segregation and recycling, there is no guarantee that the collected waste will be processed correctly. To address this, a LoRaWAN system is explored to monitor waste composition information in smart bins. This data helps to monitor waste bin fullness levels in real time, optimize garbage collection routes, and remotely manage collection schedules. We assessed the LoRaWAN system in the context of an actual city scenario using field measurements and simulation within a framework to incorporate LoRaWAN-enabled sensors and devices into existing infrastructure, aiming for operational efficiency, cost savings, and the overarching goal of enhanced environmental sustainability.**

Keywords: **LoRaWAN, Smart Bins, Intelligent Waste Management.**

1. INTRODUCTION

Smart cities are emerging as innovative and efficient solutions to improve quality of life, optimize resources, and offer sustainable development. These cities use databases, IoT sensors, and AI-driven algorithms to address urban population growth, mitigating traffic congestion, reducing pollution, improving public safety, and creating more sustainable environments. Efficient data transmission is crucial for these solutions. LoRaWAN, a promising and cost-effective technology, is used in smart cities to transmit data over long distances with minimal power consumption. This project aims to evaluate the performance and scope of data transmission using LoRaWAN to optimize planning in smart cities. The focus will be on improving the waste collection system and applying garbage data collection to automate the process. A prior analysis of LoRaWAN[1] and its various aspects is necessary to comply with the required specifications, such as efficient data transmission over long distances within urban environments and the scalability of LoRaWAN networks to accommodate the growing number of connected devices and sensors in the environment.

Since real-time waste identification systems using images and deep neural networks [2] are accessible, it is crucial to investigate economical methods to determine waste composition for smart cities. One possibility is the use of radio signals to estimate the composition of materials, for example, using WiFI channel state information (CSI) [3, 4]. This information can be obtained in a reliable manner using nodes equipped with WiFi + LoRa communication interfaces. That can be done directly in instrumented waste bins, the so-called smart bins. Considering the quantity of those units, around tens of thousands are required to cover the surface of modern cities, high-end, computationally intensive solutions are not a viable option. Therefore, the analysis of waste using a garbage bin augmented with a customized IoT device that will monitor the waste composition using CSI from WiFi radio transceivers. The system will use sensor data to estimate waste recycling potential. This information can be useful for smart cities to implement opportunistic and targeted recycling and clearing planning, such as specialized carriers for high content of biowaste, plastic or metal, among others.

2. ROLE OF IOT IN SMART CITY DEVELOPMENT

Smart City IoT deployments are driving innovation in low-power long-range wireless communication networks, leading to the development of LPWANs [5], such as LoRaWAN, in this project. Smart bins are being used in waste management, providing real-time monitoring of capacity and content. The geographic positioning of these containers on city maps provides valuable information about the dimensions and patterns of deployment of the waste network. The project focuses on Madrid and Krakow as smart cities, exploring their innovative approaches to urban environmental challenges. Madrid, a European capital with advanced technological infrastructure, is a reference for sustainability initiatives. Krakow has also integrated IoT solutions into urban life, improving efficiency and quality of life. Examples include the use of IoT sensors and smart traffic management systems to reduce traffic congestion, smart grids and metering systems to regulate energy use, and improved public safety through surveillance and emergency response sys-

tems. These cities demonstrate their dedication to adapting to the new changes and needs that come with being a smart city. By incorporating IoT solutions into various aspects of urban life, cities such as Madrid and Krakow can improve the lives of their citizens and contribute to a more sustainable future.

A numerical analysis of both cities will determine the scalability of the project, going over different aspects of the city that are decisive for the study and simulated implementation of the LoRaWAN network in it. This quantitative evaluation includes a topographic and demographic analysis of both cities, as well as their behavior with respect to garbage consumption, is shown in Table 1. Based on the Statistics Institute of the Community of Madrid, the demographic and topographical study concludes with the figures extracted from the last population census carried out in 2021. Similarly, data are collected from the statistics section by region of the Krakow Statistics Office. Subsequently, the behavior of citizens and their annual waste production habits in the past years in both cities are analyzed, obtaining very similar results. The citizens of Madrid annually produce 370 kg of trash per person, while Krakow's annual generation of municipal waste per capita is 325 kg.

Gathered Data	MADRID	KRAKOW
City surface (km^2)	604.56	320
Population	3277451	806200
Trash production kg per person per year	370	325
Population density (person/km^2)	5421.22	2519.37
Total trash production (kg/year)	1212656.87	262015
Containers per km^2	72.18	33.43

Table 1: Comparison of Demographic and Waste Behavior

The garbage containers constitute the end devices deployed in the LoRaWAN network. Therefore, these numbers allow us to evaluate the scalability of the project and to know the number of devices required to efficiently cover all specified spots, in order to simulate the deployment of the LoRaWAN network using real-world parameters.

3. PROPOSED LORAWAN NETWORK DEPLOYMENT FOR TRANSMITTING DATA

The schematic representation of the LoRaWAN network shown in Figure 1 is the one that will be displayed for the further simulation of Madrid and Krakow as smart cities. As a whole, it makes up a developed IoT waste management system whose end nodes are the individual battery powered LoRa modules installed in the garbage cans. These implemented LoRa modules will operate in the EU868 region, as the project is designed to operate in both the European countries, Spain and Poland. The gateways will then be strategically and equally distributed across our implemented smart city scenario, transmitting information to a central server, where the data will be processed and analyzed for the later proposal of optimization of waste collection routes, thereby improving efficiency and reducing operational costs in both cities.

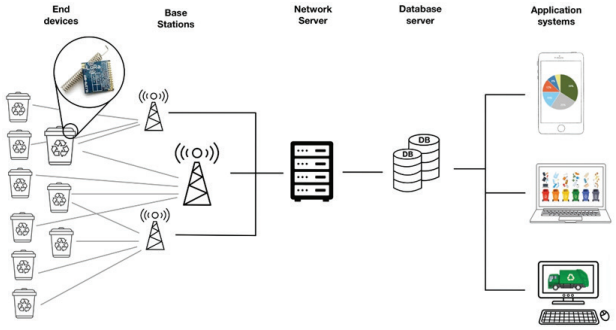


Figure 1: General Scheme

Several LoRa parameters are configured for optimal transmission. The RF frequency in which we operate is set to 865 MHz, a frequency within the ISM band commonly used in Europe EU868, allowing its operation without a license. Sequentially transmission power is set at 5dBm and bandwidth of the transmission is configured to 125 kHz. This corresponds to a suitable choice for most IoT applications and ensures a good balance between range and sensitivity [6].

Figure 2 illustrates the state machine used for transmission. **STATE_TX** represents a transmit state in which a message is sent, **STATE_RX** represents a receive state in which a device is listening for incoming messages, and **LOWPOWER** describes a low-power state where the device handles interrupts. We have the following event handling functions:

- OnTxDone: If a transmission is completed.
- OnTxTimeout: If there is a transmission timeout.
- OnRxDone: If a message is received.

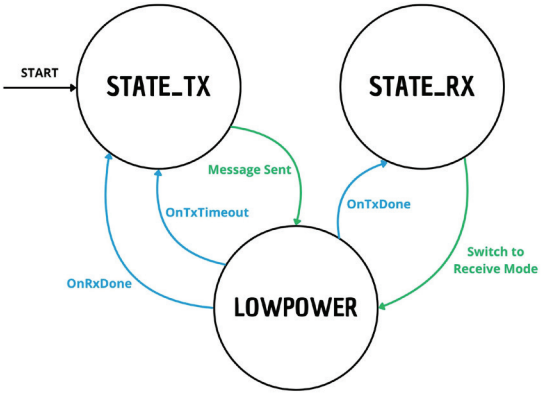


Figure 2: State Machine used for transmission.

There are the following transitions:

1. **STATE_TX** transitions to **LOWPOWER** after sending a message.
2. **STATE_RX** transitions to **LOWPOWER** after switching to the receiving mode.

3. In **LOWPOWER**, state transitions are driven by interrupt processing and events:

   - OnTxDone: it transitions to **STATE_RX**.
   - OnTxTimeout: it transitions to **STATE_TX**.
   - OnRxDone: it also transitions to **STATE_TX**.

4. EXPERIMENTAL RESULTS

We perform several LoRa transmissions between our two devices with configurations of three different parameters; spread factor, coding rate, and reception timeout. We vary the distance between the two devices in all cases. The operational parameters of the experiments are shown in Table 2.

Parameter	Case 1	Case 2	Case 3	Min Val.	Max Val.
Spreading factor	7	9	12	7	12
Coding rate	1	4	4	1	4
Timeout value	3000	5000	5000	-	-

Table 2: Parameters for the three measurement cases.

A base module was installed in an office room and measurements were taken at various points, including the office, cafeteria, second floor, first floor, hallway, and outside the front door. The aim was to evaluate the impact of indoor settings and obstacles on RSSI and the overall transmission quality of LoRa signals. The results are plotted on a graph in Figure 3.

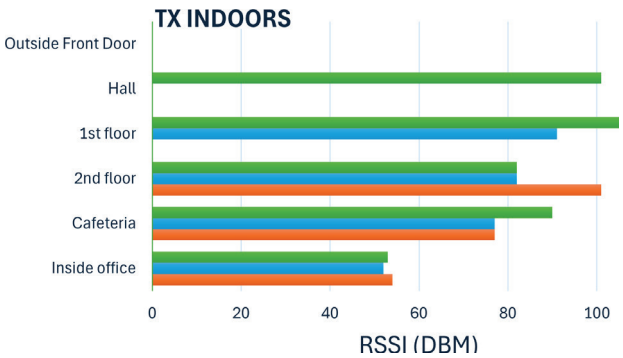


Figure 3: Gathered RSSI Data for Indoor Scenario.

The outdoor test field around campus uses a LoRa device as a base and measures points without urban obstacles. A plane is created to locate the base and measurement points, as shown in Figure 4. The chart in Figure 5 illustrates the transmission and reception outcomes for both LoRa Modules, indicating analogous performance between the stationary and mobile nodes. An investigation into the performance of these two LoRa devices uncovered considerable differences in signal strength across various conditions.

For indoor measurement, both devices transmitted and received signals with excellent results when communicating short distances, with RSSI values ranging from -52 to -54 dBm. However, as distance and barriers increase, signal strength drastically decreases. For example, in a separate room, RSSI readings are around -80 dBm, indicating a small deterioration of the signal but still good quality transmission. The signal weakens significantly as we move one floor below the base node, with RSSI values as low as -101 dBm in Case 1, while Cases 2 and 3 perform slightly better at -82 dBm. Case 3 maintains a weak signal, suggesting slightly superior resilience in interior situations. No signal is received in any of the cases where we move outside the front door of the Faculty building.

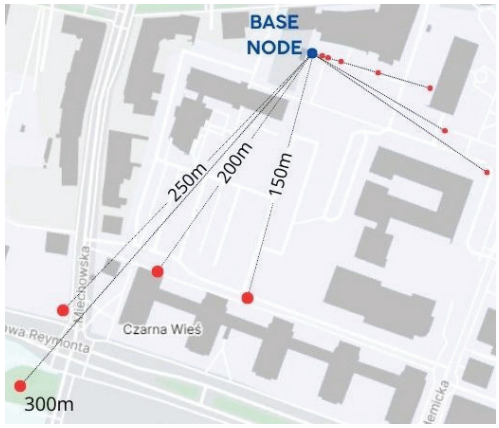


Figure 4: Measurement points located on campus map with their respective distances to Base Node Test Fields.

In outdoor measurements the signal transmission is more effective due to fewer obstructions compared to faculty buildings. Devices maintain strong signal strength over longer distances. RSSI values are between -54 to -61 dBm at 5 meters, -67 to -76 dBm at 30 meters, and -98 to -101 dBm at 150 meters. However, at 250 meters, the transmission weakens significantly, with Cases 1 and 2 losing signal entirely, whereas Case 3 retains a weak signal at -109 dBm. Beyond 300 meters, all devices fail to maintain a signal, showing the maximum effective range of the devices.

5. SIMULATION RESULTS

The LWN-Simulator is used for network deployment simulations in Madrid and Krakow scenarios. This user interface facilitates the deployment of a LoRaWAN network with its associated node and gateway configuration. It allows for precise location establishment using latitude and longitude coordinates and can simulate socket connections, enabling a more realistic simulation of LoRaWAN network design before real-world deployment. In order to function, the ChirpStark LoRa server is deployed. It enables users to confirm the network connectivity of the configured nodes and to assess the performance of the gateways. End devices are created with an identifier (Device EUI) and gateways with a MAC address. The nodes have several automatically created keys, and a message of choice is added to the payload.

Once the LoRaWAN network is deployed and simulation is completed for the designated area for both cities, some quantitative calculations are obtained on the total number of gateways necessary for the correct functioning of the network. Based on an existing physical simulation

approach [7], a quantitative estimate of the gateways required for the waste management network can be carried out. The number of gateways is estimated to be higher than the theoretical requirement to ensure reliable and consistent LoRaWAN coverage. However, the garbage level monitoring system does not require absolute reliability, unlike critical transmissions such as fire alarms or healthcare systems. The system aims to improve a sector and optimize waste collection, so occasional data loss is acceptable. Missing a few transmissions will not significantly affect the overall effectiveness of the system. The final number of gateways required for the system will exceed the theoretically calculated number by 50%, estimating that theoretically one gateway per square kilometer provides enough coverage. A quantitative analysis of the final gateways required for both cities is shown in Table 3.

6. CONCLUSIONS

LoRa devices show great performance in open spaces and across short distances, and physical barriers and increased distance have a major impact on signal strength and transmission reliability. Therefore, when using these devices inside, infrastructure obstacles should definitely be taken into account. In both test fields, Case 3 shows the highest resistance and the best performance at longer distances, leading to the best option to optimize performance indoors and outdoors. It is also important to emphasize the importance of strategic device placement and environmental conditions to achieve optimal performance in real-world applications. As far as this project is concerned, the conditions will be those taken outdoors, using the maximum transmission distance between the end nodes and the gateways.

The LWN-Simulator is useful for accurate modeling of LoRaWAN networks, particularly in dense urban environments like Madrid and Krakow. The simulation results confirm that the maximum range of a LoRaWAN network is restricted by the distance between the node and the gateway, with a maximum effective range of 7.5 km achieved. The successful deployment of the network in one square kilometer in both scenarios demonstrated the feasibility of LoRaWAN deployment in large urban environments.

The theoretical framework indicating that only one gateway per square kilometer is needed was validated. The network exhibited robust connectivity with the server, effectively handling urban signal obstructions. To improve the network robustness, the final number of gateways required for both cities was increased 1.5 times, resulting in the final number of gateway cities required. For Madrid, 908 gateways are required, while for Krakow, 491. The network also demonstrated good connectivity when all nodes transmitted data using the DR0 data rate and frequencies within the EU868 region, demonstrating its ability to maintain reliable communication even under high node density and data traffic conditions. The simulation indicates that LoRaWAN can provide extensive coverage and facilitate scalability with additional nodes and gateways, supporting urban growth and guaranteeing reliable communication in urban environments. This confirms that LoRaWAN is a suitable option for IoT applications in smart cities, especially in the waste management sector.

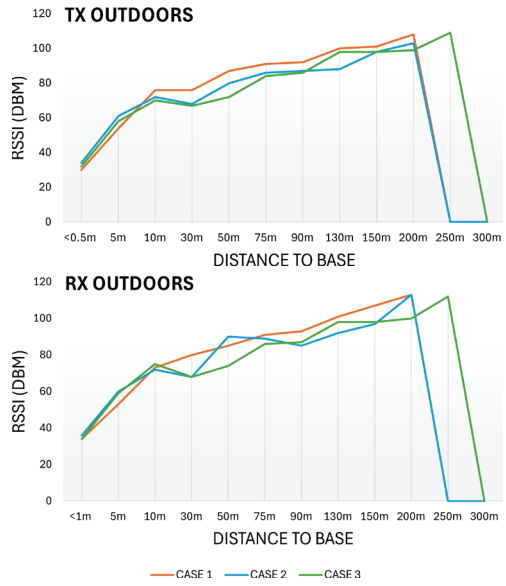


Figure 5: Gathered RSSI Data for Outdoor Scenario.

Parameters	MADRID	KRAKOW
City Surface (km^2)	604.56	326.80
Total number of end nodes	43636	10731
Required Gateways		
Theoretical	605	327
Practical	908	491

Table 3: Calculations of gateways required for Madrid and Krakow.

REFERENCES

[1] Muhammad Ayoub Kamal, Muhammad Mansoor Alam, Aznida Abu Bakar Sajak, and Mazliham Mohd Su'ud. Requirements, deployments, and challenges of lora technology: A survey. *Computational Intelligence and Neuroscience*, 2023:1–15, January 2023.

[2] Daniel Octavian Melinte, Ana-Maria Travediu, and Dan N Dumitriu. Deep convolutional neural networks object detector for real-time waste identification. *Applied Sciences*, 10(20):7301, 2020.

[3] Steven M. Hernandez, Deniz Erdag, and Eyuphan Bulut. Towards dense and scalable soil sensing through low-cost wifi sensing networks. In *2021 IEEE 46th Conference on Local Computer Networks (LCN)*, pages 549–556, 2021.

[4] Sheng Tan and Jie Yang. Object sensing for fruit ripeness detection using wifi signals. *arXiv e-prints*, pages arXiv–2106, 2021.

[5] Bharat S. Chaudhari, Marco Zennaro, and Suresh Borkar. Lpwan technologies: Emerging application characteristics, requirements, and design considerations. *Future Internet*, 12(3), 2020.

[6] Andres Vazquez-Rodas, Fabian Astudillo-Salinas, Christian Sanchez, Braulio Arpi, and Luis I. Minchala. Experimental evaluation of rssi-based positioning system with low-cost lora devices. *Ad Hoc Networks*, 105:102168, 2020.

[7] Nadège Varsier and Jean Schwoerer. Capacity limits of lorawan technology for smart metering applications. In *2017 IEEE International Conference on Communications (ICC)*, pages 1–6, 2017.

WPŁYW METOD AKTYWACJI URZĄDZEŃ PRACUJĄCYCH W SIECI LORAWAN NA BEZPIECZEŃSTWO SYSTEMU

THE IMPACT OF ACTIVATION METHODS FOR DEVICES OPERATING IN THE LORAWAN NETWORK ON SYSTEM SECURITY

Joanna Szewczyk[1]; Piotr Remlein[2]; Mariusz Nowak[1]; Wojciech Cieślak[2];

[1]Politechnika Poznańska, Instytut Informatyki, Poznań, Joanna.Szewczyk@put.poznan.pl, Mariusz.Nowak@put.poznan.pl,
[2] Politechnika Poznańska, Instytut Radiokomunikacji, Poznań, Piotr.Remlein@put.poznan.pl; Wojciech.Cieślak@student.put.poznan.pl

DOI: 10.15199/59.2024.4.34

Streszczenie**: W artykule poruszono wpływ metody aktywacji urządzeń końcowych pracujących w sieci LoRaWAN na ich podatność na ataki. Przeprowadzone eksperymenty miały na celu zidentyfikowanie słabości systemu oraz ocenę skuteczności potencjalnych ataków i ich wpływu na bezpieczeństwo komunikacji w sieci LoRaWAN. Komunikacja w sieci LoRaWAN, mimo zastosowania metod uwierzytelnienia i szyfrowania, jest podatne na ataki.**

Abstract**: The article addresses the impact of the method of activation of end devices operating in a LoRaWAN on their vulnerability to attacks. The experiments conducted were aimed at identifying system vulnerabilities and assessing the effectiveness of potential attacks and their impact on the security of LoRaWAN communications. LoRaWAN communications, despite the use of authentication and encryption methods, are vulnerable to attacks.**

Słowa kluczowe**: DoS, LoRaWAN, LPWAN, Szyfrowanie, Bezpieczeństwo Komunikacji.**

Keywords**: DoS, LoRaWAN, LPWAN, Encryption, Communication Security.**

1. WSTĘP

W obecnych czasach, w których Internet Rzeczy (IoT-Internet of Things) zyskuje na znaczeniu, sieci bezprzewodowe o dalekim zasięgu (LPWAN-Low Power Wide-Area Network) zdobywają dużą popularność. Technologie takie jak LoRaWAN, Sigfox oraz NB-IoT stanowią fundamenty dla innowacyjnych aplikacji, włączając w to inteligentne budynki, inteligentne miasta (Smart City), inteligentne systemy pomiarowe czy zdalne monitorowanie. Te technologie mają potencjał do poprawy jakości życia codziennego poprzez usprawnianie działań, optymalizację procesów i podnoszenie skuteczności działania systemów. Pomimo swoich licznych zalet, technologie LPWAN narażone są na potencjalne ataki cybernetyczne. W systemie LoRaWAN, każda transmisja może trwać od 900 milisekund do 1,2 sekundy, co daje atakującym czas na przeprowadzenie skutecznego ataku [1].

Zapewnienie bezpieczeństwa sieci LPWAN jest utrudnione z powodu ograniczonych zasobów dostępnych w urządzeniach końcowych, takich jak pamięć i bateria, co uniemożliwia stosowanie zaawansowanych algorytmów kryptograficznych wymagających dużej mocy obliczeniowej. W sieci LPWAN stosowane jest symetryczne szyfrowanie AES128, które wykorzystuje dwa klucze: klucz sesji sieciowej (NwkSKey) i klucz sesji aplikacji (AppSKey) do szyfrowania i deszyfrowania. Stała długość szyfrowanych wiadomości może ułatwić atakującym złamanie klucza. Urządzenia w sieci LPWAN są także podatne na zakłócenia, naruszenia bezpieczeństwa urządzeń oraz kluczy sieciowych, a także ataki typu Replay. Ponadto, moduły radiowe zintegrowane z mikrokontrolerami w urządzeniach LPWAN często nie obsługują algorytmów kryptograficznych, co może utrudnić weryfikację autentyczności przesyłanych komunikatów przez urządzenie końcowe [1], [8].

2. ARCHITEKTURA SIECI LORAWAN

Zarządzanie pasmem w LoRaWAN realizowane jest poprzez podział 2 MHz dostępnego pasma na trzy pseudolosowo wybierane kanały w zakresie łącza w górę i dół oraz jeden stały kanał w łączu w dół. Aby zminimalizować ryzyko zakłóceń międzykanałowych, stosuje się interwały ochronne. Skalowalność systemu jest ograniczona z powodu dostępności tylko 2 MHz nieobjętego licencją pasma. Ponadto, intensywne wykorzystanie tego nieobjętego licencją pasma przez urządzenia LoRa prowadzi do rywalizacji między użytkownikami. W celu zarządzania obciążeniem sieci i obsługą licznej grupy urządzeń, w sieci LoRaWAN stosuje się ograniczenia cyklu pracy transmisji danych, co ogranicza częstotliwość nadawania pakietów. W zależności od długości trwania pakietu w sieci i mocy transmisyjnej, stosowane są trzy różne limity cyklu pracy: 1%, 0,1% oraz 0,01%. Limity te określają czas między kolejnymi transmisjami, co pozwala na efektywniejsze zarządzanie dostępnym spektrum. Poniższy wzór opisuje ilość czasu pomiędzy kolejnymi transmisjami [2]:

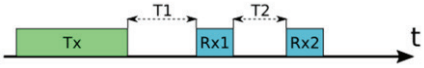
$$T_{next} = \frac{T_{packet}}{DutyCycle} - T_{packet} \quad (1)$$

gdzie:
- T_{next} – czas, który musi upłynąć przed następną transmisją w systemie LoRaWAN, który działa na zasadzie ograniczenia cyklu pracy;
- T_{packet} – czas trwania jednej transmisji pakietu;
- $DutyCycle$ – stosunek czasu, w którym system jest aktywny (wysyła dane), do całkowitego czasu [2].

2.1. Podział urządzeń końcowych w sieci

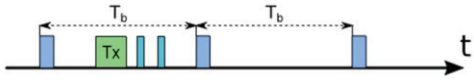
W LoRaWAN urządzenia mogą pracować w 3 różnych klasach. Klasy urządzeń w sieci LoRaWAN definiują różne poziomy funkcjonalności i efektywności energetycznej, dostosowane do specyficznych zastosowań i wymagań aplikacji IoT.

Urządzenia klasy A zapewniają dwukierunkową komunikację, tylko w określonych oknach czasowych (Rys. 1). Po każdej wysłanej wiadomości, urządzenie odczekuje czas zdefiniowany jako dwa krótkie okna czasowe. Po określonym czasie możliwość odbioru wiadomości zostaje zablokowana do czasu ponownego wystąpienia okien, urządzenie przechodzi w tryb uśpienia, co przyczynia się do oszczędności energii. Urządzenia klasy A są najbardziej wydajne energetycznie i najlepiej sprawdzają się w rozwiązaniach o niskiej częstotliwości komunikacji, takich jak systemy pomiarowe wykorzystujące czujniki temperatury lub wilgotności [7].



Rysunek 1. *Transmisja w klasie A.*

Urządzenia klasy B, działają podobnie jak urządzenia klasy A, z tym, że okna czasowe przeznaczone do komunikacji są otwierane w regularnych odstępach czasu. Czas jest synchronizowany z siecią za pomocą sygnałów "beacon" wysyłanych przez bramę, która pełni funkcję centralnego punktu komunikacyjnego, zarządzającego przesyłaniem danych między urządzeniami końcowymi a siecią. Taka funkcjonalność zwiększa zużycie energii i wymaga dokładnej synchronizacji między wszystkimi elementami systemu [7].



Rysunek 2. *Transmisja w klasie B.*

Urządzenia klasy C pozostają w ciągłym stanie nasłuchu i są zdolne do odbierania danych w dowolnym czasie, kiedy nie odbywa się transmisja danych. Ta kategoria urządzeń charakteryzuje się wysokim zużyciem energii, jednak zapewnia niemal błyskawiczną komunikację pomiędzy urządzeniami a bramą. Urządzenia klasy C są szczególnie odpowiednie dla aplikacji wymagających częstych lub natychmiastowych przekazów danych, jak na przykład systemy sterowania oświetleniem ulicznym [7].

Rysunek 3. *Transmisja w klasie C.*

2.2. Aktywacja urządzenia końcowego

Aktywacja urządzenia końcowego w sieci LoRaWAN umożliwia integrację i funkcjonowanie urządzenia w sieci. Istnieją dwie główne metody aktywacji: Over the Air Activation (OTAA) i Activation by Personalization (ABP). Metoda OTAA rozpoczyna się podczas instalacji urządzenia, po resecie lub utracie informacji sesji. Proces ten wymaga wcześniejszej personalizacji urządzenia przy użyciu unikalnych identyfikatorów urządzeń końcowych (DevEUI) i identyfikatorów aplikacji (AppEUI) oraz klucza aplikacji (AppKey) dostarczonego przez właściciela aplikacji. W ramach OTAA, urządzenie końcowe inicjuje połączenie przez wysłanie wiadomości Join-Request do serwera sieciowego, który w odpowiedzi wysyła wiadomość Join-Accept, zawierającą kluczowe informacje do generowania kluczy sesji, jeśli uzyska autoryzację do dołączenia. Ta metoda wspiera roaming między różnymi implementacjami systemu LoRaWAN. Metoda ABP jest alternatywą dla metody OTAA, eliminuje potrzebę wymiany wiadomości Join-Request/Join-Accept. W metodzie ABP, adresy urządzeń końcowych (DevAddr) oraz klucze NwkSKey i AppSKey są bezpośrednio programowane w urządzeniu, co umożliwia szybsze włączanie się urządzenia do sieci i rozpoczęcie bezpiecznej komunikacji. Metoda ABP nie umożliwia roamingu sieciowego i potencjalnie naraża na ryzyko naruszenia bezpieczeństwa. Jeśli klucze zostaną ujawnione lub przechwycone, może to prowadzić do możliwości ich odtworzenia przez osoby nieuprawnione w innych urządzeniach w sieci [10].

3. BADANIA I OCENA PODATNOŚCI NA ATAKI

Do zaimplementowania sieci LoRaWAN, w której możliwe jest przeprowadzenie symulacji ataków wykorzystano bramę Lorank8, platformę Arduino Uno Rev 3 oraz nadajnik RFM95W, serwer sieciowy LORIOT, oprogramowanie Whireshark, oprogramowanie Chirpstack (lokalny serwer LoRa), oprogramowania do przekazywania pakietów zainstalowanego w bramce, odbiornik RTL-SDR.

3.1. Symulacja ataków

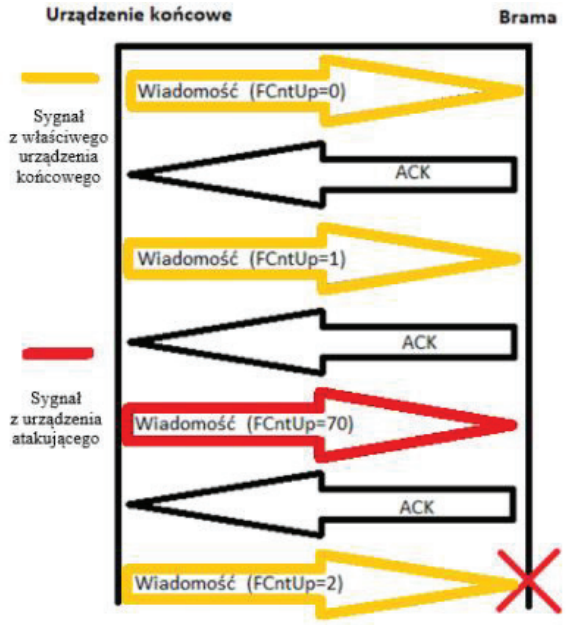
Przeprowadzone eksperymenty miały na celu zidentyfikowanie słabości systemu oraz ocenę skuteczności potencjalnych ataków i ich wpływ na bezpieczeństwo komunikacji w sieci LoRaWAN.

Atak odmowy dostępu (DoS) w trybie ABP (Activation By Personalization)

Atak ten miał na celu zablokowanie komunikacji między urządzeniem końcowym, a bramą sieci LoRaWAN. Atakujący podszywał się pod prawdziwe urządzenie, przechwytując i modyfikując jego ostatni pakiet danych, który odpowiadał wartości licznika wiadomości (FCntUP). Rysunek 4 ilustruje schemat tego ataku. Komunikacja między urządzeniem końcowym a bramą przebiega zgodnie z ustalonym protokołem. Urządzenie końcowe wysyła pierwszą wiadomość do bramy z wartością licznika wiadomości FCntUp=0. Brama odbiera wiadomość, weryfikuje poprawność licznika wiadomości i wysyła potwierdzenie odbioru (ACK). Następnie urządzenie końcowe wysyła kolejną wiadomość z wartością licznika FCntUp=1, którą brama również odbiera, weryfikuje i wysyła potwierdzenie odbioru (ACK). Atakujący przechwytuje kolejną wiadomość wysyłaną przez urządzenie końcowe. Zamiast wysłać ją dalej, modyfikuje wartość licznika wiadomości na FCntUp=70, co jest znacznie wyższą wartością niż oczekiwana przez bramę. Następnie atakujący wysyła zmodyfikowaną wiadomość do bramy. Brama odbiera wiadomość z licznikiem FCntUp=70, weryfikuje ją i traktuje jako poprawną, ponieważ licznik

wiadomości jest wyższy od wcześniej zapisanego stanu, a następnie wysyła potwierdzenie odbioru (ACK) dla wiadomości z licznikiem FCntUp=70. Skutkiem takiej manipulacji jest sytuacja, w której urządzenie końcowe, nieświadome ataku, wysyła kolejną wiadomość z licznikiem FCntUp=2. Brama, która zapisała stan licznika wiadomości jako 70, odrzuca wszystkie wiadomości z niższymi wartościami licznika, w tym wiadomość z FCntUp=2. W efekcie, urządzenie końcowe nie otrzymuje potwierdzenia odbioru (ACK) od bramy, a komunikacja między urządzeniem końcowym a bramą zostaje zakłócona.

Rysunek 5 przedstawia rzeczywiste dane przechwyconych ramek, ilustrując zmiany w wartościach licznika FCntUp oraz czas odbioru poszczególnych pakietów. Dane te demonstrują, jak manipulacja licznikiem wiadomości przez atakującego prowadzi do zablokowania komunikacji. Przechwycone ramki danych pokazują, że wiadomości z licznikiem FCntUp=0, 1, 2 i 3 były wysyłane w odstępach czasowych, jednak po manipulacji, brama przyjęła wiadomość z licznikiem FCntUp=20. Kolejna wiadomość z licznikiem FCntUp=21 została prawidłowo odebrana przez bramę dopiero po przekroczeniu wartości licznika fałszywej ramki.



Rysunek 4. *Przykład ataku powtórzeniowego dla trybu ABP.*

FCntUp	Time	Port	Data
21	a few seconds ago	1	48 65 6c
20	21 minutes ago	1	48 65 6c
3	21 minutes ago	1	48 65 6c
2	22 minutes ago	1	48 65 6c
1	23 minutes ago	1	48 65 6c
0	24 minutes ago	1	48 65 6c

Rysunek 5. *Odebrane ramki po wykonaniu ataku powtórki w trybie ABP*

Przeprowadzony atak wykazał, że manipulacja licznikiem wiadomości może skutecznie zablokować komunikację między urządzeniem końcowym a bramą, co może prowadzić do poważnych zakłóceń w działaniu sieci, zwłaszcza w aplikacjach wymagających niezawodnej transmisji danych. Zabezpieczenie takich systemów przed tego typu atakami jest kluczowe dla utrzymania ich integralności i bezpieczeństwa.

Eavesdropping - nasłuchiwanie ruchu sieciowego między warstwą serwera a bramą

Ten atak był skoncentrowany na przechwyceniu transmisji danych między serwerem a bramą w celu uzyskania dostępu do wrażliwych danych, takich jak fizyczna lokalizacja bramy. Po przechwyceniu ramki danych, atakujący mógł uzyskać informacje o lokalizacji bramy oraz innych krytycznych parametrach transmisji, takie jak moc sygnału (RSSI), identyfikator bramy i stosunek sygnału do szumu (SNR). Na rysunku 6 przedstawiono przechwyconą ramkę łącza w górę. Przypadek ten dotyczy naruszenia bezpieczeństwa komunikacji w sieci IP pomiędzy serwerem a bramą. Jest to atak na poziomie sieciowym, który można skutecznie utrudnić stosując zwykłe mechanizmy zabezpieczające, takie jak VPN. VPN szyfruje całą komunikację między serwerem a bramą, co znacznie utrudnia przechwycenie i analizę danych przez atakującego.

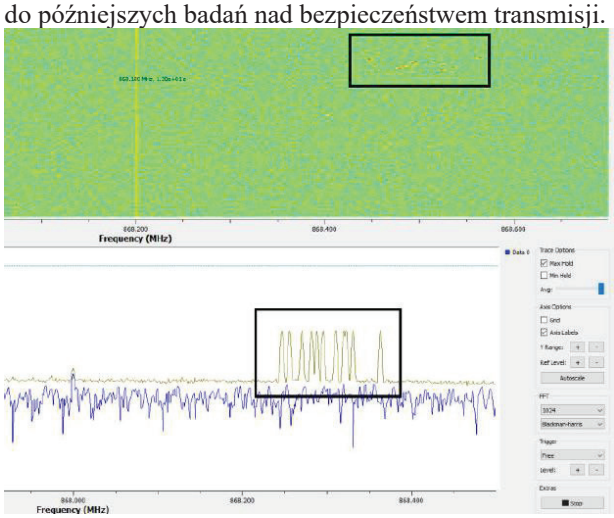
Rysunek 6. *Przechwycona ramka łącza w górę*

Przechwycenie informacji lokalizacyjnych umożliwiło atakującym określenie fizycznego położenia bramy. Znajomość lokalizacji infrastruktury sieciowej może prowadzić do fizycznych ataków na urządzenia, sabotażu oraz naruszeń infrastruktury krytycznej. Dane dotyczące częstotliwości, szerokości pasma i współczynnika rozpraszania dostarczają informacji do analizy wydajności sieci, umożliwiając atakującym przeprowadzanie zakłóceń sygnałów (jamming) lub ataków typu DoS. Znajomość identyfikatorów bram oraz parametrów technicznych może

pozwolić na podszywanie się pod legalne urządzenia lub bramy. Przechwycone ramki mogą być także użyte do ataków typu replay, gdzie stare komunikaty są retransmitowane w celu wywołania niepożądanych operacji w sieci. Przechwycenie ramki transmisji danych w sieci LoRaWAN może poważnie wpłynąć na bezpieczeństwo i integralność całego systemu. Wykorzystanie tych danych przez atakujących może prowadzić do wielu możliwości ataków, od zakłócania sygnału po naruszenia prywatności i bezpieczeństwa infrastruktury.

Eavesdropping - nasłuchiwanie ruchu sieciowego między urządzeniem końcowym a bramą

Do realizacji tego ataku użyto zestawu narzędzi składającego się z oprogramowania GNURadio oraz urządzenia odbierającego RTL-SDR. Wykorzystując te narzędzia, atakujący był w stanie przechwycić ramki danych emitowane przez urządzenia końcowe. Dzięki temu, możliwe było monitorowanie komunikacji i potencjalne zdobywanie wrażliwych informacji przesyłanych przez urządzenia końcowe. Rysunek 7 przedstawia analizę jednej z przechwyconych ramek danych. Wykorzystując przechwycone informacje, atakujący mógłby potencjalnie zagłuszyć daną częstotliwość, co uniemożliwiłoby prawidłową komunikację między urządzeniem końcowym, a bramą. Takie zagłuszenie może skutkować nie tylko przerwaniem komunikacji, ale również potencjalną utratą ważnych danych. Ten przypadek dotyczył zarejestrowania widma zawierającego fragment transmisji LoRa. Jest to bardziej pasywne nasłuchiwanie niż aktywny atak. Warto zauważyć, że mimo iż samo zarejestrowanie transmisji nie jest atakiem, to może stanowić to podstawę do późniejszych badań nad bezpieczeństwem transmisji.



Rysunek 7. *Przechwycone pakiety danych za pomocą oprogramowania GNURadio oraz urządzenia RTL-SDR.*

4. PODSUMOWANIE

Sieci LPWAN dla zabezpieczenia komunikacji wykorzystują różne metody uwierzytelnienia i szyfrowania, w tym standard AES. Mimo to, są one podatne na słabości. Ataki typu "Eavesdropping", "Replay" i "Man-in-the-Middle" stanowią poważne zagrożenia poufności, integralności lub dostępności danych. Dużym wyzwaniem jest utrzymanie równowagi między zużyciem energii, a poziomem bezpieczeństwa urządzeń systemu LPWAN, zwłaszcza w urządzeniach zasilanych bateryjnie o ograniczonych możliwościach obliczeniowych. Choć sieci LPWAN są stosunkowo bezpieczne, wymagają odpowiedniego zarządzania i zabezpieczeń. Prostsza w implementacji metoda aktywacji ABP w systemie LoRaWAN, która eliminuje potrzebę dołączania i wymiany kluczy, ułatwia wdrożenie. Kolejnym etapem w realizacji badań będzie przeprowadzenie symulacji ataków, które zostały opisane w artykule, z zastosowaniem metody aktywacji OTAA (Over the Air Activation). W ramach tego etapu zostaną zrealizowane symulacje ataku odmowy dostępu (DoS), nasłuchiwania ruchu sieciowego między warstwą serwera a bramą oraz nasłuchiwania ruchu sieciowego między urządzeniem końcowym a bramą, przy użyciu metody OTAA. Ponadto, badania będą koncentrować się na rozwoju i ocenie zaawansowanych protokołów zarządzania kluczami dla sieci LPWAN, w celu oceny ich skuteczności w kontekście zwiększonej intensywności ataków.

Praca finansowana ze środków przyznanych przez Ministerstwo Nauki i Szkolnictwa Wyższego w ramach zadania nr 0311/SBAD/0757.

LITERATURA

[1] Adefemi Alimi, K.O.; Ouahada, K.; Abu Mahfouz, A.M.; Rimer, S. 2020. “A Survey on the Security of Low Power Wide Area Networks: Threats, Challenges, and Potential Solutions”. Sensors, 20, 5800. https://doi.org/10.3390/s20205800

[2] Chaudhari, Bharat S.; Zennaro, Marco. 2020. “LPWAN Technologies for IoT and M2M Applications”. Academic Press

[3] Eef van Es, Harald Vranken, and Arjen Hommersom. 2018. “Denial of Service Attacks on LoRaWAN”. Proceedings of the 13th International Conference on Availability, Reliability and Security (ARES 2018). Association for Computing Machinery, New York, NY, USA, Article 17, 1 6. https://doi.org/10.1145/3230833.3232804

[4] Ingham, M., Marchang, J., & Bhowmik, D. 2020. Iot security vulnerabilities and predictive signal jamming attack analysis in lorawan”. IET Information Security, 14(4), 368 379. https://doi.org/10.1049/iet ifs.2019.0447

[5] K. L. Tsai, F. Y. Leu, L. L. Hung and C. Y. Ko, 2020. "Secure Session Key Generation Method for LoRaWAN Servers," in IEEE Access, vol. 8, pp. 54631 54640. doi: 10.1109/ACCESS.2020.2978100.

[6] Nowak, M.; Derbis, P.; Kurowski, K.; Różycki, R.; Waligóra, G. 2021. “LPWAN Networks for Energy Meters Reading and Monitoring Power Supply Network in Intelligent Buildings”. Energies, 14, 7924. https://doi.org/10.3390/en14237924

[7] Nowak, M.; Różycki, R.; Waligóra, G.; Szewczyk, J.; Sobiesierski, A.; Sot, G. 2023. „Data Processing with Predictions in LoRaWAN”. Energies, 16, 411. https://doi.org/10.3390/en16010411

[8] Pathak, G.; Gutierrez, J.; Rehman, S.U. 2020. “Security in Low Powered Wide Area Networks: Opportunities for Software Defined Network Supported Solutions”. Electronics, 9, 1195. https://doi.org/10.3390/electronics9081195

[9] Szewczyk, J.; Nowak, M.; Remlein, P.; Głowacka, A. 2022. “LoRaWAN communication implementation platforms”, International Journal of Electronics and Telecommunications, vol. 68, no. 4, s. 841 854

[10] The ThingNetwork website. Available online: https://www.thethingsnetwork.org/docs/lorawan/end-device-activation/ (dostęp online: 13.05.2024).

MANIANA: URZĄDZENIE IOT DO TESTOWANIA QOS I QOE APLIKACJI W SIECI DOMOWEJ

MANIANA: IOT DEVICE IN TESTING QOS AND QOE APPLICATIONS IN HOME NETWORK CONDITIONS

Dominika Wanat[1]; Dawid Juszka[2]; Mikołaj Leszczuk[3]; Lucjan Janowski[4];

[1] AGH Akademia Górniczo-Hutnicza, Kraków, dominika.wanat@agh.edu.pl
[2] AGH Akademia Górniczo-Hutnicza, Kraków, dawid.juszka@agh.edu.pl
[3] AGH Akademia Górniczo-Hutnicza, Kraków, mikolaj.leszczuk@agh.edu.pl
[4] AGH Akademia Górniczo-Hutnicza, Kraków, lucjan.janowski@agh.edu.pl

DOI: 10.15199/59.2024.4.35

Streszczenie: Metryki QoS (Quality of Service) w testach aplikacji internetowych dostarczają wielu informacji o zachowaniu danej aplikacji w warunkach niestabilnego połączenia sieciowego. Pomiar ten jest jednak często dokonywany w hermetycznych warunkach laboratoryjnych. Skutkuje to uproszczeniem ich analizy kosztem niepełnego obrazu zachowania i środowiska aplikacji oraz zachowań użytkownika. Aby zaadresować ten problem, opracowano urządzenie nazwane MANIANA (Mobile Appliance for Network Interrupting, Analysis & Notorious Annoyance). Umożliwia ono przeprowadzanie testów aplikacji w sieci domowej użytkownika. Oparte jest ono o platformę Raspberry Pi 4 oraz otwartoźródłowe komponenty, umożliwiające testy aplikacji w sposób bezpieczny, wydajny oraz uniwersalny.

Abstract: QoS (Quality of Service) metrics obtained during network application testing provide lots of information about the application's behaviour in conditions of unstable network connections. However, obtaining these values is done in a hermetic lab environment. That simplifies its analysis and provides an incomplete picture of applications' behaviour, environment and user behaviour. To address this problem, a device called MANIANA (Mobile Appliance for Network Interrupting, Analysis & Notorious Annoyance) was developed. It allows for conducting application tests in a home environment. The device is made based on Raspberry Pi 4 minicomputer and Open Source solutions and allows safe, robust, and universal testing applications.

Słowa kluczowe: Quality of Service (QoS), Quality of Experience (QoE), niezawodność aplikacji, Internet of Things (IoT), zakłócenia sieciowe

Keywords: Quality of Service (QoS), Quality of Experience (QoE), applications reliability, Internet of Things (IoT), network distortions

1. WSTĘP

Pomiar podstawowych parametrów łącza, takich jak: przepustowość, strata pakietów, jitter oraz opóźnienie jest kluczowy dla zrozumienia zachowania oraz optymalizacji współczesnych aplikacji internetowych oraz usług. Szczególnie wrażliwe na niestabilne połączenie internetowe są aplikacje multimedialne tj. serwisy streamingowe, aplikacje do wideokonferencji, gry online. W zależności od typu aplikacji stosuje się różne metody rozwiązywania problemów z łączem na warstwie aplikacji. Na przykład serwisy streamingowe implementują protokoły strumieniowania adaptacyjnego, takie jak Dynamic Application Adaptive Streaming over HTTP (DASH), HTTP Dynamic Streaming (HDS) oraz HTTP Live Streaming (HLS). Aplikacje korzystające z interfejsu WebRTC najczęściej wykorzystują algorytmy takie jak Congestion Control, Network-Assisted Dynamic Adaptation (NADA) czy Self-Clocked Rate Adaptation for Multimedia (SCReAM)[1].

Testowanie tego typu aplikacji oraz wykorzystywanych przez nich optymalizacji może odbywać się w warunkach laboratoryjnych poprzez symulację zarówno łącza sieciowego, jak i scenariuszy korzystania z usługi. Rozwiązanie to umożliwia jednak osiągnięcie wyników, które mogą nie mieć odzwierciedlenia w rzeczywistych warunkach: zarówno technicznych, takich jak posiadane przez użytkownika łącze sieciowe, jak i związanych z QoE (Quality of Experience), takich jak zachowanie i sposób korzystania z aplikacji przez użytkownika [7]. Aby rozwiązać ten problem, opracowano kompletny system IoT korzystający z zasobów sieci domowej użytkownika oparty o platformę Raspberry Pi 4B działające w trybie Access Point (AP).

Minikomputery te są chętnie wykorzystywane w wielu dziedzinach teleinformatyki. Mają one zastosowanie w edukacji, na przykład do nauki trasowania w sieciach komputerowych [4] czy przetwarzania sygnałów [5]. Popularność platformy sprawia, że autorzy podejmują się także algorytmicznych rozważań dotyczących trybu AP oferowanego przez wbudowaną kartę sieciową [2]. Równie chętnie powstają praktyczne implementacje systemów dla urządzeń działających w modelu mgły obliczeniowej (ang. *fog computing*) [3], a także bezpiecznego udostępniania połączenia WAN urządzeniom lokalnym[6]. Prezentowany autorski system, składa się z urządzeń MANIANA oraz infrastruktury towarzyszącej. Wykorzystuje on podstawowe możliwości karty sieciowej WLAN oraz trybu AP rozszerzając je o możliwość dodatkowego sterowania i monitorowania urządzeń MANIANA.

2. ANALIZA WYMAGAŃ

Możliwość przeprowadzania i kontrolowania testów w rzeczywistych warunkach zapewnią urządzenia MANIANA połączone z domowymi sieciami użytkowników. Pozwolą one na wprowadzanie zakłóceń i zbieranie informacji na temat zachowań zarówno aplikacji, jak i użytkowników z nich korzystających. Na etapie prototypowania systemu stanowiącego rozwiązanie problemu ustalono następujące wymagania gwarantujące uzyskiwanie wiarygodnych wyników:

1. niezawodność – urządzenie umożliwia badanie codziennych zachowań i schematów użycia aplikacji, a więc jest przystosowane do pracy w trybie ciągłym i przewidywalnie reaguje na przerwy w zasilaniu, brak połączenia internetowego oraz niewłaściwą konfigurację. Dane zbierane przy pomocy urządzeń powinny być buforowane w przypadku braku możliwości komunikacji z centralnym serwerem i wysyłane do niego niezwłocznie po przywróceniu połączenia.

2. skalowalność – pojedyncze urządzenie systemu może zostać rozbudowane z użyciem powszechnie znanych i dostępnych interfejsów (skalowanie pionowe, ang. *scale-up*), a system umożliwia łatwe dodanie dowolnej liczby urządzeń i zwiększenie liczby aktywnych użytkowników (skalowanie poziome, *scale-out*).

3. niski koszt pojedynczego urządzenia – pojedynczy element systemu jest możliwie tani i składa z popularnych, łatwo wymiennych komponentów, aby niskim kosztem umożliwić testowanie niezawodności aplikacji jak największej grupie użytkowników.

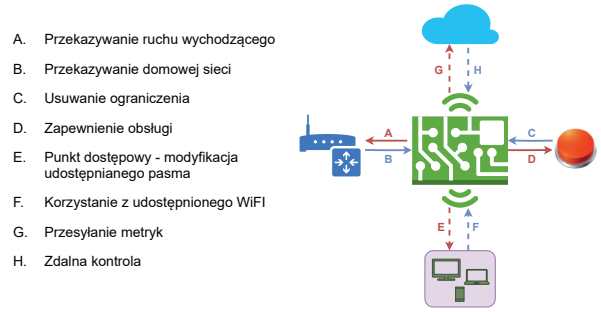
4. powtarzalność – urządzenia są zbudowane w identycznej konfiguracji sprzętowej oraz konfiguracji oprogramowania.

5. bezpieczeństwo – proces przesyłania danych odbywa się za pomocą szyfrowanych kanałów komunikacji, a pozyskane od użytkownika są przechowywane w sposób uniemożliwiający jego identyfikację.

6. prostota konfiguracji – urządzenie jest gotowe do pracy natychmiast po wykonaniu prostej procedury instalacyjnej.

3. IMPLEMENTACJA

Na podstawie przyjętych założeń zbudowano zestaw urządzeń MANIANA, oparty o minikomputery Raspberry Pi 4B. Urządzenia podłączone przewodowo do domowego routera sieciowego działają w trybie Access-Point (AP). Rozgłaszają one sieć bezprzewodową zmodyfikowaną o dowolnie konfigurowane scenariusze testowe oraz możliwość manualnego wyzwolenia akcji poprzez wciśnięcie przycisku. Uniwersalna implementacja daje szerokie możliwości tworzenia scenariuszy testowych dla różnego rodzaju aplikacji. Metryki z urządzenia są zbierane na centralnym serwerze bazy danych przystosowanych do przechowywania szeregów czasowych (ang. time-series database, TSDB), skąd mogą być agregowane, wizualizowane oraz przetwarzane. Na rysunku 1 przedstawiono schemat komunikacji pomiędzy komponentami systemu, które szczegółowo omówiono w sekcjach 3.1–3.4.



Rys. 1. Diagram akcji między urządzeniem MANIANA a pozostałymi komponentami systemu. Przerywane strzałki symbolizują komunikację bezprzewodową.

3.1 Warstwa sprzętowa

Minikomputer Raspberry Pi jest urządzeniem opartym o układ BCM2711 typu System on a chip (SoC), zintegrowanym z 64-bitowym czterordzeniowym procesorem Cortex-A72 o taktowaniu 1.8 GHz oraz kartą WLAN obsługującą WiFi w standardzie IEEE 802.11ac oraz Gigabit Ethernet. Jest to konfiguracja sprzętowa zbliżona do konsumenckich urządzeń sieciowych, otrzymywanych od operatorów telekomunikacyjnych. Wbudowana karta WLAN posiada tryb Access Point (AP). Konstrukcyjnie zadbano także o mechanizmy niwelujące obniżenie taktowania procesora przy wykryciu przegrzania CPU przez system operacyjny (ang. *CPU throttling*). Jest to dość często spotykaną wada tego układu, która mogłaby wpłynąć niekorzystnie na przeprowadzenie testów.

Sterowanie testem umożliwia fizyczny podświetlany przycisk podłączony do urządzenia poprzez piny GPIO, który można swobodnie programować korzystając z rozwiązań udostępnianych przez system operacyjny. W zaimplementowanym scenariuszu przycisk przywraca sieć do stanu początkowego, zdejmując ograniczenia pasma w rozgłaszanej sieci WiFi.

3.2 System operacyjny oraz podstawowy software

Warstwy: programowa oraz systemu operacyjnego są kluczowymi elementami zaprojektowanego systemu. Aby spełnić wymagania opisane w punkcie 2, postawiono na uniwersalnie stosowane rozwiązania, umożliwiające swobodną integrację i rozbudowę. Sercem każdego urządzenia MANIANA jest OpenWrt 22.03.2 – system operacyjny dla urządzeń wbudowanych oparty na jądrze Linux w wersji 5.10.146. Taki wybór systemu umożliwił uniwersalne opracowanie komunikacji z urządzeniami peryferyjnymi, takimi jak zastosowany przycisk, poprzez moduły jądra systemu Linux.

Konfigurację sieciową urządzania stanowi most sieciowy (bridge) interfejsu LAN (podłączonym jako klient do domowego routera) z interfejsem typu WLAN skonfigurowanym w trybie AP z innym, niż domowe, zdefiniowanym SSID oraz hasłem. Było to możliwe dzięki zastosowaniu oprogramowania `hostapd` i pozwoliło na separację środowiska testowego od domowej sieci WiFi. Dzięki temu testowanie zachowania aplikacji może odbyć się wyłącznie na urządzeniu, które świadomie przyłączono do sieci testowej.

Za uruchomienie scenariusza testowego odpowiada skrypt `limiter.sh`, wykorzystujący oprogramowanie `tc` do sterowania parametrami rozgłaszanej sieci, który co określony czas (domyślnie dwadzieścia minut) limituje dostępne pasmo o zadane wartości (domyślnie o 2500 Kib/s, 1000 Kib/s, 600 Kib/s, 300 Kib/s oraz 50 Kib/s). Oprogramowanie to uruchamia się przy starcie urządzenia, wykorzystując mechanizm flock (zapewniający uruchomienie dokładnie jednej instancji). Zadbano także o

asynchroniczne sterowanie wykonaniem programu poprzez obsługę sygnałów `SIGKILL` oraz `SIGUSR1`, który umożliwia czasowe (również dwudziestominutowe) zniesienie nałożonych ograniczeń.

Za obsługę przycisku odpowiada prekompilowany dla Raspberry Pi 4B moduł `gpio-keys`, ładowany dynamicznie przy starcie urządzenia poprzez system, używający automatycznego nakładania (ang. *overlay*) kodu modułów poprzez drzewa urządzeń (Device Tree)[1].

Utworzona konfiguracja dołącza moduł `gpio-key` przy starcie i aktywuje urządzenie wejściowe podłączonego do pinu wraz z jego specyfikacją. Następnie definiuje przypisanie akcji wciśnięcia tego przycisku na kod zdarzenia 85[2], co odpowiada naciśnięciu przycisku klawiatury odpowiadającemu za konwersję w japońskim alfabecie `KEY_ZENKAKUHANKAKU`. Powodem wyboru takiego znaku, było małe prawdopodobieństwo nadpisania go w systemie operacyjnym poprzez konieczność przyporządkowania go do jakiejkolwiek domyślnej kombinacji klawiszy.

Powiązanie asynchronicznego sygnału ze skryptu `limiter.sh` oraz akcji naciśnięcia przycisku umożliwiło oprogramowanie `triggerhappy`, które pozwala na zaprogramowanie wykonania dowolnej komendy poprzez wciśnięcie pojedynczego klawisza lub ich kombinacji. Aby poprawnie wykonać scenariusz testowy, przypisano zdefiniowanemu klawiszowi `KEY_ZENKAKUHANKAKU` możliwość wysłania sygnału `SIGUSR1` do procesu wykonującego skrypt limitujący.

3.3 Monitorowanie rozwiązania

Do obsługi metryk z urządzeń wykorzystano oprogramowanie `collectd`. Umożliwia ono zbieranie predefiniowanych metryk, bądź załączenie i napisanie własnego kolektora w językach: C, Python lub Perl. `Collectd` pozwala także na wybór dowolnego formatu wyjściowego, ma także możliwość przesłania metryk do centralnej instancji `collectd` z opcjonalnym szyfrowaniem. Metryki przesyłane są do bazy wraz z wysokorozdzielczościowym znacznikiem czasowym (timestampem). W metrykach nie są zbierane adresy IP użytkowników – urządzenia są identyfikowane poprzez ich nazwę, będącą nazwą hosta – urządzenia MANIANA. Podczas implementacji napotkano jednak na problem braku możliwości szyfrowania metryk w dostarczonym przez OpenWRT pakiecie `collectd`, który rozwiązano poprzez konfigurację Virtual Personal Network (VPN) z urządzeń do centralnego serwera metryk używając Wireguard. Zapewnia on asymetryczne szyfrowanie komunikacji. Jest też także rozwiązaniem lekkim i nie zużywającym zasobów sprzętowych, co jest istotne w przypadku urządzeń wbudowanych. Ostatnim wyzwaniem była możliwość spakowania opisanej konfiguracji oraz jej prostego reprodukowania na wiele urządzeń w taki sposób, aby jedyną ręcznie wykonywaną konfiguracją były klucze prywatne tunelu Wireguard oraz nazwy hostów. Istotne było także, aby oprogramowanie uruchamiało się automatycznie po wgraniu oprogramowania na nośnik danych Raspberry Pi i podłączeniu urządzenia MANIANA do zasilania oraz domowego routera. W tym celu wykorzystano narzędzie OpenWRT Image Builder z przygotowaną listą oprogramowania oraz bazową konfiguracją. Pozwoliło to na otrzymanie spersonalizowanego, spakowanego obrazu karty pamięci dla Raspberry Pi 4B z wszystkimi potrzebnymi komponentami uruchamianymi w czasie startu urządzenia MANIANA.

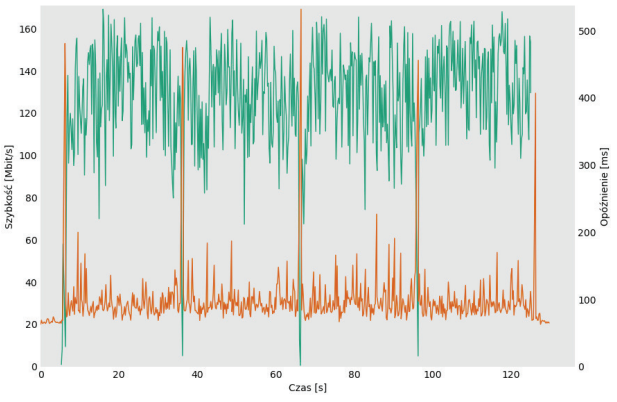
3.4 Część serwerowa

Oprócz konfiguracji samych urządzeń, niezbędna była także infrastruktura umożliwiająca komunikację oraz monitorowanie stanu urządzeń MANIANA zainstalowanych w domach testerów. Aby tego dokonać, skonfigurowane dwie maszyny wirtualne: pierwszą, nazwaną `iotgw`, pełniącą funkcję centralnego serwera metryk oraz serwera VPN, oraz drugą `movinglab` – komunikującą się z serwerem `iotgw`, przyjmującą zaszyfrowaną postać metryk, odpowiadającą za gromadzenie i wyświetlanie wyników. Podział zadań na dwa hosty miał na celu separację zadań w taki sposób, aby maszyna przechowująca oraz wyświetlająca metryki nie miała możliwości kontroli urządzeń IoT. Obydwa hosty komunikują się ze sobą jedynie w celu wysłania i odebrania metryk szyfrowanym kanałem i są odizolowane sieciowo.

4. TESTY ROZWIĄZANIA

Aby sprawdzić, czy zaimplementowane rozwiązanie skutecznie limituje dostępne pasmo w ramach AP urządzenia MANIANA, wykonano serię testów przy pomocy narzędzia *flent*[3]. Scenariusz testowy obejmował jednowątkowe pobranie treści używając protokołu TCP z europejskiego serwera `netperf-eu.bufferbloat.net` na hoście połączonym bezprzewodowo do AP urządzenia MANIANA. Najpierw przetestowano szybkość pobierania z sieci domowej, bez użycia pośrednika w postaci AP. Jak można zobaczyć na Rysunku 2, średnia prędkość wynosiła 135 Mbit/s.

Podczas testu można było zauważyć zwiększone opóźnienia, co powodowane jest przez nadmierne buforowanie pakietów (ang. *bufferbloat*) w domowym routerze testowym. Może to powodować dodatkowe problemy z buforowaniem oraz odtwarzaniem multimediów. Stanowi to dodatkowy czynnik, wpływający na wyniki, w szczególności jeśli są one testami subiektywnymi – opierają się na doświadczeniach użytkownika.



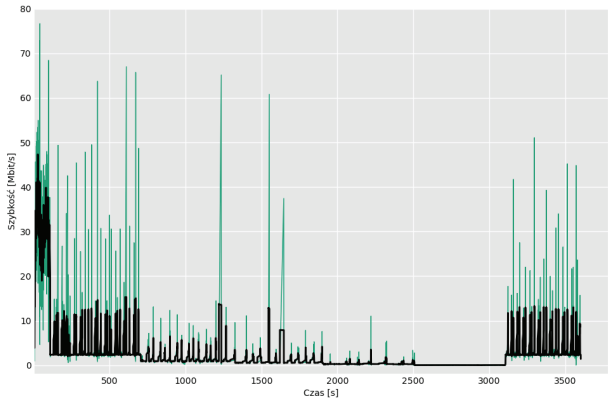
Rys. 2. Prędkość pobierania strumienia TCP z sieci domowej bez urządzenia MANIANA (kolor zielony) wraz z opóźnieniami (kolor pomarańczowy).

Następnie zmodyfikowano skrypt limitujący pasmo (`limiter.sh`) poprzez skrócenie domyślnego czasu trwania pojedynczego limitu o połowę (z 20 minut do 10 minut) oraz uruchomiono jednowątkowe pobranie TCP, a po 10 sekundach, nieprzerwany akcją użytkownika, scenariusz limitowania pasma opisany w sekcji 3.2. Na Rysunku 3 przedstawiono wyniki.

[1] Więcej informacji o mechanizmie overlayów jest dostępna pod adresem: https://github.com/raspberrypi/firmware/blob/master/boot/overlays/README

[2] Wykaz wszystkich dostępnych znaków można znaleźć pod adresem: https://github.com/torvalds/linux/blob/master/include/uapi/linux/input-event-codes.h.

[3] Więcej informacji o narzędziu *flent* znajduje się pod adresem https://flent.org



Rys. 3. Szybkość pobierania strumienia TCP z AP urządzenia MANIANA z zastosowanymi limitami. Wartości uśrednione przedstawione są kolorem czarnym.

Można z nich wywnioskować, że dość wyraźnie widoczny jest zarówno moment rozpoczęcia limitowania pasma w okolicy setnej sekundy, jak i poszczególne progi, przy czym są one zdecydowanie niższe niż te zdefiniowane w skrypcie. Odpowiadają za to w głównej mierze: jakość wbudowanego modułu karty WiFi, odległość od AP, obciążenie procesora, zastosowane dyscypliny kolejkowania.

5. PODSUMOWANIE

W niniejszym artykule zaproponowano sposób tworzenia urządzeń IoT do testów aplikacji wraz z infrastrukturą towarzyszącą. Bezpieczeństwo, niezawodność, powtarzalność, skalowalność i prostota konfiguracji to niewątpliwe atuty zaproponowanego rozwiązania. Gotowe rozwiązanie w formie kompletnego obrazu znajduje się pod adresem: https://github.com/doma2203/maniana-image. Przeprowadzone testy pokazały możliwości limitowania pasma, które można ulepszyć używając dodatkowych fizycznych komponentów (dodatkowych kart WiFi) i ulepszając sposób limitowania pasma o inne dyscypliny kolejkowania. Istotnym problemem wydajności rozwiązania jest identyfikacja dotychczasowych parametrów sieci domowej testera, aby uzyskać pełny obraz problemów, z jakimi do tej pory mógł zmagać się użytkownik, korzystając z multimediów. Należy podjąć próby niwelowania opóźnień spowodowanych przez nadmierne buforowanie pomiędzy buforem routera domowego a AP tworzonym przez urządzenie MANIANA, poprzez zastosowanie SQM (Smart Queue Management) i konfigurowanie go indywidualnie dla każdego testera, w zależności od parametrów łącza.

6. PODZIĘKOWANIA

Badania prowadzące do tych wyników zostały sfinansowane z Norweskiego Mechanizmu Finansowego 2014-2021 w projekcie 2019/34/H/ST6/00599.

Referat został sfinansowany przez Ministerstwo Nauki i Szkolnictwa Wyższego w ramach subwencji dla Wydziału Informatyki, Elektroniki i Telekomunikacji AGH.

LITERATURA

[1] „What is RMCAT congestion control, and how will it affect webrtc?". Advancing WebRTC. https://blog.mozilla.org/webrtc/what-is-rmcat-congestion-control/, Dostęp: 5.05.2024

[2] Al Mamun, Md Selim, et al. 2017. „A channel assignment extension of active access-point configuration algorithm for elastic WLAN system and its implementation using Raspberry Pi." International Journal of Networking and Computing 7.2 (2017): 248-270. http://www.ijnc.org/index.php/ijnc/article/view/153

[3] Hong, Cheol-Ho, Kyungwoon Lee, Minkoo Kang, and Chuck Yoo. 2018. „qCon: QoS-Aware Network Resource Management for Fog Computing" Sensors 18, no. 10: 3444. https://doi.org/10.3390/s18103444.

[4] Laki, Sándor, Stoyanov Radostin, Kis Dávid, Soulé Robert, Vörös Péter, and Zilberman Noa. 2021. „P4Pi: P4 on Raspberry Pi for networking education". SIGCOMM Comput. Commun. Rev. 51, 3 (July 2021), 17–21. https://doi.org/10.1145/3477482.3477486.

[5] J. Marot and S. Bourennane, Raspberry Pi for image processing education,2017 25th European Signal Processing Conference (EUSIPCO), Kos, Greece, 2017, pp. 2364-2366, https://doi.org/10.23919/EUSIPCO.2017.8081633.

[6] Hu, Shuting, Suzuki Hironobu, Kitaguchi Yoshiaki, Ohno Hiroyuki, Sampalli Srinivas. 2019. „Design, Implementation and Performance Measurement of Raspberry Gate in the IoT Field". Proceedings of the 2019 4th International Conference on Cloud Computing and Internet of Things (CCIOT '19). 82–89. https://doi.org/10.1145/3361821.3361827.

[7] Wehner, Nikolas, Seufert Michael, Egger-Lampl Sebastian, Gardlo Bruno, Casas Pedro i Schatz Raimund. 2020. „Scoring High: Analysis and Prediction of Viewer Behavior and Engagement in the Context of 2018 FIFA WC Live Streaming". Proceedings of the 28th ACM International Conference on Multimedia (MM '20), 807–815. https://doi.org/10.1145/3394171.3414016

OCENA WYDAJNOŚCI MECHANIZMU RETRANSMISJI TS011_1.0.0 W SIECIACH LORA WAN.

PERFORMANCE EVALUATION OF RELAY MECHANIZM TS011_1.0.0 IN LORA WAN NETWORKS.

Artur Frankiewicz[1]; Krzysztof Grochla[2]

[1] AIUT sp. z o. o, Gliwice, artur.frankiewicz@aiut.com
[2] Instytut Informatyki Teoretycznej i Stosowanej, Gliwice, kgrochla@iitis.pl

DOI: 10.15199/59.2024.4.36

Streszczenie:
W grudniu 2022r organizacja LoRa Alliance wydała specyfikację mechanizmu retransmisji danych dla sieci działających w protokole LoRa WAN. Propozycja mechanizmu retransmisji podąża w kierunku minimalizacji konieczności rekonfiguracji sieci i maksymalizacji uniwersalności rozwiązania kosztem nieco niższej wydajności energetycznej. Do tej pory publikacja standardu nie uzyskała szerszej debaty w środowisku naukowym. Niniejsza publikacja jest oceną wydajności energetycznej mechanizmu retransmisji jak również ewaluacją możliwości komercyjnego zastosowania urządzeń retransmitujących w sieciach zdalnego opomiarowania zużycia wody.

Abstract**: On December 2022 LoRa Alliance organization published extension TS011_1.0.0 for LoRa WAN specification. Extension covers retransmission aspects in LoRa WAN networks. Proposed algorithm offers very simplified configuration process at the cost of higher energy consumption. This paper presents performance evaluation of proposed relay mechanism with evaluation of possible use case in remote water meter readout systems.**

Słowa kluczowe**: LoRa WAN, LPWAN, sieci bezprzewodowe, retransmisja**

Keywords**: LoRa WAN, LPWAN, Radio relay, retransmission.**

1. WSTĘP

Sieć LoRa WAN (Long Range Wide Area Network) zyskuje coraz większą popularność w wielu obszarach zastosowań, np.: w systemach inteligentnych miast, rolnictwa lub w opiece zdrowotnej. Jej popularność i szybki rozwój zawdzięcza między innymi: dużej skalowalności, łatwości instalacji, pracy w nielicencjonowanym paśmie radiowym, dostępności gotowych komponentów software'owych i sprzętowych do budowy urządzeń i systemów.

Sieć LoRa WAN zalicza się do klasy sieci LPWAN. Odmiennie niż większość sieci w tej grupie, LoRa WAN działa w technologii rozproszonego widma z modulacją CSS (Chirp Spread Spectrum), dzięki której jest częściowo odporna na zakłócenia wzajemne jak również na zajęcie pasma przez inne nadajniki wykorzystujące to samo otwarte pasmo.

Rozwój Internetu Rzeczy (Internet of Things) w ujęciu światowym szacuje się na około 35% rocznie. Niniejsza analiza skupia się głównie na zastosowaniach w obszarze Smart City, stanowiącym globalnie około 20% rynku IoT. Tak szybki rozwój technologii Smart City budzi duże zainteresowanie i uwagę również ośrodków naukowych.

Pomimo publikacji standardu TS011_1.0.0 z końcem 2022r, ocena jego wydajności nie uzyskała debaty w środowisku naukowym. Na stronie organizacji LoRa Alliance, która zrzesza producentów sprzętu dostępna jest usługa Marketplace pozwalająca na prezentację oferowanych urządzeń. Do tej pory na stronie nie pojawiło się żadne komercyjne urządzenie realizujące funkcje retransmitera. Fakt ten skłania do głębszej analizy możliwości zastosowania retransmitera w realnych wdrożeniach komercyjnych w obszarze urządzeń typu Smart City.

Rozszerzenie TS011 dedykowane jest urządzeniom bateryjnym działającym w klasie A. Zatem aspekt energetyczny samego retransmitera jest jednym z głównych elementów analizy oceny możliwości jego praktycznego zastosowania.

Mechanizm retransmisji w sieciach bezprzewodowych analizowany był w wielu publikacjach. Jeden z najbardziej znanych i powszechnie stosowanych to mechanizm opisany w normie wMbus w piątym jej rozszerzeniu EN13757-7 [5]

Rozszerzenie TS011 uzyskało wzmianki w zaledwie 5 publikacjach naukowych od czasu jego publikacji z końcem 2022r. Są to publikacje: [1][2][3][7][8]. Publikacje skupiały się głównie na wyborze miejsc instalacji retransmiterów, topologii sieci, schematach retransmisji, wpływu retransmisji na sieć. Jednak żadna z tych publikacji nie podejmuje analizy energetycznej sieci.

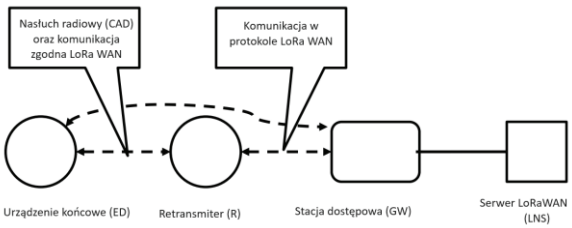
2. Definicja problemu

2.1. Motywacja

Niniejsza analiza skupia się na zastosowaniu sieci LoRa WAN w systemach inteligentnych miast (Smart Cities), gdzie sieć radiowa służy do przekazywania danych pomiarowych w systemie zdalnego odczytu liczników wody. Sieci takie instalowane są w obrębie miasta lub gminy a w ich zasięgu znajduje się od paru tysięcy do paruset tysięcy liczników. W wielu wdrożeniach systemu odczytowego po instalacji sieci liczników oraz sieci stacji dostępowych (Gateway) istnieje pewna liczba punktów końcowych będących poza zasięgiem sieci. W niektórych przypadkach taka liczba może przekraczać nawet 5% instalacji. Dzieje się tak zwłaszcza w strefie klimatycznej Europy środkowowschodniej gdzie liczniki wody w większości instalowane są w specjalnych studzienkach pod ziemią a w budynkach mieszkalnych często w głębokich piwnicach. Takie umieszczenie liczników wody wynika z konieczności prowadzenia rur poniżej granicy przemarzania gruntu. Warunki te powodują, iż planowanie sieci staje się trudne a sytuacje gdzie pojedyncze punkty odczytowe pozostają poza zasięgiem jakiejkolwiek stacji bazowej są częste.

W przypadku gdzie po instalacji urządzenia odczytowego nie ma z nim stabilnej komunikacji radiowej, można albo zagęszczać sieć stacji dostępowych - co jest zadaniem bardzo kosztowym lub próbować zastosować urządzenia z zewnętrzną anteną kablową - co poza samą konieczności podmiany urządzenia końcowego na inny model stwarza konieczność użycia wykwalifikowanego personelu będącego w stanie dobrać antenę zewnętrzną, zabezpieczyć połączenie przewodowe oraz całość okablowania. W każdym przypadku rozwiązanie z anteną zewnętrzną jest bardziej podatne na przypadkowe uszkodzenia, akty wandalizmu i kradzieże.

Rozwiązaniem problemu mogłaby być instalacja dodatkowych węzłów pośrednich sieci mogących służyć jako retransmiter „Relay" (oznaczony jako R) pomiędzy urządzeniem końcowym „End Device" (oznaczony jako ED) a stacją dostępową „Gateway" (oznaczoną jako GW). Poniższa ilustracja przedstawia proponowaną topologie sieci.



Rys. 1. Model retransmisji. Źródło TS011.

Warunkiem koniecznym stosowalności retransmitera w zastosowaniach Smart City jest jego autonomiczna praca zasilana wyłącznie bateryjnie. Uzyskanie podłączania do sieci energetycznej wymaga stosownych pozwoleń przez co jest porównywalne kosztowo z dodaniem dodatkowej stacji dostępowej.

Retransmiter powinien być instalowany stosunkowo blisko od istniejących urządzeń, lecz w takim miejscu, które pozwalać będzie na komunikację ze stacją dostępową przy użyciu stosunkowo niskiego parametru Spreading Factor, co powoduje znacznie mniejszą zajętość eteru jak również ograniczone zużycie energii przez retransmiter. W idealnym zastosowaniu retransmiter powinien być instalowany np. na słupie nad studzienką, w której zainstalowane urządzenia pozostają bez komunikacji ze stacjami bazowymi.

Dalsze obliczenia skupiają się na typowych przypadkach instalacji stosowanych w komercyjnych wdrożeniach w systemach zdalnego opomiarowania liczników zużycia wody.

2.2. Przyjęte założenia obliczeniowe

Poniższe obliczenia wykonane są dla zastosowania układu radiowego firmy Semtech SX1262 przy dodatkowych założeniach, tj.: użyty oscylator posiada funkcjonalność kompensacji temperaturowej TCXO, moc nadajnika wynosi +14dBm (zgodnie z parametrami regionalnymi dla Europy), natomiast mikrokontroler wyposażony jest w układ czasu rzeczywistego RTC i posiada możliwość wybudzenia mikrokontrolera wg zadanego wzorca czasu. W obliczaniach przyjęto, iż podtrzymanie zasilania mikrokontrolera wraz ze stabilizatorem LDO i RCT to średni koszt energetyczny 6uA. Założenie wykonane zostało na bazie realnych pomiarów projektowanych urządzeń w systemach Smart City.

Podczas estymacji przyjęto również, iż czas stabilizowania się częstotliwości na generatorze TCXO wymagany do rozpoczęcia odbioru wynosi 0,15ms a dalsza inicjalizacja odbiornika to 0.50ms zgodnie z danymi firmy Semtech. Do obliczeń przyjęto maksymalny czas pomiędzy kolejnymi skanowaniami CAD Periodicty – 1 sekunda.

Wartośći Rx oraz Tx radio current jest nieznacznie wyższa niż wartości podawane przez producenta (~10%) ze względu na użycie kwarcu kompensowanego temperaturowo oraz użycie dodatkowego przełącznika antenowego poprawiającego właściwości blokujące oraz czułość układu radiowego. Wartości przyjęte w obliczeniach są realnie zmierzonymi wartościami z referencyjnych projektów.

We wszystkich obliczeniach przyjęto, iż źródłem zasilania jest bateria litowa wykonana w technologii (Li-SOCl2) o napięciu 3,6V. Bateria wspomagana jest superkondensatorem co zapobiega spadkom napięcia w czasie największego poboru prądu, czyli podczas nadawania.

MDS to minimalna liczba symboli, po której odbiornik radiowy urządzenia jest w stanie wykryć odbierany wzór preambuły i pozostać w dalszym nasłuchu w przypadku poprawnej detekcji. Wg producenta minimalna wartość to 6 symboli natomiast w schemacie aplikacyjnym rekomendowana wartość to 8 symboli. W tym zakresie nie została wykonana realna ocena wpływu parametru

MDS na czułość lub powtarzalność detekcji preambuły. Do obliczeń została przyjęta rekomendowana wartość.

Tab. 1. Przyjęte parametry obliczeń nasłuchu

Parametr	*Wartość*	*Jednostka*
Demodulation MDS	8	Symbole
Rx radio current	6,5	[mA]
Tx radio current	37	[mA]
CAD Periodicity	1000	[ms]
Xtal start time	0,15	[ms]
Rx Init time	0,5	[ms]

W przypadku konfiguracji urządzenia końcowego został przyjęty model powszechnie stosowany w obszarze zdalnego opomiarowania liczników wody. Długość ramki radiowej (ED self-frame len) zawiera 51 bajtów danych zawierających: 24 różnicowe wartości zużycia godzinowego, adres urządzenia, słowo statusowe oraz informacje o aktualnych i historycznych alarmach. Ramki rozgłoszeniowe typowo wysyłane są w trybie 2 lub 4 dziennie. Do wyliczeń przyjęto 2 ramki dziennie.

Maksymalny dryft zgaru czasu rzeczywistego przyjęty został na poziomie 40ppm oznacza to skrajnie niekorzystny przypadek, gdzie zegar nadajnika przyśpiesza o 20ppm a odbiornika zwalnia o 20ppm lub odwrotnie. Dryft częstotliwości wynika głównie z temperatury pracy i jest naturalnym zjawiskiem generatora kwarcowego. W przypadku urządzeń bateryjnych sprzętowa kompensacja temperaturowa nie jest możliwa ze względów energetycznych. W przypadku pracy urządzenia ED oraz R w tych samych warunkach temperaturowych ich dryft temperaturowy powinien być mniej więcej zgodny a rozsynchronizowanie mniejsze. Jednak w przypadku opomiarowania wodomierzy, z którym to przypadkiem mamy do czynienia, urządzenie końcowe zawsze montowane jest pod ziemią lub w budynku a urządzenie retransmitujące na zewnątrz. Stąd przyjęty został znaczący wpływu dryftu zegarów RTC obu urządzeń. Problem ten i sposoby przeciwdziałania szczegółowo opisane są w [4].

Tab. 2. Przyjęte parametry obliczeń

Parametr	*Wartość*	*Jednostka*
ED self-frame len	51	Bajtów
Liczba ramek / dzień	2	-
Max clock drift	40	[ppm]

3. Mechanizm retransmisji pakietów

Retransmisja pakietów wg TS011_1 odbywa się na zasadzie cyklicznego nasłuchu przez urządzenie (R) na kanale radiowym dedykowanym do retransmisji. Urządzenie wykonuje cykliczne skanowanie kanału nazwane „Carrier Sense Detection" (CAD). Skanowanie wykonywane jest co okres CAD Periodicity wg specyfikacji nie rzadszej niż 1 sekunda. W przypadku wykrycia w czasie skanowania wzoru preambuły, urządzenie zostaje w stanie odbioru aż do zdekodowania kompletnego nagłówka wiadomości. Po tym następuje weryfikacja czy przesyłana wiadomość powinna zostać retransmitowana. W pozytywnym przypadku odbiornik zostaje włączony do zakończenia odbioru kompletnej wiadomości. W przeciwnym przypadku odbiornik jest wyłączany a odebrany nagłówek ignorowany.

Urządzenie końcowe (ED) w przypadku braku łączności z którąkolwiek z bram dostępowych (GW) przechodzi w tryb poszukiwania urządzenia retransmitującego. W tym stanie zaczyna cyklicznie wysyłać wiadomości Relay Join Request. Wiadomości te charakteryzują się długością preambuły dłuższą niż CAD Periodicity, w wyniku czego urządzenie R w oknie CAD jest w stanie odebrać i poprawnie zdekodować wiadomość.

Po nawiązaniu połącznia urządzenie (R) wykonuje rejestrację urządzenia retransmitowanego (ED) w sieci i przechodzi do cyklicznej komunikacji z urządzeniem końcowym. Urządzenie końcowe zostaje zautoryzowane przez LoRaWAN Network Server (LNS) i w dalszym czasie obsługiwane jest poprzez retransmiter.

Dalsza komunikacja między (ED) a (R) odbywa się w sposób synchroniczny. (ED) zaczyna nadawać kolejną wiadomość zaraz przed przewidywanym włączeniem nasłuchu przez (R). Synchronizacja czasu odbywa się na warstwie informacji sieciowych przesyłanych w protokole LoRa WAN i wykonywana jest przy każdej wiadomości. Z czasem następuje stopniowe rozsynchronizowanie zegarów co kompensowane jest poprzez odpowiednie zwiększenie preambuły oraz wcześniejsze rozpoczęcie nadawania przez (ED). W ten sposób komunikacja nadal może zostać wykonana poprawnie.

Urządzenie (R) nie powinno odbierać danych będących transmisją pomiędzy (ED) a (GW), ponieważ powodowałoby to niepotrzebny wzrost poboru energii. Wiadomości wysyłane bezpośrednio do (GW) poprzez (ED) lub (R) nie posiadają długiego sygnału preambuły, przez co zostaną odrzucone przez (R) w oknie nasłuchowym CAD.

4. Wyniki

Wartość poboru prądu dla cyklicznego nasłuchu, przy założeniu braku retransmisji, została wyliczona na podstawie minimalnego czasu aktywności radia (ustabilizowanie nasłuchu + inicjalizacja odbiornika + odbiór pierwszych 8 symboli + wyłączenie odbiornika) oraz średniego poboru prądu poprzez odbiornik w odpowiednich stanach urządzenia. Poniższa tabela przedstawia wyniki wyliczonego rocznego poboru prądu przez retransmiter wynikającego wyłącznie z nasłuchu (gotowości do retransmisji).

Tab. 3. Koszt energetyczny nasłuchu CAD

Spreading Factor	*Roczny koszt*	*Jednostka*
7	138	[mAh]
8	214	[mAh]
9	354	[mAh]
10	607	[mAh]
11	1309	[mAh]
12	2278	[mAh]

Kolejna tabela zawiera wyliczony koszt energetyczny retransmisji na który na który składa się odbiór ramki radiowej od urządzenia (ED) oraz jego transmisja do (GW). Po każdej transmisji następuje otwarcie okien RX1, RX2 zgodnie z protokołem.

Tab 4. Koszt energetyczny retransmisji pojedynczego urządzenia

Spreading Factor	*Roczny koszt*	*Jednostka*
7	80	[mAh]
8	100	[mAh]
9	137	[mAh]
10	202	[mAh]
11	385	[mAh]
12	647	[mAh]

Analiza nie zawiera możliwości występowania problemu koegzystencji różnych sieci na jednym obszarze. W takim przypadku urządzenie Relay może cyklicznie odbierać ramki Relay Join Request od urządzenia końcowego, które nie zakończą się nigdy poprawną rejestracją w sieci, gdyż urządzenie końcowe pozostanie niezdefiniowane dla LNS.

5. Wnioski

Na podstawie przeprowadzonych wyliczeń autorzy twierdzą, iż da się zaprojektować następujące przykłady wersje bateryjnego retransmitera:

Małe urządzenie wyposażone w dwie baterie litowe rozmiaru AA (14500) o łącznej użytecznej pojemności 3,2h. Urządzanie może retransmitować do 2 urządzeń końcowych komunikując się ze stacją bazową z użyciem max SF 10. W takiej konfiguracji urządzenie będzie działać 5,38 roku na baterii.

Urządzenie wyposażone w jedną baterię rozmiaru D (32650) o użytecznej pojemności 12,5Ah, retransmitujące do 5 urządzeń końcowych (SF7) komunikując się ze stacją bazową z użyciem do SF 10. W takiej konfiguracji urządzenie będzie działać 10,4 roku na baterii co równe jest czasowi działania powszechnie stosowanych nakładek na wodomierze lub elektronicznym licznikom zużycia wody.

Retransmiter powinien być zawsze montowany w lokalizacjach, gdzie istnieje możliwość uzyskania poprawnej komunikacji ze stacją bazową z użyciem parametru Spreading Factor maksymalnie na poziomie 10. Komunikacja z SF12 dałaby odpowiednio 2,16 roku oraz 3,65 roku dla pierwszego i drugiego przypadku, co wydaje się wartościami nieakceptowalnymi komercyjnie. Dodatkowo w 2 przypadku doszłoby do przekroczenia maksymalnego czasu nadawania Duty Cycle przez co komunikacja nie mogłaby zostać prawidłowo zrealizowana. Konieczność instalacji retransmitera w miejscu gdzie jego jakość komunikacji ze stacją bazową jest na odpowiednio dobrym poziomie wydaje się akceptowalna w przypadku stosunkowo małego urządzenia zasilanego bateryjnie.

Pobór prądu retransmitera jest zbyt wysoki, by jego funkcje pełniły inne urządzenia końcowe sieci. W takim przypadku nasłuch urządzeń końcowych nawet z najniższym parametrem SF wymaga więcej energii niż typowa energia wydatkowana na transmisje przez urządzenie końcowe.

LITERATURA

[1] Haif, Hamza, Abdelali Arous, and Hüseyin Arslan. *"A Decentralized Dynamic Relaying-Based Framework for Enhancing LoRa Networks Performance."* IEEE Internet of Things Journal (2024).

[2] Tropeano, Angelo, et al. *"A Field Test for Maximizing Coverage through Multi-Hop D2D LoRa Transmission."* 2023 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB). IEEE, 2023.

[3] Strzoda, Anna, and Krzysztof Grochla. *"A Nature-Inspired Approach to Energy-Efficient Relay Selection in LPWAN."* (2024).

[4] Gorawski, Michał, et al. *"Energy minimization algorithm for estimation of clock skew and reception window selection in wireless networks."* Sensors 21.5 (2021): 1768.

[5] Marjasz, Rafał. *"Modele kolejkowe z mechanizmem zawieszenia obsługi typu 'multiple vacation'– analiza z wykorzystaniem SD."* https://omega.polsl.pl/info/phd/PSLb2b85353dc37445ebd94305fff79d867/

[6] Code, Price. *"Communication Systems for Meters and Remote Reading of Meters-Part 4: Wireless Meter Readout (Radio Meter Reading for Operation in the 868 MHz to 870 MHz SRD Band)."* (2005).

[7] Strzoda, Anna. *"Zastosowanie heurystyki do wyboru lokalizacji węzłów pośredniczących w sieciach LP WAN."* Przegląd Telekomunikacyjny+ Wiadomości Telekomunikacyjne (2023).

[8] SX1261/2 Long Range, Low Power, sub-GHz RF Transceiver. https://www.semtech.com/products/wireless-rf/lora-connect/sx1262#documentation

[9] Introduction to Channel Activity Detection AN1200.85 Semtech corporation. https://www.semtech.com/uploads/technology/LoRa/cad-ensuring-lora-packets.pdf

[10] Alliance, LoRa. *"LoRaWAN Relay Specification TS011-1.0. 0."* (2022). https://www.iitis.pl/pl/content/lorawan®-relay-specification-ts011-100

KRiT 2024

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

TWORZENIE WIELOWIDOKOWYCH SEKWENCJI DLA WIZJI WSZECHOGARNIAJĄCEJ

CREATING MULTIVIEW SEQUENCES FOR IMMERSIVE VIDEO

Dawid Mieloch; Dominika Klóska; Adrian Dziembowski; Błażej Szydełko; Adam Grzelka; Jakub Stankowski

Instytut Telekomunikacji Multimedialnej, Politechnika Poznańska, Poznań
{dawid.mieloch,dominika.kloska,adrian.dziembowski,blazej.szydelko,adam.grzelka,jakub.stankowski}@put.poznan.pl

DOI: 10.15199/59.2024.4.37

Streszczenie: **Wielowidokowe sekwencje wizyjne są niezbędne do rozwoju technologii wykorzystywanych dla celów wizji wszechogarniającej. Szczególnie ważną częścią tych badań są prace nad nowymi normami kompresji. W tym artykule przedstawiono przegląd metod, których użycie w procesie tworzenia sekwencji ma korzystny wpływ na ich końcową jakość, co potwierdzają wyniki przeprowadzonych badań eksperymentalnych. Proces tworzenia wielowidokowej sekwencji został przedstawiony na przykładzie sekwencji Choreo, obecnie wykorzystywanej w pracach ISO/IEC MPEG Implicit Neural Visual Representation.**

Abstract: **Multiview sequences are essential to develop technologies utilized for immersive video purposes. Especially important part of this research involves work on the new compression standards. This article provides an overview of methods that positively influence the final result of the test sequence creation process, as confirmed by results of conducted experiments. The process of multiview sequence acquisition was presented using the example of the Choreo sequence which is now used in the work of ISO/IEC MPEG Implicit Neural Visual Representation.**

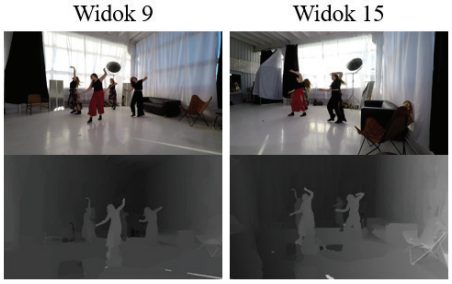
Słowa kluczowe: **system wielokamerowy. wizja wszechogarniająca.**

Keywords: **multicamera system, immersive video.**

1. WSTĘP

Wizja wszechogarniająca (immersyjna) jest tematem cieszącym się dużym zainteresowaniem ze strony mediów społecznościowych [9], branży turystycznej [8], edukacyjnej [24] i wielu innych. Użytkownik systemu wykorzystującego wizję wszechogarniającą może oglądać zarejestrowaną scenę z dowolnego punktu widzenia i nie musi ograniczać się do położeń, w których usytuowano wcześniej kamery. Taka możliwość zapewniona jest dzięki zareprezentowaniu całej trójwymiarowej sceny. Typowo, wizja wszechogarniająca wykorzystuje format MVD (Multiview Video + Depth) [18], na który składają się widoki zarejestrowane z różnych punktów widzenia oraz odpowiadające im mapy głębi, które zawierają informacje o geometrii sceny (Rys. 1).

W systemach wielokamerowych, które służą do rejestracji wizji wszechogarniającej, można stosować dowolne ułożenie kamer, zarówno perspektywicznych jak i dookólnych [2]. Największym ograniczeniem jest wymóg, aby kamery posiadały wejście synchronizacyjne, co znacznie ogranicza liczbę modeli które można z powodzeniem wykorzystać w takim systemie [5].



Rys.1 Sekwencja wielowidokowa Choreo [13], zarejestrowane widoki i odpowiadające im mapy głębi.

Proces pozyskiwania informacji o geometrii sceny różni się w zależności od typu sekwencji. W przypadku sekwencji generowanych komputerowo, geometria sceny zazwyczaj może być wyeksportowana z programu użytego do ich stworzenia (jest to możliwe np. w powszechnie używanym programie Blender [23]). Informację o geometrii sceny sekwencji naturalnej można pozyskać wykorzystując kamery głębi, które najczęściej mierzą odległość do obiektów znajdujących się w scenie za pomocą promieni podczerwonych [12]. Rozwiązanie to ma swoje wady, m.in. ograniczoną rozdzielczość nagranych map głębi czy możliwe interferencje wzbudzane przez sygnał sąsiadujących ze sobą kamer [26]. W związku z tym, w celu pozyskania informacji o geometrii sceny, najczęściej wykorzystywana jest estymacja głębi, przeprowadzana na podstawie parametrów kamer (czyli ich parametrów optycznych oraz ich położenia) oraz widoków nagranych przez system wielokamerowy [17].

Rosnąca popularność wizji immersyjnej spowodowała zwiększenie liczby badań prowadzonych w ramach rozwoju tej technologii ze szczególnym uwzględnieniem zagadnień kompresji [25] – pojawiła się pierwsza norma kompresji takiej wizji, czyli MPEG Immersive Video [2]. Ilość materiałów testowych stworzonych na potrzeby badawcze wizji wszechogarniającej jest niestety stosunkowo niewielka w porównaniu z ogromem danych testowych, wykorzystywanych na przykład do trenowania sieci neuronowych. Stworzenie wielowidokowej sekwencji, która może zostać wykorzystana do testów nie należy jednak do łatwych zadań, szczególnie w przypadku sekwencji naturalnych. Pod uwagę należy wtedy wziąć wiele czynników, które mogą negatywnie wpłynąć na końcową jakość sekwencji jeżeli nie zostaną one odpowiednio zaadresowane.

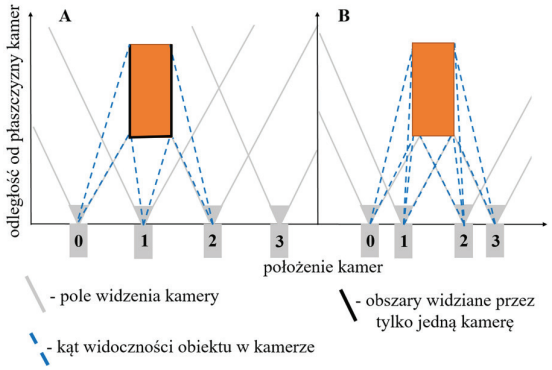
Aby ułatwić proces tworzenia sekwencji testowych, w niniejszym artykule przedstawiono przegląd metod, których zastosowanie zalecane jest w procesie akwizycji wielowidokowej sekwencji na potrzeby wizji wszechogarniającej. W kolejnym rozdziale opisano kolejne etapy rejestracji i przetwarzania sekwencji na przykładzie procesu akwizycji sekwencji testowej Choreo [13]. W rozdziale trzecim zaprezentowano wyniki eksperymentalne obrazujące wpływ poprawności przeprowadzenia kolejnych etapów tworzenia sekwencji na jej końcową jakość, a co za tym idzie, jej użyteczność w badaniach.

2. REJESTRACJA SEKWENCJI

2.1. Wybór kamer i ich rozstawienie

Proces nagrywania sekwencji testowej należy rozpocząć od ustalenia ułożenia oraz liczby kamer w systemie wielokamerowym. Aby możliwe było wyznaczenie głębi punktu w scenie konieczne jest użycie minimum dwóch kamer, jednak zwiększenie liczby kamer w systemie będzie miało pozytywny wpływ na jakość estymacji głębi [17]. Należy podkreślić, że wszystkie kamery muszą być ze sobą zsynchronizowane aby akwizycja w każdej z nich rozpoczynała się w tym samym momencie [20].

Najczęściej spotykanym sposobem rozstawienia kamer jest umieszczenie ich w równomiernych odstępach [5]. Odległość między poszczególnymi kamerami musi być jednak tak dopasowana, by nie dopuścić do sytuacji w której dany obiekt jest widziany tylko przez jedną kamerę, ponieważ wtedy estymacja głębi tego obiektu będzie niemożliwa (Rys. 2 A).

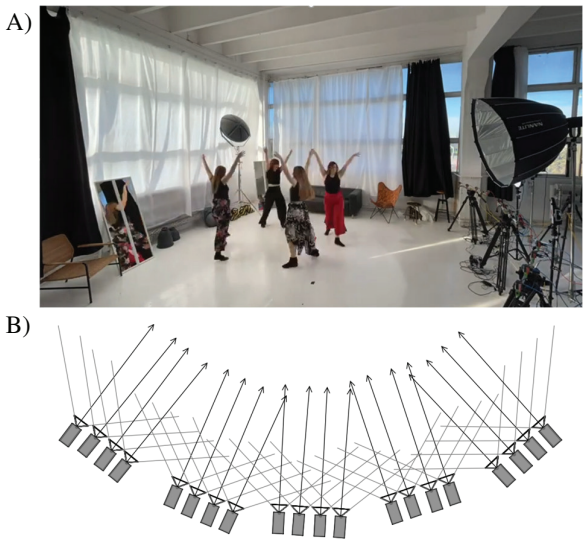


Rys. 2 (A) System wielowidokowy z niepoprawnym rozstawieniem kamer. (B) Ten sam system po zmniejszeniu odległości i zastosowaniu grupowania w pary.

Jeżeli scena zawiera przesłonięte elementy, znalezienie odpowiedniej odległości między kamerami w rozstawieniu równomiernym może okazać się niemożliwe. Odpowiedzią na powyższe problemy jest łączenie kamer w równomiernie rozstawione grupy [20] (Rys. 2 B). Takie też rozstawienie zastosowano przy tworzeniu sekwencji Choreo. Rys. 3 obrazuje dokładne ułożenie systemu wielokamerowego podczas akwizycji tej sekwencji.

2.2. Kalibracja systemu wielokamerowego

Nieodłącznym elementem procesu tworzenia sekwencji wielowidokowej jest pozyskanie danych o położeniu i ustawieniu każdej z kamer w scenie. Aby pozyskać te dane należy przeprowadzić kalibrację systemu wielokamerowego, która obejmuje zarówno wyznaczenie parametrów wewnętrznych kamer, jak i relacji przestrzennych między nimi, czyli kalibracji zewnętrznej.



Rys. 3 System wielokamerowy wykorzystany do akwizycji sekwencji Choreo. A) System, widoczny po prawej stronie nagrywanej sceny. B) Wizualizacja rozstawienia kamer, strzałki symbolizują osie optyczne.

Parametry wewnętrzne opisują układ optyczny kamery, czyli długość ogniskowej, współrzędne punktu środkowego matrycy i współczynniki zniekształceń soczewkowych. Parametry zewnętrzne definiują natomiast współrzędne położenia kamery w trzech płaszczyznach oraz wartości kąta obrotu wokół każdej osi, czyli kierunek w którym zwrócona jest oś optyczna kamery [21].

Bez względu na rodzaj parametrów, problem kalibracji można w ogólności można rozłożyć na dwa etapy. Pierwszy etap polega na zebraniu danych odniesienia poprzez ekstrakcję punktów charakterystycznych z zarejestrowanych przez kamery widoków. W drugim etapie uzyskane dane stanowią podstawę do rozwiązania problemu optymalizacji funkcji. Proces ten polega na minimalizacji różnicy między wyliczanymi a rzeczywistymi pozycjami punktów charakterystycznych w obrazie [11].

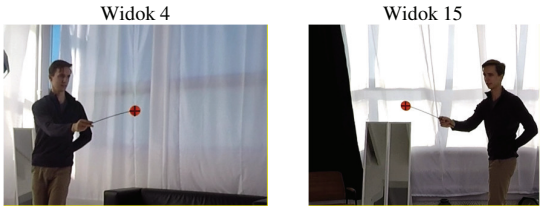
Najprostszą i najbardziej niezawodną metodą kalibracji parametrów wewnętrznych jest metoda z użyciem znacznika w formie czarno-białej szachownicy [28], przeprowadzana niezależnie dla każdej kamery. Istotnym aspektem tej metody jest to, że szachownica musi być płaska i sztywna co zapewnia dokładność detekcji punktów charakterystycznych. Kalibracja wewnętrzna nie musi odbywać się w miejscu akwizycji sekwencji, co umożliwia przeprowadzenie jej w bardziej dogodnych warunkach oświetleniowych i przy większej kontroli pozycjonowania znacznika w kadrze.

Większą uwagę należy zwrócić na kalibrację parametrów zewnętrznych, która przeprowadzana na miejscu rejestracji sekwencji i obejmuje wszystkie kamery. Klasyczne metody wyznaczania parametrów zewnętrznych obejmują kalibrację z wykorzystaniem punktów charakterystycznych otrzymanych za pomocą tablicy kalibracyjnej ze wzorem. Są to przede wszystkim: omawiana przy kalibracji wewnętrznej czarno-biała szachownica [21] lub

znaczniki ArUco [19]. Należy jednak zaznaczyć, że wzorce wykorzystywane w tych metodach powinny być dobrze widoczne z każdej perspektywy i w różnych orientacjach, co w przypadku dowolnego ustawienia kamer może być niemożliwe do osiągnięcia.

Odmiennym podejściem charakteryzują się metody bezznacznikowe, określane również jako automatyczne [1]. Bazują one na ekstrakcji punktów charakterystycznych z lokalnych cech obrazu, np. krawędzi i narożników. Popularnym algorytmem stosowanym do ekstrakcji cech i wyznaczania relacji między obrazami jest SIFT (ang. *Scale-Invariant Feature Transform*), jednak zazwyczaj jest on stosowany do kalibracji stereopar [15]. Bardziej zaawansowane techniki autokalibracji, takie jak [4], [16], charakteryzują się estymacją parametrów zewnętrznych na podstawie ruchu kamer, co wyklucza ich stosowanie przy tworzeniu sekwencji wielowidokowej na potrzeby wizji wszechogarniającej.

Jednym z podejść, które zapewnia wymaganą praktyczność i wygodę stosowania jest metoda wykorzystująca prosty znacznik kalibracyjny w formie kolorowej kuli [10], [23]. Taki znacznik jest prosty w obsłudze, a jego kształt sprawia, że z każdej perspektywy jest widziany tak samo i łatwo znaleźć jego położenie (Rys. 4).



Rys. 4 Wyznaczanie parametrów zewnętrznych sekwencji Choreo z użyciem prostego znacznika.

2.3. Estymacja głębi

Najczęstsze wymagania stawiane metodom estymacji głębi dla systemów wizji wszechogarniającej to estymacja spójna międzykamerowo oraz międzyramkowo, możliwa do przeprowadzenia dla różnego typu kamer (w tym dookólnych) [17].

Do najważniejszych metod estymacji głębi należą metody klasyczne, działające najczęściej poprzez minimalizację funkcji celu, a najczęściej wykorzystywany algorytm cięcia grafu (ang. *Graph Cut*) [14]. Drugą silnie rozwijaną w ostatnich latach grupą metod są algorytmy wykorzystujące uczenie maszynowe [27], [29]. Niestety, zapotrzebowanie na pamięć oraz moc obliczeniową w takich algorytmach zwiększa się wraz z liczbą widoków wejściowych, co uniemożliwia estymację map głębi w pełnej rozdzielczości.

Dla sekwencji Choreo, oraz szeregu innych testowych sekwencji wielowidokowych, została wykorzystana klasyczna metoda rozwijana przez ekspertów ISO/IEC MPEG Video Coding – Immersive Video Depth Estimation (IVDE) [17]. Oprogramowanie to wykorzystuje metodę segmentacji widoków wejściowych, optymalizację przez algorytm cięcia grafu, oraz szereg narzędzi zapewniających spójność czasową i zmniejszających złożoność obliczeniową. Działa dla dowolnego rodzaju oraz liczby kamer, przy niedużym zapotrzebowaniu na pamięć.

2.4. Dodatkowe przetwarzanie

Estymacja map głębi może być poprzedzona poprzez dodatkowe kroki, które mogą zapewnić wyższą jakość wizji dostarczanej końcowemu widzowi, takie jak międzywidokowa korekcja barwna czy odszumienie wejściowych sekwencji.

Należy zaznaczyć, że kroki te nie są niezbędne, jednak pozwalają podnieść jakość stworzonej sekwencji. Przeprowadzenie korekcji barwnej znacząco zmniejsza wpływ różnych charakterystyk barwnych używanych kamer na końcową jakość widoku wirtualnego [6], co potwierdzono również w jednym z przeprowadzonych w rozdziale trzecim badań. Usunięcie szumu znacząco wpływa na poprawność międzywidokowego pasowania bloków w estymacji głębi i na jej spójność w czasie [21].

3. EKSPERYMENT

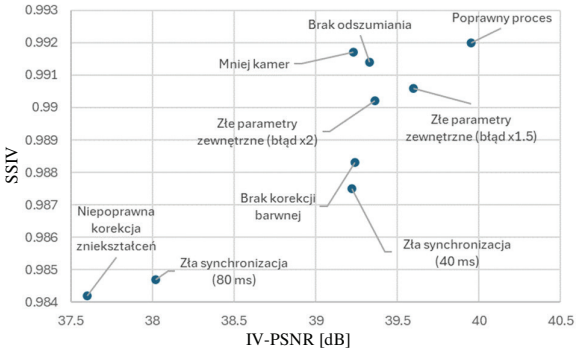
W celu zaprezentowania wpływu poszczególnych etapów przetwarzania sekwencji wielowidokowej na finalną jakość wizji wszechogarniającej, przeprowadzono eksperyment, w którym modyfikowano proces tworzenia takiej sekwencji. Badanie przeprowadzono poprzez pomijanie wykorzystywania poszczególnych narzędzi, lub też ich niepoprawne użycie.

Eksperyment przeprowadzony został dla sekwencji Choreo, dla której, w każdej testowanej konfiguracji estymowano mapy głębi dla widoków *v6, v7, v8, v10, v11, v12* i *v13*. Wybrano jedynie podzbiór widoków na środku sceny aby zmniejszyć złożoność obliczeniową. Następnie, na ich podstawie syntezowano widok wirtualny w pozycji widoku *v9*. Dzięki takiemu podejściu, możliwa była końcowa ocena jakości renderowanej treści z użyciem dwóch obiektywnych metryk jakości dedykowanych do oceny wizji wszechogarniającej: IV-PSNR [7] i SSIV [22]. W eksperymencie sprawdzono wpływ:

- przeprowadzenia odszumiania sekwencji (z użyciem algorytmu NL-means [3]),
- korekcji barwnej (algorytmem PCR [6]),
- poprawności wyznaczenia parametrów zewnętrznych (dwa testy: dokładność parametrów 1,5 i 2 razy mniejsza, niż w przypadku optymalnych, ostatecznych parametrów wyznaczonych algorytmem ECPC [23]),
- dwukrotnego zmniejszenia liczby kamer (użycie wyłącznie widoków *v6, v8, v10* i *v12*),
- niepoprawnej synchronizacji kamer [dwa testy: jedna kamera rejestrowała scenę z opóźnieniem jednej (40 ms) lub dwóch ramek (80 ms)],
- niepoprawnej korekcji zniekształceń soczewkowych (użycie parametrów zniekształcenia większych o 1 % od rzeczywistych).

Wyniki eksperymentu zaprezentowano na rysunku 5. Jak pokazano, każdy z opisanych etapów przetwarzania sekwencji wielowidokowej ma istotny wpływ na ostateczną jakość syntezowanych widoków. Dokładność synchronizacji i kalibracji systemu wielokamerowego wraz

z przeprowadzeniem niezbędnej korekcji zarejestrowanego materiału jest kluczowa z punktu widzenia ostatecznego użytkownika systemu swobodnej nawigacji.



Rys. 5. Jakość syntezowanego widoku wirtualnego dla różnych konfiguracji procesu przetwarzania sekwencji wielowidokowej.

4. PODSUMOWANIE

W niniejszym artykule zaprezentowano przegląd metod wykorzystywanych w procesie tworzenia naturalnych sekwencji wielowidokowych na potrzeby wizji wszechogarniającej. Tak przygotowana sekwencja wizyjna jest gotowa do przedstawienia końcowemu użytkownikowi na ekranie telewizora czy też z wykorzystaniem nagłownych okularów do rzeczywistości wirtualnej [2]. Wyniki przeprowadzonych eksperymentów pokazują wpływ poszczególnych kroków procesu akwizycji wizji wszechogarniającej na końcową jakość.

Proces akwizycji i tworzenia testowej sekwencji wielowidokowej z wykorzystaniem przytoczonych metod został przedstawiony na przykładzie sekwencji Choreo, która została zgłoszona przez Autorów do bazy sekwencji testowych standardu ISO/IEC MPEG Immersive Video, i która to jest obecnie wykorzystywana w testach MPEG INVR.

PODZIĘKOWANIA

Praca finansowana ze środków przyznanych przez Ministerstwo Nauki i Szkolnictwa Wyższego.

LITERATURA

[1] H. Boukamcha. i in. 2017. "Robust auto calibration technique for stereo camera". *ICEMIS 2017*.
[2] Boyce J. M. i in. 2021. “MPEG Immersive Video Coding Standard”. *Proc. of the IEEE* 109: 1521-1536
[3] Buades A. 2005. “A non-local algorithm for image denoising”. *CVPR 2005*: 60–65.
[4] Cui H. i in. 2023. “MCSfM: Multi-Camera Based Incremental Structure-From-Motion”. *IEEE Transactions on Image Processing*, 32: 6441 - 6456.
[5] Domanski M. i in. 2016. „New results in free-viewpoint television systems for horizontal virtual navigation”. *ICME 2016*.
[6] Dziembowski A. i in. 2021. „Color Correction for Immersive Video Applications”. *IEEE Access* 9: 75626-75640.
[7] Dziembowski A. i in. 2022. „IV-PSNR–the objective quality metric for immersive video applications”. *IEEE T. Circ. & Sys. Vid. Tech.* 32 (1): 7575-7591.
[8] Ghorbanzadeh D. i in. 2024. “Enhancing destination image through virtual reality technology: the role of tourists’ immersive experience”. *Curr. Psych:* 1-13.
[9] Hinds T. A. i in. 2023. “Immersive Media and the Metaverse”. *IEEE Comm. Magazine* 61: 48-54.
[10] ISO/IEC SC29/WG04, 2024, “Manual of the Extrinsic Camera Parameters Calibration framework”. *146 MPEG meeting,* Rennes, Francja.
[11] Jianzhu H. i in. 2023. “A Review and Comparative Study of Close-Range Geometric Camera Calibration Tools”. arXiv:2306.09014.
[12] Kang Y. i in. 2010. “High-quality multi-view depth generation using multiple color and depth cameras”. *ICME 2010*, pp. 1405–1410.
[13] Klóska D. i in. 2024. „Proposal of new natural content – Choreo”. *MPEG 146,* Rennes, Francja.
[14] Kolmogorov V. i in. 2004 “What Energy Functions Can Be Minimized via Graph Cuts?”. *IEEE Tr. on Pattern Analysis and Machine Int.* 26.2: 147–159.
[15] Liu R. i in. 2009. “Stereo Cameras Self-Calibration Based on SIFT”. *Int. Conf. on Measuring Tech. and Mechatronics Automation*, 1: 352-355.
[16] Meng X. i in. 2018. “Dense RGB-D SLAM with Multiple Cameras”. *Sensors*, 18 (7): 2118.
[17] Mieloch D. i in. 2020. „Depth map estimation for free-viewpoint television and virtual navigation”. *IEEE Access* 8: 5760–5776.
[18] Mueller K. i in. 2011. „3-D Video Representation Using Depth Maps”. *Proc. of the IEEE* 99: 643–656.
[19] Pengwei Z. i in. 2024. “Meta-Calib: A generic, robust and accurate camera calibration framework with ArUco-encoded meta-board”. *ISPRS Journal of Photogrammetry & Remote Sensing* 212: 357-380.
[20] Stankiewicz O. i in. 2018. “A Free-Viewpoint Television System for Horizontal Virtual Navigation”. *IEEE Transactions on Multimedia* 20 (8): 2182-2195.
[21] Stankiewicz O. i in. 2018. “Multiview video: Acquisition, processing, compression and virtual view rendering”. *Acad. Press Lib. in Sig. Proc.* 6: 3-74.
[22] Stankowski J. i in. 2024. „Miara podobieństwa strukturalnego dla wizji wszechogarniającej”. *KRiT 2024*.
[23] Szydełko B. i in. 2024. „ECPC – versatile multi-camera system calibration framework for immersive video applications”. *SoftwareX,* 26 (101669).
[24] Wei Z. i in. 2023. „Research on the Current Situation and Future Development Trend of Immersive VR in the Field of Education”. *Sustainability,* 15, 7531.
[25] Wien M. i in. 2019. “Standardization Status of Immersive Video Coding”. *IEEE J on Emerging & Selected Topics in Circuits & Systems* 9 (1): 5-17.
[26] Xiang S. i in. 2013. „A gradient-based approach for interference cancelation in systems with multiple Kinect cameras”. *ISCAS 2013*.
[27] Yao Y. i in. 2018. „MVSNet: Depth Inference for Unstructured Multi-view Stereo”. *ArXiv*1804.02505.
[28] Zhang Z. 2000. “A Flexible New Technique for Camera Calibration”. *IEEE Transactions on Pattern Analysis & Machine Intelligence.* 22 (11): 1330 - 1334.
[29] Zizhuang W. i in. 2021. „AA-RMVSNet: Adaptive Aggregation Recurrent Multi-view Stereo Network”. *ArXiv,* 2108.03824.

ANALIZA PORÓWNAWCZA WYDAJNOŚCI ARCHITEKTURY GŁEBOKIEGO UCZENIA SIE W KLASYFIKACJI OBRAZÓW PODWODNYCH

COMPARITIVE PERFORMANCE ANALYSIS OF DEEP LEARNING ARCHITECTURES IN UNDERWATER IMAGE CLASSIFICATION

Mehrunnisa[1]; Dawid Juszka[2]; Yi Zhang[3]; Mikołaj Leszczuk[4]

[1] AGH University of Science and Technology, Krakow, mehrunnisa@agh.edu.pl
[2] AGH University of Science and Technology, Krakow Dawid.Juszka @agh.edu.pl
[3] Xi'an Jiaotong University, Xi'an, China, yi.zhang.osu@xjtu.edu.cn
[4] AGH University of Science and Technology, Krakow, Mikolaj.Leszczuk@agh.edu.pl

DOI: 10.15199/59.2024.4.38

Streszczenie: W tym badaniu oceniamy wydajność modeli VGG-16, EfficientNetB0 i SimCLR w klasyfikacji 5000 podwodnych zdjeć. Zbiór danych podzielono na 75 procent do celów szkoleniowych i 25 procent do testów, przy czym reczne etykietowanie zapewniało dokładne odwzorowanie podstaw. Zastosowaliśmy grupowanie K-średnich do segmentacji zbioru danych na podstawie podobieństwa oraz PCA w celu zmniejszenia wymiarowości przy jednoczesnym zachowaniu struktury semantycznej. Zróżnicowany zbiór danych zwieksza zdolność modeli do uogólniania w różnych warunkach. Oceniliśmy grupowanie i klasyfikacje za pomoca wyniku sylwetki, wskaźnika Daviesa-Bouldina i wskaźnika Calinskiego-Harabasza. Wyniki ujawniaja mocne i słabe strony każdego modelu, dostarczajac informacji na temat przyszłych ulepszeń w analizie obrazów podwodnych.

Abstract: In this study, we assess the performance of the VGG-16, EfficientNetB0, and SimCLR models in classifying 5,000 underwater images. The dataset was split into 75 perceent for training and 25 percent for testing, with manual labeling ensuring accurate ground truth. We used K-means clustering to segment the dataset based on similarity, and PCA to reduce dimensionality while maintaining the semantic structure. The diverse dataset boosts the models' ability to generalize across various conditions. We evaluated clustering and classification using the silhouette score, Davies-Bouldin index, and Calinski-Harabasz index. The results reveal each model's strengths and weaknesses, providing insights for future improvements in underwater image analysis

Słowa kluczowe: VGG-16, Efficient, NetB0, InceptionV3, model SimCLR, PCA (analiza głównych składowych, grupowanie K-średnich, metoda łokcia, wynik sylwetki, indeks Daviesa-Bouldina, indeks Calinskiego-Harabasza.

Keywords: VGG-16, Efficient, NetB0, InceptionV3, SimCLR model, PCA (Principal Component Analysis, K-mean clustering, Elbow Method, Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index.

1. INTRODUCTION

Convolution neural networks (CNNs) have played an essential role in machine learning, especially in classifying and extracting images. Powered by artificial intelligence, CNNs imitate human thinking to bring out better decision-making. CNNs have become important tools for the decision-making process, with layers like deep neural networks in the human brain, processing rich visual inputs of color, texture, patterns, and resolution, important in the interpretation and manipulation processes. The deep learning process, which is the essence of CNNs, is particularly efficient in the extraction of features, transformation, and the analysis of patterns that are key in the recognition and processing of linear and nonlinear data. Such algorithms detect and localize patterned edges in images, hence acting as a basis for object detection using a large network [1]. Pretrained architectures, such as those witnessed with ImageNet LSVRC-2010, have the capacity to classify thousands of categories from high-resolution images [2]. Mobile-Nets are among the best in detecting facial attributes and localizing objects, while models such as ResNet50 and Inception V3 have been ranked with very high accuracy for use in disease diagnosis and COVID-19 research [3][4]. Techniques in models such as SimCLR have improved visual representation [5], which becomes valuable in the context of color distortion in underwater pictures. Models, such as VGG-16, Inception V3, and EfficientNet B0, have been employed in amplifying the accuracy of classification, with PCA and its advanced form, kPCA, employed in dimensionality reduction [6][7]. This paper explores CNNs in the task of identifying and classifying a dataset of 5000 marine images using state-of-the-art models like VGG-16, Inception V3, EfficientNet B0,

and SimCLR. With respect to this, related works on extracting features from training and testing datasets are fine-tuned to optimize their accuracy. Underwater image classification has its applications in the field of biodiversity research, marine ecosystem monitoring, and support of underwater navigation and construction. Image classification in underwater images is challenging due to light absorption and scattering, which degrade visual information. The problem is compounded by particles in water causing turbulence and changing lighting variations. The dataset consists of nearly 5,000 images from sources like scuba divers and underwater photographers, featuring flora, fish species, submerged statues, and archaeological remains. Images were resized to uniform size and resolution. The models VGG-16, Inception V3, EfficientNet B0, and SimCLR are recognized for their robustness and high performance in complex visual tasks, including underwater scenes.

2. METHODOLOGY

In this work, 5000 underwater images are classified using pre-trained CNN models. The models used RGB images for VGG-16, Inception V3, and EfficientNet B0, while SimCLR processes greyscale images. Images are segmented into five classes/categories: Statues, Underwater species, Fish, Underwater plants, and Scuba divers. These images are further divided into clusters using k-means clustering. The dataset has varied image qualities and resolutions.The tested models and their outcomes from the image classification into clusters are described below.

2.1 Visual Geometry Group (VGG-16)

• **Introduction:** This was developed by Simonyan and Zisserman from the University of Oxford. VGG-16 was the winner of the 2014 Large Scale Visual Recognition Challenge; its score even beat GoogleNet.

• **Architecture:** The VGG-16 architecture is very straightforward, with a total of 13 convolutional layers, three fully connected layers, and five pooling layers, all activated using ReLU. It works on 224 × 224 pixel images.

• **Working:** In the transfer learning theory, the loaded model is VGG-16, which has been trained with ImageNet weights. The primary convolutional and pooling layers combine the bottleneck features that strongly predict but are much compressed. Subsequently, these features are saved in a NumPy array. The architecture follows up with 5 dense layers, 1 dropout layer, and a batch normalization layer. The dense layers include 512, 256, 128, and 64 ReLU activation units, including one final Softmax activation unit in the output layer.

• **Normalization and Splitting Dataset for Training and Testing:** The 80/20 split secures data for learning and testing, avoiding overfitting, and ensuring accurate performance estimates. Normalizing VGG-16 involves resizing to 224x224, converting to arrays, expanding dimensions, and adjusting color.

• **K-mean clustering:** This is is an unsupervised method that groups data points under a pre-defined number of clusters based on features. The initial step involves random data point assignment. After that, re-estimation of centroids is done in iterative ways with the reassignment of data toward the minimization of the objective function. This analysis is done using features extracted by VGG-16, and the features are clustered into 10 data clusters.

• **Elbow method:** The method is applied to find out the optimal number of clusters while carrying out K-means clustering. It plots the sum of squared distances versus the number of clusters and shows an elbow, which is a sudden sharp decline. This shows the optimal cluster count, helping to fight underfitting and overfitting.

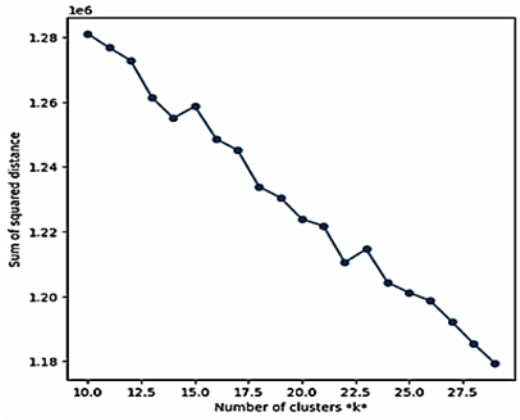


Figure 1: Elbow method showing the number of best-fit clusters to be used for the model

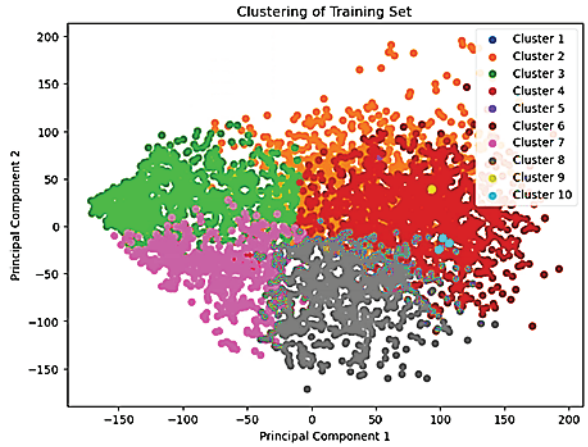


Figure 2: Clusters created using the k-mean clustering method after extraction of features by VGG-16

2.2 Inception V3

Breaking down the larger convolutions in Inception V3, which is a convolutional architecture, into smaller ones involves normal techniques, thus delivering better recognition performance. Batch normalization and label smoothing enhance asymmetrical convolutions, reducing the computational demand of the network while increasing the ability to learn complex features with fewer parameters. In deep Sea Image Analysis, for instance, we have applied some recent advanced clustering techniques in feature extraction from a dataset of 5000 underwater images to train our model, using an architecture like Inception V3. It involves clustering of features that are very similar to each other and is very much a part of the pattern recognition process in the marine environment. This powerful tool provides scientists and conservators a chance to

analyze underwater imagery by the strong union of features extracted from Inception V3 and robust clustering techniques. Training and testing sets for Inception V3 are (3913, 51200) (979, 51200).

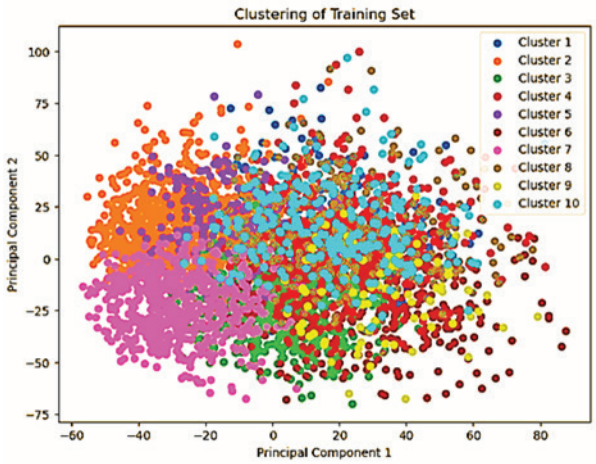


Figure 3: Clusters created using the k-mean clustering method after extraction of features by Inception V3

2.3 EfficientNet B0

EfficientNet B0 is an architecture from the EfficientNet family of convolutional neural networks. The Feature extraction uses mobile inverted bottleneck convolution (MBConv) with squeeze and excitation blocks to have the best-optimized model. This architecture applies ordered compound scaling to maintain balance in network depth, width, and resolution. The architecture begins with a standard convolution and contains several MBConv layers. EfficientNet B0 is a powerful feature extractor. Extract the features from the set of data by using the model and then applying it for K-means clustering. It fits within unsupervised learning, where it groups the data into clusters with minimum intra-cluster variance, hence effectively utilizing the high-dimensional features in EfficientNet B0. The training and testing sets for Efficient Net B0 are (3913, 131712) (979, 131712).

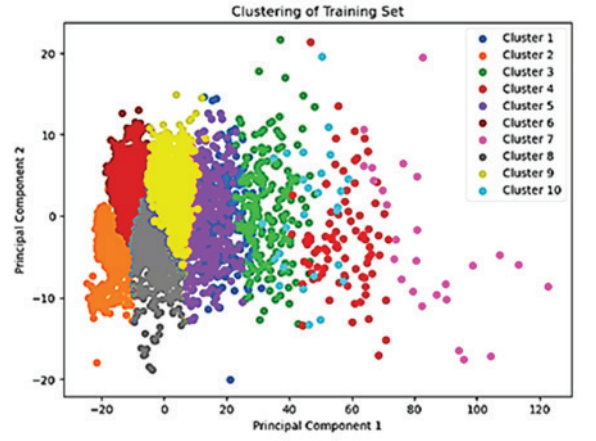


Figure 4: Clusters created using the k-mean clustering method after extraction of features using Efficient NetB0

2.4 SimCLR

SimCLR is a framework for unsupervised visual representation learning for machine learning. SimCLR is trained by sampling different views of the same image, usually with random transformations. A view consists of two different augmentations of the same image. The contrastive loss function is used to minimize the distance between two augmentations of the same image and maximize the minimum distance between two features derived from different views. It is particularly effective when used with greyscale images, for which color is meaningless, as its focus is on texture and structure. For these reasons, we decided to use it in our study, where we evaluate the performance of SimCLR on a dataset of 5,000 underwater images. More importantly, the SimCLR was very helpful for significant feature extraction in such a scenario due to difficulty in labeling, hence making it more understandable and classifiable for underwater environments. The dimensions of training and testing the SimCLR model are (3913, 128) and (979, 128), respectively.

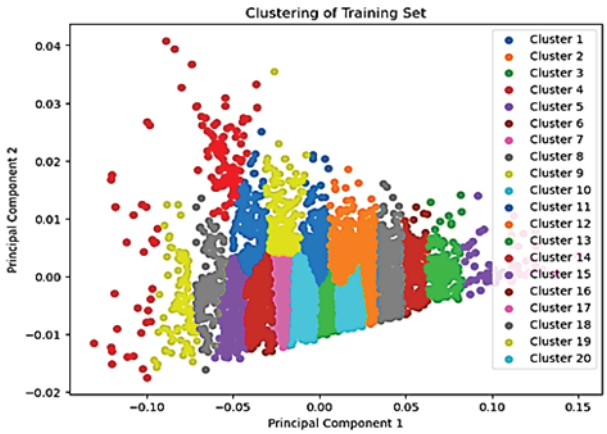


Figure 5: Clusters created using the k-mean clustering method after the extraction of features by SimCLR

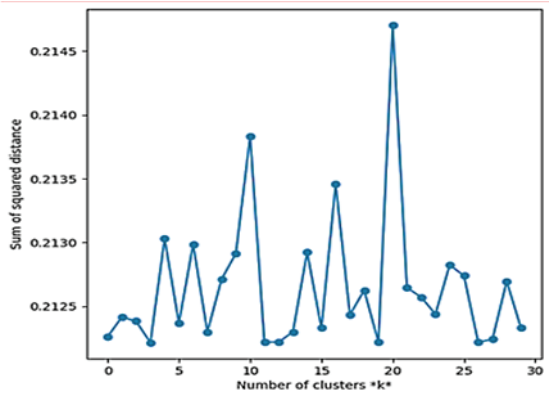


Figure 6: Elbow method showing the number of best-fit clusters to be used for SimCLR model.

3. RESULTS

The research is executed on a dataset consisting of 5,000 underwater images and divided by the Convolutional Neural Network into 80 percent for training and 20 for testing.

Through four comparative models, The K-means clustering is applied to effectively structure the data into some number of best fit clusters. It also applies Principal Component Analysis to reduce the data into important features. Model evaluation is performed with the help of silhouette values, the Calinski-Harabasz index, and the Davies-Bouldin index. The following is a summary of the interpretation for each score.

3.1 Silhouette score:

- Ranges from -1 to 1.
- The closer to 1, means better clustering.
- Around zero, the score indexes overlap clusters.
- Score less than 0 indicates that the samples might have been assigned to the wrong clusters

3.2 Davies-Bouldin index:

- Lower values indicate better clustering.
- The index is minimized when the clusters are well separated and compact.
- Values closer to 0 show better separation and compactness of clusters.

3.3 Calinski-Harabasz Index:

- The higher the better for good clustering.
- The index is maximized to return clusters that are dense and well separated.
- Measures the between-cluster to within-cluster dispersion ratio

The reference values should be taken from the index of Calinski-Harabasz or the index of silhouette, to have an approximate idea of the quality of a model of clustering: high values will mean good clustering. Evaluation of the separation and compactness is carried out with the Davies-Bouldin index, where good fit corresponds to smaller values. The scores for the three models are:
The scores for the VGG-16 model are as follows:
Silhouette score: -0.059330277.
Davies-Bouldin index: 5.560785953586021
Calinski-Harabasz Index: 13.334879549006542
The scores for the InceptionV3 model are as follows:
Silhouette score: -0.028977072.
Davies-Bouldin index: 4.854192245778805
Calinski-Harabasz Index: 11.098701993261846
The scores for EfficientNet B0 model are as follows:
Silhouette score: -0.049650136.
Davies-Bouldin index: 4.9296345644870625
Calinski-Harabasz Index: 31.858081489144915
The scores for the SimCLR model are as follows:
Silhouette score: 0.31599832.
Davies-Bouldin index: 0.9032863892093467
Calinski-Harabasz Index: 7224.351089701941

4. CONCLUSION

This paper presents computer science techniques that improve both efficiency and accuracy of models across wide range of image classification, object detection, image enhancement, semantic segmentation, and transfer learning tasks. Estimation of effectiveness and correctness of models by clustering images with similar features. Next more complex architectures are VGG-16, Inception V3, EfficientNet B0, and SimCLR. VGG-16, deep but easy to interpret architecture, Inception V3, Process images at different scales, EfficientNet B0, Balance between network scale and image resolution, SimCLR, Enhances the feature extraction with contrastive learning. SimCLR compares well to VGG-16, InceptionV3, and EfficientNet B0 for clustering underwater image datasets. It has the highest values for all three of the (Silhouette score, Davies-Bouldin index, and Calinski-Harabasz index) metrics, meaning that it creates the most cohesive, well differentiated, and distinct clusters. PCA makes sure that how well a model performs on observation is consistent with how it behaves. Recent advances in underwater imaging technology can considerably improve marine research, robotic operations, and the ability to maintain and conserve the marine environment.

5. FUTURE WORK

The models are developed for the images taken underwater using deep learning framework, which can enhance the autonomous underwater vehicle for marine life classification and damage detection in real time. This allows ecologists to monitor health and pollution at coral reefs and industries to inspect underwater infrastructure. This method can replace the traditional training approach, which alleviates the need for costly labeled data. This paper presents large and fine models using specially designed neural architectures for self-supervised learning in resource-constrained environments. These models can be used for the autonomous underwater vehicle (AUV) in immediate site assessments and for health monitoring of underwater infrastructures

REFERENCES

[1] Qayyum, W.; Ehtisham, R.; Bahrami, A. R. Assessment of Convolutional Neural Network pre-trained in 2023 materials, MDPI pages 1-16.

[2] Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet classification with deep neurons proceedings in NeuraIPS 2024.pages 1-9

[3] Tan, M.; Le, Q. V. EfficientNet: Rethinking model scaling for Convolutional Neural Networks in 2020 in arxiv journal , pages 1-11

[4] Howard, A. G.; Zhu, M.; Chen, B. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision in MobileNets 2017, page 1-9

[5] Shah, S. R.; Qadri, S.; Bibi, H.; Shah, S. M. W.; Sharif, M. I.; Marinello, F. Comparing inception v3, VGG-16, VGG 19, CNN, and ResNet 50: A case study on early detection of a rice disease in agronomy, MDPI in 2023,pages 1-13.

[6] Katari, R.; Kaza, S. A Comparative Analysis of Variant Deep Learning Models for COVID-19 Protective Face Mask Detection in 2021 in Turkish Journal of Computer and Mathematics Education, page 1-8

[7] Shi, C.; Wei, B.; Wei, S.; Wang, W.; Liu, H.; Liu, J. A quantitative discriminant method of elbow point for the optimal number of clusters in Clustering Algorithm - EURASIP Journal on Wireless Communications and networking in 2021 in EURASIP Journal on wireless communication and networking, pages 1-16

INTELIGENTNY WATERMARKING WIDEO: WYZWANIA TECHNICZNE I PERSPEKTYWY ROZWOJU Z WYKORZYSTANIEM SZTUCZNEJ INTELIGENCJI

INTELLIGENT VIDEO WATERMARKING: TECHNICAL CHALLENGES AND DEVELOPMENT PROSPECTS USING ARTIFICIAL INTELLIGENCE

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: 10.15199/59.2024.4.39

Streszczenie: **W artykule przedstawiono nowoczesne podejścia do watermarkingu wideo wspieranego sztuczną inteligencją, koncentrując się na kluczowych wyzwaniach technicznych oraz perspektywach rozwoju tej technologii. W pierwszej części omówiono rolę AI w poprawie odporności na ataki, wydajności i skalowalności oraz w podnoszeniu jakości znaków wodnych wideo. W drugiej części zaprezentowano perspektywy rozwoju, obejmujące integrację ze strukturami sieci neuronowych, takimi jak Vision Transformers, relacyjne głębokie uczenie i uczenie ze wzmocnieniem.**

Abstract: **The article presents modern approaches to video watermarking supported by artificial intelligence, focusing on key technical challenges and development prospects for this technology. The first part discusses the role of AI in improving attack resistance, performance, and scalability, and in improving the quality of video watermarking. The second part presents development prospects, including integration with neural network structures such as Vision Transformers, relational deep learning, and reinforcement learning.**

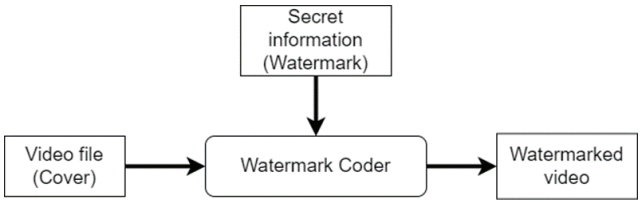
Słowa kluczowe: **odporność znaku wodnego, sieci neuronowe, sztuczna inteligencja, watermarking wideo.**

Keywords: **artificial intelligence, neural networks, robust watermark, video watermarking.**

1. WSTĘP

Watermarking wideo polega na oznakowaniu kontentu wideo widocznym lub niewidocznym znakiem wodnym. Głównym celem znakowania jest ochrona praw autorskich oraz zabezpieczenie integralności materiału wideo, czyli ochrona przed modyfikacją i dalszą nieautoryzowaną redystrybucją materiału, inaczej nazywaną piractwem multimedialnym [1]. Piractwo multimedialne szkodzi nie tylko branży filmowej, powodując starty finansowe, ale jego skutki wpływają na całą gospodarkę światową, powodując redukcję miejsc pracy i likwidację przedsiębiorstw. Jest to wielomiliardowy problem dotykający zarówno kraje rozwinięte jak i rozwijające się [2][3]. Problem dotyka także uczciwych użytkowników końcowych, którzy muszą liczyć się z licznym restrykcjami i ograniczeniami [4].

Skuteczny watermarking powinien charakteryzować się odpornością na celowe ataki, mające za zadanie usunąć zakodowany znak wodny z materiału lub zmodyfikować go. Znakowanie odbywa się poprzez zastosowanie kodera i dekodera, schemat kodera został zaprezentowany na rysunku - Rysunek 1.



Rysunek 1. Schemat kodera wideo [5].

Cyfrowy znak wodny ma na celu osadzenie określonej informacji (znaku wodnego) w pliku multimedialnym. W klasycznych metodach znak wodny ulega transformacji tak, aby mógł zostać osadzony w wybranej domenie, np. we współczynnikach wybranej transformacji. Treść z osadzonym znakiem wodnym jest następnie przywracana do pierwotnej domeny [6].

W literaturze opisano wiele technik i algorytmów znakowania wodnego plików multimedialnych [7]-[10], uwzględniając również problem kompresji stratnej [11], [12]. W klasycznym podejściu do watermarkingu wyróżnia się 3 domeny, w których może być osadzony znak wodny: domena przestrzenna [13]-[15], domena transformat [16]-[20], domena kompresji [21],[22].

Coraz bardziej istotną rolę w technologii watermarkingu zaczyna odgrywać sztuczna inteligencja. Metody klasyczne są coraz częściej wypierane przez podejścia wykorzystujące algorytmy sztucznej inteligencji [23], które charakteryzują się większą wydajnością, ale również większą złożonością obliczeniową. Często spotykane są metody kombinowane, łączące klasyczny watermarking i sztuczną inteligencję, co pozwala połączyć zalety obu podejść [24]. Zawansowane techniki AI znacznie poprawiają skuteczność, odporność i efektywność ochrony treści multimedialnych. Watermarking staje się bardziej dynamiczny, inteligentny i odporny na różnorodne zagrożenia, co stanowi istotny krok naprzód w dziedzinie ochrony praw autorskich i bezpieczeństwa cyfrowego.

Dalsza cześć artykułu została ułożona w następujący sposób: w rozdziale 2 opisano i omówiono główne wyzwania techniczne w watermarkingu wideo z wykorzystaniem algorytmów AI, w rozdziale 3 przedstawiono per-

spektywy rozwoju technologii watermarkingu z wykorzystaniem AI, w rozdziale 4 zaprezentowano wnioski z rozważań.

2. WYZWANIA TECHNICZNE DLA INTELIGENTNEGO WATERMARKINGU

Watermarking wideo, jako technologia zabezpieczania treści cyfrowych, musi stawić czoła szeregowi wyzwań technicznych, które wynikają z dynamicznie zmieniającego się środowiska i postępu technologicznego. Szybki rozwój technologii cyfrowych, rosnąca dostępność narzędzi do edycji wideo oraz coraz bardziej zaawansowane metody ataków na znaki wodne sprawiają, że utrzymanie skuteczności i niezawodności watermarkingu staje się coraz bardziej skomplikowane, a metody klasyczne przestają być wystarczające. Wykorzystanie sztucznej inteligencji, zarówno jako „silnika" algorytmu znakującego, jak i jako wsparcia metody klasycznej pozwala osiągnąć dużo większą efektywność, ale też wprowadza dodatkowe wyzwania techniczne, które należy wziąć pod uwagę przy projektowaniu finalnego rozwiązania. Główne wyzwania, które zostaną omówione w podrozdziale to odporność na ataki i manipulacje, skalowalność i wydajność systemów AI, wpływ AI na jakość materiału wideo.

2.1. Odporność na ataki

Ataki i manipulacje przeprowadzane na znaku wodnym możemy podzielić na niecelowe, które wynikają z konieczności przetwarzania materiału wideo w kanale telekomunikacyjnym oraz celowe, które mają za zadanie usunięcie znaku lub uszkodzenie go. Efektywny algorytm znakujący powinien charakteryzować się wysoką odpornością na zakłócenie, tak aby znak pozostał nienaruszony po odkodowaniu materiału wideo.

Ze względu na coraz większą popularność urządzeń mobilnych jako odbiorników wykorzystywanych do odtwarzania i udostępniania wideo, jednym z najczęstszych zakłóceń znaku wodnego jest proces kompresji wideo. Znak wodny może być odsadzony przed kodowaniem (w dziedzinie nieskompresowanej) lub podczas kodowania. Wykorzystanie sieci neuronowych, które w procesie nauki bazują na dużych ilościach danych, pozwala tworzyć bardziej odporne algorytmy watermarkingu. Odpowiednio zaprojektowana sieć uczy się analizy wzorca przy danym typie ataku np. kompresji stratnej z wykorzystaniem danego rodzaju kodeka, tworząc bardziej odporny algorytm znakujący [25]. Ponadto sieci neuronowe charakteryzują się wysoką zdolnością adaptacji, dynamicznie analizując zawartość wideo i dostosowując parametry znaku wodnego w zależności od specyfiki sceny mogą znacząco zwiększyć trwałość i odporność na manipulacje. W tym celu wykorzystywane są adaptacyjne funkcje aktywacji [26] oraz mechanizm atencji, popularny w modelach językowych [27]. Wykorzystanie mechanizmu atencji pozwala również na optymalizację miejsca umiejscowienia znaku wodnego w ramce wideo poprzez identyfikację najbardziej optymalnego miejsca zapewniającego wyższą odporność przy zachowaniu wymaganej transparentności.

Ponadto systemy AI wykorzystuje się do monitorowania sygnałów wideo pod kątem nietypowych zmian, które mogą wskazywać na próby usunięcia znaku wodnego, co umożliwia szybkie wykrycie i reakcję na takie działania [28].

2.2. Wydajność i skalowalność systemów opartych na sztucznej inteligencji

Algorytm watermarkingu wideo, zwłaszcza oparty na sztucznej inteligencji, musi radzić sobie z przetwarzaniem bardzo dużych ilości danych, często o wysokiej rozdzielczości, przy zachowaniu wymaganej szybkości działania. Obecnie poza trzema standardowymi paradygmatami znakowania tj. odporność, transparentność i pojemność [5] coraz częściej pojawia się czwarte kryterium – zdolność do pracy w czasie rzeczywistym lub zbliżonym do rzeczywistego. Stanowi to duże wyzwanie zwłaszcza przy osadzaniu znaku wodnego w materiałach o bardzo wysokiej rozdzielczości, rzędu 4K lub 8K.

Wykorzystywanie ogromnych zbiorów danych w procesie trenowania algorytmów bazujących na AI skutkuje zapotrzebowaniem na duże zasoby obliczeniowe, w tym głównie procesory graficzne. Ponadto w przypadku wielu systemów, również w procesie ewaluacji (tj. podczas normalnej pracy), dostęp do zasobów obliczeniowych będzie kluczowy dla osiągniecia maksymalnej efektywności działania, możliwej dla danego systemu. Zwiększenie efektywności jest możliwe dzięki zastosowaniu przetwarzania równoległego, które wymaga złożonej architektury systemowej oraz efektywnego zarządzania dostępnymi zasobami [29]. Systemy watermarkingu muszą być elastyczne i zdolne do skalowania w zależności od bieżących potrzeb. Oznacza to umiejętność szybkiego dostosowywania się do zmieniających się wymagań, takich jak np.: wzrost liczby plików wideo do przetworzenia.

Rozwiązaniem wielu z powyższych problemów jest wsparcie klasycznych metod watermarkingu, które charakteryzują się prostotą i szybkością działania, przez rozwiązania AI w zakresie optymalizacji dystrybucji obciążenia, co dodatkowo przyspiesza działanie systemu. Systemy oparte na sztucznej inteligencji mogą być wykorzystywane do zarządzania zasobami [30], dystrybucji obciążenia [31] oraz predykcji przyszłych obciążeń na podstawie danych historycznych.

2.3. Jakość i percepcja w watermarkingu wideo

Jakość i percepcja są kluczowymi aspektami, które decydują o skuteczności i akceptowalności danej technologii w kontekście watermarkingu wideo. Znak wodny musi być trwały, jednak osadzony w taki sposób by pozostał niewidoczny dla widza. Konieczność modyfikacji oryginalnego obrazu, wynikająca z umieszczenia w nim znaku wodnego może w różnym stopniu wpływać na jego jakość, powodując m.in.: artefakty wizualne takie jak szumy czy zmiany kolorów, degradację wideo (zwłaszcza o wysokiej rozdzielczości) zaburzenia w efektywności kompresji wideo. Jakość materiału wideo jest bardzo istotna np. w przypadku reklam, gdzie od jakości spotu reklamowego zależy skuteczność i rentowność kampanii [32].

Wykorzystanie sztucznej inteligencji pozwala na uzyskanie wyższej jakości przy zachowaniu pozostałych

parametrów na satysfakcjonującym poziomie. Popularnym podejściem jest wykorzystanie modeli typu super-resolution [33], [34] w celu poprawy jakości obrazu po osadzeniu znaku wodnego. Znak wodny może zostać osadzony w materiale o niższej rozdzielczości, co pozwoli na redukcje artefaktów oraz przyspieszy procedurę osadzania, a następnie przetworzony do rozdzielczości docelowej z wykorzystaniem modelu AI. Ponadto wykorzystanie wspomnianych wcześniej algorytmów bazujących na atencji pozwala lepiej analizować treść wideo i identyfikować najlepsze miejsca do umieszczenia znaku wodnego, które są mniej zauważalne dla ludzkiego oka. Na przykład, znaki wodne mogą być umieszczane w obszarach o mniejszym znaczeniu wizualnym lub w miejscach z dużą ilością ruchu, gdzie są mniej widoczne.

Ocena jakości wizualnej materiału wideo jest subiektywna i może różnić się w zależności od widza. To, co jedna osoba uznaje za akceptowalne, inna może uważać za zupełnie nieakceptowalne. Dobrym rozwiązaniem może okazać zastosowanie personalizacji z wykorzystaniem AI – na podstawie analizy preferencji użytkownika, algorytm może dostosować sposób znakowania do indywidulanych potrzeb, co pozwoli na stworzenie bardziej spersonalizowanych i akceptowalnych rozwiązań. Implementacja takiego rozwiązania może dotyczyć bardzo subtelnych zmian, jak modyfikacja wartości współczynnika wagowego w funkcji straty, co pozwala położyć odpowiednio mniejszy lub większy nacisk na widoczność danego typu zakłócenia.

3. PERSPEKTYWY ROZWOJU

Rozwiązania watermarkingu wideo bazujące na sztucznej inteligencji to stosunkowo nowe technologie, jedno z pierwszych takich podejść, bazujące na architekturze autoenkodera pojawiło się w 2017 roku [35]. Prowadzonych jest wiele badań testujących nowe trendy i podejścia oraz ich wpływ na efektywność watermarkingu wideo.

Sieci neuronowe są ciągle w fazie rozwoju, po ostatniej rewolucji dokonanej przez modele typu Transormers [36] pojawiły się również modele typu Vision Transformers, które zaadaptowały architekturę do rozwiązywania zadań związanych z obrazami. Modele dzielą obraz na małe fragmenty i przetwarzają je jako sekwencję [37], co może znaleźć zastosowanie w watermarkingu wideo.

Nowatorskim podejściem, które zadebiutowało w zeszłym roku jest relacyjne głębokie uczenie. które łączy głębokie sieci neuronowe z analizą danych relacyjnych. W tradycyjnych bazach danych informacje są często zorganizowane w relacyjny sposób, co oznacza, że dane są przechowywane w tabelach powiązanych ze sobą za pomocą kluczy. Relacyjne głębokie uczenie umożliwia bezpośrednie wykorzystanie tych struktur danych w procesie uczenia maszynowego, eliminując konieczność ręcznego przekształcania danych do formy płaskiej [38]. Podeście jest aktualnie wykorzystywane w analizie biznesowej, zarządzaniu łańcuchem dostaw oraz w analizie danych medycznych. Jednak może znaleźć zastosowanie również w problemach telekomunikacyjnych przy optymalizacji przepustowości i zarządzaniu zasobami, co może znacznie przyspieszyć działanie złożonych algorytmów watermarkingu.

Efektywne może okazać się również wsparcie technologii watermarkingu przez uczenie ze wzmocnieniem. Jest to rodzaj uczenia maszynowego w którym agent uczy się podejmować decyzje poprzez interakcję z otoczeniem. Uczenie ze wzmocnieniem pozwala na dynamiczne dostosowywanie strategii osadzania znaku wodnego do wymagań użytkownika oraz rodzaju kontentu w czasie rzeczywistym .

4. PODSUMOWANIE

W artykule omówiono nowoczesne podejścia do watermarkingu wideo wspieranego sztuczną inteligencją, koncentrując się na wyzwaniach technicznych i perspektywach rozwoju. Przedstawiono kluczową rolę AI w poprawie odporności na ataki, wydajności i skalowalności oraz jakości i percepcji znaków wodnych w wideo. Watermarking wideo znajduje zastosowanie w wielu obszarach, głównie jest to przemysł rozrywkowy, gdzie znakowanie wodne jest wykorzystywane do ochrony praw autorskich i integralności materiału wideo. Udział watermarkingu wideo w rynku ciągle rośnie stąd potrzeba rozwoju i udoskonalenia tej technologii. Uczenie ze wzmocnieniem, oraz inne opisane techniki umożliwiają dynamiczne dostosowanie się do zmieniających się warunków i zagrożeń, co jest kluczowe w zapewnianiu integralności i bezpieczeństwa treści multimedialnych. Pozwalają również na stosowanie watermarkingu w systemach działających w czasie rzeczywistym, nawet w bardzo wysokich rozdzielczościach.

Perspektywy rozwoju w tej dziedzinie obejmują dalsze badania nad integracją zaawansowanych struktur sieci neuronowych. Prace w tym obszarze powinny skupić się na optymalizacji tych technologii, aby sprostać rosnącym wymaganiom w zakresie ochrony praw autorskich i zarządzania treściami cyfrowymi, bez wpływu na komfort oglądania użytkownika końcowego.

LITERATURA

[1] Danaher Brett, Smith Michael D., Telang Rahul, 2014. „Piracy and Copyright Enforcement Mechanisms”, *Innovation Policy and the Economy* vol 14.

[2] Ipsos MediaCT. 2011. “Economic consequences of movie piracy – Australia”.

[3] DGA. 2010. “Piracy Statistics – Piracy by the Numbers”

[4] Greenberg Matthew. 2015. “The Economics of Video Piracy” *PIT Journal*, 6.

[5] Bistroń M, Piotrowski Z. Efficient Video Watermarking Algorithm Based on Convolutional Neural Networks with Entropy-Based Information Mapper. Entropy (Basel). 2023 Feb 2;25(2):284

[6] Li Yue, Wang Hongxia, Barni Mauro. 2021. “A survey of deep neural network watermarking techniques”, *Neurocomputing*. 461, C, pp. 171-193.

[7] Podilchuk Christine, Delp Edward J. 2001. “Digital watermarking: algorithms andapplications,” *IEEE Signal Processing Magazine*, vol. 18, no. 4, pp.33–46.
[8] Langelaar Gerrit C., Setyawan Iwan, Lagendijk Reginald L. 2000. “Watermarking digital image and video data. A state-of-the-art overview,” *IEEE Signal Processing Magazine*, vol. 17, no. 5, pp. 20–46.
[9] Do¨err Gwenaël, Dugelay Jean-Luc. 2003. “A guide tour of video watermarking”, *Signal Processing: Image Communication*, vol. 18, no. 4, pp. 263–282.
[10] Lin Eugene T. i inni. 2005. “Advancesin digital video content protection” *Proceedings of the IEEE*, vol. 93, no. 1, pp. 171–183.
[11] Stutz Thomas, Uhl Andreas. 2012. “A survey of H.264 AVC/SVC encryption”, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 3, pp. 325–339.
[12] Tew Yigi, Wong Koksheik. 2014. “An overview of information hiding in H.264/AVC compressed video”, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 2, pp. 305–319.
[13] Lancini Rosa, Mapelli Francesca, Tubaro Stefano. 2002. “A robust video watermarking technique in the spatial domain”, *International Symposium on VIPromCom Video/Image Processing and Multimedia Communications 2002*, pp. 251-256.
[14] Jobin Abraham, Varghese Paul. 2019. “An imperceptible spatial domain color image watermarking scheme”, *Journal of King Saud University - Computer and Information Sciences*, 31, 1, pp. 125-133.
[15] Chen Guangxi i inni 2018. “A Robust Video Watermarking Algorithm Based on Spatial Domain”, *Proceedings of the 2018 7th International Conference on Energy and Environmental Protection (ICEEP 2018)*, pp. 412-419.
[16] Carli Marco, Mazzeo Rafe, Neri Alessandro. 2006. “Video watermarking in 3D DCT domain”, *14th European Signal Processing Conference*, pp. 1-5,
[17] Campisi Patrizio, Neri Alessandro. 2005. “Video watermarking in the 3D-DWT domain using perceptual masking”, *IEEE International Conference on Image Processing 2005*, pp. I-997. https://doi.org/10.1109/ICIP.2005.1529921.
[18] Lee Young-Yoon, Jung Han-Seung, Lee Sang-Uk. 2003. “3D DFT-based video watermarking using perceptual models”, *2003 46th Midwest Symposium on Circuits and Systems*, pp. 1579-1582.
[19] Kulkarni Tejas, Dewan Jaya. 2016. “Digital video watermarking using Hybrid wavelet transform with Cosine, Haar, Kekre, Walsh, Slant and Sine transforms”, *2016 International Conference on Computing Communication Control and automation (ICCUBEA)*, pp. 1-5.
[20] Panyavaraporn Jantana, Horkaew Paramate. 2018. “DWT/DCT-based Invisible Digital Watermarking Scheme for Video Stream”, *2018 10th International Conference on Knowledge and Smart Technology (KST)*, pp. 154-157.
[21] Ding Hao i inni. 2021. “A Compressed-Domain Robust Video Watermarking Against Recompression Attack”, *IEEE Access* 2021, 9, pp. 35324-35337.
[22] Lee Min-Jeong i inni. 2012. “Real-time video watermarking system on the compressed domain for high-definition video contents: Practical issues”, *Digital Signal Processing 2012*, 22, 1, pp. 190-198.
[23] El'arbi Maher, Amar Chokri Ben, Nicolas Henri. 2006. “Video Watermarking Based on Neural Networks”, *2006 IEEE International Conference on Multimedia and Expo*, pp. 1577-1580.
[24] Hu Xuelong i inni. 2008. “Robust blind watermark algorithm of color image based on neural network”, *2008 International Conference on Neural Networks and Signal Processing*, pp. 430-433.
[25] Zhang Kevin A. i inni. 2019. “Robust Invisible Video Watermarking with Attention”, *ArXiv:* 1909.01285.
[26] Palnitkar Roshan, Cannady James. 2004. "A Review of Adaptive Neural Networks", *IEEE SoutheastCon, 2004. Proceedings*, pp. 38-47.
[27] Vaswani Ashish i inni. 2017. “Attention is All you Need”, *arXiv:*1706.03762.
[28] Alzahrani Ali, 2022. “Detecting Digital Watermarking Image Attacks Using a Convolution Neural Network Approach*”, Security and Communication Network* 2022.
[29] Molina-Garcia Javier i inni. 2018. "Parallel scheme in multicore processors for video authentication based on watermarking", *2018 IEEE 9th International Conference on Dependable Systems, Services and Technologies (DESSERT)*, pp. 625-630.
[30] Jeong Byeonghui i inni. 2023. “Stable and efficient resource management using deep neural network on cloud computing”, *Neurocomputing* 521, C, pp. 99-112.
[31] Pokuta Waldemar, Zatwarnicki Krzysztof. 2022. "Application of Neural Networks in Distribution of the Load in Cluster-Based Web Systems" *Applied Sciences* 12,79.
[32] Moon Jang Ho. 2004. “The Impact of Video Quality and Image Size on the Effectiveness of Online Video Advertising on YouTube”, *Informs Journal on Computing* 10, pp. 23-29.
[33] Zamfir Eduard, Conde Marcos V., Timofte Radu. 2023. “Towards Real-Time 4K Image Super-Resolution”, *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1522-1532.
[34] Hassan Moiz, Kandasamy Illanko, Xavier N. Fernando. 2024. "Single Image Super Resolution Using Deep Residual Learning" *AI* 5, no. 1, pp. 426-445.
[35] Baluja Shumeet. 2017. “Hiding Images in Plain Sight: Deep Steganography”, *Advances in Neural Information Processing Systems 30 (NIPS 2017)*.
[36] Lin Tianyang i inni. 2021.“A survey of transformers”, *ArXiv:* 2106.04554.
[37] Dosovitskiy Alexey i inni. 2020. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”, *arXiv*:2010.11929.
[38] Fey Matthias i inni. 2023, “Relational Deep Learning: Graph Representation Learning on Relational Databases”, *arXiv*:2312.04615.

Zastosowanie atencji w sieciach neuronowych do zaawansowanego znakowania wodnego obrazów w wysokiej rozdzielczości: Analiza efektywności i metryk

Application of Attention in Neural Networks for Advanced Watermarking of High-Resolution Images: Efficiency and Metrics Analysis

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: 10.15199/59.2024.4.40

Streszczenie: **W artykule prezentujemy innowacyjny algorytm znakowania wodnego obrazów, wykorzystujący głębokie sieci neuronowe z modułem atencji. Proponowane rozwiązanie pozwala na wysoce transparentne zakodowanie ukrytego znacznika w obrazach o wysokiej rozdzielczości, zachowując ich oryginalną jakość. Dzięki zastosowaniu miękkiej atencji, nasz system osiąga doskonałe wyniki w zakresie odporności znaków wodnych. Demonstrujemy efektywność metody na przykładach wraz z wykazem metryk.**

Abstract: **In the article, we present an innovative image watermarking algorithm that utilizes deep neural networks with an attention module. The proposed solution allows for highly transparent encoding of a hidden watermark in high-resolution images, while preserving their original quality. Thanks to the use of soft attention, our system achieves excellent results in terms of watermark durability. We demonstrate the effectiveness of the method with examples and a list of metrics.**

Słowa kluczowe: **znakowanie wodne, uczenie głębokie, przetwarzanie obrazu, miękka atencja.**

Keywords: **watermarking, deep learning, image processing, soft attention.**

1. WSTĘP

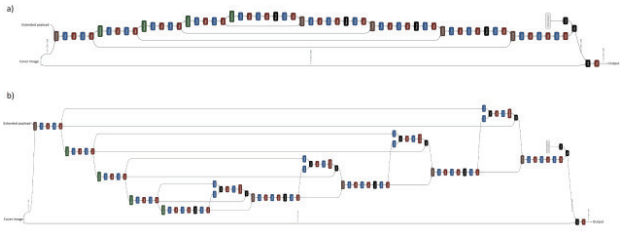
W erze cyfrowej charakteryzującej się dynamicznym rozwojem treści multimedialnych generowanych przez twórców na różnorodnych platformach społecznościowych oraz usługach streamingowych typu Video on Demand (VoD) istotnie wzrosło zapotrzebowanie na skuteczne metody ochrony praw autorskich. Znakowanie wodne, jako jedna z kluczowych technologii zabezpieczających odgrywa coraz większą rolę w zapewnianiu integralności i autentyczności cyfrowych mediów. Tym samym rozwój nowoczesnych technik znakowania, które są zarówno dyskretne, jak i trwałe, staje się niezbędny w kontekście ciągłej walki z naruszeniami praw autorskich w środowisku cyfrowym.

W odpowiedzi na rosnące zapotrzebowanie na skuteczne metody ochrony praw autorskich prezentujemy nowatorski algorytm znakowania wodnego obrazów, wykorzystujący głębokie sieci neuronowe wspomagane nowoczesnymi rozwiązaniami z tej dziedziny, takimi jak moduł atencji. Omawiane rozwiązanie charakteryzuje się odpornością na standardowy zestaw zakłóceń występujących w obrazach, takich jak szum czy kompresja. Zdaniem autorów dodatkowo wyróżnia się wysoką pojemnością, wystarczającą do zakodowania dowolnego ukrytego znacznika. Proponowane rozwiązanie opiera się na klasycznym koncepcie kodera i dekodera znaku wodnego, co zapewnia jego skuteczność i efektywność.

2. OPIS ALGORYTMU

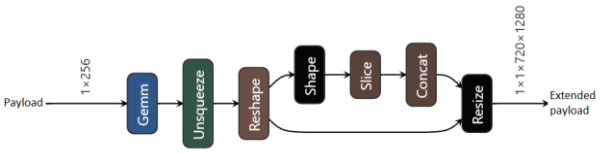
2.1. Architektura

Podstawą omawianego rozwiązania jest sieć Unet [1] wraz z modułem atencji, znana jako Attention Unet [2], będąca rozwinięciem standardowej architektury Unet. Attention Unet wprowadza mechanizmy uwagi, które pozwalają na skupienie się na istotnych cechach obrazu, zwiększając precyzję działania sieci. Różnice w złożoności obu modeli ilustruje Rys. 1.



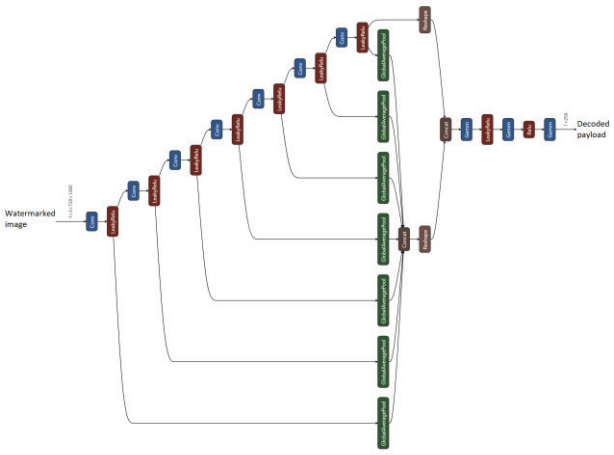
Rys. 1. Przedstawienie architektur: a) Unet b) Attention Unet

Jak przedstawiono na powyższym rysunku, architektura wzbogacona o moduł atencji staje się nieco bardziej złożona. W każdej z warstw dodatkowy mechanizm uwagi jest odpowiedzialny za dynamiczne wagowanie cech, co pozwala sieci skoncentrować się na istotnych obszarach obrazu w trakcie przetwarzania. W przeciwieństwie do tradycyjnych perceptronów, te mechanizmy pozwalają na adaptacyjne dostosowywanie się do zmieniającej się istotności cech, co przyczynia się do zwiększenia precyzji w wykrywaniu i kodowaniu ukrytych znaczników.



Rys. 2. Przedstawienie architektury sieci pre-processingowej

Tak jak przedstawiono na *Rys. 1* przed wprowadzeniem do sieci kodującej obraz bazowy jest łączony z ukrytą informacją. W prezentowanym systemie ukrytą wiadomością jest wektor zawierający sekwencję binarną. W związku z tym konieczne jest rozszerzenie wymiarów wektora do postaci macierzy, co umożliwia jego konkatenację z obrazem bazowym jako dodatkowy czwarty kanał. Tak przygotowana macierz jest następnie wprowadzana do wejścia kodera. W celu realizacji wspomniane rozszerzenie wymiarów niezbędna jest sieć przetwarzająca zaprezentowana na Rys. 2.

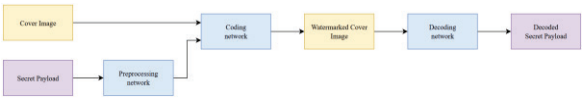


Rys. 3. Przedstawienie architektury sieci dekodującej

Produktem sieci kodującej jest obraz z nałożonym wysoce transparentnym znakiem wodnym. Jest to macierz trzykanałowa o 8-bitowej głębokości koloru. Zadaniem sieci przedstawionej na *Rys. 3* – sieci dekodującej jest przekształcenie obrazu (macierzy) z powrotem do formy wektora i zdekodowanie zawartości znaku wodnego. Proces ten wykorzystuje zaawansowane techniki przetwarzania cech obrazu na różnych poziomach abstrakcji, co umożliwia efektywne wydobywanie ładunku znaku wodnego, nawet jeśli opiera się on na subtelnych zmianach w obrazie.

2.2. Przepływ danych

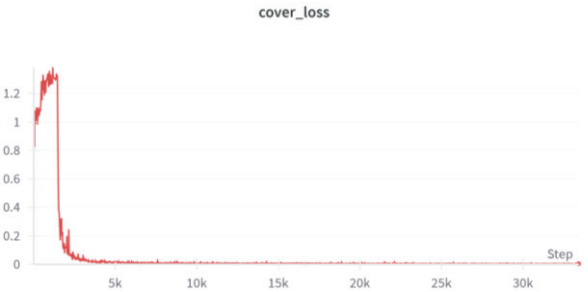
Wszystkie trzy wymienione sieci neuronowe tworzą łańcuch przepływu danych przedstawiony na *Rys. 4*. Jak wspomniano, system operuje na obrazach nośnych o wysokiej rozdzielczości przetwarzając cały obraz, jako jeden spójny element w przeciwieństwie do wielu obecnych systemów znakowania wodnego, które z powodu wysokich kosztów obliczeniowych lub skomplikowania często używają tylko fragmentu obrazu lub jego znacznie zmniejszonej wersji. Zdecydowano się na użycie obrazów o rozdzielczości 1280x720 pikseli, co odpowiada klatce filmu HD w formacie 16:9. Taki system przetwarzania umożliwia rozproszenie artefaktów znakowania wodnego po znacznie większym obszarze, co znacząco zwiększa transparentność rozwiązania. Skryty znacznik jest stemplem binarnym o pojemności wynoszącej 256 bitów (taka jest również długość wektora wejściowego do systemu), co przekłada się na gęstość informacji na poziomie 0.000278 bita na piksel



Rys. 4. Przepływ danych w opisywanym systemie

2.3. Proces treningu

Przedstawione sieci neuronowe były trenowane razem w cyklu treningowym, aby wzajemnie poprawiać swoje wyniki. Dekoder uczył się rozpoznawać i dekodować coraz to nowe schematy wprowadzane przez koder. Trening bazował na własnym zbiorze danych zawierającym ponad 9 tysięcy próbek klatek filmowych o różnej zawartości, co miało na celu jak najlepsze uogólnienie działania sieci. Modele trenowano przez 20 epok na komputerze z 48 GB RAM oraz 48 GB pamięci GPU i procesorem Intel o 9 rdzeniach.
W celu optymalizacji zbieżności modelu podczas treningu zastosowano zróżnicowany system wag. W początkowej fazie treningu podczas pierwszej epoki wagi funkcji kosztu dla transparentności zostały ustawione na zero, co skupiło uwagę modelu wyłącznie na poprawnej ekstrakcji ukrytego ładunku. W miarę postępu treningu waga odpowiedzialna za transparentność rosła znacznie szybciej niż waga dotycząca sekretnego ładunku. Dzięki temu, sieć, która nauczyła się już ekstrahować ukryty ładunek zaczynała coraz bardziej skupiać się na minimalizowaniu widoczności swoich działań. Takie podejście umożliwia modelowi stopniowe i efektywne przyswajanie poszczególnych umiejętności, co przekłada się na lepsze rezultaty w bardziej złożonych etapach treningu.



Rys. 5. Przebieg wartości funkcji kosztu dla transparentności znaku wodnego

Szczególnie interesujące jest zilustrowanie momentu, kiedy waga transparentności przyjmuje wartość większą od zera, co można zaobserwować na Rys. 5. Wykres ten prezentuje wartości funkcji kosztu w zależności od kroku algorytmu (przejścia przez trening jednego batcha danych). Po zakończeniu pierwszej epoki (przejścia wszystkich batchy ze zbioru treningowego) funkcja kosztu dla transparentności zaczyna gwałtownie maleć, ponieważ jej waga staje się większa od zera. Tymczasem wartość funkcji kosztu związanej z ukrytym stemplem

binarnym (Rys. 6) maleje od samego początku treningu, ponieważ przypisano jej niezerową wagę już na starcie procesu uczenia. W trakcie procesu uczenia wartość ta rośnie jednak od pewnego moment z powodu wzrostu wagi transparentności.

Rys. 6. Przebieg uśrednionych wartości funkcji kosztu dla ustalonej pojemności danych znaku wodnego podczas treningu z zakłóceniami nośnika

Jak wcześniej wspomniano w procesie treningu wykorzystano dwie funkcje kosztu, które były odpowiedzialne za transparentność znaku wodnego oraz spójność zdekodowanego znacznika. Pierwsza funkcja kosztu opierała się na trzech miarach. Pierwszą z nich jest MSE (Mean Squared Error), który mierzy średnią kwadratową różnic między oryginalnymi, a przewidywanymi wartościami, służąc do oceny dokładności rekonstrukcji obrazu. Drugą miarą jest SSIM (Structural Similarity Index Measure), który ocenia jakość obrazu na podstawie percepcyjnego podobieństwa do oryginału, biorąc pod uwagę zmienne takie jak tekstura, kontrast i luminancja. Trzecią miarą jest LPIPS[6] (Learned Perceptual Image Patch Similarity), który wykorzystuje głębokie sieci neuronowe do oceny podobieństwa perceptualnego między obrazami, co pozwala na bardziej zniuansowane porównania niż tradycyjne metody statystyczne. Miarą dla drugiej z wymienionych funkcji kosztu była binarna entropia krzyżowa, która oceniała, jak skutecznie algorytm przewiduje przynależność danej wartości do jednej z dwóch klas, 0 lub 1, odpowiadających wartościom binarnym.

2.4. Odporność algorytmu

W celu dostosowania algorytmu do potencjalnego zastosowania w warunkach produkcyjnych, czy też przynajmniej realistycznych do procesu treningowego wprowadzono elementy symulujące to środowisko w celu przygotowania algorytmu do rozpoznawania i odpowiedniego reagowania na te zniekształcenia. Wykorzystano różne rodzaje zakłóceń: szum gaussowski, rozmycie gaussowskie, rotację obrazu, przysłonięcie części obrazu oraz kompresję JPEG. Niezbędne było zapewnienie różniczkowalności wszystkich tych operacji, co osiągnięto dzięki wykorzystaniu odpowiednich do tego celu bibliotek [7][8]. Szansa na wystąpienie danego zakłócenia w iteracji wynosiła 50%, dzięki czemu różne etapy treningu były obciążone różnymi kombinacjami przekształceń.

2.5. Uzyskane rezultaty

Wytrenowane modele, które opisano wcześniej zostały poddane ewaluacji na zbiorze testowym składającym się z 1000 obrazów, które nie były częścią zestawu treningowego. Ewaluacja przeprowadzona została w dwóch scenariuszach: z zakłóceniami oraz bez zakłóceń, aby stwierdzić, jak model radzi sobie w różnych warunkach. Wyniki tej oceny zostały przedstawione w Tab. 1 oraz Tab. 2 gdzie zebrano metryki oceniające wydajność sieci. Test przeprowadzano na komputerze wyposażonym w dedykowaną kartę graficzna NVIDIA GTX 1650 z 4 GB pamięci VRAM, 16 GB pamięci RAM oraz 8 rdzeniowym procesorem Intel.

Tab. 1. Otrzymane metryki dla testu bez zakłóceń

PSNR (dB)	*SSIM*	*BER (%)*	Czas inferencji (s)
51.8725	0.9951	1,2%	Enkoder 0.007 Dekoder 0.0018

Tab. 2. Otrzymane metryki dla testu z zakłóceniami

PSNR (dB)	*SSIM*	*BER (%)*	Czas inferencji (s)
51.8725	0.9951	24,86%	Enkoder 0.007 Dekoder 0.0018

Ponieważ zakłócenia są wprowadzane na etapie już oznakowanego obrazu (po przetworzeniu przez koder), nie mają one wpływu na transparentność obrazu, ale mogą wpływać na poprawność odczytu zakodowanego ładunku. Do pomiaru tego aspektu wykorzystano współczynnik błędu bitowego (BER). Wprowadzenie zakłóceń znacząco zwiększa błąd odczytu danych, podobnie jak to miało miejsce podczas fazy treningu, gdzie jest to efekt kombinacji różnych rodzajów zniekształceń. Tab. 3 przedstawia wpływ zniekształceń na wyniki, prezentując wartości BER w kontekście zastosowanych zakłóceń.

Tab. 3. Wpływ poszczególnych zakłóceń na BER

Nazwa zakłócenia	*BER (%)*
Random Rotation	2,77
Random Erasing	47,26
JPEG compression	49,6
Gaussian Noise	49,4

Prezentowane rozwiązanie osiąga bardzo dobre wyniki w porównaniu do czołowych algorytmów znakowania wodnego opartych na głębokich sieciach neuronowych. Uzyskane metryki obrazujące efektywność metody w warunkach z dodanymi zniekształceniami przedstawiono w Tab. 4. Opisywany algorytm wyróżnia się bardzo wysokim poziomem transparentności, co jest efektem zastosowania innowacyjnej sieci neuronowej oraz możliwości umieszczenia znaku wodnego na obrazie o wysokiej rozdzielczości. Współczynnik SSIM jest tylko nieznacznie niższy niż w przypadku zaawansowanej architektury HiNet. Współczynnik błędu bitowego

(BER) również jest akceptowalny, o ile zakłócenia nie występują w dużym natężeniu.

Tab. 4. Porównanie otrzymanych wyników transparentnosci z innymi metodami głębokiego znakowania

Metoda	*PSNR (dB)*	*SSIM*	*Odporność na szum*
4bit-LSB	33.19	0.9453	NIE
HiDDeN[3]	35.21	0.9691	TAK
Weng et al. [4]	39.75	0.9765	NIE
HiNet[5]	48.99	0.9971	NIE
Opisana w tym artykule	51.8725	0.9951	TAK

3. Podsumowanie

Podsumowując, omawiana metoda stanowi innowacyjne podejście do problemu znakowania wodnego obrazu opartego o głębokie sieci neuronowe. Dzięki zastosowaniu sieci z mechanizmem uwagi udało się uzyskać bardzo dobre wyniki w porównaniu do innych metod. W ramach ewaluacji przeprowadzonej na zbiorze testowym liczącym 1000 zdjęć osiągnięto zadowalające rezultaty. Algorytm cechuje się znaczną odpornością, co pozwala na efektywne zaadresowanie sytuacji, gdzie nośnik jest uszkodzony. Oprócz tego została zapewniona zadowalająca pojemność systemu, co podkreśla jego efektywność i potencjalne zastosowanie w praktycznych scenariuszach znakowania wodnego.

LITERATURA

[1] Ronneberger, Olaf & Fischer, Philipp & Brox, Thomas. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. LNCS. 9351. 234-241. 10.1007/978-3-319-24574-4_28.

[2] Oktay, Ozan & Schlemper, Jo & Folgoc, Loic & Lee, Matthew & Heinrich, Mattias & Misawa, Kazunari & Mori, Kensaku & McDonagh, Steven & Hammerla, Nils & Kainz, Bernhard & Glocker, Ben & Rueckert, Daniel. (2018). Attention U-Net: Learning Where to Look for the Pancreas.

[3] Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In Proceedings of the European Conference on Computer Vision (ECCV), pages 657–672, 2018

[4] Xinyu Weng, Yongzhi Li, Lu Chi, and Yadong Mu. High- capacity convolutional video steganography with temporal residual modeling. In Proceedings of the 2019 on Interna- tional Conference on Multimedia Retrieval, pages 87–95, 2019.

[5] J. Jing, X. Deng, M. Xu, J. Wang and Z. Guan, "HiNet: Deep Image Hiding by Invertible Network," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 4713-4722, doi: 10.1109/ICCV48922.2021.00469. keywords: {Training;Learning systems;Computer vision;Wavelet domain;Codes;Neural networks;Receivers;Low-level and physics-based vision;Vision applications and systems}

[6] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

[7] E. Riba, D. Mishkin, D. Ponsa, E. Rublee and G. Bradski, "Kornia: an Open Source Differentiable Computer Vision Library for PyTorch," 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass, CO, USA, 2020, pp. 3663-3672, doi: 10.1109/WACV45572.2020.9093363.

[8] C. Reich, B. Debnath, D. Patel and S. Chakradhar, "Differentiable JPEG: The Devil is in the Details," in 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2024 pp. 4114-4123. doi: 10.1109/WACV57701.2024.00408

SYSTEM AUTOMATYCZNEGO MIKSOWANIA SYGNAŁÓW MUZYCZNYCH Z WYKORZYSTANIEM JEDNOWYMIAROWYCH AUTOENKODERÓW WAVE-U-NET

AUTOMATIC AUDIO SIGNAL MIXING SYSTEM BASED ON ONE-DIMENSIONAL WAVE-U-NET AUTOENCODERS

Damian Koszewski[1]

[1] Politechnika Gdańska, Gdańsk, damian.koszewski@gmail.com

DOI: 10.15199/59.2024.4.41

Streszczenie: **Celem niniejszego artykułu jest zaprezentowanie systemu miksowania muzyki, który w sposób automatyczny miksuje surowe nagrania z dobrą jakością, niezależnie od gatunku muzycznego. Zaproponowano architekturę opartą na jednowymiarowych autoenkoderach Wave-U-Net oraz wykonano szereg testów subiektywnych i obiektywnych. Uzyskane wyniki dowodzą, że miksy stworzone automatycznie przez zaproponowany system mogą być oceniane równie wysoko jak miksy przygotowane profesjonalnie.**

Abstract: **The purpose of this article is to develop an automatic song-mixing system that is capable of automatically mixing a song with good quality in any music genre. An architecture based on one-dimensional Wave-U-Net autoencoders is proposed for automatic music mixing. Both objective and subjective tests were conducted. The achieved results prove that mixes created automatically by Wave-U-Net can be evaluated as highly as mixes created professionally.**

Słowa kluczowe: **autoenkoder Wave-U-Net, automatyczne miksowanie muzyki, parametryzacja sygnału muzycznego, testy odsłuchowe.**

Keywords: **automatic music mixing, listening tests, music signal parametrization, Wave-U-Net autoencoder.**

1. WSTĘP

Proces produkcji utworu muzycznego może być podzielony na następujące etapy: kompozycja, nagranie, edycja, miks oraz mastering. Wzrastająca potrzeba automatyzacji poszczególnych kroków procesu produkcji muzycznej zainspirowała tę pracę, która skupia się na fazie miksowania utworu muzycznego. Definicyjnie miks można określić jako proces obróbki pojedynczych ścieżek dźwiękowych prowadzący do połączenia nagrania wielośladowego w jeden, stereofoniczny plik muzyczny [4]. Wybór tego właśnie etapu w badaniach został podyktowany tym, że miksowanie jest istotnym etapem produkcji muzycznej [12] i prawdopodobnie najważniejszym krokiem w zapewnieniu tego, że bez względu na sposób, w jaki muzyka została nagrana, będzie ona oddziaływać z odpowiednią estetyką i przekazywać artystyczną wypowiedź zgodną z intencją muzyka [11]. Jest to etap zarówno techniczny, jak i kreatywny. Ponadto ten właśnie etap jest obecnie najczęściej poddawany próbom automatyzacji [3, 10, 14].

Celem pracy jest stworzenie systemu do automatycznego miksu utworów, który otrzymując na wejściu nagrane i wyedytowane ścieżki, jest w stanie w sposób automatyczny (bez ingerencji użytkownika) zmiksować utwór z dobrą jakością w dowolnym gatunku muzycznym.

2. METODOLOGIA GŁÓWNEGO EKSPERYMENTU

2.1. Kompletowanie danych

W celu uzyskania odpowiedniej ilości danych treningowych wykorzystano bazę MUSDB18-HQ [13], która zawiera grupy instrumentów: perkusyjne, basowe, wokalne, inne oraz gotowy miks. Baza ta jest oparta na bazie Cambridge [6], która zawiera pojedyncze ścieżki do utworów. W ten sposób zebrano 150 utworów (około 10 godzin) z różnych gatunków muzycznych.

Dodatkowo, pięć nowych utworów zostało nagranych oraz zmiksowanych przez autora w celu powiększenia bazy danych. Wszystkie utwory zostały nagrane w audytorium Wydziału Elektroniki i Telekomunikacji Politechniki Gdańskiej.

2.2. Trening i walidacja

Zaproponowano architekturę sieci neuronowej zbudowaną z jednowymiarowych autoenkoderów Wave-U-Net. System złożony jest z pięciu modeli. Wszystkie modele zostały wytrenowane osobno, a następnie połączone w jeden system. Modele różnią się liczbą wejść i wyjść. System imituje sposób, w jaki miksy są tworzone przez inżyniera dźwięku, tzn. pojedyncze ścieżki są miksowane razem do grupy instrumentów (np. ścieżki wokali do grupy wokali itd.), a następnie grupy te są miksowane ze sobą, tworząc końcowy efekt.

Pojedynczy model został wytrenowany wykorzystując następujące parametry:

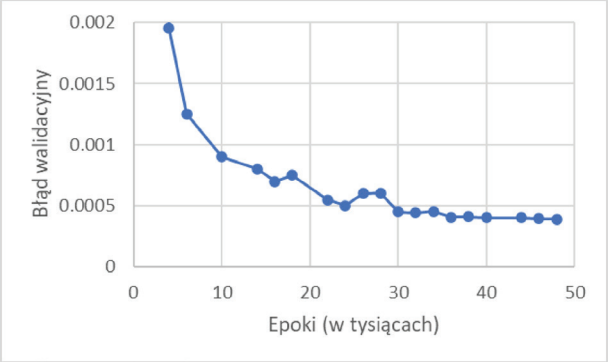
- Liczba warstw U-Net: 10;
- Maksymalna ilość wejść: 15;
- Upsampling: liniowy;
- Typ warstwy wyjściowej: liniowa bez aktywacji;
- Szybkość uczenia: 1e-4;
- Rozszerzenie bazy danych: false;

- Wielkość pojedynczej partii: 16;

- Liczba aktualizacji na epokę: 200;

- Optymalizator: Adam;

100 utworów zostało zgrupowane do treningu a 55 do walidacji. Do dalszych badań model z najmniejszym błędem był wybierany. Wykres funkcji błędu walidacyjnego podczas treningu modelu *stem-to-mix* przedstawiono na Rysunku 1.



Rys. 1. Funkcja błędu walidacyjnego modelu stem-to-mix

2.3. Przygotowanie miksów

Do dalszych testów przygotowano cztery różne miksy tej samej piosenki. Profesjonalny miks (”Pro”), amatorski miks (”Amateur”), miks stworzony przy pomocy najnowocześniejszego dostępnego komercyjnie oprogramowania (”Izotope”) oraz miks stworzony przy pomocy zaproponowanego systemu (”Unet”). W sumie przygotowano osiem utworów z wybranych czterech gatunków muzycznych (Pop, Alternatywa, Elektronika oraz Rock).

Profesjonalne miksy były przygotowane przez znanych i doświadczonych inżynierów dźwięku. Amatorskie miksy przygotowała osoba mająca słuch muzyczny i doświadczenie z teorią muzyki natomiast bez doświadczenia w miksowaniu. Miksy ”Izotope” zostały przygotowane półautomatycznie przy pomocy wtyczek firmy Izotope w dwóch krokach: automatyczne ustawienie balansu pomiędzy wszystkimi ścieżkami i stworzenie indywidualnych ustawień efektów dla każdego kanału. Miksy ”Unet” uzyskano przy wykorzystaniu systemu zaprezentowanego wcześniej. Pojedyncze ścieżki zostały automatycznie zmiksowane do odpowiednich grup instrumentów a te zostały zmiksowane w końcowy plik stereofoniczny.

3. ANALIZA JAKOŚCI UZYSKANYCH MIKSÓW

Podczas prowadzenia badań poddano obiektywnej analizie próbki sygnałów muzycznych, które nie zostały znormalizowane, tj. zbadano przebiegi poziomu takie jak: poziom średni głośności (*ang. RMS level*), zintegrowaną głośność (*ang. Integrated Loudness*), zakres głośności (*ang. Loudness Range*) i poziom szczytowy (*ang. True Peak level*) oraz wyznaczono wartości niskopoziomowych parametrów, m.in. deskryptorów MPEG-7 [8] (*ang. Odd-to-Even Harmonic Ratio, RMS-Energy Envelope, Harmonic Energy*). Na podstawie uzyskanych wyników przeprowadzono analizę statystyczną i porównano uzyskane wyniki z parametrami wyznaczonymi dla profesjonalnych miksów.

Dodatkowo przeprowadzono testy odsłuchowe na znormalizowanych próbkach miksów, w których słuchacze oceniali każdą próbkę w wielu kategoriach oceny jakości dźwięku: balans (*ang. balance*), przejrzystość (*ang. clarity*), panoramowanie (*ang. panning*), przestrzeń (*ang. space*) i dynamika (*ang. dynamics*). Wyniki testów subiektywnych poddano analizie statystycznej.

3.1. Obiektywna analiza

Nieznormalizowane próbki zostały wykorzystane do obiektywnej analizy. W pierwszej kolejności obliczono parametry postaci czasowej dla wszystkich próbek. Następnie, dla wszystkich próbek zostały obliczone wybrane deskryptory MPEG-7 [8]. Do tego celu wykorzystano narzędzie timbre toolbox [5] w środowisku Matlab. Wybrane deskryptory charakteryzują się możliwością interpretacji związanej z percepcją słyszenia. Dla każdego wspominanego deskryptora została przeprowadzona analiza statystyczna, aby ocenić czy różnice pomiędzy próbkami są istotne statystycznie. Do tego celu wykorzystano rozkład ANOVA [15] wraz z testem Tukey’a-Kramera [8]. Poziom istotności statystycznej założono na poziomie $\propto = .05$. Wyniki dla przykładowego deskryptora przedstawiono w Tabeli 1.

Tab. 1. Analiza istotności statystycznej wyników deskryptora energii harmonicznych dla utworu ”Secretariat – Over the top”

Porównane próbki	*Estymata*	*Istotność statystyczna p*
“Amateur”/”Izotope”	0.00	0.00
“Amateur”/”Unet”	0.00	0.00
“Amateur”/”Pro”	0.00	0.00
“Izotope”/”Unet”	-0.01	0.00
”Izotope/”Pro”	-0.01	0.00
”Unet”/”Pro”	0.00	0.37

Uzyskane wyniki analizy statystycznej potwierdzają, że możliwe jest automatyczne miksowanie utworu muzycznego składającego się z wielu ścieżek z użyciem jednowymiarowej adaptacji autoenkodera Wave-U-Net, które może być obiektywnie oceniane jako porównywalne jakościowo do miksu stworzonego przez profesjonalnego inżyniera dźwięku.

3.2. Subiektywna analiza

Przed testem odsłuchowym wszyscy uczestnicy zostali poproszeni o wypełnienie formularza. Pytania były tak ukierunkowane, aby uzyskać informację o wieku, doświadczeniu w miksowaniu i znajomości gatunków muzycznych uczestników. Rodzaj muzyki najczęściej słuchany przez danego użytkownika był zależny od osoby odpowiadającej. Natomiast większość uczestników testu była zaznajomiona z takimi gatunkami jak rock, alternative, hiphop czy jazz. 85% słuchaczy była muzykami, 60% inżynierami miksu. 20% słuchaczy miało więcej niż 10 lat doświadczenia w miksowaniu, 15% pomiędzy 6 a 10 lat, 25% pomiędzy 1 a 5 lat a 25% mniej niż rok.

Po odpowiednim przygotowaniu nagrań (po znormalizowaniu ich do jednego poziomu -14 LUFS) słuchacze zostali poproszeni o ocenienie wszystkich 32 próbek, które zostały im przedstawione. Słuchacze mieli za zadanie ocenić próbki w pięciostopniowej skali (1 = najniżej, 5 = najwyżej), z wykorzystaniem metodologii zaproponowanej przez Zacharova i Huopaniemiego [17]. Taki test może zostać zinterpretowany jako MOS (*ang. Mean Opinion Score*) i jest stososowany w ewaluacji tego typu próbek [16]. Test przeprowadzono w pomieszczeniu odsłuchowym R1 w Uniwersytecie w Hamburgu.

Następnie została przeprowadzona analiza statystyczna otrzymanych wyników. Wykonano ją w programie IBM SPSS Statistics 25 [1]. Podczas analizy wykonano test ANOVA wraz z testem Shapiro-Wilka oraz obliczono korelację, wykorzystując współczynnik Pearsona (r) [7]. Za poziom istotności statystycznej przyjęto poziom $\propto = 0.05$, zaś wyniki na poziomie $0.05 < p < 0.1$ zinterpretowano jako istotne statystycznie na poziomie trendu. Porównanie parami z poprawką Sidaka pokazuje to, że miksy "Pro" były najwyżej oceniane przez słuchaczy, a następnie − w kolejności miksy „Unet", „Amateur" oraz „Izotope".

W ostatnim kroku obliczono korelację pomiędzy odpowiedziami słuchaczy (ich oceną każdego typu miksu) a ich doświadczeniem w miksowaniu (ile lat doświadczenia mają). W tym celu posłużono się analizą z wykorzystaniem współczynnika korelacji Pearsona (Tab. 2). Analiza ta udowania silną korelację pomiędzy doświadczeniem w miksowaniu a ocenianiem miksów „Pro" oraz „Amateur". Oznacza to, że im więcej lat doświadczenia ma słuchacz, tym wyżej będzie oceniać miks profesjonalny i tym niżej miks amatorski. Korelacja miksu „Unet" jest na poziomie istotności statystycznej.

Tab. 2. Korelacja pomiędzy doświadczeniem w miksowaniu a oceną

Rodzaj miksu		*Doświadczenie w miksowaniu*
"Amateur"	(r)	-0.31
	istotność	0.186
"Izotope"	(r)	-0.51
	istotność	0.018
„Unet"	(r)	0.38
	istotność	0.098
„Pro"	(r)	0.69
	istotność	0.001

4. PODSUMOWANIE ORAZ DALSZE KIERUNKI BADAWCZE

Głównym celem artykułu było opracowanie i przetestowanie systemu miksowania plików audio, który w sposób automatyczny umożliwia tworzenie miksów z surowych sygnałów muzycznych, których jakość dorównywałaby miksom wykonanym profesjonalnie. W ramach pracy zaprojektowano architekturę opartą na jednowymiarowych autoenkoderach typu Wave-U-Net. Zaimplementowany system składa się z pięciu modeli, które zostały wytrenowane. Do treningu wykorzystano bazę danych MUSDB18-HQ, którą wzbogacono o utwory z bazy Cambridge oraz pięć oryginalnych kompozycji autora.

Aby sprawdzić słuszność hipotez przeprowadzono kilka eksperymentów. Pierwszy z nich dotyczył porównania obiektywnych cech uzyskanych miksów. W drugiej kolejności uzyskane miksy porównano w sposób subiektywy w testach odsłuchowych. Grupa słuchaczy ekspertów oceniała próbki w kilku kategoriach. Opracowany system powinien tworzyć miksy, które byłyby obiektywnie oraz subiektywnie lepsze niż miksy wykonane przez amatora oraz porównywalne z miksami stworzonymi przez profesjonalnego inżyniera dźwięku. Na podstawie uzyskanych wyników wykazano, że możliwe jest automatyczne miksowanie ścieżek wejściowych (dostarczonych przez użytkownika) przy użyciu wcześniej wytrenowanych modeli, tak aby finalny miks był bardzo blisko miksowi przygotowanemu przez profesjonalnego inżyniera dźwięku.

Wszystkie miksy stworzone przy użyciu zaproponowanego systemu były wolne od zniekształceń czy artefaktów. Ogólną jakość można ocenić jako dobrą lub nawet bardzo dobrą (szczególnie w porównaniu z miksami amatorskimi). Zaproponowany system zachowuje się podobnie niezależnie od gatunku muzycznego.

W planach rozwoju proponowanej metody przewiduje się uwzględnienie w proponowanym systemie dodatkowego modułu, jakim jest integracja modułu automatycznej klasyfikacji instrumentów na wejściu systemu. W ten sposób użytkownik nie musiałby przypisywać odpowiednich ścieżek muzycznych do odpowiednich wejść w systemie. W obecnej formie, aby system działał poprawnie, użytkownik musi wskazać ścieżki basowe do modelu5 basu, ścieżki perkusyjne do modelu perkusji, itp. Automatyczna klasyfikacja instrumentów jest możliwa i poprawiłaby wydajność systemu. Takie próby zostały przeprowadzone przez autora rozprawy na wstępnym etapie badań [2, 9]. Poprawiłoby to również wrażenia użytkownika i łatwość użytkowania dla początkujących użytkowników, którzy nie są profesjonalnymi inżynierami dźwięku.

Zaproponowany przez autora system można by dodatkowo rozbudować, wprowadzając automatyczny moduł wykrywania gatunku muzycznego utworu. Może być zaimplementowany na wejściu systemu (podczas predykcji), jak i na wyjściu (podczas treningu). Takie rozwiązanie przyniosłoby dodatkowy zysk podczas uczenia modeli, które są później wykorzystywane do predykcji (miksu). System mógłby pozyskiwać informacje o gatunku wejściowym podczas uczenia bezpośrednio z bazy danych, dzięki czemu modele

uczyłyby się miksowania utworów w danym gatunku. Podczas predykcji automatyczna klasyfikacja gatunkowa pozwoliłaby na lepsze miksowanie utworów w różnych gatunkach, ponieważ modele byłyby trenowane w analogiczny sposób. Co więcej, możliwe byłoby miksowanie tych samych zestawów utworów w różnych gatunkach muzycznych. Interesującym rozwiązaniem byłaby możliwość jednoczesnego miksowania utworu w wielu różnych gatunkach, gdzie użytkownik mógłby wybrać preferowany miks lub łączenia gatunków (np. 60% Rock i 40% Electronica). Obecnie modele są trenowane na bazie danych składającej się z czterech gatunków muzycznych – Pop, Alternative, Rock, Electronica.

Kolejnym proponowanym kierunkiem dalszych prac badawczo-rozwojowych jest dodatkowy moduł umożliwiający edycję poszczególnych utworów. Taki moduł pozwalałby na automatyczną synchronizację utworów ze sobą (np. w wielościeżkowych nagraniach perkusji) oraz automatyczne usuwanie (lub zmniejszanie głośności) niepożądanych dźwięków (takich jak oddech wokalisty czy przypadkowe uderzenia mikrofonu pomiędzy wartościowym sygnałem). Taki moduł mógłby zostać zaimplementowany na wejściu systemu, aby wszystkie ścieżki można było edytować przed miksowaniem. Obecnie użytkownik musi ręcznie synchronizować wszystkie utwory i edytować niepożądane lub przypadkowe dźwięki.

Szczególnie interesującym kierunkiem badawczym, będącym zakresem odrębnym od badań prowadzonych na potrzeby niniejszej rozprawy, jest automatyczne rozpoznawanie rytmu i tempa. Znajomość tempa danego utworu w jednostkach uderzeń na minutę (ang. beats per minute) jest kluczowa przy korzystaniu z efektów takich jak pogłos czy echo (ang. delay). Jeśli pogłos nałożony na ścieżkę jest zbyt długi, może wystąpić efekt maskowania. Znajomość rytmu i tempa utworu może również pozwolić na ustawienie idealnego tempa echa. Synchronizowanie efektu echa z tempem utworu to bardzo powszechna procedura stosowana przez inżynierów miksu. Na przykład, wiedząc, że tempo utworu wynosi 120 BPM, mikser jest w stanie ustawić tempo odtwarzania kolejnych odbić echa jako ćwierćnuty. Większość popularnych wtyczek programowych, które oferują ten efekt, pozwala użytkownikowi wybrać opcję „1/4” lub w przypadku konieczności podania wartości w milisekundach użytkownik może ją obliczyć (w podanym przykładzie będzie to 500 ms).

Automatyzacja procesu miksowania sygnałów fonicznych może być szczególnie przydatna w obszarze produkcji gier komputerowych czy tworzenia tzw. sygnatury muzycznej firm czy sieci sklepowych (ang. audio/music branding).

LITERATURA

[1] Aldrich J.O. 2018. “Using IBM SPSS statistics: An interactive hands-on approach”. *Sage Publications*.

[2] Blaszke Maciej, Koszewski Damian. 2020. “Determination of Low-Level Audio Descriptors of a Musical Instrument Sound Using Neural Network”. *Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA) Proceedings*.

[3] De Man B., Reiss J.D. 2013. "A knowledge-engineered autonomous mixing system". *Audio Engineering Society Convention 135*.

[4] De Man B., Stables R., Reiss J.D. 2019. “Intelligent Music Production”. *Focal Press*.

[5] https://github.com/mondaugen/timbretoolbox

[6] https://www.cambridge-mt.com/ms/mtk/

[7] Kim H. Y. 2015. “Statistical notes for clinical researchers: post-hoc multiple comparisons”. *Restorative dentistry & endodontics* 40 (2) : 172-176.

[8] Koenen R., Pereira F. 2000. “MPEG-7: A standardized description of audiovisual content”. *Signal Processing: Image Communication* 16 (1-2) : 5-13.

[9] Koszewski Damian, Kostek Bożena. 2020. “Musical instrument tagging using data augmentation and effective noisy data processing”. *Journal of Audio Engineering Society,* 68 (1/2) : 57-65.

[10] Lefford M. NY., Bromham G., Fazekas G., Moffat D. 2021. “Context-Aware Intelligent Mixing Systems”. *J. Audio Eng. Soc* 69 (3) : 128-141.

[11] Moffat D., Sandler M. B. 2019. "Approaches in intelligent music production," *Arts* 8 (5) : 14.

[12] Pachet F., Deleure O. 2000. “On-the-fly multi-track mixing”. *Audio Engineering Society Convention* 109.

[13] Rafii Z., Liutkus A., Stoter F. R., Mimilakis S. I., Bittner R. 2019. “MUSDB18-HQ – an uncompressed version of MUSDB18”.

[14] Reiss J.D. 2011. “Intelligent systems for mixing multichannel audio”. *17th International Conference on Digital Signal Processing (DSP)* : 1-6.

[15] Ross A., Willson V.L. 2017. “One-way ANOVA”. *Basic and advanced statistical tests :* 21-24.

[16] Unehara M., Yamada K., Shimada T. 2014. “Subjective evaluation of music with brain wave analysis for interactive music composition by IEC”. *Soft Computing and Intelligent Systems (SCIS)* : 66-70.

[17] Zacharov N., Huopaniemi, J. 1998. “Results of a Round Robin Subjective Evaluation of Virtual Home Theatre Sound Systems”. *107th International Audio Eng. Soc. Convention*.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

CELOWE ZANIECZYSZCZANIE PILOTÓW W ŁĄCZU W GÓRĘ W INTERFEJSIE 5G NR

INTENTIONAL PILOT CONTAMINATION IN 5G NR UPLINK

Jarosław Magiera, Piotr Rajchowski

Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk
jaroslaw.magiera@pg.edu.pl, piorajch@pg.edu.pl

DOI: 10.15199/59.2024.4.42

Streszczenie: **Referat poświęcono zagadnieniu zakłócania sygnałów pilotowych w interfejsie radiowym 5G NR . Przedstawiono charakterystykę sygnału referencyjnego DMRS oraz uwarunkowania możliwości jego selektywnego zakłócenia. Opisano schemat transmisji w kanale fizycznym PUSCH, zaimplementowany w oprogramowaniu Sionna. Zaprezentowano model symulacyjny oraz założenia badań wpływu zanieczyszczenia pilotów na jakość transmisji. Przedstawiono wyniki badań w zależności od mocy źródła zakłócenia oraz od jego lokalizacji względem stacji bazowej.**

Abstract: **The paper concerns the issue of contaminating pilot signals in the 5G NR radio interface. The characteristics of the DMRS reference signal and the conditions for the possibility of its selective jamming are presented. The transmission scheme in the PUSCH physical channel, implemented in the Sionna software, is described. A simulation model and assumptions for testing the influence of pilot contamination on transmission quality are described. The results of tests depending on the power of the interference source and its location in relation to the base station are presented.**

Słowa kluczowe: **5G NR, DMRS, zakłócanie selektywne, zanieczyszczenie pilotów.**

Keywords: **5G, DMRS, pilot contamination, selective jamming.**

1. WSTĘP

Podstawowym celem transmisji sygnałów referencyjnych (pilotowych) w łączu radiokomunikacyjnym jest umożliwienie estymacji stanu kanału radiowego po stronie odbiorczej. Współczynniki odpowiedzi kanału, wyznaczone w oparciu o znane sygnały pilotowe, są następnie wykorzystywane do korekcji kanału, w wyniku której otrzymuje się sygnał stanowiący aproksymację sygnału nadawanego.

Poprawność estymacji parametrów kanału ma szczególne znaczenie w łączach MIMO, gdzie dokładna znajomość macierzy kanału jest niezbędna aby zminimalizować wzajemne interferencje sygnałów transmitowanych równocześnie przez wiele kanałów przestrzennych [1]. Co więcej, przy transmisji z dupleksem czasowym (TDD), charakterystyka kanału, wyznaczona na podstawie sygnału odbieranego przez stację bazową, może posłużyć do uformowania wiązek antenowych dla sygnału nadawanego zwrotnie w łączu w dół, w tym samym kanale częstotliwościowym.

Aby zachować czystość sygnałów referencyjnych, transmisja w komórce jest zorganizowana tak, aby uniemożliwić wzajemne interferencje sygnałów pilotowych skojarzonych z różnymi transmisjami. Oznacza to, że piloty są nadawane w różnych zasobach czasowo-częstotliwościowych, albo też ich sekwencje są wzajemnie ortogonalne. Zaburzenie tego warunku jest nazywane zanieczyszczeniem pilotów.

Do zanieczyszczenia pilotów może dochodzić m.in. w wyniku interferencji wewnątrzsystemowej, gdy terminale znajdujące się w sąsiednich komórkach mają przypisane wzajemnie zakłócające się sekwencje pilotów. Dotyczy to głównie terminali znajdujących się na obrzeżach komórki. Prawdopodobieństwo wystąpienia takiej sytuacji może być zminimalizowane poprzez odpowiednio dobrany mechanizm zarządzania interferencjami w sieci.

Należy jednak zauważyć, że zanieczyszczenie pilotów może mieć także charakter celowy, tzn. przyjąć formę zamierzonej emisji zakłócenia, mającego na celu zaburzyć lub uniemożliwić proces estymacji kanału [2]. Z uwagi na selektywny charakter takiego zakłócenia, ma ono większy zasięg niż zakłócenie ciągłe o takiej samej energii. Ponadto, zakłócenie takie można wykorzystać aby oddziaływać na kształt formowanej charakterystyki kierunkowej anteny stacji bazowej w trybie TDD.

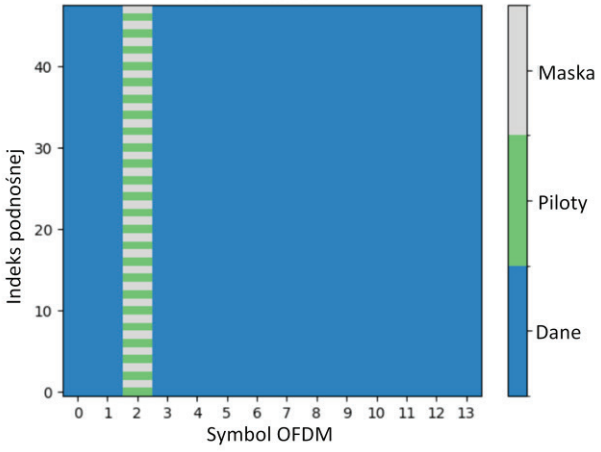
Niniejszy artykuł dotyczy celowego zanieczyszczenia pilotów w łączu w górę interfejsu 5G New Radio (5G NR). Przedstawiono wyniki badań symulacyjnych, prezentujących wpływ celowego zanieczyszczenia pilotów na jakość transmisji kanale fizycznym PUSCH.

2. SYGNAŁ REFERENCYJNY DMRS

Interfejs radiowy 5G NR, podobnie jak wiele innych aktualnych standardów interfejsów radiowych, stosuje technikę transmisji OFDM. Oznacza to, że na potrzeby transmisji sygnałów pilotowych dokonuje się alokacji określonych zasobów w siatce czasowo-częstotliwościowej.

Schemat transmisji pilotów w sieci 5G istotnie różni się od tego stosowanego w sieci 4G/LTE, w szczególności w łączu w dół, gdzie dotychczas stosowano ciągłą transmisję pilotów w ustalonych pozycjach siatki zasobów. Sieć 5G jest pod tym względem bardziej elastyczna. Jako podstawowy sygnał pilotowy, dla potrzeb estymacji kanału w celu demodulacji danych w kanale fizycznym,

stosuje się sygnał DMRS (Demodulation Reference Signal) [3]. Sygnał ten jest zawsze skojarzony z transmisją kanału fizycznego, np. PDSCH, PUSCH (Physical Downlink/Uplink Shared Channel), tzn. nie jest nadawany przy braku transmisji w tych kanałach, w przeciwieństwie do sygnału CRS (Cell Specific Reference Signal) znanego z LTE. Ponadto, w przypadku sygnału DMRS, możliwa jest elastyczna alokacja zasobów. Przykładowo, dla typowych warunków propagacyjnych, DMRS będzie nadawany w jednym z czternastu symboli OFDM w szczelinie czasowej, a dla kanału o krótszym czasie koherencji można nadawać go z częstotliwością np. 2 symboli na szczelinę. Dodatkowo, możliwe jest zdefiniowanie zbioru podnośnych w ramach bloku zasobów PRB, które zostaną przeznaczone na transmisję pilotów. Dzięki ortogonalizacji sekwencji DMRS, możliwe jest wykorzystanie tych samych zasobów OFDM do transmisji pilotów różnych użytkowników lub pilotów skojarzonych z różnymi portami antenowymi jednego terminala lub stacji bazowej. Możliwe jest także zdefiniowanie maski transmisji, czyli indeksów siatki zasobów niedostępnych dla transmisji, aby nie zakłócać transmisji pilotów innych użytkowników lub warstw.



Rys. 1. Alokacja zasobów w kanale PUSCH

Na rys. 1 przedstawiono przykładową alokację zasobów w szczelinie kanału PUSCH dla czterech bloków zasobów (48 podnośnych). Kolorem zielonym oznaczono zasoby DMRS, kolor szary to maska (brak transmisji), natomiast kolor niebieski określa zasoby do transmisji danych PUSCH.

3. ZAKŁÓCANIE PILOTÓW

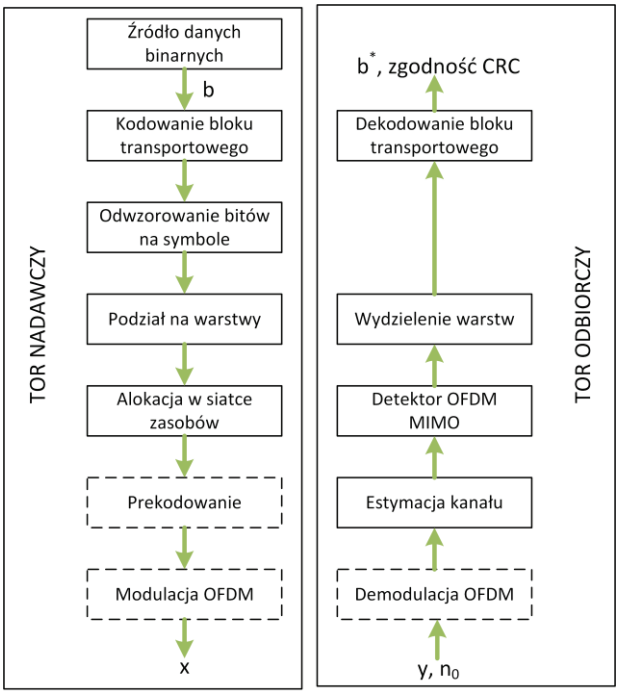
Celowe zanieczyszczenie pilotów polega na emisji zakłócenia w sposób selektywny w dziedzinach czasu i częstotliwości, tak aby pokrywało się ono z zasobami odbieranych sygnałów referencyjnych występujących w zakłócanej transmisji. Aby było to możliwe, źródło zakłócenia musi być zsynchronizowane czasowo z nadajnikiem sygnału użytecznego. Określenie chwili rozpoczęcia najbliższej transmisji przez wybrany terminal w łączu w górę jest możliwe w oparciu o analizę ruchu w łączu dół, gdzie stacja bazowa przesyła informacje o przydziale zasobów w tzw. uplink grant.

Ponadto, znaczenie dla efektywności zakłócenia ma dobór sekwencji pilota i jego alokacji. Sygnał DMRS bazuje na sekwencji pseudolosowej Golda o długości 2^{31}-1, której procedura generowania jest przedstawiona w specyfikacji [3]. Parametry określające postać tej sekwencji i jej ulokowania w siatce zasobów są uzależnione od identyfikatora komórki oraz określone w informacjach konfiguracyjnych przesyłanych w łączu w dół. Uzyskanie takich informacji jest możliwe np. z użyciem skanera radiowego, który dekoduje zbiór informacji systemowej [4].

Zastosowanie identycznej postaci pilotów w sygnale użytecznym i zakłócającym stanowi najbardziej niekorzystny przypadek z punktu widzenia odbiornika. Po pierwsze, wysoki stopień wzajemnej korelacji powoduje duży poziom interferencji i brak możliwości separacji sygnału zakłócającego od użytecznego. Po drugie, obecność identycznego pilota może nie zostać zidentyfikowana jako zakłócenie, a raczej jako zmiana w odpowiedzi impulsowej kanału, co utrudnia wykrycie takiego ataku.

4. MODEL TRANSMISJI W KANALE 5G-NR PUSCH

W celu zbadania wpływu zanieczyszczenia pilotów na jakość transmisji w kanale fizycznym PUSCH, posłużono się otwartoźródłową biblioteką Sionna [5], umożliwiającą przeprowadzanie badań symulacyjnych warstwy fizycznej 5G/6G. W ramach tej biblioteki zaimplementowano m.in. funkcje torów nadawczo-odbiorczych dla kanału PUSCH. Schemat przetwarzania przez poszczególne bloki funkcjonalne przedstawiono na rys. 2.



Rys. 2. Schematy torów nadawczego i odbiorczego

Transmisja pseudolosowych danych binarnych jest zorganizowana w bloki transportowe. Bity w bloku są segmentowane i podlegają kodowaniu kanałowemu LDPC (Low-Density Parity Check), a następnie są odwzorowywane na symbole zespolone wybranej modulacji. Sprawność kodu i rząd modulacji są uwarunkowane

obowiązującym schematem modulacji i kodowania (MCS – modulation and coding scheme). Symbole modulacyjne mogą być rozdzielone na warstwy, jeśli nadajnik wykorzystuje więcej niż jeden port antenowy. Następnie, symbole z każdej warstwy są odwzorowywane w siatce zasobów OFDM, zgodnie z przyjętym schematem alokacji. Dalej występują opcjonalne bloki prekodowania oraz modulacji OFDM. Ten ostatni przekształca sygnał z dziedziny częstotliwość-czas na sygnał w dziedzinie czasu, przy czym konieczność jego użycia jest uzależniona od rodzaju przyjętego modelu kanału radiowego. Po stronie odbiorczej sygnał, po konwersji do dziedziny czas-częstotliwość podlega procesowi estymacji parametrów kanału w oparciu o piloty DMRS. Stosowany jest estymator bazujący na metodzie najmniejszych kwadratów z interpolacją liniową. Następnie, przeprowadza się korekcję kanału, która, w przypadku transmisji wieloantenowej, jest połączona z detekcją warstw. Symbole z poszczególnych warstw są wyodrębniane i szeregowane w celu odtworzenia słowa kodowego. Słowo to podlega dekodowaniu kanałowemu, w wyniku którego otrzymuje się odtworzony blok transportowy.

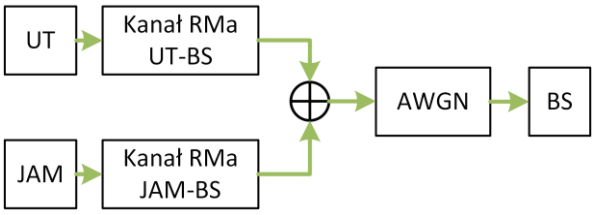
Istotnym elementem dla badań jakości transmisji jest sposób modelowania kanału radiowego. Oprogramowanie Sionna udostępnia szereg modeli kanałów, w tym ten zdefiniowany w specyfikacji 3GPP dla 5G [6]. Można w nim wyróżnić scenariusze propagacji dla środowisk miejskich (UMi - urban micro, UMa - urban macro) oraz obszarów wiejskich (RMa – rural macro). Model ten, w przeciwieństwie do modeli typu TDL (tapped delay line), uwzględnia wzajemną orientację nadajnika i odbiornika przy ustalonych charakterystykach kierunkowych anten. Możliwe jest również zamodelowanie odbioru sygnału przez stację bazową wyposażoną w układ antenowy o zadanej liczbie i geometrii elementów. W takim wypadku, współczynniki kanału są wyznaczane odrębnie dla każdej anteny w szyku, a następnie wykorzystywane w korektorze kanału LMMSE, realizującym funkcje odbioru zbiorczego.

5. BADANIA SYMULACYJNE ZANIECZYSZCZENIA PILOTÓW

5.1. Założenia modelu symulacyjnego

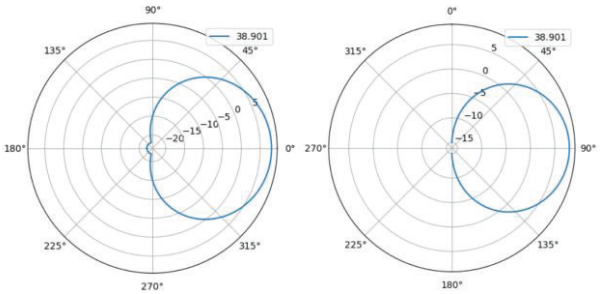
Aby przeprowadzić badania zanieczyszczania pilotów w kanale 5G PUSCH, zaimplementowano model systemu przedstawiony na rys. 3. Zdefiniowano dwa źródła sygnału w łączu w górę – terminal użytkownika (UT) i jammer (urządzenie zakłócające, JAM). Oba te źródła bazują na modelu toru nadawczego, opisanego w poprzedniej sekcji. Jedyną różnicą pomiędzy nimi jest wygaszenie transmisji symboli danych przez jammer – nadaje on jedynie w czasie trwania symbolu DMRS, podczas gdy terminal nadaje pełny sygnał, przenoszący również dane. Sekwencje DMRS generowane przez UT i JAM są identyczne.

Sygnały podlegają zjawiskom w kanale radiowym modelowanym zgodnie z [6] dla scenariusza RMa, przy czym realizacje współczynników tego kanału są generowane niezależnie dla terminala i jammera. Łączny sygnał odbierany przez stację bazową (BS) jest poddany działaniu szumu AWGN o mocy zależnej od zadanej wartości parametru E_b/N_0.



Rys. 3. Model symulacyjny

Terminal i jammer są wyposażone w takie same anteny o dookólnej charakterystyce promieniowania i polaryzacji pionowej, umieszczone na wysokości 1,5 m. Z kolei stacja bazowa korzysta z, zawieszonego na wysokości 35 m, liniowego szyku antenowego złożonego z czterech elementów oddalonych o połowę długości fali. Charakterystyka pojedynczej anteny stacji bazowej jest zgodna ze specyfikacją [6] przy założeniu polaryzacji pionowej. Wykres tej charakterystyki przedstawiono na rys. 4. Dodatkowo, przyjęto pochylenie charakterystyki anteny w dół o 5 stopni.



Rys. 4. Charakterystyka kierunkowa elementu antenowego stacji bazowej (przekrój poziomy i pionowy)

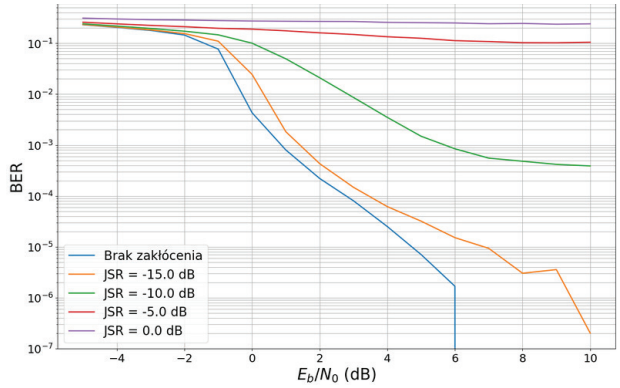
Transmisja jest realizowana na nośnej o częstotliwości 3,6 GHz, która leży pośrodku pasma przeznaczonego do równoczesnej obsługi dużej liczby terminali 5G w Polsce [7].

W modelu założono, że stacja bazowa jest idealnie zsynchronizowana z sygnałem docierającym z terminala. Badania nie mają na celu dokonania oceny jakości synchronizacji, lecz określenie wpływu obecności zakłócenia na jakość transmisji w kanale fizycznym.

5.2. Wyniki badań

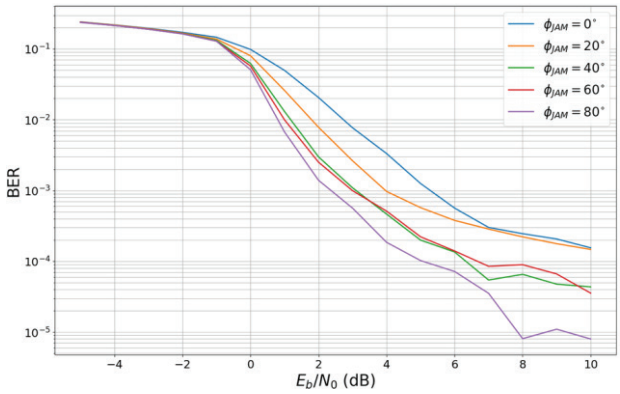
W pierwszej kolejności przeprowadzono badania charakterystyki szumowej dla różnych poziomów mocy zakłócenia do mocy sygnału użytecznego (JSR, jamming-to-signal ratio). Przez wartość JSR należy rozumieć stosunek mocy symboli DMRS odbieranych odpowiednio od jammera i terminala. Z kolei wartość bitowej stopy błędów (BER) wyznaczono poprzez porównanie sekwencji źródłowych i odtworzonych bitów danych w kanale PUSCH (odpowiednio b i b^* na Rys. 2).

Przyjęto jednakowe lokalizacje terminala i jammera, ustalone w odległości 2 km od stacji bazowej w kierunku prostopadłym do osi szyku antenowego.



Rys. 5. Wyniki badań jakości transmisji dla różnych wartości parametru JSR

Wyniki przedstawiono na rys. 5. Jak można zauważyć, już przy mocy zakłócenia niższej o 15 dB względem mocy sygnału DMRS, następuje niewielkie pogorszenie jakości transmisji. Przy wartości JSR na poziomie -10 dB jest widoczne wyraźne pogorszenie jakości transmisji, a bitowa stopa błędów stabilizuje się na poziomie ok. 4×10^{-4} dla wysokich wartości E_b/N_0. Po zwiększeniu JSR do poziomu -5 dB i 0 dB charakterystyki szumowe stają się płaskie, jako że zakłócenie staje się czynnikiem determinującym jakość transmisji, niezależnie od stosunku mocy sygnał-szum.



Rys. 6. Wyniki badań jakości transmisji dla różnych lokalizacji źródła zakłóceń

W kolejnej symulacji przeanalizowano wpływ lokalizacji źródła sygnału zakłócającego na jakość transmisji, przy założeniu, że terminal znajduje się na wprost anteny stacji bazowej. Zmieniano lokalizację jammera aby jego odległość od stacji bazowej była niezmienna (2 km) lecz zmieniał się jego azymut względem niej (Φ_{JAM}). Ustalono moc nadajnika zakłócenia na poziomie -10 dB względem mocy nadawanego sygnału DMRS. Wyniki zaprezentowano na rys. 6.

Krzywa dla kąta $\Phi_{JAM} = 0°$ odpowiada przypadkowi JSR = -10 dB z rys. 5. Widocznym jest, że w miarę oddalania kierunku źródła zakłócenia od głównego kierunku wiązki anteny stacji bazowej, wpływ zakłócenia na jakość transmisji jest coraz mniejszy. Wyjątkiem są kąty 40° i 60°, dla których uzyskano podobne charakterystyki szumowe.

6. PODSUMOWANIE

Wyniki badań wskazują, że celowe zanieczyszczenie pilotów, nawet emisją o względnie niewielkiej mocy, może spowodować istotne pogorszenie jakości transmisji. Takie zakłócenie stanowi rodzaj tzw. zagłuszania inteligentnego (smart jamming), w którym wykorzystuje się wiedzę o sygnale użytecznym [8]. Wiadomym jest, że takie działanie jest znacznie skuteczniejsze np. od ciągłego zakłócania zaporowego w całym paśmie sygnału [9]. Jak wspomniano, wymagane informacje o sygnale użytecznym mogą zostać w dużej mierze pozyskane poprzez analizę transmisji przy użyciu odpowiedniego skanera.

Uzyskane wyniki umożliwiły zweryfikowanie użyteczności oprogramowania Sionna dla potrzeb badań symulacyjnych interfejsu radiowego 5G. W dalszym etapie planowane jest przeprowadzenie bardziej zaawansowanych i innowacyjnych badań, związanych m.in. z poniższymi aspektami.

Zastosowanie modelu propagacyjnego UMi, UMa lub RMa, zaimplementowanego w bibliotece Sionna, umożliwia przeprowadzenie symulacji dla dowolnej topologii sieci, uwzględniającej różne lokalizacje i orientacje anten terminala i jammera względem układu antenowego stacji bazowej. W dalszych badaniach planowane jest zastosowanie modelu kanału bazującego na technice śledzenia promieni dla zadanej mapy cyfrowej, co również jest wspierane przez oprogramowanie Sionna.

Stacja bazowa, dysponująca układem antenowym ma możliwość separacji składowych przestrzennych, w tym składowej użytecznej i niepożądanej. Docelowo planowane jest zbadanie algorytmów detekcji i identyfikacji źródła zakłócenia w oparciu o analizę estymat współczynników kanału dla poszczególnych anten odbiorczych. Przy niewielkiej liczbie składowych przestrzennych i odpowiednio dużej liczbie elementów szyku antenowego, stacja bazowa powinna umożliwiać wykrycie i skompensowanie zanieczyszczenia pilotów w łączu w górę.

Innym kierunkiem dalszych prac może być badanie zanieczyszczenia sygnału referencyjnego NR SRS (new radio sounding reference signal), który jest używany m.in. dla potrzeb formowania wiązek w łączu w dół w trybie TDD [3].

LITERATURA

[1] Li, Ye. *Optimum training sequences for OFDM systems with multiple transmit antennas.* Globecom'00-IEEE. Global Telecommunications Conference. Conference Record (Cat. No. 00CH37137). Vol. 3. IEEE, 2000.

[2] Akgun, Berk, Marwan Krunz, O. Ozan Koyluoglu. *Pilot contamination attacks in massive MIMO systems.* 2017 IEEE Conference on Communications and Network Security (CNS). IEEE, 2017.

[3] 3GPP TS 38.211. *NR; Physical channels and modulation.* 3rd Generation Partnership Project; Technical Specification, Group Radio Access Network.

[4] Rowell Corbett, *5G mobile network testing using a passive network scanner*, Rohde & Schwarz, 2022 (online)
[5] Hoydis Jakob, et al. *Sionna: An open-source library for next-generation physical layer research*. arXiv preprint arXiv:2203.11854, 2022.
[6] 3GPP TR 38.901. *5G; Study on channel model for frequencies from 0.5 to 100 GHz.* 3rd Generation Partnership Project; Technical Report, Group Radio Access Network.
[7] https://www.gov.pl/web/5g/czestotliwosci
[8] Arjoune Youness, Saleh Faruque. *Smart jamming attacks in 5G new radio: A review.* 2020 10th annual computing and communication workshop and conference (CCWC). IEEE, 2020.
[9] Skokowski Paweł, et al. *Practical Trial for Low-Energy Effective Jamming on Private Networks With 5G-NR and NB-IoT Radio Interfaces*. IEEE Access, 2024.

KRiT 2024
KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

FORMOWANIE KLASTRÓW OBSŁUGUJĄCYCH W SIECI 6G OPEN RAN O ARCHITEKTURZE ZORIENTOWANEJ NA UŻYTKOWNIKA

SERVING CLUSTER FORMULATION IN OPEN RAN USER-CENTRIC CELL-FREE NETWORK

Marcin Hoffmann [1];

[1] Politechnika Poznańska, Poznań, marcin.hoffmann@put.poznan.pl

DOI: 10.15199/59.2024.4.43

Streszczenie: Jednym z najważniejszych wyzwań związanych z implementacją sieci Open RAN o architekturze zorientowanej na użytkownika jest odpowiednie formowanie tzw. klastrów obsługujących. Problem ten może być efektywnie rozwiązany za pomocą zaproponowanego algorytmu wykorzystującego współpracę pomiędzy xApp'em i rApp'em. Wyniki symulacji pokazały, że zaproponowane rozwiązanie pozwala na zwiększenie mediany z rozkładu przepływności użytkowników o 16% względem standardowej architektury zorientowanej na sieć.[1]

Abstract: One of the main challenges associated with the implementation of an Open RAN User-Centric Cell-Free network is the appropriate formulation of serving clusters. This problem can be effectively solved using a proposed algorithm that leverages cooperation between xApp and rApp. Simulation results have shown that the proposed solution allows for a 16% increase in the median of the user throughput distribution compared to the state-of-the-art network-centric architecture.

Słowa kluczowe: 6G, Massive MIMO, sieć zorientowana na użytkownika, Open RAN

Keywords: 6G, Massive MIMO, User-Centric networks, Open RAN

1. WSTĘP

Jedną z kluczowych cech sieci mobilnych szóstej generacji (6G) jest zastosowanie tzw. architektury zorientowanej na użytkownika (*ang. User-Centric Cell-Free*) w połączeniu z techniką M-MIMO (*ang. Massive Multiple-Input Multiple-Output*) [4]. Poprzednie generacje sieci mobilnych stosowały tzw. architekturę zorientowaną na sieć, gdzie użytkownik typowo obsługiwany jest przez jedną stację bazową od której odbiera najwyższą moc sygnału referencyjnego RSRP (*ang. Reference Signal Received Power*). W nowej architekturze zorientowanej na użytkownika może być on obsługiwany przez wiele zsynchronizowanych stacji bazowych (nazywanych w tym kontekście punktami dostępu) koordynowanych przez tzw. centralną jednostkę przetwarzania CPU (*ang. Central Processing Unit*). Wykorzystanie takiej architektury pozwala przede wszystkim na zmniejszenie dysproporcji w mocy odbieranej w różnych punktach sieci. Jest to szczególnie istotne w przypadku użytkowników znajdujących się w podobnej odległości od kilku stacji bazowych [6]. W tzw. kanonicznej implementacji sieci zorientowanej na użytkownika zakłada się, że wszystkie stacje bazowe będą obsługiwały wszystkich użytkowników [3]. W praktyce jednak, zarówno stacje bazowe jak i CPU mają ograniczone zasoby obliczeniowe i mogą obsłużyć jedynie skończoną liczbę użytkowników, np. ze względu na konieczność estymacji kanału i obliczania wag prekodera. Ponadto, część stacji bazowych znacząco oddalonych od użytkownika będzie miała marginalny wpływ na poprawę jego przepływności, "marnując" w ten sposób cenne zasoby radiowe. Z tej perspektywy jednym z kluczowych zagadnień dla sieci zorientowanej na użytkownika jest procedura tzw. formowania klastrów obsługujących. Ma ona na celu określenie zbioru stacji bazowych, które powinny obsługiwać konkretnego użytkownika [1]. Klastry obsługujące mogą być dynamicznie formowane np. na bazie RSRP, z wykorzystaniem różnych algorytmów np. uczenia maszynowego ML (*ang. Machine Learning*). Ponadto, formowanie klastrów obsługujących może uwzględniać różne funkcje celu, takie jak: maksymalizacja efektywności energetycznej lub maksymalizacja przepływności osiąganych przez użytkowników.

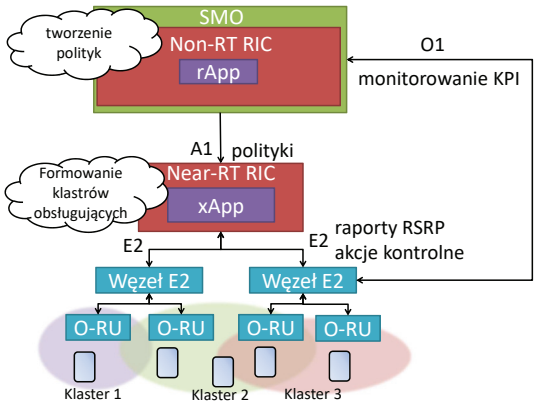
Potencjalną szansą na praktyczną implementację sieci zorientowanej na użytkownika jest wykorzystanie koncepcji otwartej sieci dostępowej Open RAN (*ang. Open Radio Access Network*) [7]. W ramach koncepcji Open RAN zaproponowany został tzw. podział 7.2, pozwalający na rozdzielenie funkcji stacji bazowej na zwirtualizowaną jednostkę rozproszoną O-DU (*ang. Open RAN Distributed Unit*) realizującą m.in., funkcje warstwy fizycznej wyższego poziomu i warstwy dostępu do łącza, oraz fizyczną jednostkę radiową O-RU (*Open RAN Radio Unit*), odpowiedzialną za funkcje warstwy fizycznej niższego poziomu i bezpośrednio transmisje radiową. Korzystając z podziału 7.2, tzw. węzeł E2 (tutaj O-DU) może służyć jako CPU i kontrolować transmisję z wielu jednostek radiowych O-RU wykorzystujących technikę M-MIMO. Jedną z kluczowych cech sieci Open RAN jest możliwość kontrolowania sieci dostępowej poprzez implementację dedykowanych algorytmów w formie tzw. xApp'ów i rApp'ów, które są wdrażane w ramach tzw. inteligentnego kontrolera sieci dostępowej pracującego w czasie bliskim rzeczywistemu Near-RT RIC (*ang. Near Real-Time RAN Intelligent Controller*), oraz tzw. inteligentnego kontrolera sieci dostępowej pracującego w czasie nierzeczywistym Non-RT RIC. Near-RT RIC pracuje

[1]Praca powstała w ramach projektu PRELUDIUM finansowanego przez Narodowe Centrum Nauki nr 2022/45/N/ST7/01930.

w pętli kontroli o skali czasu od 10 ms do 1 s, podczas gdy Non-RT RIC w pętli kontroli o skali czasu powyżej 1 s. Wykorzystanie xApp'u do formowania klastrów obsługujących w sieci Open RAN zorientowanej na użytkownika zostało zaproponowane w [2]. Autorzy jednak rozważali jedynie wykorzystanie Near-RT RIC'a, bez uwzględnienia dodatkowych danych, które mogłyby być przesyłane przez rApp w Non-RT RIC'u. Ponadto, analiza symulacyjna przeprowadzona w tej pracy, zakłada prosty model kanału radiowego, oraz wąskopasmową transmisję na pojedynczej nośnej, podczas gdy rzeczywiste sieci 5G, a prawdopodobnie także 6G, wykorzystują wielodostęp oparty na modulacji wielotonowej OFDMA (*ang. Orthogonal Frequency Division Multiple Access*).

W tej pracy proponowany jest algorytm formowania klastrów obsługujących w sieci Open RAN o architekturze zorientowanej na użytkownika z wykorzystaniem dwóch współpracujących ze sobą aplikacji: xApp'a w Near-RT RIC'u i rApp'a w Non-RT RIC'u. Zaproponowane rozwiązanie zostanie zbadane za pomocą zaawansowanego symulatora sieci M-MIMO o architekturze zorientowanej na użytkownika, wykorzystującego trójwymiarowe śledzenie promieni (*ang. 3D Ray Tracing*), zaproponowanego we wcześniejszej pracy autora [5]. W dalszych częściach tej pracy, w Rozdziale 2 przedstawiona zostanie rozważana sieć Open RAN o architekturze zorientowanej na użytkownika. Rozdział 3 zawiera opis zaproponowanego algorytmu formowania klastrów obsługujących. Wyniki symulacji przedstawione są w Rozdziale 4, a wnioski zostały sformułowane w Rozdziale 5.

2. SIEĆ OPEN RAN O ARCHITEKTURZE ZORIENTOWANEJ NA UŻYTKOWNIKA



Rysunek 1: Sieć Open RAN o architekturze zorientowanej na użytkownika.

W tej pracy rozważane jest łącze w dół w sieci Open RAN zorientowanej na użytkownika przedstawionej na Rys. 1. W tej sieci O-RU jest jednostką radiową wyposażoną w macierz antenową M-MIMO. O-RU odpowiada bezpośrednio z transmisję danych użytkownika w interfejsie radiowym. Węzeł E2 to połączona zwirtualizowana jednostka rozproszona O-DU i zwirtualizowana jednostka centralna O-CU (*ang. Open RAN Centralized Unit*). Pełni ona rolę CPU, tzn., odpowiada za przydział zasobów radiowych, a także koordynację transmisji do użytkowników z kontrolowanych przez nią fizycznych jednostek radiowych O-RU. Jednym z najistotniejszych zadań węzła E2 w ramach koordynacji transmisji, jest zestawienie klastra obsługującego dla każdego z użytkowników. Wymaga ono konfiguracji jednostek radiowych O-RU tworzących ten klaster w celu zapewnienia, m.in., wymiany danych użytkownika, synchronizacji transmisji, oraz wymiany informacji o współczynnikach kanałów radiowych poszczególnych użytkowników. W szczególności, potencjalnym wyzwaniem jest formowanie klastrów obsługujących wymagających koordynacji pomiędzy dwoma węzłami E2, tj. fragmentami sieci obsługiwanymi przez różne CPU. Formowanie klastrów obsługujących może być kontrolowane z użyciem interfejsu E2 przez xApp znajdujący się w Near-RT RIC'u. W rozważanej sieci taka kontrola polega na wskazaniu węzłowi E2 które jednostki radiowe O-RU powinny obsługiwać danego użytkownika. Oprócz przesyłania danych kontrolnych, interfejs E2 umożliwia również przesyłanie raportów dotyczących tzw. kluczowych wskaźników wydajności KPI (*ang. Key Performance Indicators*) sieci dostępowej i użytkowników, np. wartości RSRP, które xApp może wykorzystać w procesie formowania klastrów obsługujących. Z drugiej strony węzeł E2 połączony jest z tzw. SMO (*ang. Service Management and Orchestration*) interfejsem O1, który umożliwia m.in., przekazywanie raportów zawierających KPI, np. rozkład przepływności użytkowników. Tego typu dane mogą być używane przez rApp do formowania tzw. polityk (*ang. policies*) mających na celu kierowanie xApp'em i docelowo poprawę wydajności sieci dostępowej. Polityki przekazywane są poprzez interfejs A1 z Non-RT RIC'a do Near-RT RIC'a. W rozważanej sieci polityka może zawierać parametry algorytmu formowania klastrów obsługujących ustalone na podstawie analizy obserwowanych KPI w dłuższej skali czasowej.

3. ALGORYTM FORMOWANIA KLASTRÓW OBSŁUGUJĄCYCH

W tej pracy proponowany jest algorytm formowania klastrów obsługujących w sieci Open RAN zorientowanej na użytkownika z wykorzystaniem współpracy pomiędzy xApp'em i rApp'em. W dalszej części tego rozdziału opisana zostanie najpierw rola xApp'a, a następnie rola rApp'a w zaproponowanym algorytmie.

3.1 Rola xApp'a

Rolą xAppa w zaproponowanym podejściu jest dynamiczne formowanie klastrów obsługujących w oparciu o wartości RSRP poszczególnych użytkowników raportowane poprzez interfejs E2, oraz parametry algorytmu otrzymywane z Non-RT RIC'a za pośrednictwem interfejsu A1. W pierwszym kroku xApp wyznacza zbiór jednostek radiowych O-RU, które powinny obsługiwać danego użytkownika. Dla k-tego użytkownika, niezależnie spośród L jednostek radiowych O-RU indeksowanych przez zmienną l, wybieranych jest N_k o największej wartości RSRP, które zapewnią spełnienie warunku [9]:

$$\frac{\sum_{l\in\mathcal{M}_k} p_{k,l}}{\sum_{l=1}^{L} p_{k,l}} \geq \Delta, \tag{1}$$

gdzie $\mathcal{M}_k$ jest klastrem obsługującym k-tego użytkownika, $p_{k,l}$ oznacza RSRP odebrane przez k-tego użytkownika

od l-tej jednostki radiowej O-RU, a Δ jest parametrem algorytmu przekazywanym przez rApp. Po wyznaczeniu zbioru jednostek radiowych O-RU obsługujących k-tego użytkownika - $\mathcal{M}_k$, jest on przesyłany przez xApp jako akcja kontrolna do węzła E2.

3.2 Rola rApp'a

W zaproponowanym algorytmie, rolą rApp'a znajdującego się w Non-RT RIC'u jest tworzenie polityk, które kierują procesem formowania klastrów obsługujących w xApp'ie. Polityka wysyłana jest do Non-RT RIC'a za pomocą interfejsu A1 i zawiera wartość parametru Δ, która powinna być zastosowana przez xApp. W tej pracy zaproponowano podejście, w którym wybór parametru Δ dokonywany jest z użyciem ML i zdefiniowany jako tzw. problem wielorękiego bandyty (*ang. Multi-Arm Bandit*) [8]. W tym przypadku każda "ręka" wielorękiego bandyty odpowiada jednej wartości parametru Δ. W równych odstępach czasowych T rApp podejmuje akcję - wybiera jedną z wartości parametru Δ i przekazuje ją do xApp'a za pośrednictwem interfejsu A1. Zaproponowane jest, żeby algorytm podejmował akcje w sposób zachłanny tzn. zawsze wybierał wartość parametru Δ związaną z najwyższą spodziewaną nagrodą, oznaczoną jako $Q(\Delta)$. Następnie, obserwuje związaną z nią nagrodę - proponujemy, żeby była nią mediana z rozkładu przepływności użytkowników otrzymanego w okresie T, następującym po wyborze akcji. W tym celu rApp monitoruje odpowiednie KPI za pośrednictwem interfejsu O1. Po obliczeniu aktualnej nagrody rApp aktualizuje wartość spodziewanej nagrody $Q(\Delta)$ według wzoru:

$$Q(\Delta) \leftarrow \alpha r + (1-\alpha)Q(\Delta), \tag{2}$$

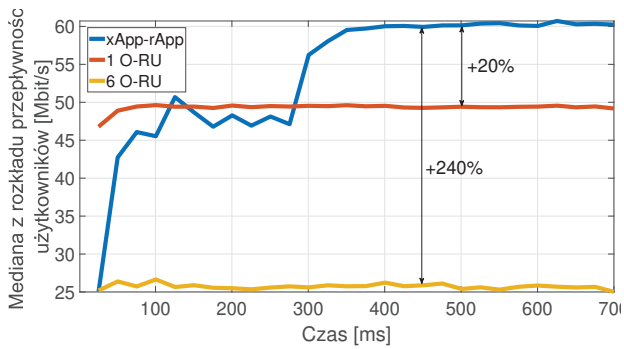
gdzie r oznacza zaobserwowaną nagrodę, a α jest arbitralnie dobieranym parametrem w zakresie $< 0 : 1 >$. Po aktualizacji nagrody rApp wybiera kolejną akcję i cykl się powtarza.

4. WYNIKI SYMULACJI

Zaproponowany algorytm formowania klastrów obsługujących w sieci Open RAN zorientowanej na użytkownika został zbadany za pomocą zaawansowanego symulatora zaproponowanego przez autora w poprzedniej pracy [5]. Symulator modeluje sieć M-MIMO o architekturze zorientowanej na użytkownika i wykorzystującej OFDMA. Zastosowano w nim realistyczny model kanału radiowego oparty o trójwymiarowe śledzenie promieni firmy Wireless InSite, jak również podział na wiele bloków funkcjonalnych takich jak: formowanie klastrów obsługujących, dedykowany algorytm przydziału zasobów radiowych, prekoder ZF (*ang. Zero-Forcing*) działający niezależnie dla każdej jednostki radiowej O-RU i 15 schematów modulacji i kodowania. W ramach jednej jednostki radiowej O-RU, podział mocy w obrębie pojedynczego bloku zasobów odbywa się równo pomiędzy wszystkich użytkowników, którym został on przydzielony w rezultacie multipleksacji przestrzennej. Pozostałe parametry przeprowadzonych symulacji zostały zestawione w Tabeli 1. Przeprowadzono 15 przebiegów symulacyjnych, trwających 700 ms (1400 slotów czasowych), w których decyzja o zmianie parametru Δ podejmowana była przez w odstępach czasowych $T = 25$ ms. Każdy przebieg symulacyjny związany był z niezależnym wylosowaniem 40 użytkowników. Dodatkowo zaproponowane rozwiązanie porównano z architekturą zorientowaną na sieć (1 O-RU), gdzie każdy użytkownik obsługiwany jest przez jedna jednostkę radiową O-RU, oraz z tzw. kanoniczną implementacja architektury zorientowanej na użytkownika (6 O-RU), gdzie wszystkie jednostki radiowe O-RU obsługują wszystkich użytkowników. Na Rys. 2 przedstawiono medianę z rozkładu przepływności użytkowników uśrednioną po 15 przebiegach symulacyjnych w funkcji czasu. Widać, że w przypadku zaproponowanego algorytmu bazującego na xApp'ie i rApp'ie, przez około 300 ms (12 cykli decyzyjnych) rApp uczy się odpowiedniego doboru parametru Δ, by później ustabilizować osiąganą medianę przepływności na poziomie około 60 Mbit/s, zapewniając 20% zysku względem architektury zorientowanej na sieć i 240% względem kanonicznej implementacji architektury zorientowanej na użytkownika. Wyniki te wskazują na konieczność odpowiedniego formowania klastrów obsługujących, gdyż w przypadku gdy wszystkie jednostki radiowe obsługują wszystkich użytkowników, często duża część mocy nadawana jest do bardzo odległych użytkowników.

Tabela 1: Parametry Symulacji

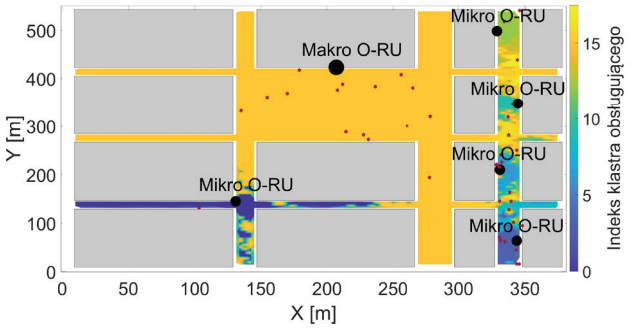
Parametr	*Wartość*
Częstotliwość nośna	3.6 GHz
Liczba bloków zasobów	69
Liczba węzłów E2	1
Liczba jednostek radiowych O-RU	1 makro, 5 mikro
Wysokość anten - makro O-RU	45 m
Wysokość anten - mikro O-RU	6 m
Moc nadawana - makro O-RU	46 dBm
Moc nadawana - mikro O-RU	30 dBm
Liczba anten - makro O-RU	128
Liczba anten - mikro O-RU	32
Czas symulacji	700 ms
Liczba użytkowników	40
Wireless InSite - liczba odbić	15
Wireless InSite - liczba dyfrakcji	1



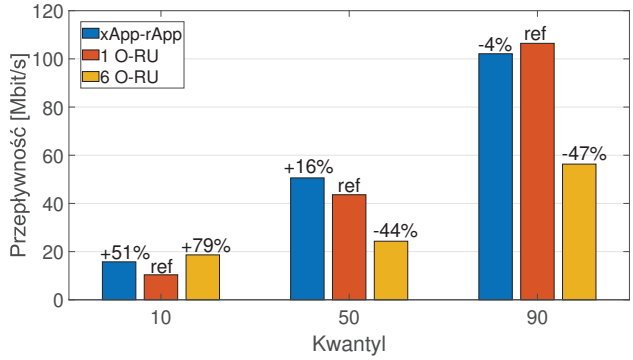
Rysunek 2: Mediana z rozkładu przepływności użytkowników w funkcji czasu, uśredniona po 15 przebiegach symulacyjnych.

Na Rys 3 przedstawione zostało rozmieszczenie jednostek radiowych O-RU, przykładowe rozmieszczenie użytkowników, oraz podział rozważanej sieci na klastry obsługujące dla wartości parametru $\Delta = 0.8$, która była

najczęściej wybierana przez rApp podczas przeprowadzonych symulacji. Kolor odpowiada zbiorowi jednostek radiowych O-RU obsługujących użytkownika znajdującego się w danym punkcie. Duża część rozważanej sieci jest pokrywana przez najsilniejszą jednostkę radiową makro O-RU, stąd największy obszar zajmuje klaster pomarańczowy, w którym użytkownicy obsługiwani są wyłącznie przez tę jednostkę. Dzięki temu jednostki radiowe mikro O-RU nie zużywają mocy na obsługę użytkowników znacząco od nich oddalonych lub oddzielonych budynkami. Można natomiast zaobserwować, że w sąsiedztwie jednostek radiowych mikro O-RU pojawia się wiele klastrów obsługi. Jest to szczególnie widoczne w alejce po prawej stronie. Sformowane klastry pozwalają na obsługę użytkowników przez sąsiadujące ze sobą jednostki radiowe O-RU i tym samym zwiększenie mocy odbieranej przez znajdujących się tam użytkowników, oraz koordynacje interferencji.



Rysunek 3: Rozmieszczenie jednostek radiowych O-RU, przykładowe rozmieszczenie użytkowników (ciemnoczerwone gwiazdy), oraz podział sieci na klastry obsługujące dla $\Delta = 0.8$.



Rysunek 4: Kwantyle rzędu 10, 50 i 90 z rozkładu przepływności użytkowników.

Na Rys. 4 przedstawione zostały kwantyle rzędu 10, 50 i 90 z rozkładu przepływności użytkowników otrzymanego z łącznie 15 przebiegów symulacyjnych, dla zaproponowanego algorytmu formowania klastrów obsługujących bazującego na współpracy pomiędzy xApp'em i rApp'em, architektury zorientowanej na sieć i kanonicznej implementacji architektury zorientowanej na użytkownika. Widać, że dla kanonicznej implementacji architektury zorientowanej na użytkownika występuje największy zysk względem architektury zorientowanej na sieć w przypadku kwantyla rzędu 10 reprezentującego użytkowników o najgorszych warunkach propagacyjnych. Niestety w przypadku mediany i kwantyla rzędu 90 reprezentującego użytkowników o najlepszych warunkach propagacyjnych można zaobserwować znaczące pogorszenia odpowiednio aż o 44% i 47%. Z drugiej strony, zaproponowany algorytm poprzez odpowiednie formowanie klastrów obsługujących, pozwala na zwiększenie zarówno kwantyla rzędu 10 jak i mediany z rozkładu przepływności użytkowników odpowiednio o 51% i 16% względem architektury zorientowanej na sieć. Zysk osiągnięty jest kosztem minimalnego pogorszenia kwantyla rzędu 90, tj., o 4%.

5. WNIOSKI

Jednym z najważniejszych wyzwań związanych z implementacja sieci Open RAN o architekturze zorientowanej na użytkownika jest odpowiednie formowanie klastrów obsługujących. Problem ten może być efektywnie rozwiązany za pomocą zaproponowanego algorytmu wykorzystującego współpracę pomiędzy xApp'em i rApp'em. Wyniki symulacji pokazały, że zaproponowane rozwiązanie pozwala na zwiększenie mediany z rozkładu przepływności użytkowników odpowiednio o 16% i 52% względem architektury zorientowanej na sieć i kanonicznej implementacji architektury zorientowanej na użytkownika.

LITERATURA

[1] Ammar, Hussein A., Adve, Raviraj, Shahbazpanahi, Shahram, Boudreau, Gary, Srinivas, Kothapalli Venkata, 2022, "User-Centric Cell-Free Massive MIMO Networks: A Survey of Opportunities, Challenges and Solutions". *IEEE Communications Surveys & Tutorials*, 24 (1): 611–652.

[2] Beerten, Robbert, Girycki, Adam, Pollin, Sofie, 2022, "User Centric Cell-Free Massive MIMO in the O-RAN Architecture: Signalling and Algorithm Integration". *2022 IEEE Conference on Standards for Communications and Networking (CSCN)*, 181–187.

[3] Buzzi, Stefano, D'Andrea, Carmen, 2017, "Cell-free massive mimo: User-centric approach". *IEEE Wireless Communications Letters*, 6 (6): 706–709.

[4] He, Jianhua, Yang, Kun, Chen, Hsiao-Hwa, 2021, "6g cellular networks and connected autonomous vehicles". *IEEE Network*, 35 (4): 255–261.

[5] Hoffmann, Marcin, Kryszkiewicz, Paweł, 2024, "Evaluation of User-Centric Cell-Free Massive Multiple-Input Multiple-Output Networks Considering Realistic Channels and Frontend Nonlinear Distortion". *Applied Sciences*, 14 (5).

[6] Ngo, Hien Quoc, Ashikhmin, Alexei, Yang, Hong, Larsson, Erik G., Marzetta, Thomas L., 2015, "Cell-free massive mimo: Uniformly great service for everyone". *2015 IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 201–205.

[7] Ranjbar, Vida, Girycki, Adam, Rahman, Md Arifur, Pollin, Sofie, Moonen, Marc, Vinogradov, Evgenii, 2022, "Cell-free mmimo support in the o-ran architecture: A phy layer perspective for 5g and beyond networks". *IEEE Communications Standards Magazine*, 6 (1): 28–34.

[8] Sutton, Richard S., Barto, Andrew G., 2018, *Reinforcement Learning: An Introduction*. The MIT Press, second edition.

[9] Özlem Tugfe Demir, Björnson, Emil, Sanguinetti, Luca, 2021, "Foundations of user-centric cell-free massive mimo". *Foundations and Trends® in Signal Processing*, 14 (3-4): 162–472.

MODELOWANIE DOKŁADNOŚCI RADIOLOKALIZOWANIA W RÓŻNYCH WARUNKACH ŚRODOWISKOWYCH PRZY WYKORZYSTANIU INTERFEJSU RADIOWEGO 5G-NR

EXPERIMENTAL ANALYSIS OF 5G-NR BASED POSITIONING IN DIFFERENT ENVIRONMENTS

Piotr Rajchowski[1]; Luis M. Correia[2]; Krzysztof K. Cwalina[1]

[1] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, piorajch, kkcwalina@pg.edu.pl
[2] Instituto Superior Técnico – INESC-ID/INOV, University of Lisbon, Lizbona, luis.m.correia@tecnico.ulisboa.pt

DOI: 10.15199/59.2024.4.44

Streszczenie: W artykule przedstawiono wyniki eksperymentalnych badań dokładności estymacji położenia terminala użytkownika korzystającego z interfejsu radiowego 5G-NR. W środowisku miejskim dokonano rejestracji rzeczywistych sygnałów sieci 5G, a następnie przeprowadzono badania numeryczne. Celem było zweryfikowanie różnic dokładności estymacji położenia w trzech różnych środowiskach: wewnątrz- i zewnątrzbudynkowym oraz tzw. deep-indoor.

Abstract: In this paper, an experimental analysis of the localization accuracy in a 5G-NR network operating in various urban environments is presented. Measurements of 5G-NR downlink signals were performed in several urban-type environments to investigate the user terminal localization process. The main goal of the research is to analyze the differences in the estimated positions accuracy in three different environments, i.e., outdoor, light-indoor and deep-indoor.

Słowa kluczowe: 5G-NR, Charakterystyka kanału radiowego, Radiolokalizacja, SDR.

Keywords: 5G-NR, Channel characteristic, Radiolocalization, SDR.

1. WSTĘP

Radiolokalizowanie jest jedną z najbardziej pożądanych usług dodatkowych w sieciach bezprzewowodowych. Wiedza o położeniu terminala użytkownika UE (User Equipment) może być kluczowa np. dla świadczenia usługi innej niż telekomunikacyjna, dostosowanej do potrzeb użytkownika. Ponadto w sieciach 5G-NR przyjęto, że estymowane położenie UE służy do dostosowania parametrów interfejsu radiowego, co ma na celu m.in. minimalizowanie zakłóceń wewnątrzsystemowych.

Rozwój sieci 5G-NR pracujących w trybie standalone (SA) umożliwił m.in. realizację sieci prywatnych, dając możliwość przeprowadzania eksperymentów reprezentujących realistyczne scenariusze. W pracach badawczych związanych z problemem estymowania położenia UE szeroko analizowane są różne rodzaje środowiska i przypadki aplikacyjne. Niemniej jednak nadal istnieje potrzeba analizy dokładności i skuteczności lokalizowania nie tylko poprzez badania symulacyjne, ale także pomiary przeprowadzane w rzeczywistych środowiskach propagacyjnych, ze znaczącym wpływem kanału radiowego. W niektórych pracach autorzy badają możliwość lokalizowania UE przy użyciu tylko jednej stacji bazowej (gNodeB), co wiąże się z kompleksową analizą odpowiedzi impulsowej kanału, gdzie znaczący wpływ mają zmiany warunków propagacji, w tym propagacja wielodrogowa. Zatem w większości przypadków zakłada się, że UE wykryje dedykowane sygnały lokalizacyjne z co najmniej trzech gNodeB w celu estymowania położenia w wariancie dwuwymiarowym [1, 2, 3, 4].

W niniejszym artykule przedstawiono wyniki badań eksperymentalnych lokalizowania UE w oparciu o sygnały radiowe interfejsu 5G-NR zarejestrowane w środowisku miejskim, na zewnątrz i wewnątrz budynku (ZB, WB) oraz w tzw. środowisku deep-indoor (D-I). Podczas badań numerycznych zastosowano metodę TDOA (Time Difference of Arrival) wraz z algorytmem Foya, a uzyskana średnia dokładność estymacji położenia UE zawierała się w zakresie od 5 do 295 m, zależnie od rodzaju środowiska i pasma pracy interfejsu 5G-NR.

2. RADIOLOKALIZOWANIE W SIECI 5G-NR

Estymacja położenia UE w sieciach komórkowych może być realizowana przy wykorzystaniu dedykowanych sygnałów lokalizacyjnych, np. PRS (Position Reference Signal) proponowanego dla sieci 4G-LTE, a także w zmodyfikowanej formie dla interfejsu radiowego 5G-NR.

Obecnie w większości komercyjnych sieci 5G-NR sygnały PRS nie są transmitowane ze względów biznesowych, co można zaobserwować także w sieciach 4G-LTE [5, 6]. Eliminuje to możliwość estymacji położenia UE z dokładnością założoną w standardzie 3GPP, niezależnie od scenariusza. Niemniej jednak w łączu w dół (Downlink - DL) interfejsu radiowego 5G-NR dostępne są inne sygnały, które mogą być przetwarzane przez terminal jako tzw. sygnały okazjonalne. W niniejszym artykule przyjęto wykorzystanie sygnału synchronizacyjnego PSS (Primary

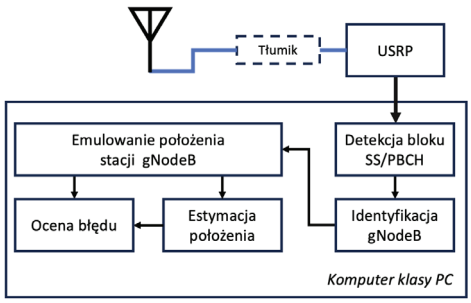
Synchronization Signal) do wyznaczania estymaty położenia UE, czyli sygnału przesyłanego okresowo w ramkach radiowych w DL, konfigurowanych przez operatora sieci. Sygnał PSS jest przesyłany wspólnie z drugim sygnałem synchronizacyjnym SSS (Secondary Synchronization Signal) i elementami zasobów kanału rozgłoszeniowego PBCH (Physical Broadcast Channel) w postaci tzw. bloku SS/PBCH. Szerokość pasma niniejszego bloku jest większa, a czas transmisji krótszy w porównaniu do sygnałów synchronizacyjnych i rozgłoszeniowych przesyłanych w interfejsie radiowym LTE. Dzięki tej odmienności możliwe jest zwiększenie dokładności estymacji położenia względem sieci 4G, ale także zmniejszeniu ulega prawdopodobieństwo wykrycia niniejszych sygnałów [5, 6].

3. METODYKA PRAC BADAWCZYCH

W niniejszym rozdziale opisano zaprojektowane i wykonane stanowisko badawcze oraz metodykę zastosowaną do analizy dokładności estymacji położenia UE w środowisku miejskim.

3.1 Stanowisko pomiarowe

W celu przeprowadzenia badań numerycznych dokładności estymacji położenia UE bazując na sygnałach sieci komercyjnej zaprojektowano i wykonano uniwersalne stanowisko testowe. Do pozyskania próbek sygnałów radiowych, i ich dalszego przetwarzania przy pomocy opracowanego oprogramowania, wykorzystano platformę SDR (Software Defined Radio), bazując na urządzeniu USRP (Universal Software Radio Peripheral) X310 i B200 mini [7] wraz z komputerem klasy PC. Sygnał interfejsu radiowego 5G-NR próbkowano z częstotliwością 30,72 MHz, czyli czterokrotnie wyższą niż minimum wymagane do analizy bloku synchronizacyjnego SS/PBCH. Schemat blokowy stanowiska testowego przedstawiono na Rys. 1.



Rysunek 1. Schemat stanowiska badawczego

Sygnał radiowy interfejsu radiowego 5G-NR odbierano z użyciem anteny o charakterystyce dookólnej Cobham AO2-0,3-10,0V/1667 [8] podłączonej niskostratnym kablem Huber-Suhner Sucoflex. W scenariuszu ZB zastosowano dodatkowy tłumik, ze względu na wysoką moc sygnału wynikającą z bliskości stacji gNodeB. Częstotliwości środkowe wynosiły 783,75 MHz (P1), 2,59455 GHz (P2) i 3,62064 GHz (P3), co odpowiada aktualnemu rozmieszczeniu w Polsce i Portugalii sieci komercyjnych 5G. Przed właściwą rejestracją sygnałów zweryfikowano częstotliwość środkową, na której znajdował się blok synchronizacyjny, tak aby zarejestrować tylko odpowiednią część pasma systemu oraz dodatkowo sprawdzano, czy przesyłane są sygnały PRS [5, 6].

Sygnały DL interfejsu 5G-NR rejestrowano stacjonarnie w trzech lokalizacjach odpowiadających różnym rodzajom środowiska miejskiego, jako typowe lokalizacje użytkowników, tj. ZB, WB oraz w tzw. środowisku D-I. Metrykami zastosowanymi do rozróżnienia tych środowisk był stosunek składowej bezpośredniej sygnału (LoS - Line of Sight) do pozostałych składowych (innych niż LoS), reprezentowany przez współczynnik Rice'a K i wartość skuteczną rozmycia opóźnienia (RMS delay spread) σ [9, 10, 11]. Wyznaczone na podstawie zarejestrowanych sygnałów wartości parametrów dla analizowanych środowisk i pasm przedstawiono w Tab. 1.

Tabela 1. Parametry statystyczne analizowanych środowisk

Rodzaj środowiska	*Współczynnik Rice'a K [dB]*	*RMS delay spread σ [ns]*
ZB **P1**	15.8	38
WB **P1**	5.6	39
D-I **P1**	-19.3	68
ZB **P2**	14.0	41.2
WB **P2**	-0.7	102.6
D-I **P2**	-13.2	60.5
ZB **P3**	12.8	22
WB **P3**	-2.6	23
D-I **P3**	-17.3	41

3.2 Metodyka lokalizowania terminala użytkownika

Położenie terminala użytkownika wyznaczano z użyciem metody TDOA (Time Difference of Arrival) wraz z algorytmem Foya [12, 13]. Głównym celem przeprowadzonych badań było zweryfikowanie wpływu typu środowiska na dokładność wyznaczanych estymat położenia (tj. nie wpływu geometrii gNodeB). W celu emulowania procesu radiolokalizowania zarejestrowane sygnały z różnych gNodeB zostały przetworzone poprzez dodanie próbek opóźniających [13]. Operacja ta pozwoliła wyznaczać położenie UE na obszarze testowym z optymalną, trójkątną geometrią stacji referencyjnych gNodeB. W związku z tym możliwe było zbadanie dokładności wyznaczanego położenia, na którą wpływa jakość detekcji sygnału PSS.

Wyznaczono kwadratowy obszar badawczy o wymiarach 700×700 m^2, na którym umieszczono 400 punktów badawczych w rozstawie osiowym 35 m. Obszar ten znajdował się wewnątrz trójkąta równobocznego, którego wierzchołki stanowiły stacje gNodeB. Obszar testowy przedstawiono na Rys. 2, gdzie niebieskie znaki + reprezentują emulowane punkty testowe, zaś czerwone wyznaczone położenie UE. Stacje gNodeB oznaczono czarnymi, kwadratowymi punktami.

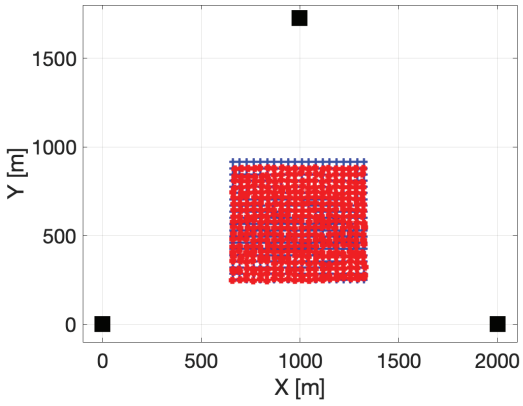
Dokładność estymacji położenia oceniano poprzez wyznaczenie wartości błędu średniokwadratowego RMSE (Root Mean Square Error) pomiędzy pozycjami oszacowanymi a pozycjami referencyjnymi oraz wskazując odchylenie standardowe σ_{RMSE} występujących błędów estymat położenia. Ponadto sprawdzono dla ilu punktów testowych algorytm Foya nie osiągnął zbieżności lub wartość błędu RMSE przekroczyła połowę odległości pomiędzy gNodeB (tj. 1000 m), czyli odsetek punktów, dla któ-

rych nie udało się wyznaczyć prawidłowego położenia.

4. ANALIZA UZYSKANYCH WYNIKÓW

Wybrane wyniki estymacji położenia UE przedstawiono na Rys. 2, 3, odpowiednio dla środowiska zewnątrz- i wewnątrzbudynkowego dla sygnału pasma P2. Na rysunkach przedstawiono także rzeczywiste położenie punktów testowych wraz z wyznaczonymi estymatami położenia. Ponadto na Rys. 4 pokazano ujęcie statystyczne dla wszystkich analizowanych przypadków w postaci dystrybuant błędów, zaś w Tab. 2 zawarto wartości numerycznej oceny dokładności estymacji położenia.

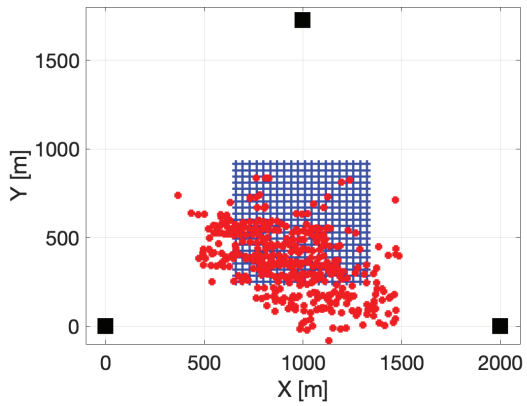
Rys. 2 przedstawia niemal całkowite pokrycie się wyznaczonych estymat położenia (ZB, P2) z punktami referencyjnymi (RMSE równe 22 m i wartość 39 m dla 90% przypadków, wraz z σ_{RMSE} wynoszącym 11 m), gdzie najlepszym uzyskanym wynikiem (P3) jest średnia błędu RMSE równa 5 m i wartość 10 m dla 90% przypadków, wraz z odchyleniem standardowym wynoszącym 2 m. Warto zauważyć, iż przyjęte podejście do rozmiarów siatki punktów testowych wydaje się być słuszne, ponieważ punkty testowe są od siebie oddalone o 35 m w kierunku pionowym i poziomym. Ponadto można zaobserwować, że kształt obszaru testowego jest zachowany, bez znaczących deformacji geometrycznych spowodowanych np. niewłaściwą detekcją sygnałów PSS, co pozwala na osiągnięcie zbieżności algorytmu Foya we wszystkich przypadkach. Zatem analizowany przypadek jest środowiskiem o silnej, dominującej składowej bezpośredniej i stosunkowo niskim parametrze σ.



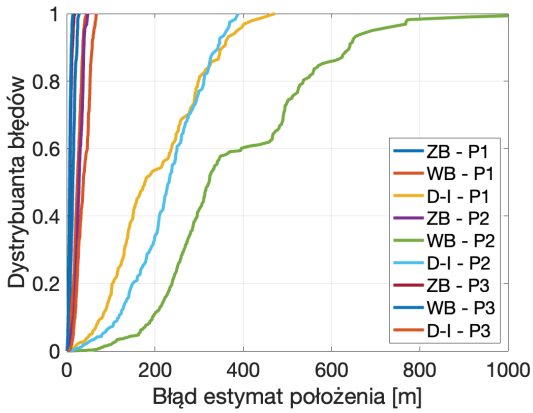
Rysunek 2. Punkty testowe (kolor niebieski) oraz wyznaczone estymaty położenia (kolor czerwony) w środowisku ZB dla pasma P2

Dla środowiska wewnątrzbudynkowego, Rys. 3, można zaobserwować znaczący wzrost błędu wyznaczonych estymat położenia. Dla P2 (przypadek najgorszy) błąd RMSE wynosi 640 m dla 90% procenta przypadków przy wartości średniej 322 m i σ_{RMSE} 165 m, co znacznie przekracza rozdzielczość siatki wynoszącej 35 m.

Z analizy procesu obliczeniowego wynika, że zjawisko wyraźnie większego błędu RMSE wyznaczonych estymat położenia dla środowiska wewnątrzbudynkowego jest spowodowane niewłaściwą detekcją sygnału PSS. Gdy składowa bezpośrednia nie jest dominująca, podczas detekcji sygnału PSS wybierana jest jedna ze składowych propagacji wielodrogowej, co ostatecznie nie jest procesem ciągłym i jednolitym we wszystkich punktach testowych. Skutkuje to wysokim błędem wyznaczanych estymat położenia i jego dużą zmiennością. Drugą, bezpośrednią przyczyną występujących błędów jest opóźnienie propagacyjne wybranej, dominującej składowej propagacji wielodrogowej. W wielu przypadkach powoduje to alternatywną selekcję różnych składowych wielodrogowych, co można zaobserwować także na wykresie dystrybuant (Rys. 4). Warto wspomnieć, że mały rozrzut opóźnienia nie powoduje nierealistycznych danych wejściowych dla metody TDOA i algorytmu Foya. Algorytm uzyskuje zbieżność w znaczącym odsetku przypadków, zatem wyznaczone estymaty położenia należy poddać dalszej analizie, aby można je było odrzucić w przypadku np. dużego przemieszczenia pomiędzy pobliskimi punktami testowymi.



Rysunek 3. Punkty testowe (kolor niebieski) oraz wyznaczone estymaty położenia (kolor czerwony) w środowisku WB dla pasma P2



Rysunek 4. Dystrybuanty błędów wyznaczonych estymat położenia

W wynikach dla środowiska typu D-I (P2) można zaobserwować wartości błędu RMSE mniejsze niż dla środowiska wewnątrzbudynkowego, tj. 342 m i 229 m odpowiednio dla 90% i 50% procenta przypadków, przy σ_{RMSE} 102 m. Zauważalny jest jednak silny wzrost liczby punktów testowych, w których algorytm Foya nie uzyskał zbieżności, tj. w 33% przypadków. Główną przyczyną tego zjawiska jest dodatkowe tłumienie propagacyjne obserwowane w tym środowisku, gdyż wszystkie odbierane składowe sygnału radiowego są składowymi propagacji wielodrogowej NLoS (Non-Line of Sight), co przekłada się pogorszenie jakości detekcji sygnału PSS [13]. Niemniej jednak więk-

szość składowych wielodrogowych została wytłumiona, co eliminuje problem naprzemiennego wykrywania sygnałów odbitych jako dominujących (jak w przypadku środowiska wewnątrzbudynkowego).

Tabela 2. Parametry statystyczne wyznaczonych estymat położenia

Rodzaj środowiska	*RMSE [m]*	σ_{RMSE} *[m]*	*Brak estymat [%]*
ZB **P1**	10	6	0
WB **P1**	49	29	0
D-I **P1**	128	99	< 18
ZB **P2**	22	11	0
WB **P2**	295	165	< 2
D-I **P2**	168	102	< 33
ZB **P3**	5	2	0
WB **P3**	6	4	0
D-I **P3**	31	16	< 36

5. PODSUMOWANIE

W artykule przedstawiono badania eksperymentalne, których celem było zweryfikowanie dokładności estymacji położenia UE w sieci 5G-NR działającej w różnych środowiskach miejskich. Rejestrację sygnałów PSS w łączu w dół interfejsu 5G-NR przeprowadzono w trzech różnych środowiskach (zewnątrz- i wewnątrzbudynkowym oraz typu deep-indoor) z niezależnej sieci komercyjnej na obszarze miejskim.

Z przedstawionych analiz wynika, że nawet jeśli użytkownik zmieni swoje położenie w promieniu do 30 m, np. wchodząc do budynku, warunki lokalizacji i osiągnięta dokładność wyznaczanych estymat położenia znacznie się różni. Zmiana środowiska powoduje znaczącą degradację dokładności lokalizowania UE, gdzie błąd RMSE może zawierać się zakresie od 6 m do 295 m. Skutkiem są nierealistyczne przemieszczenia użytkownika na obszarze świadczenia usługi lokalizacyjnej, np. do innego pobliskiego budynku.

W ramach przyszłych prac badawczych autorzy planują rozbudować metodę detekcji sygnału PSS w łączu DL w celu zwiększenia dokładności estymacji położenia, m.in. poprzez analizę zależności czasowych mającą na celu kompensację negatywnego wpływu propagacji wielodrogowej, przeanalizować wpływ synchronizacji gNodeB na dokładność estymacji położenia czy też wpływ zwiększonej częstości emisji bloku SS/PBCH przez głowice radiowe stosujące technikę Massive MIMO (Multiple Input, Multiple Output).

6. FINANSOWANIE

Wsparcie finansowe przeprowadzonych badań zostało przyznane przez Politechnikę Gdańską w formie grantu DEC-03/2023/IDUB/II.1/AMERICIUM w ramach programu AMERICIUM – „Inicjatywa Doskonałości – Uczelnia Badawcza".

LITERATURA

[1] Kottkamp M., A. Pandey, A. Roessler, R. Stuhlfauth, D. Raddino. 2022. „5G New Radio - Fundamentals, Procedures, Testing Aspects" Rohde & Schwarz GmbH & Co. KG, ISBN 9783939837152.

[2] Traboulsi S. 2022. „Overview of 5G-oriented Positioning Technology in Smart Cities" Procedia Computer Science, Wol. 201, 368-374, doi: https://doi.org/10.1016/j.procs.2022.03.049.

[3] Del Peral-Rosado J. A., R. Raulefs, J. A. López-Salcedo, G. Seco-Granados. 2018. „Survey of Cellular Mobile Radio Localization Methods: from 1G to 5G" IEEE Communications Surveys & Tutorials, Wol. 20, 1124 - 1148, doi: 10.1109/COMST.2017.2785181.

[4] 3GPP - 3rd Generation Partnership Project. 2018. „Technical Specification Group Services and System Aspects, Study on positioning use cases," 3GPP TR 22.872, V16.1.0.

[5] 3GPP - 3rd Generation Partnership Project. 2023. „NR, Physical channels and modulation," 3GPP TS 38.211, V17.6.0.

[6] Liberg O., Mårten Sundberg, E. Wang, J. Bergman, J. Sachs, G. Wikström. 2020. „Cellular Internet of Things, From Massive Deployments to Critical 5G Applications," Academic Press.

[7] Ettus Research. 2023. „USRP X3x0 Series, UHD and USRP Manual", Ver. 4.6.0.0-7-gece7c4811, https://files.ettus.com/manual/pageusrpx3x0.html. Data dostępu: 10.05.2024

[8] Cobham. 2011. „OA2-0.3-10.0V/1505, Ultra Wide Band Omni Antenna" Ver. 060510, https://www.european-antennas.co.uk/media/1638/ds1505-060510.pdf. Data dostępu: 10.05.2024

[9] ITU – International Telecommunication Union. 2021. „Recommendation, Multipath Propagation and Parametrization of its Characteristics", ITU-R P.1407-8, 9, 2021. https://www.itu.int/ dms-pubrec/itu-r/rec/p/R-REC-P.1407-8-202109-I!!PDF-E.pdf, Data dostępu: 10.05.2024

[10] Wang J., Y. Shi, C. Yang, S. Ji, H. Su. 2020. „Research on Fading Characteristics of Ultrahigh Frequency Signals in Karst Landform Around Radio Quiet Zone of FAST" Radio Science, Wol. 55, Wyd. 10, 1 - 10, doi: 10.1029/2019RS007048

[11] Adegokeet E. I., E. Kampert, M. D. Higgins. 2021. „Channel Modeling and Over-the-Air Signal Quality at 3.5 GHz for 5G New Radio" IEEE Access, Wol. 9, 11183 - 11193, doi: 10.1109/ACCESS.2021.3051487

[12] Sadowski J. 2018. „Badanie i analiza systemów radiolokalizacyjnych do zastosowań specjalnych", Politechnika Gdańska, ISBN 978-83-7348-745-1

[13] Rajchowski P., K. K. Cwalina, J. Sadowski. 2023. „The Influence of the Radio Channel on Precision of Position Estimation of the User Terminal Using the NB-IoT Radio Interface" XXXVth General Assembly and Scientific Symposium of the International Union of Radio Science (URSI GASS), 1-6, doi: 10.23919/URSIGASS57860.2023.10265561

NOWA METODA WYDZIELANIA SYGNAŁU INTENCJONALNYCH ZAKŁÓCEŃ W TRANSMISJI RADIOWEJ MIMO

A NEW METHOD OF JAMMING SIGNAL ISOLATION IN MIMO RADIO TRANSMISSION

Krzysztof Wesołowski

Politechnika Poznańska, Poznań, krzysztof.wesolowski@put.poznan.pl

DOI: 10.15199/59.2024.4.45

Streszczenie: W artykule przedstawiono metodę separacji sygnału intencjonalnych zakłóceń z ważonej sumy sygnału użytecznego i zakłócającego. Metoda jest oparta na zasadzie ślepej separacji źródeł sygnałów wymagającej jedynie określonej liczby próbek sygnału odebranego i odpowiedniej liczby anten odbiornika MIMO. Jako kryterium optymalizacji algorytmu odbioru zastosowano kryterium adaptacji zaproponowane wcześniej przez autora do ślepej korekcji adaptacyjnej sygnału danych. Przedstawiono również wyniki symulacji modelowanego systemu.

Abstract: In this paper, we present a method of jamming signal separation from the mixture of useful and jamming signals based on the blind source separation principle that requires a sufficient number of received signal samples and MIMO the receiver. The optimization criterion used in the receiver is based on the adaptation criterion proposed earlier by the author for application in blind adaptive equalization. The simulation results for the proposed algorithm are also presented.

Słowa kluczowe: Ślepa separacja źródeł sygnałów, systemy MIMO, intencjonalne zakłócanie.

Keywords: Blind source separation, jamming, MIMO systems.

1. WPROWADZENIE

Intencjonalne zakłócanie transmisji radiowej, tzw. *jamming* jest jedną z metod ataków na systemy radiowe, w tym systemy komórkowe. Ataki te mogą obniżyć jakość działania a nawet uniemożliwić funkcjonowanie takich systemów. W przypadku sieci 5G jest to szczególnie istotne dla jej usług typu URLLC (ang. *ultra-reliable low-latency communications*). Minimalizacja wpływu jammingu może być realizowana w dwóch krokach: jego detekcji (np. według metody opisanej w [1]), a następnie jego eliminacji za pomocą wybranej metody.

Autorzy pracy [2] zidentyfikowali szereg metod zwalczania jammingu w systemach komórkowych podając ich bogaty przegląd i odniesienia do literatury. Z powodu ograniczeń objętości artykułu nie mogą być tu zaprezentowane. Większość z nich jest związana z transmisją wielotonową dominującą w obecnie działających systemach komórkowych. Metody zwalczania jammingu dotyczą najczęściej warstwy fizycznej i algorytmów przetwarzania sygnałów. Oprócz nich istnieją również propozycje zmiany trasy przesyłanych pakietów do innej stacji bazowej, która może obsłużyć terminal, gdy jego dotychczasowa stacja jest zagłuszana. Techniki kodowania i skramblingu to kolejne metody walki z zagłuszaniem relacjonowane w literaturze.

W ostatnich latach można znaleźć w literaturze prace na temat metod zwalczania zagłuszania bazujących na tzw. ślepej separacji źródeł sygnałów (ang. *blind source separation* (BSS)) [3]. Była one jednak najczęściej stosowana do separacji sygnałów sinusoidalnych lub dźwiękowych. W metodach tych kryterium optymalizacji ma charakter nieliniowy. Raportowano więc zastosowanie statystyk wyższych rzędów (ang. *kurtosis*), miar różnicy między entropią procesu gaussowskiego i niegaussowskiego (ang. *negentropy*), czy kryterium maksimum wiarygodności. Ślepa separacja źródeł sygnałów w swoim działaniu wykorzystuje wyłącznie odebrane próbki sygnału, który jest ważoną sumą sygnałów z każdego źródła. Ostatnio BSS zaprezentowano w [4] jako metodę zintegrowanego odbioru realizowanego przez odbiornik wieloantenowy sygnałów radarowych, telekomunikacyjnych i zakłócających.

2. ŚLEPA SEPARACJA ŹRÓDEŁ SYGNAŁÓW

Podstawowy model BSS jest opisany następującym równaniem:

$$\mathbf{x}(t) = \mathbf{H}\mathbf{s}(t) + \mathbf{n}(t) \tag{1}$$

gdzie $\mathbf{H}$ jest macierzą pełnego rzędu o wymiarach $[N \times M]$ z nieznanymi współczynnikami kanałów, $\mathbf{s}(t) = [s_1(t), s_2(t), ..., s_M(t)]^T$ jest wektorem sygnałów z M źródeł, $s_i(t)$, $s_j(t)$, $i, j = 1, 2, ..., M$, które są statystycznie niezależne. Wektor $\mathbf{x}(t) = [x_1(t), x_2(t), ..., x_N(t)]^T$ oznacza sygnały otrzymane na N antenach odbiornika, zaś $\mathbf{n}(t) = [n_1(t), n_2(t), ..., n_N(t)]^T$ opisuje addytywny szum gaussowski na wejściach odbiornika. Zmienna t oznacza chwile czasowe, zazwyczaj dyskretne. Głównym zadaniem odbiornika jest odseparowanie sygnałów źródeł $\{s_i(t)\}$, $i = 1, ..., M$ na podstawie wyłącznie obserwowanego wektora próbek $\mathbf{x}(t)$. Każdy składnik tego wektora $x_j(t)$, $j = 1, .., N$ jest ważoną sumą sygnałów ze wszystkich źródeł i na podstawie prawa wielkich liczb jego rozkład dąży do rozkładu gaussowskiego, chociaż sygnały składowe mają inne rozkłady prawdopodobieństwa. Istotnym narzędziem separacji źródeł sygnałów jest tzw. analiza niezależnych składników (ang. *independent component analysis* (ICA)) [5]. Działa ona przy założeniu, że wśród separo-

wanych składników co najwyżej jeden z nich ma rozkład gaussowski a wszystkie składniki są statystycznie wzajemnie niezależne. Niestety ICA nie jest w stanie zidentyfikować dokładnie liczby źródeł sygnałów, ich uporządkowania i określić ich skali i znaku.

Zadaniem metody ICA jest transformacja obserwowanego ciągu wektorów próbek $\mathbf{x}(t)$ w wektor $\mathbf{z}(t)$ odseparowanych sygnałów z zastosowaniem liniowej statycznej transformacji $\mathbf{Q}$ takiej, że

$$\mathbf{z}(t) = \mathbf{Q}\mathbf{x}(t). \tag{2}$$

Wektor możliwie statystycznie niezależnych składników $\mathbf{z}(t)$ jest otrzymywany za pomocą wybranej funkcji kosztu (ang. *contrast function*) mierzącej ich niezależność. Zazwyczaj odbywa się to w dwóch krokach. W pierwszym kroku obserwowany ciąg wektorów próbek podlega wybieleniu za pomocą macierzy filtru wybielającego $\mathbf{W}$.

$$\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t). \tag{3}$$

W drugim kroku wektor $\mathbf{y}(t)$ jest mnożony przez macierz separacji $\mathbf{B}$, czyli

$$\mathbf{z}(t) = \mathbf{B}\mathbf{y}(t). \tag{4}$$

Jest ona zwykle uzyskiwana w sposób iteracyjny z zastosowaniem wybranej funkcji kosztu mierzącej niezależność składników otrzymanego wektora $\mathbf{z}(t)$. Jak widać, $\mathbf{Q} = \mathbf{BW}$.

Znaczna liczba algorytmów ICA bazuje na tym, że przedmiotem separacji sygnałów z różnych źródeł są takie, które nie mają rozkładu gaussowskiego a funkcja kosztu mierzy jak dalece różnią się one od sygnału o rozkładzie Gaussa. Kilka lat temu zaproponowano funkcję kosztu bazującą na technice Bussganga.

3. TECHNIKA BUSSGANGA

Proces stochastyczny $z(n)$ jest nazwany procesem Bussganga, jeśli spełnia następujący warunek [6]:

$$E[z(n)z(n-k)] = E\left[z(n)g(z(n-k))\right].$$

Taki sam warunek może być spełniony w przypadku odbiornika MIMO dla wektora odebranych próbek $\mathbf{z}_n = \mathbf{z}(n\Delta t)$, przy czym Δt jest odstępem próbkowania. Warunek ten ma wtedy postać:

$$E\left[g(\mathbf{z}_n)\mathbf{z}_n^H\right] = E\left[\mathbf{z}_n\mathbf{z}_n^H\right]. \tag{5}$$

Funkcja $g(.)$ opisuje bezpamięciową nieliniowość, a wtedy

$$g(\mathbf{z}_n) = \left[g(z_{n,1}), g(z_{n,2}), ..., z_{n,M}\right]^T.$$

Technika Bussganga była stosowana w teorii filtrów adaptacyjnych. Wśród funkcji nieliniowych $g(.)$ jest również taka, którą zaproponował autor niniejszego artykułu w 1992 [7]. Dla pojedynczego sygnału ma ona postać:

$$\begin{aligned} g(z_{k,i}) = &\left(\left(\mathrm{Re}(z_{k,i})\right)^2 - A_2 + \mathrm{Re}(z_{k,i})\right) \\ &+ j\left(\left(\mathrm{Im}(z_{k,i})\right)^2 - A_2 + \mathrm{Im}(z_{k,i})\right) \end{aligned} \tag{6}$$

gdzie stała A_2 zależy od zastosowanej modulacji. W przypadku QPSK/QAM (por. [7]), dla których symbole danych sygnału użytecznego i zakłócającego mają postać $d_n = d_n^I + jd_n^Q$, A_2 jest dane wzorem $A_2 = \frac{E[(d_n^I)^4]}{E[(d_n^I)^2]} = \frac{E[(d_n^Q)^4]}{E[(d_n^Q)^2]}$. Dlatego funkcja stosowana w kryterium adaptacji jest następująca:

$$|g(z_{k,i}) - z_{k,i}|^2 = \left(\left(\mathrm{Re}(z_{k,i})\right)^2 - A_2\right)^2 + \left(\left(\mathrm{Im}(z_{k,i})\right)^2 - A_2\right)^2 \tag{7}$$

Funkcja ta może zostać łatwo zaadaptowana do zastosowania w ślepej separacji źródeł sygnałów.

4. SEPARACJA SYGNAŁÓW ZA POMOCĄ BSS

Jak pamiętamy, algorytm ICA najpierw wybiela sekwencję wektorów próbek wejściowych $\mathbf{x}(t)$ za pomocą transformacji (3). Zatem konieczne jest obliczenie estymaty macierzy wybielającej $\mathbf{W}$, Najpierw estymuje się macierz autokorelacji $\mathbf{R} = E[\mathbf{x}(t)\mathbf{x}(t)^H]$, gdzie $(.)^H$ oznacza hermitowską transpozycję. W praktyce estymacja macierzy $\mathbf{R}$ jest realizowana przez uśrednianie w czasie $n\Delta t$ kolejnych odstępów próbkowania:

$$\mathbf{R} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^H \tag{8}$$

Jeśli charakterystyka kanału zmienia się w czasie, wtedy estymata funkji autokorelacji może być wyliczona ze wzoru

$$\mathbf{R} = \frac{1-\lambda}{1-\lambda^n}\sum_{i=1}^{n}\lambda^{n-i}\mathbf{x}_i\mathbf{x}_i^H \tag{9}$$

gdzie $\lambda \leq 1$ jest współczynnikiem okna geometrycznego. W wersji rekursywnej można zastosować wzór:

$$\mathbf{R}_i^{'} = \lambda\mathbf{R}_{i-1}^{'} + \mathbf{x}_i\mathbf{x}_i^H, \quad i = 1, 2, ..., n \quad \mathbf{R} = \frac{1-\lambda}{1-\lambda^n}\mathbf{R}_n^{'}. \tag{10}$$

Z kolei dokonujemy dekompozycji estymaty macierzy $\mathbf{R}$ na macierze wektorów własnych i wartości własnych:

$$\mathbf{R} = \mathbf{U}\Lambda\mathbf{U}^H \tag{11}$$

gdzie $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_N]$ jest macierzą wektorów własnych a wektor $\mathbf{u}_i$ jest stowarzyszony z wartością własną λ_i, $i = 1, 2, ..., N$, zaś $\mathrm{diag}(\Lambda) = [\lambda_1, \lambda_2, ..., \lambda_N]^H$. Z kolei wartości własne zostają uporządkowane od największej do najmniejszej. M największych wartości własnych jest skojarzonych z M źródłami, zaś pozostałe są związane z szumem addytywnym. Gdy $M = 2$, (tj. mamy jeden sygnał użyteczny i jeden zakłócający), można dowieść, że macierz wybielająca jest dana równaniem:

$$\mathbf{W} = \left[(\lambda_1 - \tilde{\sigma}^2)^{-1/2}\mathbf{u}_1 \quad (\lambda_2 - \tilde{\sigma}^2)^{-1/2}\mathbf{u}_2\right]^H \tag{12}$$

gdzie λ_1 oraz λ_2 są dwiema największymi wartościami własnymi a $\tilde{\sigma}^2$ jest średnią pozostałych $N - M$ wartości własnych.

W kolejnym kroku realizowany jest właściwy algorytm separacji źródeł. Tak więc wybielony wektor danych wejściowych $\mathbf{y}(t)$ jest pomnożony przez macierz separacji $\mathbf{B}$, dając w wyniku wektor $\mathbf{z}(t)$. W tym celu należy wyznaczyć macierz separacji $\mathbf{B}$. Proponuje się, aby dokonać minimalizacji następującego kryterium:

$$J(\mathbf{B}_n) = E\left[\left\|g(\mathbf{z}_k) - \mathbf{z}_k\right\|^2\right] = E\left[\sum_{i=1}^{M}\left|g(z_{k,i}) - z_{k,i}\right|^2\right] \tag{13}$$

gdzie $\mathbf{z}_k = [z_{k,1}, z_{k,2}, ..., z_{k,M}]^T$ jest wektorem próbek wektora $\mathbf{z}(t)$ w chwilach $t = k\Delta t$. Jak wynika ze wzoru (13) $J(\mathbf{B}_n)$ jest miarą błędu średniokwadratowego pomiędzy wektorem $\mathbf{z}_k$ wyjść układu separacji a wektorem jego nieliniowego przekształcenia $g(\mathbf{z}_k)$. Zazwyczaj macierz separacji otrzymuje się iteracyjnie za pomocą algorytmu gradientowego.

$$\mathbf{B}_{n+1} = \mathbf{B}_n - \alpha \cdot \underset{\mathbf{B}_n}{\mathrm{grad}}\, J(\mathbf{B}_n) \tag{14}$$

gdzie α jest krokiem adaptacji algorytmu. W przypadku zastosowanej funkcji (7) gradient miary $J(\mathbf{B}_n)$ jest określony wyrażeniem

$$\begin{aligned} \underset{\mathbf{B}_n}{\mathrm{grad}}\, J(\mathbf{B}_n) = 4E\Big[\Big[&((\mathrm{Re}(\mathbf{z}_n))^2 - A_2)\mathrm{Re}(\mathbf{z}_n) \\ &+ j((\mathrm{Im}(\mathbf{z}_n))^2 - A_2)\mathrm{Im}(\mathbf{z}_n)\Big]\mathbf{y}_n^*\Big] \end{aligned} \tag{15}$$

Zazwyczaj jednak stosuje się jego stochastyczną wersję i wtedy zastosowane jest wyrażenie znajdujące się w nawiasach operacji wartości oczekiwanej (15).

Zastosowanie stochastycznego algorytmu gradientowego wymaga zazwyczaj wielu iteracji aby osiągnąć wartość kryterium bliską minimum. Zamiast niego, często w adaptacji sieci neuronowych stosuje się tzw. *gradient naturalny* [8] charakteryzujący się znacznie wyższą szybkością zbieżności. W przypadku ogólnym zwykły gradient jest uzupełniony lewostronnie o odwrotność macierzy zwanej *macierzą informacji Fishera*. Z powodu ograniczeń w objętości niniejszego artykułu autor proponuje zainteresowanym czytelnikom studia literaturowe na ten temat. Poniżej podajemy jedynie wersję stochastyczną gradientu oznaczaną jako gradNat:

$$\underset{\mathbf{B}_n}{\text{gradNat}}\, J(\mathbf{B}_n) = \left(\mathbf{e}_n \mathbf{z}_n^H - \mathbf{z}_n \mathbf{e}_n^H\right) \mathbf{B}_n \tag{16}$$

W której wektor błędu $\mathbf{e}_n$ jest dany wzorem

$$\begin{aligned}\mathbf{e}_n = &\left[\left((\mathrm{Re}(\mathbf{z}_n))^2 - A_2\right) \mathrm{Re}(\mathbf{z}_n) - \mathrm{Re}(\mathbf{z}_n)\right] \\ &+ j\left[\left((\mathrm{Im}(\mathbf{z}_n))^2 - A_2\right) \mathrm{Im}(\mathbf{z}_n) - \mathrm{Im}(\mathbf{z}_n)\right].\end{aligned} \tag{17}$$

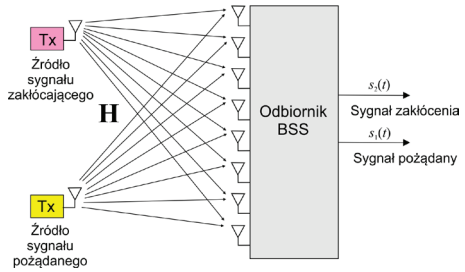
Ostatecznie zastosowanie algorytmu z gradientem naturalnym prowadzi do następującego algorytmu iteracyjnego wyznaczającego macierz separacji $\mathbf{B}_n$:

$$\mathbf{B}_{n+1} = \mathbf{B}_n - \alpha \cdot \left(\mathbf{e}_n \mathbf{z}_n^H - \mathbf{z}_n \mathbf{e}_n^H\right) \mathbf{B}_n \tag{18}$$

Zwróćmy uwagę, że algorytm minimalizujący wyrażenie (7) może dokonać zbieżności do jednego z czterech równoważnych minimów, co prowadzi do różnicy faz, które są wielokrotnością 90 stopni. Zatem należy zidentyfikować pożądaną sekwencję danych na podstawie znanych pilotów lub sygnatur. Ten krok pozwala również na identyfikację sygnału pożądanego w porównaniu z sygnałem jammingu. Pomimo tych niedogodności, zastosowane kryterium pozwala nie tylko na separację sygnałów z różnych źródeł, ale również ich detekcję.

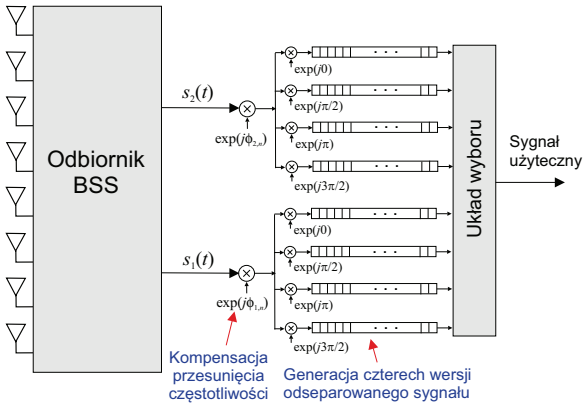
5. MODEL SYSTEMU ZASTOSOWANEGO W SYMULACJI I JEJ WYNIKI

Zastosowany model systemu (rys. 1) odpowiada transmisji na pojedynczej podnośnej systemu OFDM, dlatego każdy kanał składowy dla pary anten nadawczej i odbiorczej może być scharakteryzowany pojedynczym współczynnikiem, a wszystkie kanały składowe macierzą współczynników **H**. W badaniach zastosowano modulację QPSK dla sygnału pożądanego oraz jammingu.



Rysunek 1: Model transmisji MIMO z odbiornikiem BSS

W odbiorniku należy wziąć pod uwagę dwa dodatkowe czynniki: fakt istnienia czterech równoważnych minimów zastosowanej funkcji kosztu oraz to, że możliwy jest offset częstotliwości pomiędzy sygnałem pożądanym a sygnałem jammingu. Z tego powodu odbiornik powinien zostać uzupełniony o dodatkowe bloki funkcjonalne, co pokazano na rys. 2.



Rysunek 2: Odbiornik BSS z kompensacją przesunięcia częstotliwości i wyborem wersji sygnału

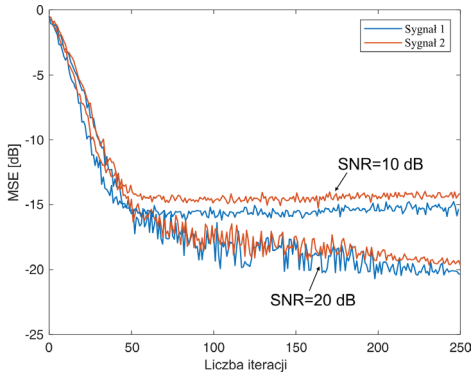
W celu kompensacji przesunięcia fazy i częstotliwości należy zastosować cyfrową pętlę synchronizacji fazy drugiego rzędu (DPLL) opisaną równaniami:

$$\varepsilon_{n,i} = \mathrm{Im}\left(e_{n,i}^* z_{n,i}\right), \quad \phi_{n,i} = \phi_{n-1,i} + \mu(\varepsilon_{n,i} + \rho S_{n-1,i}) \tag{19}$$

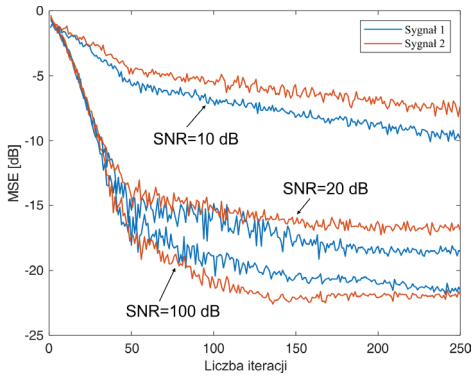
$$S_{n,i} = \lambda S_{n-1,i} + (1-\lambda)\varepsilon_{n,i} \tag{20}$$

gdzie kroki adaptacji μ oraz ρ determinują możliwości śledzenia przesunięcia fazy i częstotliwości przez pętlę DPLL. W celu oceny efektywności działania ślepej separacji sygnałów z zaproponowaną funkcją kosztu (13) zbudowano prosty model symulacyjny, w którym nadawane są z dwóch źródeł (pożądanego oraz zakłócającego) dwa sygnały QPSK ($M = 2$) odbierane przez odbiornik z N antenami. Założono, że każdy kanał składowy jest scharakteryzowany przez zespolony współczynnik o rozkładzie normalnym $N(0,1)$. Losowane na początku przebiegu symulacyjnego współczynniki kanału pozostają niezmienne w jego trakcie. Zastosowano algorytm aktualizacji macierzy separacji $\mathbf{B}_n$ za pomocą stochastycznego gradientu naturalnego z krokiem adaptacji α zmniejszanym k-krotnie po L_{est} krokach. Estymaty błędu średniokwadratowego na wyjściu odbiornika są średnią z $L_{run} = 500$ przebiegów symulacyjnych, zaś w pojedynczym przebiegu transmitowanych jest $L_{iter} = 500$ symboli. Zespolone próbki szumu gaussowskiego są dodawane na wejściu każdej anteny odbiornika, zaś $SNR = 10\log_{10}(P_{\mathrm{QPSK}}/\sigma^2)$. Badano również wpływ przesunięcia częstotliwości sygnału jammingu na jakość detekcji.

Pomimo, że model symulacyjny systemu jest bardzo prosty, dobrze modeluje on sytuację na pojedynczej podnośnej systemu OFDM. Wiedząc, że podnośne w pojedynczym bloku fizycznym (PRB) poddawane są oddziaływaniu bardzo podobnych kanałów, byłoby interesujące sprawdzenie ile bloków PRB jest niezbędnych do estymacji macierzy autokorelacji i działania algorytmu separacji źródeł. Zakładając 12 podnośnych oraz 7 symboli OFDM w pojedynczym PRB, algorytm miałby do dyspozycji 84 symbole danych. Kolejne rysunki ilustrują działanie algorytmu z proponowanym kryterium adaptacji.

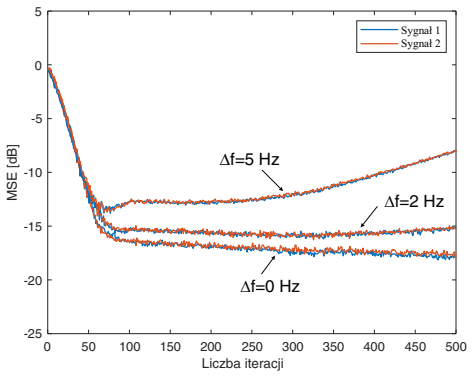


Rysunek 3: Błąd średniokwadratowy (MSE [dB]) na wyjściu odbiornika BSS ($N = 8$ anten, $k = 5$, $\alpha = 0.015$, $L_{est} = 168$, $L_{thres} = 50$, $SNR = 20$ dB lub $SNR = 10$ dB)



Rysunek 4: Błąd średniokwadratowy (MSE [dB]) na wyjściu odbiornika BSS ($N = 4$ anteny, $k = 5$, $\alpha = 0.015$, $L_{est} = 84$, $L_{thres} = 50$, $SNR = 100,\ 20$, and 10 dB)

Na rys. 3 przedstawiono krzywe zbieżności dla wartości SNR równej 20 i 10 dB, gdy odbiornik dysponuje $N = 8$ antenami, Zmniejszenie liczby anten do $N = 4$ (rys. 4) skutkuje znaczącym pogorszeniem jakości detekcji, gdy $SNR = 10$ dB. Z kolei na rys. 5 zaprezentowano krzywe MSE w funkcji liczby iteracji, gdy sygnał zagłuszający cechuje się przesunięciem częstotliwości względem sygnału pożądanego. Jest to równoważne sytuacji, gdy kanał ze źródła zagłuszającego do odbiornika jest zmienny w czasie. Z tego powodu w estymacji macierzy autokorelacji wektora sygnału odebranego i w pętli DPLL zastosowano okno geometryczne ze współczynnikiem $\lambda = 0.99$. Macierz $\mathbf{R}$ była estymowana według wzorów (9) oraz (10). Sygnały na wyjściu bloku dokonującego separacji były dodatkowo korygowane za pomocą pętli DPLL (równania (19) i (20)). Algorytm wymagał zmniejszonej wartości kroku adaptacji α i większej liczby próbek wektorów zastosowanych w estymacji macierzy autokorelacji $\mathbf{R}$. Na rys. 5 można też zaobserwować, że gdy przesunięcie częstotliwości jest równe $\Delta f = 5$ Hz, algorytm powoli traci uzyskaną zbieżność. Remedium na to zjawisko byłaby ciągła estymacja macierzy $\mathbf{R}$ i aktualizacja macierzy separacji z wykorzystaniem na bieżąco aktualizowanej macierzy filtru wybielającego (12). Będzie to przedmiotem dalszych eksperymentów.



Rysunek 5: Błąd średniokwadratowy (MSE [dB]) na wyjściu odbiornika BSS ($N = 8$ anten, $k = 5$, $\alpha = 0.012$, $\lambda = 0.99$, $L_{est} = 338$, $L_{thres} = 100$, $\mu = 0.005$, $\rho = 0.01$, $SNR = 20$ dB, $\Delta f = 0$, 2, lub 5 Hz)

6. WNIOSKI

Zaproponowana w artykule funkcja kosztu jest użyteczna w wydzieleniu sygnału użytecznego z ważonej sumy sygnałów pożądanego oraz zakłócającego za pomocą algorytmu ślepej separacji źródeł sygnałów. Zaproponowane rozwiązanie może zostać łatwo uogólnione na przypadek transmisji MIMO OFDM. Wzrasta wtedy liczba równolegle transmitowanych sygnałów użytecznych. Rozszerzenie algorytmu separacji sygnałów na przypadek sygnałów wielotonowych stosowanych w systemach 5G/6G będzie przedmiotem badań w najbliższej przyszłości.

Praca finansowana z projektu CYBERSECIDENT/487845/IV/NCBR/2021.

LITERATURA

[1] K. Wesołowski. "A Simple Algorithm for Jamming Detection in OFDM Systems," in Proc. 2023 IEEE 97th Vehicular Technology Conf., VTC 2023 Spring, June 20-23, Florence, Italy

[2] H. Pirayesh, and H. Zeng, "Jamming Attacks and Anti-Jamming Strategies in Wireless Networks: A Comprehensive Survey," *IEEE Comm. Surv. Tutorials*, vol. 24 No.2, 2022, str. 767-809

[3] M. Girolami, "Self-organizing neural networks: Independent component analysis and blind source separation," London, U.K.: Springer-Verlag 1999.

[4] W. Luo, H. Jin, X. Li, H. LI, K. Liu, R. Yang, "A novel complex-valued blind source separation and its applications in integrated reception,", *Electronics*, vol. 12, 2023, str. 3954.

[5] B. Ans, J. Hérault, C. Jutten, C. (1985). "Architectures neuromimétiques adaptatives : Détection de primitives," Cognitiva 85 Vol. 2, 1985, str. 593-597.

[6] S. Haykin, "Adaptive filter theory," 4th Ed., Prentice-Hall, Upper Saddle River, N.J., 2002

[7] K. Wesołowski, "Analysis and properties of the modified constant modulus algorithm for blind equalization," *AEÜ*, vol. 3, No. 3, May-June 1992, str. 225-230

[8] S. Amari, "Natural gradient works efficiently in learning," Neural Computation, Vol. 10, 1998, str. 251-276

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

EFEKTYWNOŚĆ SCHEMATU KOMUNIKACJI LICZNIKÓW MEDIÓW W PROCESIE ZDALNEGO ODCZYTU OBCHODZONEGO

EFFICIENCY OF THE COMMUNICATION SCHEME FOR UTILITY METERS IN THE REMOTE WALK-BY READOUT PROCESS

Zbigniew Łaskarzewski[1]; Krzysztof Grochla[2]

[1] AIUT Sp. z o.o., Gliwice, Zbigniew.Laskarzewski@aiut.com
[2] Instytut Informatyki Teoretycznej i Stosowanej PAN, Gliwice, kgrochla@iitis.pl

DOI: 10.15199/59.2024.4.46

Streszczenie: **Liczniki mediów coraz częściej wyposażane są w możliwość zdalnego odczytu wskazań. Jest on realizowany w modelu stacjonarnym lub obchodzonym. W modelu obchodzonym najpopularniejszym schematem komunikacji jest periodyczne nadawanie ramek pomiarowych, jednak jak pokazuje analiza w niniejszej pracy, schemat komunikacji oparty na energooszczędnym nasłuchu (próbkowaniu preambuły) jest bardziej energooszczędny, a także wymaga mniej czasu na realizację transmisji. Wskazane także zostały parametry mające wpływ na wybór schematu komunikacji.**

Abstract: **Utility meters are increasingly equipped with the ability to remote readouts process. It is implemented in an stationary or walk-by model. In walk-by model, the most popular communication scheme is the periodic transmission of measurement frames, but as the analysis in this work shows, the communication scheme based on low power listening (preamble sampling) is more energy-efficient and also requires less time to complete the transmission. Parameters influencing the choice of communication scheme were also indicated.**

Słowa kluczowe: **odczyt obchodzony, efektywność energetyczna komunikacji, bezprzewodowe sieci sensorowe, Internet Rzeczy, odbiór próbkowany**

Keywords: **walk-by readout process, energy efficiency of communication, Wireless Sensor Network, Internet of Things, preamble sampling**

1. WSTĘP I MOTYWACJA

Liczniki zużycia mediów komunalnych wyposażane są w układy elektroniczne, które odpowiadają za proces metrologiczny realizowany dzięki sensorom zależnym od metody pomiarowej, za prezentację wartości wskazania oraz za transmisję pomiarów do systemu informatycznego. Wodomierze, ciepłomierze i gazomierze są urządzeniami, których elektronika zasilana jest bateryjnie, co oznacza ograniczoną ilość energii, która musi wystarczyć do poprawnego funkcjonowania urządzenia przez cały okres jego eksploatacji wynoszący typowo 5-10 lat. Zapewnienie możliwości komunikacji zdalnej celem przesyłania danych pomiarowych stanowi istotne wyzwanie techniczne.

W przypadku wodomierzy, ciepłomierzy, gazomierzy komunikacja realizowana jest bezprzewodowo w jednym z dwóch modeli:

- Stacjonarnym – z wykorzystaniem urządzeń infrastruktury tj. stacji bazowych sieci komórkowych GSM/4G/5G lub innych dodatkowych, stale zainstalowanych urządzeń koncentratorów, np. bramki systemu LoRaWAN, które odbierają pomiary od urządzeń końcowych i przekazują je do systemu inkasenckiego,

- Obchodzonym – z wykorzystaniem przenośnych urządzeń odczytowych, pojawiających się w miejscu realizacji odczytu tylko w czasie jego realizacji. Urządzenia przenośne obsługiwane są przez inkasentów, którzy realizują odczyt zdalny. Urządzenie odczytowe może również być umieszczone na dronie [2], który przelatuje w zasięgu radiowym urządzeń liczników.

Model stacjonarny wykorzystujący technologie takie jak LTE-M, NB-IoT czy LoRaWAN pozwala na nawet codzienną realizację odczytów. Jednak lokalna komunikacja radiowa niewymagająca wykorzystania infrastruktury jest rozwiązaniem ekonomicznym, niejednokrotnie w pełni wystarczającym. W erze liczników bez elektroniki wskazania zużycia mediów realizowane były przez inkasentów lub samych odbiorców końcowych w cyklach miesięcznych, dwumiesięcznych lub rocznych. Obecnie odczyty w takich systemach realizowane są z tą samą częstością przez inkasentów, jednak bez potrzeby fizycznego bliskiego dostępu do licznika, a zdalnie, z miejsca pobliskiego jego zainstalowania, z wykorzystaniem urządzenia odczytowego.

Typowy schemat komunikacji liczników mediów w modelu obchodzonym opiera się na periodycznym nadawaniu pomiarów. Model taki jest łatwy w realizacji oraz zapewnia determinizm zużycia energii oraz czasu działania urządzenia. Jednakże zdecydowana większość przesyłanych pomiarów nie jest odbierana, co powoduje niepotrzebne straty energii urządzenia. Taki schemat działania wykorzystywany jest m.in. w licznikach realizujących komunikację zgodnie z popularnym protokołem WM-BUS (tryb S) [3].

Innym schematem komunikacji jest pozostawanie urządzeń liczników w nasłuchu i gotowości do odebrania i przetworzenia żądania przesłania pomiarów. Wybór modelu działania zależy od czynników takich jak: energia zużywana na nadawanie i nasłuch, częstość realizacji odczytu, zagęszczenie urządzeń w zasięgu urządzenia odczytowego.

Niniejsza praca skupia się na efektywności schematu komunikacji liczników mediów w procesie zdalnego odczytu obchodzonego. Efektywność owa jest rozumiana w dwóch kontekstach:

- efektywność energetyczna komunikacji wpływająca na czas działania urządzenia w oparciu o bateryjne źródło zasilania,

- szybkość realizacji procesu zdalnego odczytu wpływająca na efektywność pracy inkasenta.

2. PRZEGLĄD LITERATURY

Rozwiązania bazujące na pozostawaniu urządzeń liczników w nasłuchu i gotowości do odebrania i przetworzenia żądania przesłania pomiarów opisywane są w literaturze. Aktywność odbiornika radiowego powoduje duże zużycie energii, dlatego algorytmy minimalizują czas jego aktywności. Technika Low Power Listening (LPL) [1] po stronie odbiornika polega na próbkowaniu tj. cyklicznej weryfikacji stanu kanału radiowego zamiast ciągłego nasłuchu, natomiast nadajnik przesyła ramkę z preambułą o długości cyklu. W odbiorniku cykl ten obejmuje:

- Włączenie odbiornika na krótki okres czasu i weryfikację czy nadawana jest preambuła,

- Gdy wykryta jest preambuła, następuje oczekiwanie na początek ramki, jej odbiór, następnie podejmowana jest aktywność wynikająca z jej zawartości,

- Gdy preambuła lub w dalszej części ramka nie zostaje wykryta, urządzenie wraca do trybu uśpienia do końca cyklu.

Długość cyklu jest kluczowym parametrem i jest dostosowana do wymagań aplikacji i bezpośrednio przekłada się na konsumpcję energii oraz czas dostępu do komunikacji z urządzeniem.

Technika LPL jest wykorzystywana przez szereg protokołów. Protokół B-MAC rozszerza ją o adaptację okresu próbkowania celem zmniejszenia opóźnień i zwiększenia przepustowości przy zachowaniu dużej energooszczędności. Protokół X-MAC [4] proponuje zastąpienie długich preambuł wieloma ramkami z krótszymi preambułami celem skrócenia czasu odbioru ramki przez odbiorniki, w tym te, do których ramka nie jest kierowana, co ogranicza zużycie energii tych odbiorników.

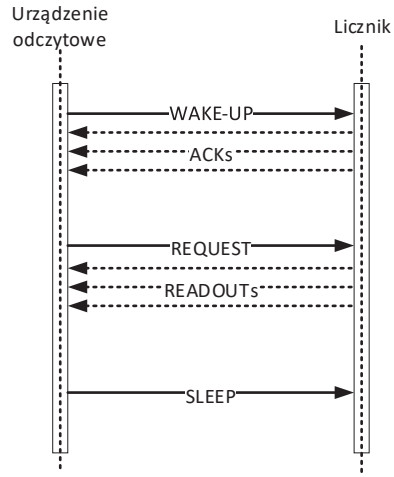
3. SCHEMAT KOMUNIKACJI ODCZYTU OBCHODZONEGO

3.1. Opis schematów komunikacji

Minimalizacja zużycia energii urządzenia oparta jest o ograniczanie czasu, w którym moduł radiowy jest w stanie nadawania lub odbioru. Optymalnym schematem komunikacji byłoby aktywowanie odbiornika tylko wówczas gdy stacja nadawcza do niego przesyła ramkę oraz aktywowanie nadajnika tylko na potrzebę wysłania ramki pomiarowej, którą urządzenie odczytowe odbierze. Założenia powyższego schematu wymagają utrzymania synchronizacji pomiędzy nadajnikiem i odbiornikiem, która może być realizowana przesyłaniem dodatkowych ramek. Jednak w procesie zdalnego odczytu obchodzonego komunikacja, a zatem obecność urządzenia odczytowego, występuje najczęściej raz w miesiącu, co powoduje, że utrzymanie synchronizacji na wymaganym poziomie jest niemożliwe, gdyż nawet dla wysokiej dokładności zegara 1 ppm wartość rozsynchronizowania po miesiącu mogłaby wynieść 2,6 sekundy.

Schemat komunikacji oparty o periodyczne nadawanie ramek z pomiarami jest deterministyczny, a jego efektywność energetyczna jest łatwa do wyznaczenia. Urządzenie może być wyposażone w uproszczony moduł radiowy realizujący wyłącznie funkcję nadawania. Wówczas czas transmisji ramki realizowany jest w oparciu o ALOHA, często z dodatkowym czynnikiem losowym, przeciwdziałającym długotrwałemu interferowaniu ramek z urządzeń nadających z tym samym okresem, zgodnie w fazie. W tym schemacie zdecydowana większość przesyłanych pomiarów nie jest odbierana, co powoduje niepotrzebne straty energii urządzenia.

Schemat komunikacji oparty o periodyczny nasłuch w myśl techniki LPL do aktywacji procesu odczytowego jest rozwiązaniem pozwalającym na oszczędność energii urządzenia, o ile wykonywanie próbek jest mniej energochłonne niż nadawanie ramek, co zazwyczaj ma miejsce. Urządzenia znajdują się w trybie uśpienia, wykonując próbkowanie polegające na krótkim włączeniu modułu radiowego do trybu nasłuchu celem detekcji preambuły nadawanej ramki.



Rys. 1. Schemat komunikacji odczytu obchodzonego

Proponowany schemat komunikacji bazuje na 3 fazach (rys. 1):

- Wybudzenie urządzeń tj. przesłanie ramki z długą preambułą wybudzającą (ramka WAKE-UP), która jest kierowana do grupy urządzeń przeznaczonych do odczytu z adresacją grupy urządzeń lub rozgłoszeniową (ang. *broadcast*). Ramka zawiera żądanie przesłania potwierdzenia wybudzenia w modelu rywalizacyjnym, z wykorzystaniem szczelin czasowych. Urządzenie odczytowe ustala liczbę szczelin, spośród których urządzenie końcowe (w ramach algorytmu szczelinowego ALOHA) losowo wybiera tę szczelinę, w której nada ramkę potwierdzenia (ACK), zawierającą co najmniej adres radiowy urządzenia.

- Odpytanie o pomiary tj. dla ramki ACK i odczytanych zeń w poprzednim kroku adresów radiowych wysyłane jest żądanie przesłania pomiarów (REQUEST). Jego przesłanie realizowane jest z wykorzystaniem krótkiej preambuły oraz adresacji multicast, co pozwala na przesłanie tylko 1 ramki żądania oraz efektywne wykorzystanie łącza. Urządzenia odpowiadają (READOUT) w modelu rezerwacyjnym – w poszczególnych szczelinach czasowych w kolejności umieszczenia na liście adresów w ramce multicast. Po odpytaniu wszystkich urządzeń te, które nie odpowiedziały, odpytywane są drugi raz.

- Uśpienie tj. po odebraniu ramek z pomiarami do urządzeń które je przesłały, wysyłane jest polecenie wyłączenia odbiornika (SLEEP) i niewykonywania próbek przez przewidywany czas realizacji procesu odczytu, celem zapobiegnięcia ponownemu wybudzaniu się urządzeń, które przesłały już pomiary. Ramka uśpienia nie powoduje przesyłania odpowiedzi przez urządzenie.

3.2. Efektywność procesu zdalnego odczytu

Efektywność energetyczna schematów komunikacji wyznaczana jest poprzez ustalenie ilości ładunku (wyrażonego w mAh) wymaganego do działania urządzenia przez okres 10 lat.

Szybkość realizacji procesu zdalnego odczytu jest określana poprzez czas wymagany do uzyskania pomiarów z prawie wszystkich urządzeń (z pominięciem jednego).

3.3. Efektywność modelu periodycznego nadawania

W przypadku periodycznego nadawania ramek ładunek wymagany do działania przez 10 lat jest równy:

$$Q = I_{tx} \cdot t_{readout} \cdot \frac{10 \cdot 365 \cdot 24 \cdot 3600\ [s]}{p_{readout}} \quad (1)$$

gdzie:
- I_{tx} – natężenie prądu podczas nadawania
- $t_{readout}$ – czas trwania ramki readout
- $p_{readout}$ – częstość nadawania ramki readout

Szybkość realizacji procesu odczytowego w modelu periodycznego nadawania wynika ze stratności ramek wynikającej z interferencji w wyniku algorytmu ALOHA oraz stratności wynikającej z pozostałych czynników. Stratność ramek w ramach pojedynczego cyklu w algorytmie ALOHA wynosi:

$$L_{aloha} = 1 - e^{-2G_{readout}} \quad (2)$$

gdzie:
- $G_{readout}$ – obciążenie łącza rozumiane jako:

$$G_{readout} = (n-1) \cdot \frac{t_{readout}}{p_{readout}} \quad (3)$$

gdzie:
- n – liczba urządzeń

Stratność łączna wynikająca z ALOHA oraz pozostałych czynników wynosi:

$$L = 1 - (1 - L_{aloha})(1 - q) \quad (4)$$

gdzie:
- q – stratność pakietów wynikająca z innych czynników

Oczekiwany czas realizacji procesu zdalnego odczytu wynosi:

$$T = -p_{readout} \cdot \log_L n \quad (5)$$

3.4. Efektywność modelu periodycznego nasłuchu

W przypadku periodycznego nasłuchu ładunek wymagany do działania przez 10 lat jest równy:

$$Q = Q_s \cdot \frac{10 \cdot 365 \cdot 24 \cdot 3600\ [s]}{p_s} \quad (6)$$

gdzie:
- Q_s – ładunek zużywany na pojedyncze próbkowanie
- p_s – czas pomiędzy nasłuchami (próbkowaniami)

Przy realizacji zdalnego odczytu z miesięczną częstotliwością lub rzadziej, realizacja procesu ma znikomy wpływ na wymagany ładunek.

Szybkość realizacji procesu odczytowego w modelu periodycznego nasłuchu wynika ze stratności wszystkich ramek aż do uzyskania ramki pomiarowej (READOUT).

$$L_{wa} = (1 - (1-q) \cdot (1-q)) \quad (7)$$

$$L_{s_aloha} = 1 - e^{-G_a} \quad (8)$$

$$G_a = (n-1) \cdot \frac{t_a}{p_a} \quad (9)$$

$$L_{rr} = (1 - (1-q) \cdot (1-q))^2 \quad (10)$$

gdzie:
- t_a – czas trwania ramki ACK
- p_a – łączny czas szczelin dla odbioru ramek ACK
- L_{wa} – łączna stratność wymiany ramek WAKE-UP i ACK
- L_{s_ALOHA} – stratność wynikająca z interferencji ramek ACK w ramach szczelinowego ALOHA
- L_{rr} – łączna stratność wymiany ramek REQUEST i READOUT, z powtórzeniem

Ostatecznie stratność w pojedynczym cyklu wynosi:

$$L = 1 - (1 - L_{wa}) \cdot (1 - L_{s_aloha}) \cdot (1 - L_{rr}) \quad (11)$$

Czas realizacji pojedynczego cyklu wynosi:

$$p_{cycle} = p_s + t_w + p_a + 2 \cdot t_{req} + p_{readout} + t_s \quad (12)$$

gdzie:
- t_w – czas trwania ramki WAKE-UP
- t_{req} – czas trwania ramki REQUEST
- $p_{readout}$ – łączny czas szczelin dla odbioru ramek READOUT, wraz z powtórzeniami
- t_w – czas trwania ramki SLEEP

Oczekiwany czas realizacji procesu zdalnego odczytu wynosi:

$$T = -p_{cycle} \cdot \log_L n \quad (13)$$

4. UZYSKANE WYNIKI

Efektywność procedury odczytu obchodzonego została zbadana z wykorzystaniem elektroniki urządzenia gazomierza wyposażonego w moduł radiowy Spirit1 firmy STMicroelectronics. Parametry komunikacji (tab. 1) przekładają się na parametry modeli (tab. 2 i 3).

Tab. 1. Parametry komunikacji

Parametr	*Wartość*	*Jednostka*
Prędkość transmisji	4800	[bit/s]
Modulacja	2-GFSK	
Częstotliwość	868	[MHz]

4.1. Model periodycznego nadawania

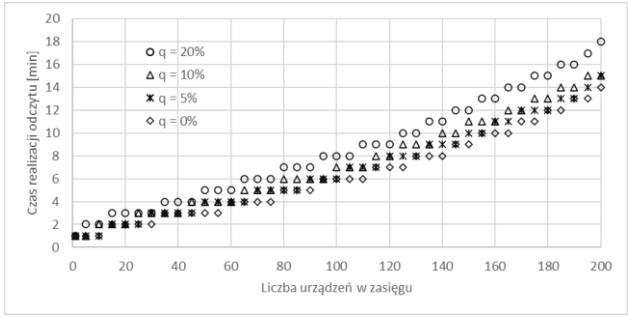
Tab. 2. Parametry modelu

Parametr	*Wartość*	*Jednostka*
I_{tx}	33	[mA]
$t_{readout}$	0,167	[s]
$p_{readout}$	60	[s]

Dla parametrów modelu (tab. 2) ładunek wymagany do działania przez 10 lat wynosi:

$$Q = 33 \cdot 0{,}167 \cdot \frac{10 \cdot 365 \cdot 24 \cdot 3600}{60} mAs = 8030 mAh \quad (14)$$

Oczekiwany czas realizacji procesu zdalnego odczytu zależy od liczby urządzeń w zasięgu oraz poziomu stratności ramek (q) (rys. 2).



Rys. 2. Czas odczytu w modelu nadawania

4.2. Model periodycznego nasłuchu

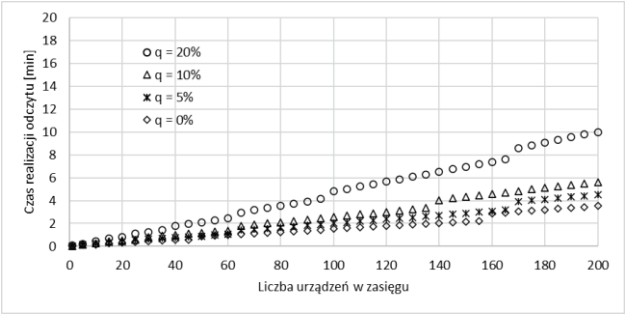
Tab. 3. Parametry modelu

Parametr	*Wartość*	*Jednostka*
Q_s	0,168	[mAs]
p_s	3	[s]
t_w	0,020	[s]
t_a	0,020	[s]
p_a	$3 \cdot t_a \cdot$(oczekiwane n)	[s]
t_{req}	0,020	[s]
$t_{readout}$	0,167	[s]
$p_{readout}$	$t_{readout} \cdot$(odebrane n)	[s]

Dla parametrów modelu (tab. 3) ładunek wymagany do działania przez 10 lat wynosi:

$$Q = 0{,}168 \cdot \frac{10 \cdot 365 \cdot 24 \cdot 3600}{3} mAs = 4096\, mAh \quad (15)$$

Oczekiwany czas realizacji procesu zdalnego odczytu zależy od liczby urządzeń w zasięgu oraz poziomu stratności ramek (q) (rys. 3).



Rys. 3. Czas odczytu w modelu nasłuchu

5. PODSUMOWANIE

Model periodycznego nasłuchu, wykorzystujący próbkowanie, pozwala na istotne ograniczenie ilości zużywanej energii oraz znaczące przyspieszenie realizacji procesu zdalnego odczytu obchodzonego. Model periodycznego odbioru przynosi szczególne korzyści dla technik transmisji, które cechują się niskim zużyciem energii na wykonanie próbki względem nadania ramki pomiarowej, w przypadku rzadko realizowanej komunikacji. Ze względu na większą liczbę wymienianych ramek, także występowanie niskiej stratności pakietów faworyzuje model periodycznego nasłuchu.

Dla przytoczonej aplikacji zastosowanie modelu periodycznego nasłuchu pozwoliło na ograniczenie zużycia energii o 39% oraz pozwoliło na przyspieszenie procesu odczytu obchodzonego od 1,8 razy (n=200, q=20%) do 5,4 razy (n=10, q=0-5%).

PODZIĘKOWANIA

Badania przedstawione w niniejszej pracy zostały współfinansowane przez Narodowe Centrum Badań i Rozwoju w ramach grantu POIR.04.01.04-00-0005/17 (Dynamiczne środowisko komunikacji bezprzewodowej dla infrastruktury internetu rzeczy inteligentnych miast – Smart City IoT).

LITERATURA

[1] Dinh Thanh, Kim Younghan, Gu Tao, Vasilakos Athanasios V. 2018. „An Adaptive Low-Power Listening Protocol for Wireless Sensor Networks in Noisy Environments". IEEE Systems Journal, vol. 12, no. 3, pp. 2162-2173.

[2] J. E. B. Cerillo Jose Eduardo B., Loquias Rizza T, Fenomeno Senen D., 2022. „Automatic Flight Plan Generation for UAV-Assisted Elevated Meter Reading using Ant Colony Optimization" IEEE 3rd Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2022, pp. 1-7.

[3] 2019. „Communication systems for meters - Part 4: Wireless M-Bus communication; English version EN 13757-4:2019".

[4] Buettner Michael, Yee Gary, Anderson Eric, Han Richard. 2006. „X-MAC: A short preamble MAC protocol for duty-cycled wireless sensor networks". ACM Sensys. 4. 307-320.

Koncepcja kognitywnej transmisji wykorzystującej rozpraszanie widmowe za pomocą szybkich przeskoków częstotliwości

The concept of cognitive transmission using spread spectrum through fast frequency hopping

Adrian Kubowicz[1]; Jerzy Łopatka[2]; Piotr Gajewski[3]

[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
[3] Wojskowa Akademia Techniczna, Instytut Systemów Łączności, Warszawa, piotr.gajewski@wat.edu.pl

DOI: 10.15199/59.2024.4.47

Streszczenie: **W artykule przedstawiono koncepcję kognitywnej transmisji metodą podkładania, z zastosowaniem rozpraszania widma za pomocą szybkich przeskoków częstotliwości FHSS (ang. Frequency Hopping Spread Spectrum) wraz z architekturą rozwiązania aplikacyjnego. Przedstawiono wyniki symulacji. Omówiono wyzwania jakie się wiążą z rozwojem tej technologii.**

Abstract: **The article presents the concept of underlay cognitive transmission using the fast Frequency Hopping Spread Spectrum (FHSS) method, with the architecture of the system application. Simulation results are also provided. The challenges associated with the development of this technology are also discussed.**

Słowa kluczowe: **filtr dopasowany, okienkowanie sygnału, Radio Kognitywne, synchronizacja, FHSS**

Keywords: **matched filter, signal windowing, Cognitive Radio, synchronisation, FHSS**

1. WSTĘP

Radio kognitywne, jako koncepcja powstało w latach 90' i stanowi szansę na bardziej elastyczne wykorzystanie zasobów widma radiowego [1]. Kognitywna transmisja danych wymaga zastosowania technik monitorowania środowiska radiowego, aby uzyskać informacje potrzebne do podjęcia decyzji o transmisji w danym paśmie częstotliwości, szczególnie jeśli dotyczy to transmisji w pasmach licencjonowanych używanych przez użytkowników „pierwotnych". Istnieją dwie główne metody dzielenia pasma. Pierwsza to „podkładanie", czyli transmisja z użyciem rozpraszania widma, które dla użytkowników pierwotnych stanowi akceptowalne tło szumowe. W drugim podejściu z „nakładaniem", wyszukujemy niezajętych chwilowo pasm częstotliwości – tzw. białych przestrzeni (*ang. Wite Spaces*) w komunikacji licencjonowanej w celu realizacji własnej transmisji, a w przypadku pojawienia się transmisji użytkowników pierwotnych, następuje możliwie szybkie zaprzestanie nadawania. Ponieważ w technikach radia kognitywnego, oprócz wykorzystywanych częstotliwości należy brać pod uwagę także czas i geolokalizację, szybkie zaprzestanie transmisji w przypadku wykrycia kolizji może nie być wystarczające [6].

W artykule zaproponowano zatem koncepcję kognitywnej transmisji z użyciem rozpraszania widmowego za pomocą szybkich przeskoków częstotliwości. Wraz ze wzrostem szybkości przełączania częstotliwości, czas jaki zajmuje transmisja symbolu na jednej częstotliwości skraca się, przez co czas występowania potencjalnych interferencji dla użytkownika pierwotnego jest również krótszy. Graniczny przypadek jaki można założyć, jest to transmisja jednego okresu fali nośnej, którego obecność, bądź brak w danym momencie, przy idealnej synchronizacji stanowi podstawę do podjęcia decyzji o odbiorze danych przez odbiornik. W praktyce należy jednak wygenerować większą liczbę okresów fali nośnej i zastosować okienkowanie, aby istotnie zmniejszyć efekt zjawiska przecieku widma [5]. Istotą koncepcji jest założenie, iż w przypadku kolizji tak krótkiego sygnału z sygnałem użytkownika pierwotnego, może nastąpić jedynie krótkotrwałe zakłócenie jego transmisji. Przyjmuje się, że będzie ono na tyle krótkie, że w ramach pakietu danych spowoduje jedynie nieznaczny wzrost bitowej stopy błędów BER (*ang. Bit Error Rate*) użytkownika pierwotnego, który może zostać skorygowany za pomocą istniejących zabezpieczeń kodowych.

Opracowanie takiego systemu nadawania jest szczególnie istotne zarówno z punktu widzenia lepszego wykorzystania widma częstotliwości radiowych, jak i zwiększenia bezpieczeństwa transmisji sygnałów szczególnie w warunkach zakłócania celowego, co jest elementem tzw. Walki Radioelektronicznej.

2. PRZEGLĄD ROZWIĄZAŃ Z SZYBKIM FHSS

FHSS jako metoda rozpraszania widma stosowana jest w wielu systemach np. w Bluetooth z 1600 skokami na sekundę [8]. Na danej częstotliwości komunikację można rozpatrywać jako wąskopasmową. Dzięki pseudolosowym zmianom częstotliwości uzyskujemy zwiększenie bezpieczeństwa i odporności na zakłócenia. Trudniej

jest podążać za takimi zmianami w transmisji w celu podsłuchu, gdzie sekwencja skoków częstotliwości znana jest tylko w nadajniku i odbiorniku. Zwiększenie szybkości zmian częstotliwości stanowi dodatkowe utrudnienie dla przeciwnika jednakże stanowi istotne wyzwanie techniczne w realizacji takiej komunikacji.

Z opracowania [2] wynika, iż podstawowym problemem przy projektowaniu szybkich układów FHSS jest zdolność układu lokalnego oscylatora do szybkiego przestrajania. Tutaj osiągnięto czas przełączania oscylatora 1,5ns, co możliwe było dzięki oscylatorowi typu „injection locked" z przestrajanym filtrem-rezonatorem LC o wysokiej dobroci. Zaprezentowany system jest zdolny, nominalnie, do 47M skoków/s, gdzie jeden bit jest kodowany na 100 skokach. Daje to przepustowość na poziomie 470kb/s. Projekt został zrealizowany w układzie scalonym CMOS 65nm, o niskim poborze mocy zintegrowanego rozwiązania.

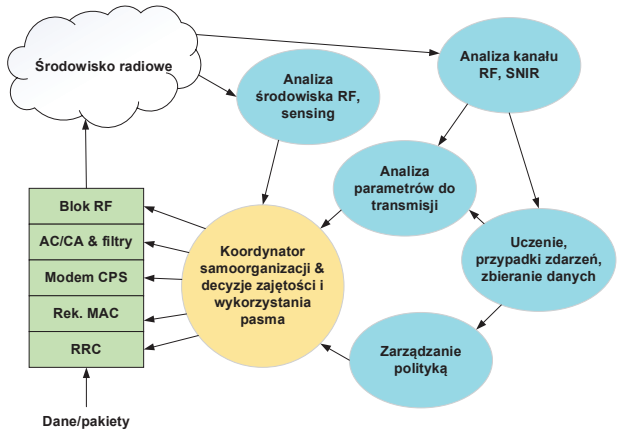
W pracy [3] przedstawiono rozwiązanie dla interfejsu BLE (ang. *Bluetooth Low Energy*). Układ taktujący oparto o tzw. „*Bulk Acoustic Wave Resonator"*. W celu zwiększenia odporności na zakłócenia selektywne zaproponowano nadajnik-odbiornik zdolny do przełączania częstotliwości w ciągu 1µs.

Istotne zwiększenie szybkości przeskoków można uzyskać w układach wykorzystujących technologię fotoniki mikrofalowej [4]. Przetwarzanie sygnału w układach optycznych umożliwiło zwiększenie szybkości zmian częstotliwości fali nośnej o częstotliwości rzędu dziesiątek GHz. Istniejące rozwiązania umożliwiają czas przełączania częstotliwości 50ps.

3. KONCEPCJA SYSTEMU KOGNITYWNEGO Z SZYBKIM FHSS

Z punktu widzenia oportunistycznego dostępu w pasmie licencjonowanym istotna jest ocena, jak system z szybkimi przeskokami FHSS będzie współegzystował z innymi systemami radiokomunikacyjnymi. Przyjęta koncepcja się potwierdzi, jeśli przy odpowiednio krótkich czasach zakłócania na danej częstotliwości powodowanych przez sygnał FHSS, 'pierwotny' system radiokomunikacyjny przy występujących kolizjach odtworzy utracone dane za pomocą własnych kodów korekcyjnych, lub występujące błędy będą poniżej określonego progu tolerancji BER, np. na poziomie <1%, co wyższe warstwy komunikacyjne systemu powinny potraktować to jako tolerowane zakłócenie. W oparciu o takie założenia, można utworzyć koncepcję „Idealnego Radia Kognitywnego", gdzie użytkownicy pierwotni mogliby kontynuować komunikację w trakcie działania FHSS. Wpływ sygnałów FHSS na użytkowników pierwotnych byłby zależny od rodzaju transmisji użytkownika pierwotnego, rodzaju modulacji, metody rozpraszania i szybkości symbolowej.

W zaproponowanym rozwiązaniu (Rys.1) urządzenie radiowe wykorzystuje platformę radia definiowanego programowo SDR (*ang. Software Defined Radio*) z oprogramowaniem wspierającym kognitywną zdolność do adaptacji z zastosowaniem procedur oportunistycznego dostępu do widma.



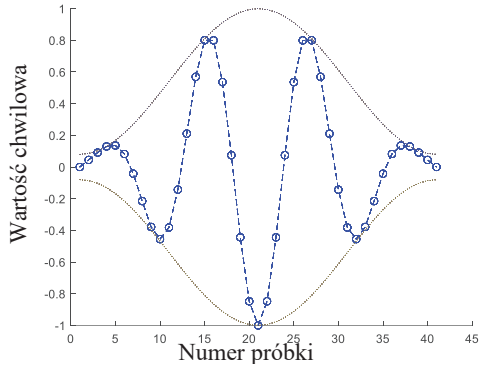
Rys.1 Schemat funkcjonalny proponowanego rozwiązania

W architekturze SDR z bezpośrednim przetwarzaniem analogowo-cyfrowym i cyfrowo-analogowym, pominięto obwody mieszaczy RF w celu kontroli nad kształtem syntezowanego sygnału w dziedzinie czasu. W torze nadawczym i odbiorczym zaproponowano dodanie przestrajanego filtra analogowego, o charakterystyce pasmowo przepustowej. Zależnie od długości sygnału wyrażonego w liczbie okresów fali nośnej, charakterystyka filtra wymaga dopasowania. W ramach badan, bez korelatorów i innych zaawansowanych technik analogowych, cyfrowy układ synchronizacji pozwoli na wyrównanie startu sekwencji skoków częstotliwości w nadajniku i odbiorniku oraz wytworzy impulsy ponownej synchronizacji. Należy przy tym zwrócić uwagę na wymagania związane z synchronizacją w tego rodzaju aplikacji. Przykładowo, przy częstotliwości fali nośnej 10 MHz, jeśli synchronizacja będzie wymagała dokładności równej czasowi jednego okresu fali nośnej, przy precyzji sygnału zegarowego +-0,5ppm (*ang. parts per milion*) w nadajniku i odbiorniku, konieczne będzie podstrajanie układu synchronizacji nie rzadziej niż co 100ms.

Architektura programowa poza procedurami szerokopasmowego, cyfrowego przetwarzania sygnałów (CPS), zawiera procedury radia kognitywnego włączające obok aplikacji metod dynamicznego zarządzania widmem i zarządzania polityką radiową, procedury nasłuchu, optymalizacji parametrów nadawania poprzez zmiany szybkości i sposobu przeskoków. W systemie zastosowane zostaną rozwiązania sztucznej inteligencji AI z uczeniem maszynowym i podejmowaniem decyzji zarówno w oparciu o identyfikację przypadków użycia, jak i bieżącej optymalizacji określonych parametrów transmisyjnych.

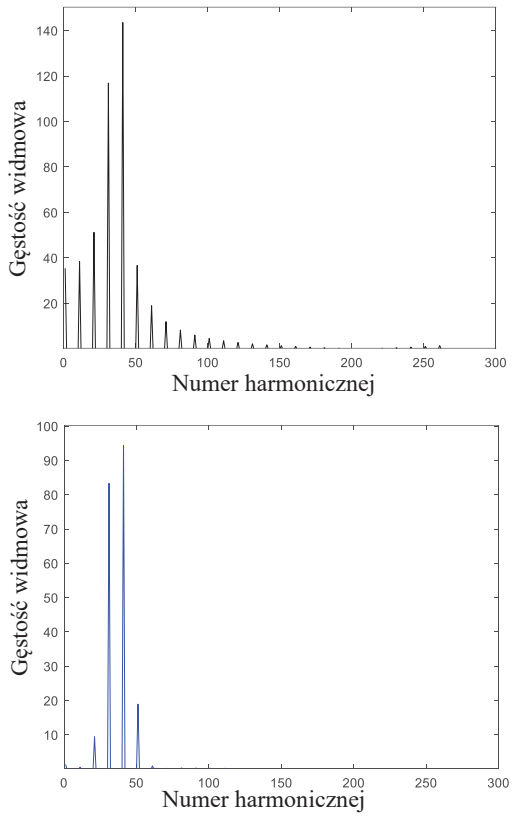
4. WYNIKI SYMULACJI

Z powodu ograniczonej szybkości przełączania układów analogowych oraz występowania przecieku widma, uznano iż minimalna liczba okresów fali nośnej powinna wynosić co najmniej 3, wraz z zastosowaniem okienkowania sygnału. Dla większej ilości okresów szybkość przeskoków jest mniejsza, ale węższe jest również pasmo sygnału. W środowisku Matlab przeprowadzono zatem symulacje okienkowanego sygnału do wytworzenia w torze nadawczym. Rys. 1 przedstawia unormowany sygnał o częstości próbkowania 2 MHz, z 4 okresami fali nośnej i wygładzaniem z użyciem okna Hanninga.



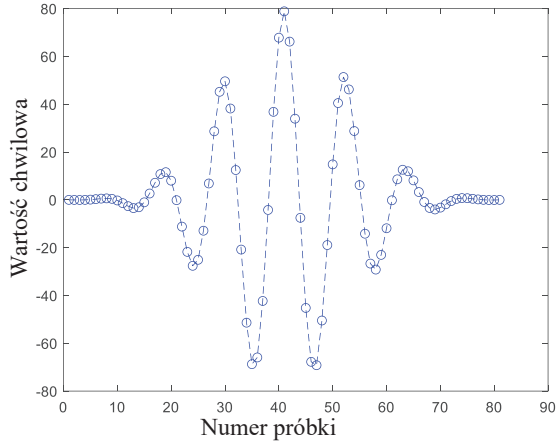
Rys. 1 Przebieg sygnału okienkowanego w dziedzinie czasu

Na Rys. 2 przedstawiono efekt okienkowania sygnału w dziedzinie częstotliwości. Jak widać występuje tu silne tłumienie wyższych składowych częstotliwości, co minimalizuje poziom interferencji pozapasmowych.



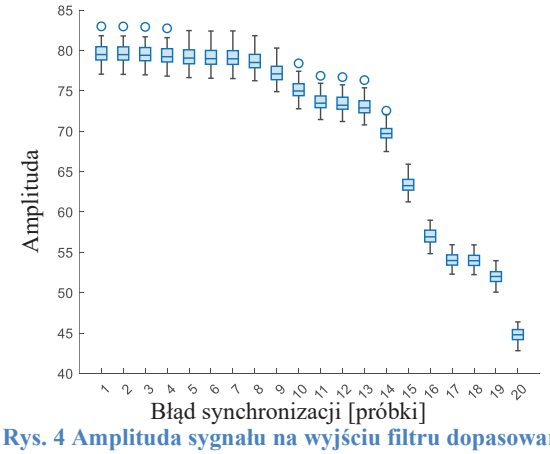
Rys. 2 Widmo ciągu 10 impulsów sygnału bez okienkowania i sygnału okienkowanego

Aby móc odebrać tego rodzaju sygnał, jednocześnie zapewniając możliwie najlepsze tłumienie zakłóceń, jako detektor sygnału w odbiorniku zaproponowano zastosowanie filtra dopasowanego. Sygnał przed detekcją został zakłócony poprzez dodanie szumu białego, otrzymując określoną wartość odstępu sygnał-szum SNR (*ang. Signal Noise Ratio*). Uzyskany wynik detekcji dla SNR=20dB przedstawia Rys. 3. Trzeba przy tym zaznaczyć, iż dotyczy to sytuacji idealnej synchronizacji odebranego sygnału.



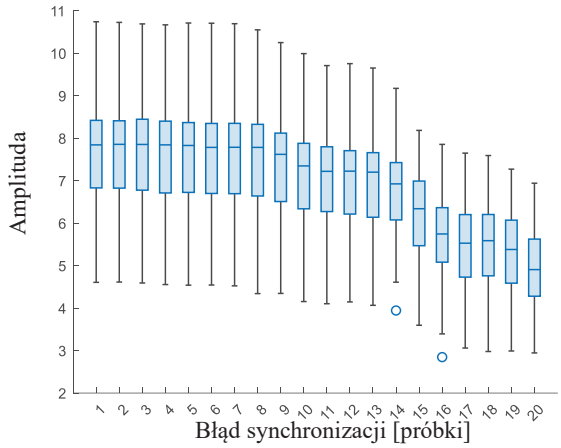
Rys. 3 Sygnał na wyjściu filtru dopasowanego przy idealnej synchronizacji dla SNR=20 dB

Pojawienie się błędu synchronizacji, można zaobserwować w postaci spadku amplitudy sygnału na wyjściu filtru dopasowanego, co przedstawia Rys. 4. Widoczne jest, że dla 100 powtórzeń, przy SNR=20 dB amplituda sygnału po filtracji ma niewielki rozrzut, a błąd synchronizacji nie powinien przekraczać 13 próbek. Dla większych błędów synchronizacji amplituda sygnału na wyjściu detektora zaczyna szybko spadać.



Rys. 4 Amplituda sygnału na wyjściu filtru dopasowanego w funkcji błędu synchronizacji dla SNR=20 dB, dla 100 powtórzeń

Przypadek dla większego poziomu szumu, SNR=0 dB obrazuje Rys. 5. Rozrzut jest znacznie większy a amplituda na wyjściu detektora jest mniejsza. Widoczna jest również zmiana amplitudy w funkcji błędu synchronizacji, a dopuszczalne błędy synchronizacji są porównywalne do poprzedniego przypadku.



Rys. 5 Amplituda sygnału na wyjściu filtru dopasowanego w funkcji błędu synchronizacji dla SNR=0 dB, dla 100 powtórzeń

5. PODSUMOWANIE

Z symulacji wynika, iż zastosowanie filtra dopasowanego w odbiorniku pozwoli poprawnie odebrać sygnał o amplitudzie porównywalnej z poziomem szumu, co umożliwia realizację kognitywnej transmisji wykorzystującej rozpraszanie widmowe za pomocą szybkich przeskoków częstotliwości FHSS. Przestawiona koncepcja wymaga dalszych badań w zakresie wpływu takich emisji na komunikację użytkowników pierwotnych, jak i obecnych możliwości implementacji w sprzęcie, docelowo na poziomie układu scalonego.

W ramach kolejnych prac, planowane jest również opracowanie układów do adaptacyjnego doboru sekwencji skakania po częstotliwościach, w celu minimalizacji liczby kolizji i unikania pasm gdzie transmisja FHSS jest silnie zakłócana. Ich działanie będzie oparte o elementy uczenia maszynowego i sztucznej inteligencji. Warstwa fizyczna będzie również adaptacyjna i będzie umożliwiała dostosowanie szybkości transmisji i pasma sygnału do istniejących warunków widmowych. Opracowanie szybkich filtrów przestrajanych umożliwi filtrowanie i kształtowanie sygnałów w torze nadawczym. W torze odbiorczym pozwoli na uodpornienie na sygnały pozapasmowe.

Finalnie system znajdzie zastosowanie w dziedzinach, w których niezbędna jest komunikacja w złożonym środowisku radiowym bez konieczności wyłączenia z komunikacji pierwotnych użytkowników. Szczególnym zastosowaniem jest zdalne sterowanie pojazdów. W zastosowaniach cywilnych potencjalnym zastosowaniem jest skuteczne przekazywanie alarmowych połączeń głosowych bądź wiadomości poza licencjonowanym systemem telefonii. Do rozważenia jest zastosowanie do komunikacji w warunkach pola walki, gdzie skuteczność komunikacji nie powinna wykluczać z działania innych własnych systemów, niezależnie od zasobów radiowych jakich używają.

Od wyniku dalszych symulacji oraz analizy przypadków zastosowań zależy wybór platformy radia zdefiniowanego programowo do testów umożliwiających praktyczną weryfikację koncepcji. Dla częstotliwości sygnału poniżej 30MHz, pomijając stopnie mieszaczy możliwe jest zastosowanie np. Ettus USSRP x310. Dla pasma 1GHz RFSoC Zynq Ultrascale z przetwarzaniem analogowo cyfrowym 4GS/s [7]. Zagadnieniem badawczym jest projekt filtrów przestrajanych które umożliwią poprawne nadawanie i odbiór krótkich okienkowanych sygnałów z użyciem mieszaczy.

LITERATURA

[1] Mitola III Joseph, Maguire Gerald Q. Jr. 1999, sierpień. "*Cognitive radio: making software radios more personal*". IEEE Personal Communications, 13-18.

[2] Mousavi Naser. 2019. "*Ultra Fast Frequency Hopping Transceiver Design and Implementation for Secure Wireless Communication*". 22-25. University of Minnesota

[3] Yazicigil Rabia Tugce. 2018. "*Ultra-Fast Bit-Level Frequency-Hopping Transmitter for Securing Low-Power Wireless Devices*". IEEE Radio Frequency Integrated Circuits Symposium (RFIC), Philadelphia, PA, USA, 176-179

[4] Liu Qidi, Fok Mable P. 2020. "Ultrafast and Wideband Microwave Photonic Frequency-Hopping Systems: A Review". Applied Sciences. 10.521:1-17

[5] Lyons Richard. 2004. "*Understanding Digital Signal Processing*". 74-82. Prentice Hall, New Jersey.

[6] Kalil Mohamed, Liers Florian, Volkert Thomas, Mitschele-Thie Andreas. 2009. "*A Novel Opportunistic Spectrum Sharing Scheme for Cognitive Ad Hoc Networks*". Electronic Communications of the EASST, 17: 1-5

[7] Advanced Micro Devices. dostęp 2024-05-24. https://docs.amd.com/v/u/en-US/zynq-usp-rfsoc-product-selection-guide

[8] Bluetooth Special Interest Group. 2014. "BLUETOOTH SPECIFICATION Version 4.2". Vol 2, Part B: 72

Model optymalizacji dla energooszczędnego routingu unicast w sieciach FANET

Optimization model for energy-efficient unicast routing in FANET networks

Tomasz Jóźwik[1]

[1] Wojskowa Akademia Techniczna, Warszawa, tomasz.jozwik@wat.edu.pl

DOI: 10.15199/59.2024.4.48

Streszczenie: Referat omawia zastosowania sieci FANET w rozwiązaniach wojskowych oraz cywilnych. Przedstawiono wyzwania związane z projektowaniem i użytkowniem sieci FANET. Omówiono klasyfikację stosowanych protokołów routingu dla sieci FANET. Zaproponowano uogólniony model optymalizacji matematycznej na potrzeby oceny efektywności energetycznej protokołów. Przedstawiono zachowanie modelu dla prostych przypadków testowych i omówiono wyzwania związane z jego zastosowaniem.

Abstract: The paper discusses applications of FANET networks in military and civilian solutions. The challenges of designing and using FANET networks are presented. A classification of the routing protocols used for FANETs is discussed. A generalized mathematical optimization model for evaluating the energy efficiency of protocols was proposed. The behavior of the model for simple test cases is presented, and the challenges associated with its application are discussed.

Słowa kluczowe: energooszczędność, optymalizacja, protokoły routingu, routing unicast, sieci FANET

Keywords: energy efficiency, FANET networks, optimization, routing protocols, unicast routing

1. WSTĘP

Sieci FANET złożone z bezzałogowych statków powietrznych (UAV) są wykorzystywane zarówno w zastosowaniach cywilnych, jak i wojskowych. Do zastosowań sieci FANET należą między innymi wymiana informacji między jednostkami wojskowymi, pozyskiwanie informacji w sytuacji katastrof naturalnych, operacje poszukiwawcze oraz zarządzanie i monitorowanie infrastruktury miejskiej [7]. W związku z zastosowaniami sieci FANET, często konieczna jest współpraca wielu węzłów w celu wykonania określonego zadania [2]. Możliwość szybkiego przemieszczania się węzłów sieci wiąże się z koniecznością reorganizacji jej topologii oraz wymaga uwzględnienia wpływu zmian w sieci na wykorzystywany routing.

W celu umożliwienia oceny jakości wykorzystywanych ścieżek routingu w sieciach typu FANET, dokonano analizy wyzwań stojących przed takimi sieciami oraz wykorzystywanych w nich protokołów routingu.

2. PROBLEMY I WYZWANIA SIECI FANET

Badacze wskazują na wyzwania związane z wykorzystaniem sieci FANET. Omówiono problemy wywołane przez węzły z niskim poziomem dostępnego źródła zasilania, dane wymagające natychmiastowego przesłania, wrogie węzły, słabą jakość sygnału radiowego, wystąpienie awarii, względną szybkość węzłów oraz użycie pesymistycznej ścieżki routingu [3].

Dokonano porównania charakterystyki sieci FANET z sieciami MANET oraz VANET. Wskazano na podobieństwo części cech związanych z komunikacją bezprzewodową. Badacze przeanalizowali różnice związane z możliwością przemieszczania się węzłów sieci FANET w dodatkowym wymiarze, możliwość rotacji węzłów oraz możliwość ich szybkiego przemieszczania się [6].

W celu zasilania węzłów sieci przeanalizowano zasilanie z użyciem baterii, ogniw wodorowych, silnika spalinowego oraz energii słonecznej. Wskazano na problemy poszczególnych źródeł zasilania – długi czas ładowania w przypadku baterii, znaczny rozmiar silnika spalinowego oraz ogniw wodorowych, a także dużą wymaganą powierzchnię i zależność od warunków atmosferycznych energii słonecznej [1].

3. ROUTING W SIECIACH FANET

Badacze dokonali podziału protokołów routingu sieci FANET ze względu na moment wyznaczania ścieżek. Wyróżnioniono protokoły routingu [11]:

- adaptacyjne – dostosowujące ścieżki routingu na podstawie aktualnego stanu sieci,
- proaktywne – wyznaczające ścieżki routingu przed nadejściem pakietu,
- reaktywne – wyznaczające ścieżki routingu w momencie nadejścia pakietu,
- hybrydowe – łączące podejścia proaktywne i reaktywne w zależności od konkretnych węzłów.

Dokonano porównania działania routingu unicast i multicast w sieciach FANET. Przedstawiono możliwości ograniczenia wykorzystania energii i ograniczenia komunikatów sterujących, wynikające z routingu multicast [8].

Przeanalizowano protokoły routingu sieci FANET pod kątem strategii wyznaczania ścieżek. Określono podstawowe grupy protokołów – bazujące na topologii, ba-

zujące na położeniu geograficznym, podejścia hybrydowe oraz strategie inspirowane zjawiskami naturalnymi [10]. Przedstawiono grupę problemów związanych z wyborem optymalnego protokołu routingu dla sieci FANET – skalowalność, przystosowanie do nagłych zmian prędkości, uwzględnienie wysokości lotu węzłów oraz ich znacznej mobilności [12].

Badacze zwracają też uwagę na konieczność uwzględnienia zużycia energii podczas projektowania protokołów routingu sieci FANET. Zaprojektowano protokół routingu wykorzystujący algorytm świetlika, uwzględniający wykorzystanie energii [4]. Zaproponowano rozwiązanie ECaD, mające na celu zrównoważenie zużycia energii w celu przedłużenia żywotności sieci i dokonano jego weryfikacji w środowisku symulacyjnym [5].

4. MODEL OPTYMALIZACJI MATEMATYCZNEJ

W celu określenia optymalnych ścieżek routingu ze względu na wykorzystanie zasobów energii, można zdefiniować model optymalizacji. Zbliżone podejście zostało zaproponowane przez badaczy dla sieci komputerowych wykorzystywanych w centrach danych [9]. Proponowany model dla sieci FANET zakłada osiągnięcie wyłącznie jak najmniejszego zużycia energii i nie uwzględnia innych czynników, takich jak opóźnienia w komunikacji między węzłami, przy wyborze optymalnego rozwiązania.

4.1 Zbiory

Niech istnieje zbiór węzłów sieci N, niezmienny w czasie. Utrata węzła sieci spowodowana wyczerpaniem źródła zasilania lub innymi warunkami zewnętrznymi rozumiana jest jako całkowita utrata mocy nadajnika węzła i nie wywołuje skutku na liczność zbioru N. Komunikacja między węzłami sieci odbywa się z wykorzystaniem przepływów – powiązanych ze sobą grup pakietów, tworzących zbiór M. Każdy przepływ posiada dokładnie jeden węzeł źródłowy i jeden węzeł docelowy, należące do zbioru N. Na potrzeby analizy założono dyskretyzację czasu. Niech istnieje zbiór próbek czasu T, oznaczonych kolejnymi liczbami naturalnymi, taki że:

$$T = \{1, 2, ..., z\} \tag{1}$$

gdzie z jest długością trwania rozważanego przypadku testowego.

4.2 Parametry

Niech każdy przepływ m zostanie wygenerowany przez węzeł źródłowy i będzie gotowy do rozpoczęcia transmisji w momencie g_m, takim że:

$$\forall m \in M : g_m \in T \tag{2}$$

Uproszczony model optymalizacji zakłada, iż całkowity rozmiar nagłówków wykorzystywanego protokołu jest przeznaczony na przesyłane dane. Rozmiar s_m przepływu m jest dodatni i wyrażony w niepodzielnych jednostkach (pakietach):

$$\forall m \in M : s_m > 0 \tag{3}$$

Każdy z przepływów m posiada zdefiniowany węzeł źródłowy i docelowy na potrzeby routingu:

$$v_{m,n} = \begin{cases} 1 - \text{jeżeli } n \in N \text{ jest węzłem źródłowym } m \in M \\ 0 - \text{w p.p.} \end{cases} \tag{4}$$

$$b_{m,n} = \begin{cases} 1 - \text{jeżeli } n \in N \text{ jest węzłem docelowym } m \in M \\ 0 - \text{w p.p.} \end{cases} \tag{5}$$

Koszt obsługi jednostki przepływu między dwoma kolejnymi węzłami na ścieżce routingu jest dodatni i może być zmienny w czasie:

$$\forall n_1, n_2 \in N, t \in T : c_{n_1,n_2,t} > 0 \tag{6}$$

W przypadku braku możliwości obsługi przepływów między węzłami $n_1, n_2 \in N$ w chwili $t \in T$, spowodowanej niedostępnością któregoś z węzłow lub brakiem bezpośredniej widoczności węzłów w środowisku radiowym:

$$c_{n_1,n_2,t} = \infty \tag{7}$$

Każdy z węzłów n posiada nadajnik o nieujemnej mocy p_n, rozumianej jako maksymalny możliwy do obsługi koszt nadawania przepływów w pojedynczym momencie analizowanego przypadku testowego. Zerowa moc nadajnika jest rozumiana jako zdolność do odbioru przepływów z innych węzłów z jednoczesnym brakiem możliwości nadawania przepływów do innych węzłów.

$$\forall n \in N : p_n \geq 0 \tag{8}$$

4.3 Zmienne decyzyjne

Na potrzeby modelu definiujemy zmienne decyzyjne określające obciążenie nadajnika danego węzła, wynikające z nadawania konkretnego przepływu. Niech zmienna decyzyjna $x_{n_1,n_2,m,t}$ określa obciążenie nadajnika na węźle $n_1 \in N$, wykorzystywane na potrzeby nadawania przepływu $m \in M$ do węzła $n_2 \in N$ w momencie $t \in T$.

4.4 Ograniczenia

Obciążenie każdego z nadajników musi być nieujemne. Ograniczenie ma na celu przeciwdziałanie pozornemu pozyskiwaniu dodatkowej mocy nadajników w przypadku bliskości komunikujących się węzłów, poprzez generowanie nadmiernego obciążenia dla wcześniej powstałych przepływów i równoważenie ich ujemnymi obciążeniami w późniejszych momentach.

$$\forall n_1, n_2 \in N, m \in M, t \in T : x_{n_1,n_2,m,t} \geq 0 \tag{9}$$

Przepływy nie mogą być nadawane przed momentem ich generacji na węźle źródłowym, jako że mogą zawierać dane zebrane lub wytworzone podczas pracy danego węzła:

$$\forall n_1, n_2 \in N, m \in M, t \in \{1, ..., g_m - 1\} : x_{n_1,n_2,m,t} = 0 \tag{10}$$

W celu nadania określonej części przepływu przez dany węzeł, musi zostań ona najpierw przez dany węzeł odebrana lub na nim wygenerowana. Dane nadawane przez węzeł muszą uprzednio znajdować się na tym węźle:

$$\forall m \in M, n_1 \in N, t \in \{g_m, ..., z\} \setminus \{1\} : \\ v_{m,n_1} s_m + \sum_{t_x \in \{g_m,...,t-1\}} \sum_{n_2 \in N} x_{n_2,n_1,m,t_x} - \\ \sum_{t_x \in \{g_m,...,t\}} \sum_{n_2 \in N} x_{n_1,n_2,m,t_x} \geq 0 \tag{11}$$

Prawo Kirchhoffa powinno zostać spełnione na koniec przypadku testowego. Oczekujemy dostarczenia wszystkich przepływów z węzłów źródłowych do docelowych, po wszystkich momentach $t \in T$:

$$\begin{gathered} \forall n_1 \in N, m \in M : \\ v_{m,n_1} s_m + \sum_{t \in T} \sum_{n_2 \in N} x_{n_2,n_1,m,t} = \\ b_{m,n_1} s_m + \sum_{t \in T} \sum_{n_2 \in N} x_{n_1,n_2,m,t} \end{gathered} \tag{12}$$

Każdy z wykorzystywanych węzłów posiada ograniczoną moc nadajnika, która nie powinna zostać przekroczona w żadnym z momentów $t \in T$:

$$\forall t \in T, n_1 \in N : \sum_{n_2 \in N} \sum_{m \in M} c_{n_1,n_2,t} x_{n_1,n_2,m,t} \leq p_{n_1} \tag{13}$$

4.5 Funkcja celu

Niech funkcja celu określa łączną wykorzystaną moc nadajników na wszystkich węzłach $n \in N$ podczas wszystkich momentów symulacji $t \in T$:

$$f_c = \sum_{t \in T} \sum_{n_1 \in N} \sum_{n_2 \in N} \sum_{m \in M} c_{n_1,n_2,t} x_{n_1,n_2,m,t} \tag{14}$$

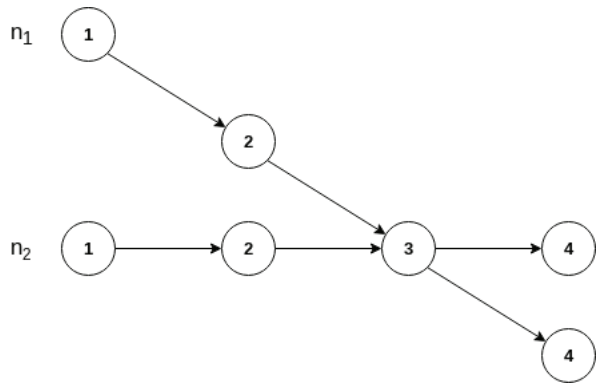
Zadaniem optymalizacji jest minimalizacja funkcji celu w celu określenia optymalnych ścieżek routingu mających na celu ograniczenie zużycia energii:

$$min(f_c) \tag{15}$$

5. PRZYPADKI TESTOWE

Na potrzeby analizy i testów zaproponowanego modelu optymalizacji dokonano jego implementacji w środowisku Python z wykorzystaniem solvera Gurobi Optimizer. Niech dla każdego z przypadków testowych koszt obsługi jednostki przepływu między kolejnymi węzłami będzie wprost proporcjonalny do kwadratu odległości między nimi w przestrzeni $\mathbb{R}^3$.

5.1 Przemieszczenie węzłów



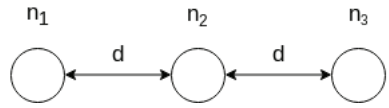
Rysunek 1: Przemieszczenie węzłów w czasie

Niech istnieją dwa węzły sieci FANET $N = \{n_1, n_2\}$. Węzły przemieszczają się w momentach symulacji $T = \{1, 2, 3, 4\}$ ruchem jednostajnym prostoliniowym, tak że w momencie $t = 3$ ich położenie jest tożsame. Dany jest przepływ $m_1 \in M$ wygenerowany w momencie $g_{m_1} = 1$, dla którego $v_{m_1,n_1} = b_{m_1,n_2} = 1$. Zakładamy, iż rozmiar przepływu $m_1 \in M$ pozwala na przesłanie go między węzłami $n_1, n_2 \in N$ w dowolnym momencie $t \in T$, o ile spełnione są pozostałe ograniczenia.

W związku ze spadkiem kosztu transmisji przepływu między węzłami n_1, n_2 wraz ze zmniejszeniem odległości między nimi, transmisja przepływu m_1 nastąpiła w chwili $t = 3$. Model optymalizacji wskazał rozwiązanie najkorzystniejsze energetycznie na podstawie określonych założeń. Należy zauważyć iż w przypadku ograniczonego rozmiaru bufora na węźle n_1 mogłoby dojść do sytuacji jego przepełnienia przy znacznej liczbie przepływów. Jednocześnie optymalne energetycznie rozwiązanie zakłada wprowadzenie opóźnienia nadawania przepływu m_1 między momentem jego generacji $t = 1$, a momentem jego nadania $t = 3$, co mogłoby nie być akceptowalne w przypadku niektórych zastosowań sieci FANET.

5.2 Wykorzystanie węzła pośredniczącego

Niech istnieją trzy węzły sieci FANET $N = \{n_1, n_2, n_3\}$. Węzły znajdują się w stałych położeniach we wszystkich momentach przypadku testowego $T = \{1, 2, 3, 4\}$. Odległości między węzłami n_1, n_2 oraz n_2, n_3 są równe i stanowią połowę odległości między węzłami n_1, n_3. Dany jest przepływ $m_1 \in M$ wygenerowany w momencie $g_{m_1} = 1$, dla którego $v_{m_1,n_1} = b_{m_1,n_3} = 1$. Zakładamy, iż rozmiar przepływu $m_1 \in M$ pozwala na przesłanie go między dowolnymi dwoma węzłami $n_x, n_y \in N$, w dowolnym momencie $t \in T$, o ile spełnione są pozostałe ograniczenia.



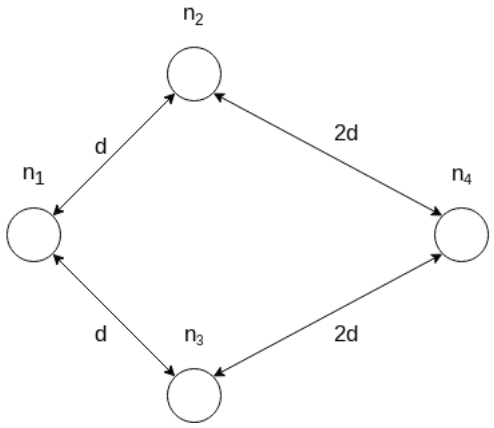
Rysunek 2: Wykorzystanie węzła pośredniczącego

Uzyskane rozwiązanie optymalne założyło przesłanie przepływu m_1 między węzłami n_1, n_2 w momencie $t = 3$, a następnie przesłanie tego przepływu między węzłami n_2, n_3 w momencie $t = 4$. Do przesłania przepływu m_1 został wykorzystany węzeł pośredniczący n_2 ze względu na rosnący koszt obsługi przepływu wprost proporcjonalnie do kwadratu odległości między komunikującymi się węzłami. Transmisja przepływu m_1 rozpoczęła się w momencie $t = 3$, z opóźnieniem wobec momentu generacji przepływu $t = 1$ ze względu na istnienie wielu równoważnych rozwiązań w kontekście zdefiniowanego modelu optymalizacji. W celu ograniczenia opóźnienia transmisji przepływu m_1 należałoby zdefiniować dodatkową funkcję kary, zawartą w funkcji celu, której wartość zwiększałaby się wraz ze wzrostem opóźnień w nadawaniu przepływów po ich generacji.

5.3 Wykorzystanie wielu ścieżek

Niech istnieją węzły sieci FANET $N = \{n_1, n_2, n_3, n_4\}$. Węzły n_1, n_2 oraz n_1, n_3 znajdują się w odległości d. Węzły n_2, n_4 oraz n_3, n_4 znajdują się w odległości $2d$. Odległość między węzłami n_1, n_4 nie pozwala na bezpośrednią komunikację powyższych węzłów. Dany jest przepływ $m_1 \in M$ wygenerowany w momencie $g_{m_1} = 1$, dla którego $v_{m_1,n_1} = b_{m_1,n_4} = 1$. Rozmiar przepływu m_1 nie pozwala

na przesłanie go w całości z wykorzystaniem wyłącznie jednego z węzłów n_2, n_3.



Rysunek 3: Wykorzystanie wielu ścieżek

Uzyskane rozwiązanie optymalne założyło przesłanie przepływu m_1 z wykorzystaniem dwóch ścieżek, wykorzystujących zamiennie węzły n_2 oraz n_3. Transmisja odbywała się podczas trwania wszystkich momentów przypadku testowego $t \in T$. Podział przepływu między obie ścieżki jest nierównomierny. W związku z jednakowym kosztem obsługi przepływu na każdej ze ścieżek, istnieje wiele rozwiązań równoważnych, różniących się poziomem wykorzystania nadajników na danej ścieżce routingu.

6. PODSUMOWANIE

Rozwój sieci FANET w zastosowaniach cywilnych i wojskowych wiąże się z powstawaniem kolejnych wyzwań związaych z funkcjonowaniem takich sieci. Proponowany model optymalizacji matematycznej pozwala na wyznaczenie najlepszych dostępnych ścieżek dla routingu unicast w uproszczonych przypadkach testowych. Dalsze perspektywy rozwoju proponowanego modelu zakładają analizę zachowania routingu multicast, zwiększenie uwzględnianych w modelu czynników celem lepszego odwzorowania rzeczywistego środowiska, a także wprowadzenie możliwości iteracyjnego rozwiązywania modelu dla mniejszych okien czasowych z uwzględnieniem zmian stanów węzłów odbywających się równolegle z przeprowadzanymi analizami.

LITERATURA

[1] Townsend Ashleigh, Jiya Immanuel N., Martinson Christiaan, Bessarabov Dmitri, Gouws Rupert. 2020. "A comprehensive review of energy sources for unmanned aerial vehicles, their shortfalls and opportunities for improvements", Heliyon, 6 (11) : e05285.

[2] Zhang Jiachi, Xueyun Wang, Liu Liu. 2023. "A Location and Velocity Prediction-Assisted FANET Networking Scheme for Highly Mobile Scenarios" Electronics, 12 (12) : 2731.

[3] Chaker Boutalbi Mohammed, Amine Riahla Mohamed, Aimad Ahriche. 2021. "A summary of the existing challenges in the design of a routing protocol in UAVs network", 2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH), Boumerdes, Algeria, 59-64.

[4] Lansky Jan, Rahmani Amir Masoud, Malik Mazhar Hussain, Yousefpoor Efat, Yousefpoor Mohammad Sadegh, Khan Muhammad Umair, Hosseinzadeh Mehdi. 2023. "An energy-aware routing method using firefly algorithm for flying ad hoc networks", Scientific Reports, 13 : 1323.

[5] Oubbati Omar Sami, Mozaffari Mohammad, Chaib Noureddine, Lorenz Pascal, Atiquzzaman Mohammed, Jamalipour Abbas. 2019. "ECaD: Energy-efficient routing in flying ad hoc networks", International Journal of Communication Systems, 32 : e4156.

[6] Guillen-Perez Antonio, Cano Maria-Dolores. 2018. "Flying Ad Hoc Networks: A New Domain for Network Communications", Sensors, 18 (10) : 3571.

[7] Singh Kuldeep, Verma Anil. 2018. "Flying Adhoc Networks: Concept and Challenges", Advanced Methodologies and Technologies in Network Architecture, Mobile Computing, and Data Analytics.

[8] Hussen Hassen Redwan, Sung-Chan Choi, Jong-Hong Park, Jaeho Kim. 2019. “Predictive geographic multicast routing protocol in flying ad hoc networks”, International Journal of Distributed Sensor Networks, 15 (7).

[9] Arabas Piotr, Jóźwik Tomasz, Niewiadomska-Szynkiewicz Ewa. 2023. "Router Activation Heuristics for Energy-Saving ECMP and Valiant Routing in Data Center Networks", Energies, 16 (10) : 4136.

[10] Lakew Demeke Shumeye, Sa'ad Umar, Dao Nhu-Ngoc, Na Woongsoo, Cho Sungrae. 2020. "Routing in Flying Ad Hoc Networks: A Comprehensive Survey", IEEE Communications Surveys & Tutorials, 22 (2) : 1071-1120.

[11] Khan Muhammad Fahad, Kok-Lim Alvin Yau, Rafidah Md Noor, Muhammad Ali Imran. 2020. "Routing Schemes in FANETs: A Survey", Sensors, 20 (1) : 38.

[12] Al-Emadi Sara, Al-Mohannadi Aisha. 2020. "Towards Enhancement of Network Communication Architectures and Routing Protocols for FANETs: A Survey", 3rd International Conference on Advanced Communication Technologies and Networking (CommNet) (CommNet'20).

ZASTOSOWANIE INTERFEJSÓW D-ROF/A-ROF W SIECIACH 802.11 NASTĘPNYCH GENERACJI

APPLICATION OF D-ROF/A-ROF INTERFACES IN THE NEXT GENERATION 802.11 NETWORKS

Zbigniew Zakrzewski [1]

[1] Politechnika Bydgoska im. Jana i Jędrzeja Śniadeckich, Bydgoszcz, zbizak@pbs.edu.pl

DOI: 10.15199/59.2024.4.49

Streszczenie**: W artykule przedstawiono propozycję wprowadzenia do sieci WLAN architektury C-AP (Centralized/Cloud Access Point), która umożliwi w przyszłości stosowanie interfejsów D-RoF (Digitized Radio-over-Fiber) lub A-RoF (Analog Radio-over-Fiber). Na podstawie przeprowadzonych obliczeń wykazano, że nawet dla interfejsów Wi-Fi 802.11 pracujących w paśmie mmWave z bardzo szerokimi kanałami częstotliwościowymi jest to możliwe. Wzorce zostały zaczerpnięte z rozwiązań stosowanych w dobrze rozpoznawalnym interfejsie CPRI (Common Public Radio Interface).**

Abstract**: The paper presents a proposal to introduce the C-AP (Centralized/Cloud Access Point) architecture into WLANs, which will enable the use of D-RoF (Digitized Radio-over-Fiber) or A-RoF (Analog Radio-over-Fiber) interfaces in the future. Based on the calculations performed, it was shown that this is possible even for Wi-Fi 802.11 interfaces operating in the mmWave band with very wide frequency channels. The patterns were taken from solutions used in the well-recognized CPRI (Common Public Radio Interface).**

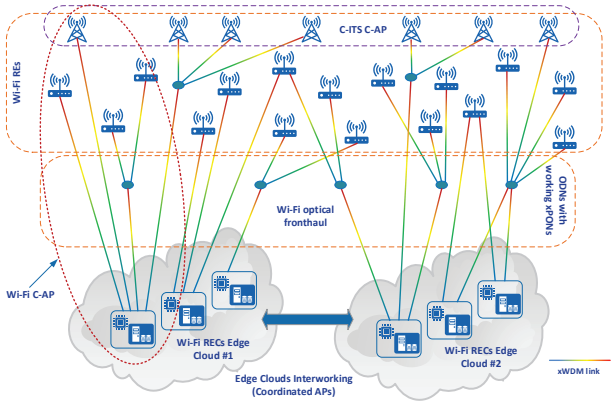
Słowa kluczowe**: architektura C-RAN, CPRI, DIPP, DSB-RFoF, radio poprzez światłowód, Wi-Fi fronthaul.**

Keywords**: CPRI, C-RAN architecture, DIPP, DSB-RFoF, radio-over-fiber, Wi-Fi fronthaul.**

1. WSTĘP

Współczesne bezprzewodowe sieci mobilnego i przenośnego dostępu do Internetu stanowią jeden z kluczowych fragmentów sieci teleinformatycznych. Standardy definiujące sieci mobilne stają się coraz bardziej ogólnoświatowe, wręcz monopolizują technologię w ruchomej komunikacji naziemnej. Poczynając od 4G konsorcjum 3GPP przyjęło rolę lidera w tworzeniu i wyznaczaniu trendów we wprowadzaniu nowatorskich rozwiązań w systemach radiowych. W trakcie powstawania pierwszych wydań 5G już było wiadomo, że w zasadzie na świecie nie będzie konkurenta w zakresie sieci WWAN (Wireless WAN). Dodatkowym atutem w stymulacji rozwoju nowej generacji systemów mobilnych było udostępnienie pasm częstotliwościowych wolnych od opłat i koncesji. Niestety tego rodzaju posunięcie powoduje potrzebę koegzystencji z innymi od dawna pracującymi systemami, zwłaszcza małego zasięgu. Systemy 5G, jak też przyszłe 6G niewątpliwie będą wchodziły w zakres usług systemów i sieci, które do niedawna nie mały konkurentów. W tym względzie najbardziej zagrożonymi wydają się być rozwiązania IEEE wspierające dostęp osobisty, sensoryczny oraz lokalny.

W niniejszym opracowaniu zwrócono szczególną uwagę na lokalne sieci teleinformatyczne bazujące na dostępie z wykorzystaniem interfejsu IEEE 802.11. Aby sieci bazujące na systemach 3GPP 5G/6G nie wyparły sieci IEEE 802.11, należy myśleć o wydajnej ich współpracy oraz dotrzymywaniu kroku w unowocześnianiu rozwiązań. By tak się stało, nowe generacje sieci WLAN 802.11 powinny bazować na rozwiązaniach podobnych do tych, które są wprowadzane przez 3GPP w sieciach mobilnych. W tym miejscu można wyciągnąć wniosek, że zasadne jest rozszerzenie architektury lokalnych sieci 802.11 na rozwiązania chmurowe i jednocześnie scentralizowane w kontekście sterowania i zarządzania zasobami radiowymi.



Rys. 1 Propozycja wprowadzenia architektury C-AP w bezprzewodowych sieciach lokalnych bazujących na interfejsach IEEE 802.11

Na rysunku 1 przedstawiono propozycję rozdzielenia funkcji związanych z przygotowaniem sygnałów radiowych w lokalizacji Wi-Fi REC (Radio Equipment Control) oraz funkcji wykonawczych dotyczących emisji i wzmocnienia tych sygnałów w lokalizacji Wi-Fi RE (Radio Equipment). W przypadku szerokopasmowego i masowego dostępu do sieci zajdzie potrzeba zwiększenia mocy obliczeniowej w przetwarzaniu sygnałów w Wi-Fi RECs, co na podobieństwo NG-RAN pracującej w architekturze O-RAN (Open Radio Access Network) może zostać zrealizowane w brzegowej chmurze.

Tab. 1 Zestawienie podstawowych parametrów kluczowych radiowych interfejsów z rodziny IEEE 802.11 [4,7,8,9]

Standard	*Pasmo RF [GHz]*	*Szerokość kanału RF [MHz]*	*Modulacja w kanale*	*Liczba podnośnych OFDM*	*Modulacja pasmowa QAM*	*FFT/ DFT*	*Technika przestrzenna maks.*
802.11be	2,4; 5; 6	do 320	OFDM	do 3920	do 4096	do 4096	MIMO 16x16
802.11bn	2,4; 4; 6; 7; 42,5 - 71	do 640	OFDM	do 7900	do 8192 (opcja 16k)	do 8192	MIMO 16x16
802.11ad/ay	57 - 71	2160 – 8640	OFDM/SC	do 1611	do 64	do 2048	MIMO 8x8 (opcja 16x16)
802.11p/bd	5,9; 57 - 71	do 20	OFDM	do 114	do 256	do 256	Sektory – do 8 (opcja MIMO 2x2)

Zważywszy, że światłowodowe sieci dystrybucyjne ODN (Optical Distribution Network) zbudowane w wersji pasywnej stają się powszechnie dostępne na osiedlach, w biurach i ośrodkach przemysłowych, można je z powodzeniem wykorzystać do transportowania sygnałów radiowych przygotowanych w formatach D-RoF, czy A-RoF [14]. Systemy xPON pracujące w tych sieciach oparte są na z góry zdefiniowanych standardach, które mogą zajmować znaczące pasmo optyczne należące do zakresu jednomodowości światłowodów ITU-T G.652D/G.657A [10,11]. Po analizie zasobów zajętych w sieci ODN wywnioskowano, że jeszcze nie zostało obsadzone przez systemy xPON optyczne pasmo E [16]. Dalsze rozważania i obliczenia będą skierowane na ten zakres długości fal optycznych.

2. INTEFESJY BEZPRZEWODOWE STOSOWANE W SIECIACH 802.11

Do najpopularniejszych interfejsów Wi-Fi zaliczamy obecnie 802.11ax (gen. 6 i 6E), jednak już jest on wypierany przez 802.11be (gen. 7) [9] mimo, że standard jeszcze nie został zatwierdzony. W interfejsie tym w paśmie 6 GHz istnieje możliwość utworzenia kanału o szerokości 320 MHz. Jego następcą będzie 802.11bn (gen. 8) [4], w którym planowane jest utworzenie kanału o podwojonej szerokości wynoszącej 640 MHz. Dodatkowym atutem ma być integracja go z pasmem mmWave (60 GHz). W rodzinie standardów 802.11 już istnieje rozwiązanie dla fal milimetrowych pod rozszerzeniem 802.11ad, które zostało udoskonalone w wersji 802.11ay [7]. Brak szerokiego zastosowania tego odosobnionego standardu spowodował, że podjęto plan jego integracji z nową już 8 generacją przyszłego rozwiązania Wi-Fi. Wyjątkowym interfejsem jest 802.11p, który bazuje na zastrzeżonym paśmie dla systemów trakcyjnych ITS 5.9 GHz, ale w wersji rozszerzonej 802.11bd [8] może także pracować w paśmie 60 GHz. Wspólną cechą wszystkich interfejsów jest to, że wykorzystują podobne lub te same techniki modulacji oraz rozwiązania przestrzenne. Różnice wynikają przede wszystkim z częstotliwościowych pasm pracy, co warunkuje możliwość tworzenia kanałów o określonej szerokości zawierających dużą liczbę podnośnych OFDM. W tabeli 1 zawarto zestawienie podstawowych parametrów wymienionych bezprzewodowych interfejsów z rodziny IEEE 802.11. Niektóre z nich, zwłaszcza te dotyczące techniki modulacji OFDM oraz wykorzystywanego pasma RF, będą miały zastosowanie przy wskazaniu parametrów zaproponowanego optycznego interfejsu Wi-Fi fronthaul bazującego na optycznej sieci ODN.

3. SIEĆ WI-FI FRONTHAUL PRACUJĄCA W FORMACIE D-ROF

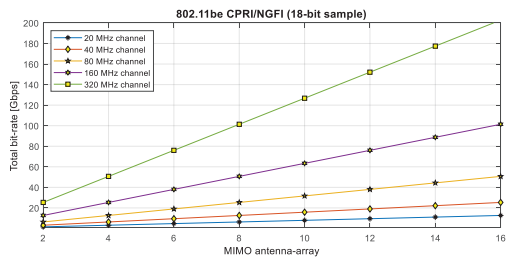
W poprzednim dziesięcioleciu największą popularność zdobył interfejs CPRI [1], który w dalszej części opracowania posłuży jako wyznacznik w tworzeniu optycznego łącza fronthaul dla Wi-Fi pracującej w formacie D-RoF. CPRI został stworzony dla domeny RAN systemów mobilnych i jest oparty na podstawowym zegarze 3,84 MHz taktującym strumień chipów w interfejsie 3G (UMTS). W konsekwencji takich uwarunkowań należałoby utworzyć nową ramkę CPRI, która będzie dopasowania do potrzeb interfejsu 802.11, którego główną cechą jest półdupleksowa praca bazująca na technice dostępu CSMA/CA. Interfejs wykorzystujący pełny dupleks Wi-Fi, określany jako MLO (Multi-Link Operation), jest planowany do wprowadzenia w ramach nowego rozszerzenia 802.11bn [4]. Dupleks ten w celu zachowania zgodności wstecz będzie bazował na równoległej pracy w co najmniej dwóch pasmach RF.

W niniejszym opracowaniu skupimy się na obliczeniu zapotrzebowania na przepustowość w sieci ODN, w celu przeniesienia ruchu Wi-Fi CPRI (D-RoF) w paśmie optycznym E [16], które nie jest jeszcze zajęte przez systemy z rodziny xPON. Do tego celu zastosujemy zależność [13]:

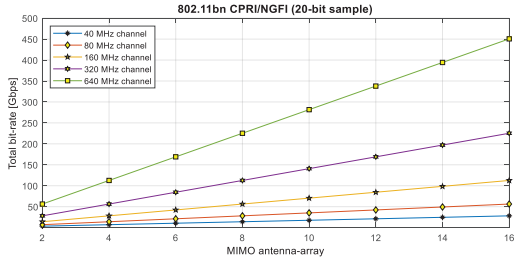
$$BR_8 = S \cdot A \cdot f_s \cdot b_s \cdot IQ \cdot HF \cdot LC \qquad (1)$$

gdzie: S – liczba sektorów, A – liczba modułów antenowych na sektor, f_s – częstotliwość próbkowania, b_s – liczba bitów na próbkę, IQ – zwielokrotnienie kwadraturowe, HF – współczynnik informacji zawartej w nagłówku protokołu CPRI, LC – sprawność alfabetowego kodu liniowego. Częstotliwość próbkowania jest zależna od odstępu między podnośnymi OFDM oraz liczby punktów FFT/DFT (tab. 1) stosowanej w kanale z pasma podstawowego BB, zaś liczba bitów przypadających na próbkę zależy od wartościowości modulacji pasmowej oraz stosowanego stopnia kompresji.

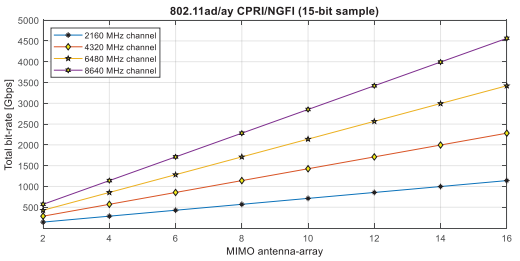
a)



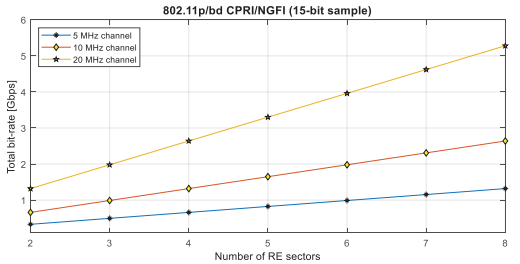
b)



c)



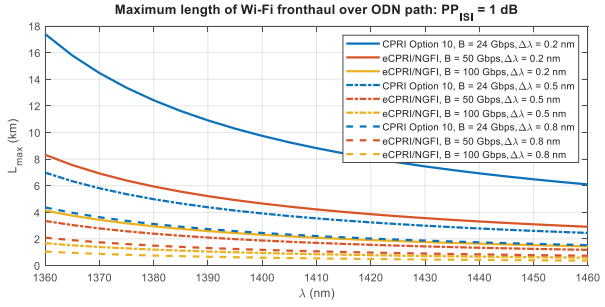
d)



Rys. 2 Wyniki obliczeń maksymalnej przepływności bitowej potrzebnej do przekazania scyfryzowanych sygnałów BB poprzez ODN dla interfejsu: a) 802.11be, b) 802.11bn, c) 802.11ad/ay, d) 802.11 p/bd

Na rysunku 2 przedstawiono wyniki obliczeń wykonanych na podstawie zależności (1). Dobrano różne wartości współczynnika b_s, w zależności od stopnia złożoności sygnału zmodulowanego. W interfejsach z pasm do 7 GHz odstępy między podnośnymi zmieniają się numerologicznie względem odstępu pierwotnego 312,5 kHz (802.11a), zaś w paśmie 60 GHz wzorzec stanowi 5156,25 kHz. Obliczenia pokazują, że głównym źródłem wysokich przepływności bitowych Wi-Fi fronthaul jest duży odstęp między podnośnymi oraz potrzeba stosowania dużej liczby bitów na próbkę.

W przypadku, gdy łączna przepływność przekracza możliwości pojedynczego kanału optycznego, należy strumień podzielić na wiele podkanałów możliwych do zestawienia w łączu ODN.

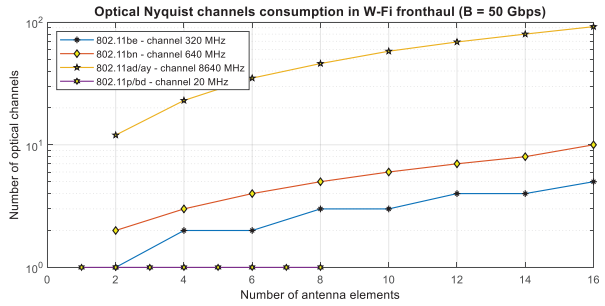


Rys. 3 Przykłady dopuszczalnych długości ścieżki optycznej dla różnych strumieni NRZ-OOK oraz różnych laserów jako źródeł optycznej nośnej

Granicę możliwości kanału dla najprostszego optycznego sygnału NRZ-OOK można wyznaczyć na podstawie zależności [12]:

$$L_{\max} = \frac{\varepsilon c}{D\lambda^2 B\sqrt{\left(\frac{B}{\pi f}\right)^2 + \sigma_v^2}} \tag{2}$$

gdzie: ε - dopuszczalne zmniejszenie czułości odbiornika dla interferencji międzysymbolowej PP_{ISI} na poziomie 1 dB, D – współczynnik dyspersji chromatycznej ścieżki optycznej, λ - długość fali optycznej nośnej, B – bitowa przepływność łącza, f – wypełnienie szczeliny czasowej, σ_v – 20-dB częstotliwościowa szerokość widma lasera. Wyniki obliczeń zostały przedstawione na rysunku 3. Uwzględniono także przepływność 100 Gbps, gdyż ścieżka optyczna fronthaul może mieć wymaganą długość nieprzekraczającą kilkuset metrów. W następnym kroku można wyznaczyć liczbę kanałów optycznych potrzebnych do zestawienia łącza fronthaul.



Rys. 4 Liczba potrzebnych kanałów optycznych xWDM do przeniesienia strumienia Wi-Fi fronthaul przy założeniu, że przepływność w kanale wynosi 50 Gb/s.

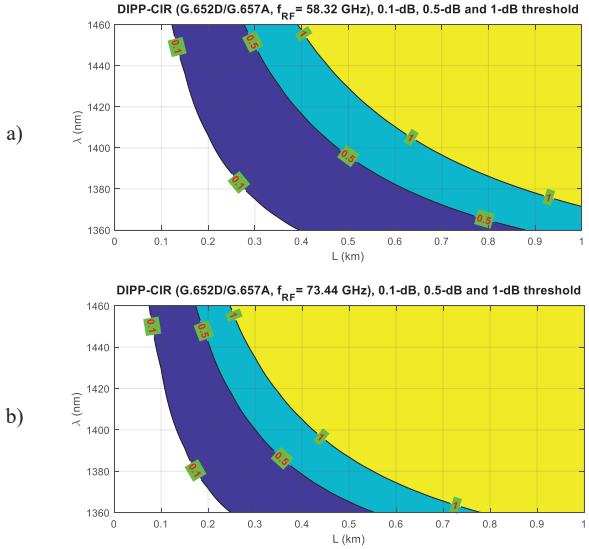
Wyniki obliczeń przedstawionych na rysunku 4 pokazują, że zrealizowanie tego zadania jest możliwe w optycznym paśmie E, jednak dla interfejsu 802.11ad/ay byłoby konieczne uruchomienie stabilnych kanałów DWDM (Dense Wavelength Division Multiplexing). W pozostałych przypadkach można zastosować rzadszą siatkę przy wykorzystaniu tańszych laserów. W przypadku interfejsu 802.11p/bd transmisja może odbywać się w jednym kanale przy zastosowaniu multipleksacji CPRI/TDM.

4. OPTYCZNE ŁĄCZE WI-FI FRONTHAUL PRACUJĄCE W FORMACIE A-ROF

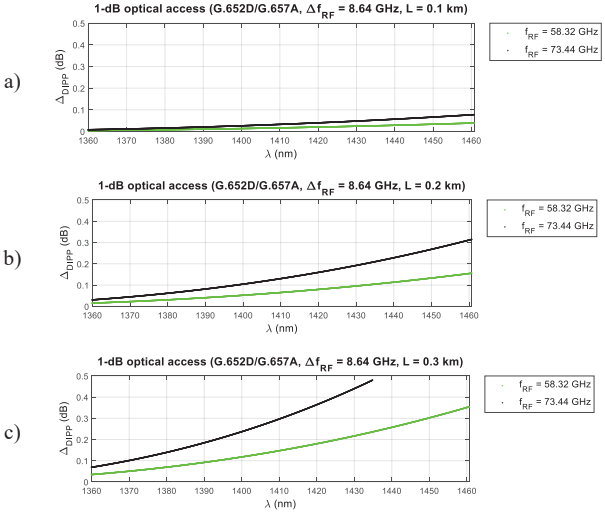
Zastosowanie formatu A-RoF do transmisji sygnałów radiowych może być użyteczne w przypadku, gdy wolne zasoby ODN są zbyt małe, aby zastosować format D-RoF. W przypadku wprowadzenia A-RoF do łącza Wi-Fi fronthaul najbardziej podatny na dyspersyjne oddziaływanie ścieżki optycznej będzie sygnał DSB-RFoF (Double Side Band Radio Frequency over Fiber) [15]. Wpływ zjawiska DIPP będzie tutaj kluczowy, co można ocenić na podstawie zależności [3]:

$$\text{DIPP-CIR}_{IM} = 20\log\cos\left(\pi D(\lambda_0) L \frac{\lambda_0^2}{c} f_{RF}^2\right) \tag{3}$$

gdzie: L – długość ścieżki optycznej, f_{RF} – częstotliwość nośnej radiowej, λ_0 – długość fali optycznej nośnej, D – współczynnik dyspersji chromatycznej ścieżki ODN.



Rys. 5 Dostępność zasobów w optycznym paśmie E przy przesyłaniu sygnału DSB-RFoF w interfejsie Wi-Fi fronthaul 802.11ad/ay: a) nośna RF 58,32 GHz, b) nośna RF 73,44 GHz



Rys. 6 Wyniki obliczeń różnicowego wygaszania podnośnych OFDM w kanale optycznym dla interfejsu Fi-Fi fronthaul 802.11ad/ay w formacie DSB-RFoF: a) L=0,1 km, b) L=0,2 km, c) L=0,3 km

Obliczenia przeprowadzono dla najbardziej szerokopasmowego interfejsu Wi-Fi fronthaul 802.11ad/ay, zaś wyniki przedstawiono na rysunkach 5 i 6. Można zaobserwować, że praktycznie dla ścieżki do 500 metrów, przy kanale radiowym o szerokości nawet 8,64 GHz, można w optycznym paśmie E prowadzić transmisję z minimalnym szumem fazowym wprowadzanym przez dyspersję chromatyczną występująca w światłowodzie G.652D/G.657A.

5. WNIOSKI I DALSZE BADANIA

Przeprowadzone obliczenia pokazują, że zastosowanie interfejsów D-RoF i A-RoF w proponowanych łączach Wi-Fi fronthaul jest możliwe. Podczas obliczeń założono, że transmisja będzie odbywała się w istniejącej w danej lokalizacji światłowodowej sieci ODN, w której będą pracowały systemy z rodziny xPON. W celu uniezależnienia transmisji, wybrano dostępne optyczne pasmo E [16]. Taki wybór był podyktowany tym, że systemy xPON wyższych generacji nie są przygotowane do transmisji sygnałów A-RoF, zaś dostępne strumienie cyfrowe bazują na zbyt niskiej liczbie optycznych kanałów xWDM.

Interfejs CPRI w założeniu zapewnia pełny synchronizm, co daje możliwość skutecznego wykorzystania techniki kształtowania wiązki w bezprzewodowym interfejsie 802.11 na poziomie Wi-Fi REC (rys. 1). Wadą tego rozwiązania jest generowanie bardzo dużych strumieni bitowych. W sytuacji, gdy część funkcji dotyczących przetwarzania informacji i sygnałów radiowych będzie mogła być przeniesiona do Wi-Fi RE, wówczas można wykorzystać łącza Ethernet wprowadzając rozwiązania eCPRI/NGFI [2,5,6], co będzie tematem kolejnych badań.

LITERATURA

[1] CPRI Industry Forum (Ericsson, Huawei, NEC, and Nokia). 2015. CPRI Specification 7.0.

[2] CPRI Industry Forum (Ericsson, Huawei, NEC, and Nokia). 2019. eCPRI Specification 2.0.

[3] Elrefaie A.F., Wagner R.E., Atlas D.A., Daut D.G. 1988. "Chromatic dispersion limitations in coherent lightwave transmission systems". *Journal of Lightwave Technology*. 6(5): 704-709.

[4] Galati-Giordano L., Geraci G., Carrascosa M., and Bellalta B. 2023. "What Will Wi-Fi 8 Be? A Primer on IEEE 802.11bn Ultra High Reliability", *Networking and Internet Architecture*, arXiv:2303.10442.

[5] IEEE, 1914.1. 2019. Standard for Packet-Based Fronthaul Transport Networks.

[6] IEEE, 1914.3. 2023. Standard for Radio Over Ethernet Encapsulations and Mappings.

[7] IEEE, 802.11ay. 2021. "Enhanced Throughput for Operations in License-exempt Bands above 45 GHz"

[8] IEEE, 802.11bd. 2022. "Enhancements for Next Generation V2X".

[9] IEEE, P802.11be/D3.0. 2023. Enhancements for Extremely High Throughput (EHT).

[10] ITU-T, G.652. 2016. Characteristics of a single-mode optical fibre and cable.

[11] ITU-T, G.657. 2016. Characteristics of a bending-loss insensitive single-mode optical fibre and cable.

[12] ITU-T, Supp. 39 to G-series recomm. 2016. Optical system design and engineering considerations.

[13] Pfeiffer T. 2015. Next Generation Mobile Fronthaul and Midhaul Architectures. *J. Opt. Commun. Netw.*, 7.

[14] Zakrzewski Zbigniew. 2020. "D-RoF and A-RoF Interfaces in an All-Optical Fronthaul of 5G Mobile Systems". *Appl. Sci*. 10 (1212).

[15] Zakrzewski Zbigniew. 2021. "Optical Channel Selection Avoiding DIPP in DSB-RFoF Fronthaul Interface". *Entropy* 23(11): 1554.

[16] Zakrzewski Zbigniew. 2023. "Integracja O-RAN oraz xPON na warstwie optycznych zasobów ODN". *PTiWT* 4/2023: 277-280.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

Implementacja rutera z mechanizmem adaptacyjnego multi-homingu

Implementation of a router with adaptive upstream connection multi-homing

Piotr Jurkiewicz[1]; Michał Rzepka[1]; Piotr Jaglarz[1]; Jerzy Domżał[1]; Robert Wójcik[1];

[1] link2net s.c., kontakt@link2net.pl

DOI: 10.15199/59.2024.4.50

Streszczenie: W artykule przedstawiono implementację rutera klasy domowej umożliwiającego jednoczesny adaptacyjny dostęp do Internetu przy użyciu wielu sieci dostępowych. Urządzenie składa się z płytki rozwojowej BPI-W2 oraz autorskiej płytki nakładkowej zapewniającej funkcje sieciowe i pracuje pod kontrolą systemów operacyjnych OpenWRT i Android. Przeprowadzone testy wykazały, że umożliwia na jednoczesny dostęp do Internetu przy użyciu dwóch sieci dostępowych i osiągnięcie co najmniej przepustowości 183% pojedynczego łącza w przypadku pobierania i 159% w przypadku wysyłania danych.

Abstract: The article presents the implementation of a home-class router that allows simultaneous adaptive Internet access using multiple access networks. The device consists of a BPI-W2 development board and a custom overlay board that provides networking functions, operating under the control of OpenWRT and Android operating systems. Tests have shown that it enables simultaneous Internet access using two access networks, achieving at least 183% of the bandwidth of a single link for downloading and 159% for uploading data.

Słowa kluczowe: ruter, przepływy, tablica przepływów, inżynieria ruchu, multi-homing

Keywords: router, flows, flow table, traffic engineering, multi-homing

1. WSTĘP

Najbardziej popularnym typem urządzeń sieciowych są domowe rutery IP. Zwykle sieć lokalna jest połączona przez ruter tylko z jednym dostawcą usług internetowych (ISP). Podłączenie do więcej niż jednego dostawcy mogłoby zwiększyć niezawodność, ponieważ gdy jedno z połączeń przestanie działać, pakiety mogłyby być przesyłane przez łącza innych operatorów. Łączenie hosta lub sieci końcowej z więcej niż jedną siecią dostępową nazywane jest *multi-homingiem*.

Niektóre z dostępnych na rynku domowych ruterów mają możliwość podłączenia wielu łączy dostępowych i pracy w tzw. trybie *fail over*. W takim przypadku transmisja odbywa się tylko przez jedno łącze dostępowe. Pozostałe interfejsy są używane jedynie w przypadku awarii głównego połączenia. Można sobie jednak także wyobrazić **jednoczesne** korzystanie z wielu łączy dostępowych. Pozwoliłoby to na zwiększenie osiąganej przepustowości. Ponadto, w zależności od obciążenia sieci dostępowej lub mocy sygnału, bardziej efektywne w danym momencie mogłoby być korzystanie z konkretnej sieci dostępowej.

W referacie [1], przedstawionym na konferencji KRiT 2023, zaprezentowaliśmy rezultaty badań **symulacyjnych** nad mechanizmem umożliwiającym równoległy i adaptacyjny dostęp do Internetu przy użyciu wielu sieci dostępowych. Jego istotą było adaptacyjne wybieranie łącza dostępowego w zależności od obciążenia sieci tak, aby by możliwe było uzyskanie zrównoleglonej transmisji i utrzymanie jak największej sumarycznej przepustowości w warunkach zmieniającej się w czasie przepustowości łączy dostępowych.

W tym artykule, który stanowi kontynuację uprzednio zaprezentowanych badań, przedstawiono implementację rutera opartego o system operacyjny OpenWRT wdrażającego opisany mechanizm. Przedstawia on także wyniki badań i testów rutera przeprowadzonych zarówno w warunkach laboratoryjnych, jak i w rzeczywistym użytkowaniu.

2. POWIĄZANE PRACE

Większość dostępnych na rynku ruterów IP do zastosowania domowego i biurowego nie zapewnia żadnego rodzaju *multi-homingu*. Wyjątkiem są niektóre rutery firmy ASUS, np. model GT-AX11000 [2]. Umożliwiają one na skonfigurowanie drugiej sieci WAN, która może być połączona zarówno poprzez kabel Ethernet, jak i bezprzewodowy modem USB. Mogą one działać w dwóch trybach: *load-balance* i *fail over*. W trybie *load-balance* użytkownik ustawia stosunek w jakim ruter będzie przydzielał nowe połączenia na obie sieci WAN. Stosunek ten jest konfigurowany przez użytkownika i nie zmienia się w czasie. W trybie *fail over* używana jest tylko jedna sieć WAN. Ruter okresowo przesyła zapytania DNS przez tą sieć i gdy wykryje, że przestała ona działać, przełącza się na drugą.

Istnieje także pakiet dla otwartoźródłowego oprogramowania OpenWRT [3] przeznaczonego dla ruterów wielu producentów, który zapewnia podobne funkcjonalności, o nazwie mwan3 [4]. Podobnie jak oprogramowanie firmy ASUS, umożliwia statyczne równoważenie ruchu pomiędzy wiele sieci oraz pracę w trybie *fail over*. Pakiet pozwala także na tworzenie bardziej zaawansowanych konfiguracji, w których sieć WAN wybierana jest w zależności od źródłowego lub docelowego adresu IP, portu źródłowego lub docelowego oraz typu protokołu warstwy transportowej (*policy-based routing*). Jednak podobnie jak w przypadku oprogramowania ASUS, stosunek w jakim dzielony jest ruch jest stały i nie zmienia się w czasie.

Brak jest dostępnych na rynku urządzeń, czy też otwartoźródłowego oprogramowania, umożliwiającego adaptacyjny *load-balancing* w ruterach domowych. Zagadnienie to nie było także dotychczas przedmiotem badań akademickich.

3. BUDOWA RUTERA

W trakcie przeprowadzonych prac badawczych opracowany został prototyp urządzenia składający się z płytki rozwojowej BPI-W2 oraz dedykowanej, autorsko zaprojektowanej płyty nakładkowej rutera oferującej odpowiednie interfejsy łączności oraz obsługę ekranu wraz z dotykowym panelem (Rysunek 1). Umożliwia ona obsługę 4 portów Ethernetowych, ekranu IPS z interfejsem LVDS oraz pojemnościowego panelu dotykowego z interfejsem I2C. Na podstawie przeprowadzonej analizy wymagań, oczekiwanych parametrów oraz docelowej budowy rutera, jako główny układ mikroprocesorowy wybrano chip Realtek RTD1296. Za udostępnienie i obsługę dodatkowych portów Ethernetowych odpowiedzialny jest układ pomocniczy RTL8367RB. Urządzenie zostało także wyposażone we wbudowany akumulator umożliwiający kontynuowanie pracy rutera w sytuacji awarii zasilania.

Płytka rozwojowa BPI-W2 nie posiada wbudowanego WiFI, z tego powodu musieliśmy użyć modułu zewnętrznego. Na podstawie informacji od producenta postanowiliśmy dodać wsparcie modułu WiFi RTL8822BE, który został podłączony za pomocą interfejsu M.2. Kolejnym wykonanym przez nas zadaniem, było dodanie wsparcia dla LTE. Płytka rozwojowa BPI-W2 posiada slot na kartę SIM oraz przygotowane gniazdo na modem LTE. Modemem, który wybraliśmy do wsparcia, był Sierra Wireless EM7455. Został on zainstalowany w gnieździe M.2, które miało połączenie z kartą SIM. W celu komunikacji z modemem wykorzystaliśmy interfejs USB, znajdującym się na tym gnieździe.

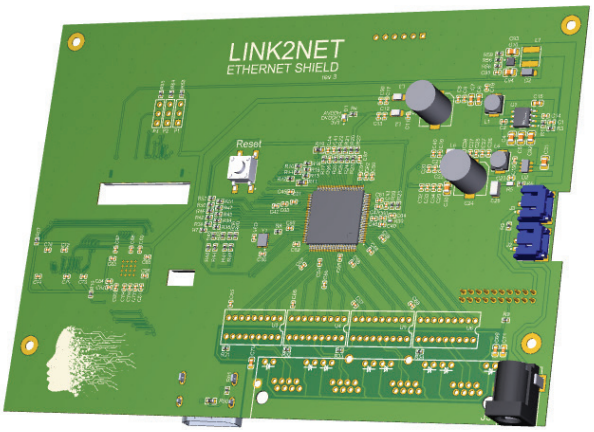
Zrealizowany przez nas ruter pracuje pod kontrolą dwóch systemów operacyjnych: OpenWRT oraz Android (Rysunek 2). OpenWrt, jako główny system operacyjny, odpowiedzialny jest za funkcje sieciowe oraz zarządzanie energią. Android, jako dodatkowy system operacyjny, odpowiadał za interfejs graficzny oraz wyświetlanie obrazu.

Zastosowanie dwóch systemów operacyjnych było możliwe dzięki wykorzystaniu wirtualizacji z pomocą hipernadzorcy Xen. Podstawowym systemem operacyjnym został OpenWrt, ponieważ odpowiadał on za kluczowe dla rutera działania sterowanie ruchem sieciowym. W naszej konfiguracji, Android jest systemem dodatkowym. Dostarcza on dla użytkownika nowoczesny interfejs graficzny. Dzięki niemu w łatwy sposób możemy tworzyć aplikacje do konfiguracji i zarządzania urządzeniem. Umożliwia on także na dalszy rozwój funkcjonalności urządzenia, na przykład integrację z systemami inteligentnego sterowania budynkami.

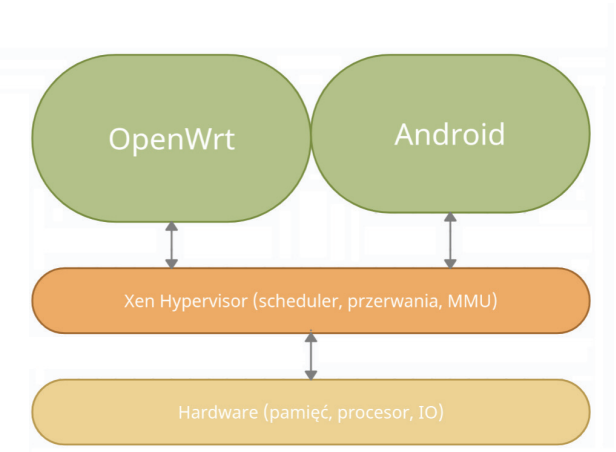
Taka architektura umożliwia rozdzielenie kluczowych funkcji urządzenia (sterowanie ruchem sieciowym, zarządzanie energią) od dodatkowych funkcji, które nie są wymagane podczas braku zasilania z sieci energetycznej. Tą unikalną cechę wykorzystaliśmy w algorytmie zarządzania energią. Umożliwia on zatrzymanie wirtualnej maszyny Androida, co powoduje zamrożenie stanów wszystkich procesów działających w systemie Android, również tych pracujących w tle. Dzięki temu procesor zostaje odciążony, co powoduje zmniejszenie zużycia energii.

Urządzenie zostało wyposażone w akumulator złożony z trzech ogniw o napięciu 3,6V oraz pojemności 3400 mAh. Przetestowaliśmy ruter pod kątem czasu działania wyłącznie na zasilaniu bateryjnym. W czasie testu urządzenie miało włączony punkt dostępowy WiFi oraz było podłączone za pomocą kabla Ethernetowego do zewnętrznej sieci WAN. Zastosowana pojemność baterii pozwoliła na osiągnięcie czasu pracy na baterii równego 4 godzin i 4 min.

Mechanizm adaptacyjnego *multi-homingu* został zaimplementowany w postaci dodatkowego demona działającego jako usługa w tle pod kontrolą systemu OpenWRT. Zastosowany został algorytm *ping*, który w badaniach symulacyjnych przedstawionych w [1] uzyskał najlepsze rezultaty. Demon okresowo przesyłał pakiety próbkujące przez każde z łączy dostępowych i na podstawie zmierzonych opóźnień oraz poziomu strat pakietów obliczał stosunek wag poszczególnych łączy. Stosunek wag łączy ustalany był jako stosunek odwrotności pierwiastków czasów odpowiedzi na pakiety ping wysłane przy użyciu poszczególnych łączy. W przypadku straty pakietu jako czas odpowiedzi przyjmowano 1 sekundę. Wagi te były następnie używane jako wagi wyboru odpowiedniej bramy domyślnej w systemowej tablicy rutingu dla nowo pojawiających się przepływów. W celu oszczędności energii pakiety próbkujące były wysyłane tylko w momencie pojawienia się nowego przepływu, ale nie częściej niż co 2 sekundy.



Rysunek 1: Płytka nakładkowa



Rysunek 2: Architektura urządzenia

4. TESTY W WARUNKACH LABORATORYJNYCH

Po zakończeniu prac implementacyjnych przystąpiliśmy do weryfikacji uzyskanych parametrów transmisji. Minimalne wymagania określone w założeniach projektu wynosiły:

- opóźnienie w transmisji dla pojedynczego pakietu mniejsze niż 10 ms,
- transmisja o prędkości min. 90% nominalnej przepustowości łącza,
- osiągnięcie min. 150% prędkości transmisji pojedynczego łącza przy podłączeniu dwóch takich samych łączy.

W celu weryfikacji ruter został podłączony do stanowiska pomiarowego, którym był komputer PC z wieloma interfejsami Ethernet (Rysunek 3). Jeden z nich został podłączony do rutera do gniazda po stronie LAN, a dwa z nich do gniazd po stronie WAN. Gniazdo LAN na ruterze działało z prędkością 1 Gbps i było jednym z gniazd wyprowadzonych bezpośrednio na płytce BPI-W2. Pierwsze z gniazd WAN było także gniazdem BPI-W2, jednak ustawione zostało na prędkość 100 Mbps. Natomiast drugie z gniazd WAN było gniazdem udostępnianym przez układ switcha RTL8367B i zostało także ustawione na 100 Mbps. Dzięki temu testy zweryfikowały także poprawność działania opracowanej płytki nakładkowej i jej połączenia do płytki bazowej BPI-W2. Na komputerze w trzech osobnych przestrzeniach nazw sieciowych (*network namespace*) emulowano hosta sieci LAN i dwa serwery podłączone do rutera po stronie WAN.

Pomiary opóźnienia transmisji (*round trip time* (RTT)) przeprowadzono przy pomocy narzędzia ping. Z emulowanego hosta znajdującego się po stronie LAN rutera przesyłano pakiety kolejno do emulowanych serwerów znajdujących się po obu stronach WAN rutera. Do każdego z serwerów po stronie WAN przeprowadzono 3 serie pomiarowe w trakcie każdej wysłano pojedynczo 10 pakietów i zmierzono opóźnienie każdego z nich.

Z przeprowadzonych testów wynika że ruter zapewnia maksymalne opóźnienie transmisji mniejsze niż 0,774 ms, a średnio wynosi ono ok. 0,5 ms. W przypadku interfejsu umieszczonego na płytce nakładkowej i udostępnianego przez dodatkowy układ switcha opóźnienie jest nieznacznie wyższe niż na interfejsach podłączonych bezpośrednio do SoC, co jest zgodne z przewidywaniami.

W celu pomiaru prędkości transmisji pojedynczego łącza użyto generatora ruchu sieciowego iperf w trybie TCP. Każdy z pomiarów trwał 10 sekund i mierzono średnią prędkości transmisji w ciągu całego pomiaru. Ze względu na specyfikę protokołu TCP, po rozpędzeniu się powinien on osiągnąć prędkość równą prędkości łącza.

W pierwszej kolejności, aby określić nominalną prędkość łącza, przeprowadzono pomiar przy połączeniu bezpośrednim interfejsów LAN i WAN pomiarowego stanowiska (bez użycia rutera). Przeprowadzono 10 pomiarów, których średni wynik wyniósł 94,68 Mbps.

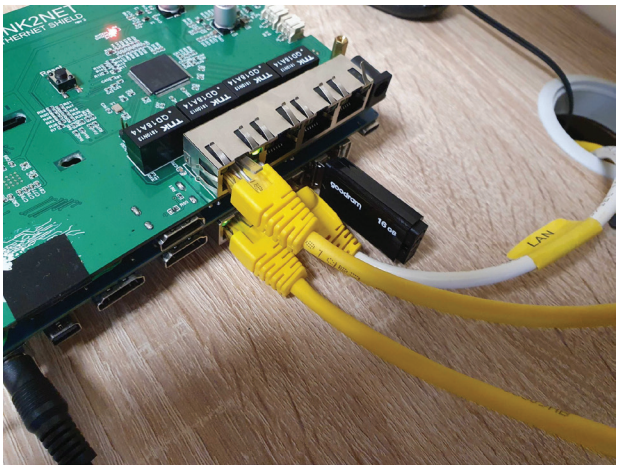
Następnie dla każdego z obu łączy WAN przeprowadzone osobno 5 pomiarów, których średnio wyniki wyniosły odpowiednio 94,66 Mbps i 92,68 Mbps. W przypadku łącza wan0 (na interfejsie podłączonym bezpośrednio do SoC) prędkość transmisji stanowi 99,98% nominalnej przepustowości łącza, natomiast w przypadku łącza wan1 (na interfejsie na płytce nakładkowej) prędkość transmisji stanowi 97,89% nominalnej przepustowości łącza.

Ostatecznie przebadany został scenariusz jednoczesnej transmisji z użyciem obu łączy. Przeprowadzono 10 pomiarów z których każdy trwał 10 sekund. Podczas każdego pomiaru jednocześnie uruchomione były dwa przepływy TCP, z których każdy używał innego łącza WAN. Mierzona była suma średniej prędkości transmisji obu z nich i wyniosła ona 181,25 Mbps.

Przyjmując za nominalną prędkość transmisji pojedynczego łącza 94,68 Mbps, średnia prędkości jednoczesnej transmisji dwóch łączy WAN wynosi 191,43% prędkości transmisji pojedynczego łącza. Wyniki testów laboratoryjnych zaprezentowano w Tabeli 1.

Tabela 1: Wyniki testów

Parametr	Wartość
Opóźnienie transmisji (RTT)	0,774 ms
Prędkość transmisji (1 łącze)	99,98%; 97,89%
Prędkość transmisji (2 łącza)	191,43%



Rysunek 3: Stanowisko testowe

5. TESTY W WARUNKACH RZECZYWISTYCH

Przeprowadziliśmy także testy w warunkach rzeczywistych. W tym przypadku od strony LAN rutera kablem Ethernetowym 1 Gbps podłączony był komputer PC. Do rutera podłączono także dwa modemy USB, zawierające karty SIM dwóch różnych operatorów (Plus i T-Mobile). W scenariuszu przeprowadzone zostały dwie serie testów. Zostały one przeprowadzone w różnych porach dnia, co oznaczało różne obciążenie sieci poszczególnych operatorów. Przy pomocy narzędzia *speedtest.pl* (certyfikowane prze UKE) mierzono opóźnienie oraz średnią prędkość pobierania i wysyłania danych. Podczas każdej serii przeprowadzono po 5 pomiarów kolejno z podłączonym jedynie pierwszym modemem, jedynie drugim modemem oraz z oboma modemami. Wyniki obu serii zaprezentowano w Tabeli 2.

Jak wynika z zaprezentowanych rezultatów, w zależności od serii testowej zastosowanie *multi-homingu* pozwoliło na osiągnięcie średnio od 183% do 204% średniej prędkości pojedynczego łącza w przypadku pobierania danych, oraz od 166% do 159% prędkości pojedynczego łącza w przypadku wysyłania danych.

Tabela 2: Wyniki testów w warunkach rzeczywistych przy użyciu speedtest.pl

Numer testu	Operator	Ping [ms]	Download [Mb/s]		Upload [Mb/s]	
Seria 1						
347518731	Plus Poland	23	46,06		23,23	
347518844	Plus Poland	23	45,64		22,01	
347518900	Plus Poland	27	48,29		20,27	
347518951	Plus Poland	22	45,53		19,31	
347518997	Plus Poland	21	45,58		21,30	
347519645	T-Mobile	25	24,35		29,04	
347519686	T-Mobile	27	25,07		28,04	
347519712	T-Mobile	26	21,08		29,63	
347519727	T-Mobile	17	24,68		28,36	
347519793	T-Mobile	24	24,11		29,91	
Średnio			35,04		25,11	
347519108	Plus Poland + T-Mobile	23	58,67		37,00	
347519242	Plus Poland + T-Mobile	21	66,76		41,32	
347519374	Plus Poland + T-Mobile	28	63,14		41,39	
347519445	Plus Poland + T-Mobile	22	67,45		43,56	
347519558	Plus Poland + T-Mobile	23	65,04		44,62	
Średnio (zysk)			64,21	(183%)	41,58	(166%)
Seria 2						
347520283	T-Mobile	24	37,15		30,76	
347520303	T-Mobile	23	38,70		30,36	
347520318	T-Mobile	24	33,93		29,57	
347520338	T-Mobile	24	34,88		30,46	
347520355	T-Mobile	19	35,84		28,96	
347520522	Plus Poland	20	14,41		9,89	
347520543	Plus Poland	18	13,01		9,80	
347520557	Plus Poland	19	11,45		9,80	
347520568	Plus Poland	25	11,84		9,22	
347520579	Plus Poland	19	12,76		9,14	
Średnio			24,40		19,80	
347520388	Plus Poland + T-Mobile	28	46,77		20,06	
347520409	Plus Poland + T-Mobile	18	51,48		34,54	
347520440	Plus Poland + T-Mobile	18	49,45		34,55	
347520471	Plus Poland + T-Mobile	23	49,69		31,34	
347520498	Plus Poland + T-Mobile	24	51,46		36,63	
Średnio (zysk)			49,77	(204%)	31,42	(159%)

6. PODSUMOWANIE

W artykule przedstawiono implementację rutera klasy domowej umożliwiającego jednoczesny adaptacyjny dostęp do Internetu przy użyciu wielu sieci dostępowych. Urządzenie zbudowane jest się z płytki rozwojowej BPI-W2 oraz autorskiej płytki nakładkowej zapewniającej funkcje sieciowe i pracuje pod kontrolą systemów operacyjnych OpenWRT i Android.

Testy przeprowadzone w warunkach laboratoryjnych wykazały, że po podłączeniu dwóch kablowych łączy WAN możliwe jest osiągniecie przepustowości równej średnio 191,43% prędkości pojedynczego łącza. Jednocześnie dodatkowe opóźnienie wprowadzane przez ruter jest zaniedbywalne.

Z kolei testy przeprowadzone w rzeczywistych warunkach sieci bezprzewodowych LTE przy użyciu certyfikowanego narzędzia do pomiaru prędkości Internetu u użytkowników końcowych wykazały, że zaimplementowane rozwiązanie umożliwia jednoczesny dostęp do Internetu przy użyciu dwóch komórkowych sieci dostępowych i osiągnięcie co najmniej przepustowości 183% pojedynczego łącza w przypadku pobierania i 159% w przypadku wysyłania danych.

PODZIĘKOWANIA

Artykuł powstał w ramach projektu "Innowacyjny ruter o zwiększonej prędkości transmisji danych" o numerze RPMP.01.02.01-12-0612/18 finansowanego w ramach Regionalnego Programu Operacyjnego Województwa Małopolskiego na lata 2014-2020.

LITERATURA

[1] P. Jurkiewicz, M. Rzepka, P. Jaglarz, J. Domżał, and R. Wójcik, “Dynamiczne dostosowywanie wag łączy dostępowych w multi-homingu,” *Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne*, 2023.

[2] “[Wireless Router] Dual WAN introduction and setup - Failover and Load Balance,” https://www.asus.com/support/FAQ/1011719/, accessed: 2023-05-26.

[3] “OpenWRT project,” https://openwrt.org/, accessed: 2023-05-26.

[4] “mwan3 (Multi WAN load balancing/failover),” https://openwrt.org/docs/guide-user/network/wan/multiwan/mwan3, accessed: 2023-05-26.

OCENA PARAMETRÓW QoS USŁUGI TRANSMISJI DANYCH Z WYKORZYSTANIEM MODUŁÓW DTC SOL8SDR DLA ROZWIĄZAŃ DRONOWYCH

QoS PARAMETERS ASSESSMENT FOR DATA TRANSMISSION SERVICE USING DTC SOL8SDR MODULES FOR UAV APPLICATION

Krzysztof Maślanka[1]; Jarosław Wojtuń[2]; Jan M. Kelner[3]

[1] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, krzysztof.maslanka@wat.edu.pl
[2] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, jaroslaw.wojtun@wat.edu.pl
[3] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, jan.kelner@wat.edu.pl

DOI: 10.15199/59.2024.4.51

Streszczenie**: Rynek platform i aplikacyjnych zastosowań bezzałogowych statków powietrznych (BSP) jest obecnie jednym z najbardziej dynamicznie rozwijających się gałęzi przemysłu i usług, zarówno w odniesieniu do segmentu cywilnego jak i wojskowego. W tym artykule przedstawiamy testy jakości usług dla radiowego modułu DTC SOL8SDR wykorzystywanego na BSP do transmisji danych. Moduł ten będzie wykorzystany w systemie rozpoznania radioelektronicznego na grupie BSP.**

Abstract**: The market for platforms and applications of unmanned aerial vehicles (UAVs) is currently one of the most dynamically developing industries and services, both in the civil and military segments. In this paper, we present quality of service (QoS) tests for the SOL8SDR DTC radio module used on a UAV for data transmission. This module is planned to be used in the radio-electronic reconnaissance system of a UAV group.**

Słowa kluczowe**: bezzałogowy statek powietrzny (BSP), jakość usług, moduł DTC SOL8SDR, system.**

Keywords**: quality of service (QoS), reconnaissance system, module DTC SOL8SDR, unmanned aerial vehicles (UAV).**

1. WPROWADZENIE

Ostatnie dziesięciolecie można uznać za okres szybkiego rozwoju platform bezzałogowych, zarówno lądowych, morskich (nawodnych i podwodnych), ale przede wszystkim latających. Bezzałogowe statki powietrzne (BSP) to nie tylko powszechnie dostępne elementy rozrywki wyposażone w kamerę. W tym sektorze najbardziej popularną marką w Polsce jest chińska DJI [1].

Obecnie rynek BSP obejmuje szereg firm oferujących różne platformy i usługi. Najszybciej rozwijające się gałęzie rynku usług BSP obejmują m.in. [2][3]: fotografię i filmowanie, rolnictwo precyzyjne (monitorowanie upraw, precyzyjne opryski), dozór (monitorowanie ruchu, inspekcje infrastruktury krytycznej – mostów, linii energetycznych, infrastruktury telekomunikacyjnej, nadzór budynków i dachów), logistykę oraz dostawy (towarów, medykamentów i materiału biologicznego), geodezję i kartografię (tworzenie map), ochronę środowiska (monitorowanie lasów i akwenów wodnych), bezpieczeństwo i monitorowanie (nadzór i ochrona mienia, operacje ratunkowe i zarządzanie kryzysowe), badania naukowe (archeologia, monitorowanie przyrody, badania atmosferyczne, propagacyjne, etc.).

Prognozy dotyczące rozwoju rynku BSP są bardzo optymistyczne. W 2023 r. wartość światowego rynku BSP i jego zastosowań szacowana była na 44,57 miliardów dolarów (USD). Prognozy wskazują, że w 2030 r. rynek ten wzrośnie w niewyobrażalny sposób, o 1684,9% w stosunku do roku 2023, osiągając kwotę 795,57 miliardów USD przy skumulowanym rocznym wskaźniku wzrostu CAGR (ang. *compound annual growth rate*) równym 50,9% [4]. Ta dynamika wzrostu przyczynia się do poszukiwania nowych zastosowań technologii BSP.

Rynek BSP należy rozpatrywać przede wszystkim w dwóch kategoriach: platform i zastosowań. Z drugiej strony rynek BSP można podzielić na powszechnie dostępne cywilne oraz wojskowe [5]. Rynek militarny, choć znacznie mniejszy (szacowany na 15,88 miliarda USD w 2023 r., prognoza na 2027 r. wynosi 20,64 miliarda USD, co oznacza 6,8% CAGR [6]) od cywilnego, jest specyficzny ze względu na dedykowane platformy i aplikacje, które zazwyczaj nie mają odpowiedników na rynku cywilnym. Z drugiej strony wraz z pojawieniem się sieci 5G i 6G, rozwój potencjału sieci nienaziemnych NTN (ang. *non-terrestrial networks*) [7] pozwoli na zabezpieczenie łączności wojskowej w trudnych warunkach [8].

2. BSP DO ZASTOSOWAŃ ROZPOZNANIA RADIOELEKTRONICZNEGO

Wojny w Górskim Karabachu oraz na Ukrainie uwidoczniły znaczenie BSP we współczesnych konfliktach zbrojnych. Dotychczasowe wojskowe zastosowania BSP obejmują przede wszystkim rozpoznanie obrazowe (z wykorzystaniem kamer multispektralnych), lidarowe lub radarowe (w szczególności SAR (ang. *synthetic aperture radar*)) oraz przenoszenie uzbrojenia (np. tzw. amunicja krążąca). Od kilku lat coraz większym zainteresowaniem cieszą się również BSP realizujące rozpoznania lub zakłócanie radioelektroniczne [9][10][11]. W zakresie rozpoznania radioelektronicznego, BSP może być wykorzystane do monitoringu widma oraz lokalizacji emiterów radiowych [12][13].

Zespół z WAT, w ramach obecnie realizowanego projektu o kryptonimie UAV-COMINT, opracowuje demonstrator systemu rozpoznania radioelektronicznego na grupie BSP [9]. Jednym z istotnych problemów, który należy rozwiązać jest kwestia przetwarzania brzegowego na BSP. Ze względu na ograniczone zasoby widmowe dla łącza radiowego pomiędzy BSP a stanowiskiem dowodzenia i kierowania (operatorem), dąży się do wykonania części obliczeń na platformie mobilnej. Wówczas w kanale radiowym przesyłane są tylko dane niezbędne do budowania świadomości sytuacyjnej na stanowisku dowodzenia i kierowania grupą BSP.

W tym artykule analizujemy możliwości w zakresie transmisji danych wybranego transceivera radiowego DTC SOL8SDR, który stanowi element BSP NEO3 [14] wykorzystywanego w projekcie. Transceiver ten będzie użyty do transmisji danych generowanych przez opracowane sensory radiowe służące do monitoringu widma i lokalizacji dopplerowskiej [9].

BSP realizujące zadania rozpoznania radioelektronicznego muszą transmitować do stanowiska dowodzenia dane informujące o zajętości monitorowanych kanałów (zasobów widmowych) oraz informację nt. wykrytych emiterów radiowych i ich lokalizacji (telemetria). W przypadku realizacji zadań przez pojedynczy BSP, przesyłanie informacji ogólnych o parametrach analizowanych zasobów widmowych czy lokalizowanych obiektów jest zwykle wystarczające. Jednakże, realizacja fuzji danych na potrzeby budowania świadomości sytuacyjnej na podstawie grupy BSP znajdujących się w różnych lokalizacjach, może jednak wymagać transmisji zarejestrowanych próbek IQ sygnałów radiowych z poszczególnych BSP do stanowiska dowodzenia. Rejestracja próbek IQ i transmisja ich do stanowiska dowodzenia może być również niezbędna przy wykonywaniu bardziej zaawansowanej analizy technicznej sygnałów (np. rozpoznanie modulacji i kodowania, demodulacja i dekodowanie sygnału, etc.). Podobna sytuacja może wystąpić w przypadku BSP realizującego lokalizację metodą SDF, gdy wykorzystywany jest dodatkowy BSP do rejestracji sygnałów referencyjnych [12]. Takie podejście może zapewnić zwiększenie dokładności pozycjonowania emiterów.

W Tab. 1 przedstawiono przykładowe przepustowości łącza radiowego niezbędne do realizacji różnych zadań BSP w systemie rozpoznania radioelektronicznego.

Tab. 1. Przepustowość łącza dla różnych zadań BSP

Przeznaczenie BSP	*Wymagana przepustowość łącza radiowego*
Telemetria	< 1 kbit/s
Monitoring widma	~ 1 Mbit/s
Rejestracja próbek IQ sygnału radiowego	> 2 Mbit/s

W tym artykule dokonujemy oceny parametrów jakości usług QoS (ang. *quality of service*) dla transmisji danych z wykorzystaniem modułów DTC SOL8SDR dla rozwiązań dronowych.

3. CHARAKTERYSTYKA MODUŁÓW RADIOWYCH DTC SOL8SDR

Bezprzewodowa łączność radiowa pomiędzy platformą bezzałogową (morską, lądową lub powietrzną) a stanowiskiem dowodzenia i kierowania (operatorem) odgrywa kluczową rolę podczas prowadzenia operacji. Operatorzy mogą przekazywać do platformy polecenia sterujące oraz odbierać od niej określone strumienie danych.

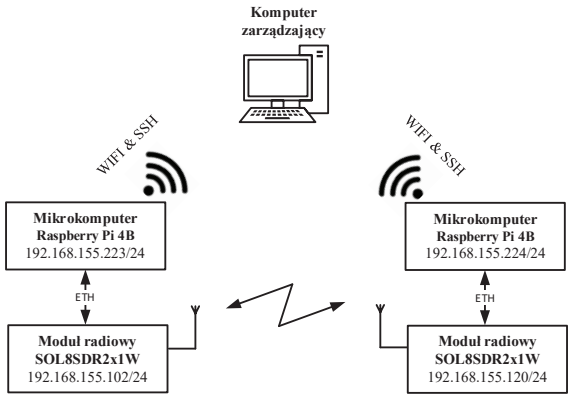
W przypadku platform latających kluczowym aspektem jest uwzględnienie maksymalnej masy startowej. Moduł komunikacyjny musi posiadać relatywnie małe rozmiary i cechować się małą masą. Ważnym czynnikiem wpływającym na czas trwania operacji jest również uwzględnienie wymagań dotyczących zasilania i poboru mocy. Przykłady układów radiowych dedykowanych do rozwiązań dronowych można znaleźć w portfolio firm Domo Tactical Communications (DTC) [15] czy Embention [16]. Na rys. 1 pokazano rozpatrywany w artykule moduł komunikacyjny DTC SOL8SDR2x1W-U.

Rys. 1. DTC SOL8SDR2x1W-U (widok góra, dół) [17]

Moduł SOL8SDR2x1W firmy DTC to zaawansowany transceiver, zaprojektowany specjalnie do integracji w systemach bezzałogowych. Urządzenie, w zależności od wersji może pracować na podzakresach częstotliwości 1,2–1,7 GHz, 1,78–2,29 GHz, 1,98–2,7 GHz, 4,4–5 GHz z pasmem od 1,25–20 MHz. Moduł obsługuje różne rodzaje modulacji, takie jak BPSK, QPSK, 16QAM i 64QAM. Urządzenie oferuje moc wyjściową 2 W (do 1 W dla każdego z kanałów). Producent deklaruje czułość odbiornika na poziomie –110 dBm. Cechuje się wysoką elastycznością dzięki liczbie złącz: Ethernet, USB-C, RS232, oraz niskim zużyciem energii (7 W w trybie IP Mesh). Transmisja może być zabezpieczona z wykorzystaniem szyfrowania AES128/256. Jego kompaktowa budowa (95 mm × 60 mm × 15 mm) i niewielka waga (110 g) sprawiają, że jest idealny do zastosowań w platformach latających [18].

4. SCENARIUSZ BADAWCZY

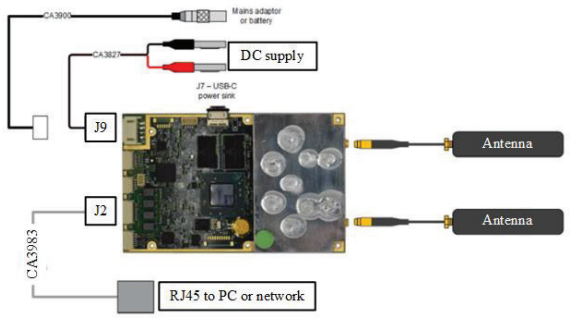
Rysunek 2 przedstawia schemat stanowiska laboratoryjnego. W rozpatrywanym scenariuszu badawczym wykorzystano dwa mikrokomputery Raspberry Pi 4B oraz dwa moduły radiowe SOL8SDR2x1W-U. Dodatkowo wykorzystano komputer zarządzający służący do połączenia się, poprzez sieć WiFi i protokół SSH (ang. *Secure Shell*), z mikrokomputerami Raspberry Pi.



Rys. 2. Schemat stanowiska laboratoryjnego

Zakłada się, że w rozwiązaniu docelowym, jeden z modułów radiowych wraz z mikrokomputerem umieszczony będzie na platformie latającej, drugi zaś umiejscowiony będzie na stanowisku dowodzenia i kierowania.

Mikrokomputer z modułem radiowym połączony jest za pomocą przewodu CA3983 (nomenklatura producenta) – złącze J2 z jednej strony oraz wtyk RJ45 z drugiej. Złącze J2 obsługuje standard Gigabit Ethernet (GbE). Umiejscowienie interfejsu J2 oraz interfejsu zasilania (J9) dla układu SOL8SDR2x1W-U pokazano na rys. 3.



Rys. 3. Rozmieszczenie interfejsów zasilania oraz GbE dla układu SOL8SDR2x1W-U [17]

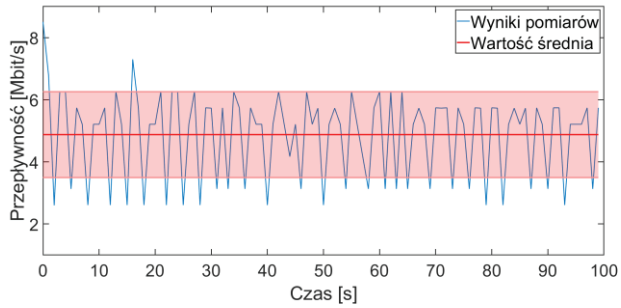
Moduł radiowy SOL8SDR2x1W-U może pracować w jednym z trzech trybów pracy Single Mesh, MIMO Mesh oraz MeshUltra. W badaniach wykorzystano tryb MeshUltra. Tryb ten może obsłużyć aż do 80 węzłów w sieci. Podczas testów urządzenia pracowały na częstotliwości 2225 MHz z pasmem o szerokości 2,5 MHz i mocą nadajnika 32 dBm.

Rysunek 4 ilustruje stanowisko badawcze podczas realizacji testów.

Rys. 4. Zdjęcie stanowiska laboratoryjnego

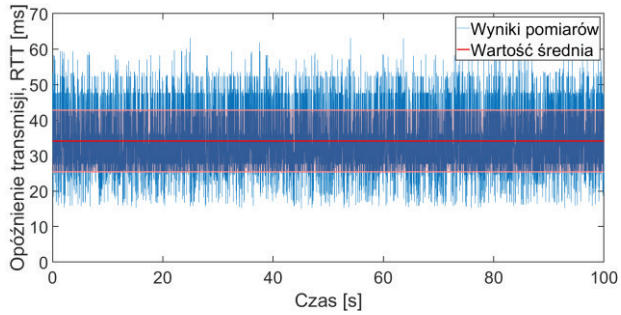
5. WYNIKI BADAŃ

Badania wykonane zostały w warunkach laboratoryjnych z wykorzystaniem środowiska testowego przedstawionego na rys. 4. Na rys. 5 przedstawiono wynik pomiaru przepływności osiąganej na łączu między dwoma punktami odwzorowującymi BSP i stację bazową umożliwiającą jego sterowanie i odbiór danych. Badanie wykonano z zastosowaniem dostępnego w narzędziu iPerf testu przepływności łącza [19]. Jak można zauważyć na rysunku przepływność łącza wykazuje się typowym dla łączy radiowych wahaniem wartości wynikającym z właściwości propagacyjnych oraz interferencji występujących w środowisku transmisyjnym. Średnia wartość przepływności osiągana jest na poziomie 4,88 Mb/s z odchyleniem standardowym wynoszącym 1,38 Mb/s. Wartości te osiągane w warunkach laboratoryjnych muszą zostać zweryfikowane w warunkach rzeczywistych pozwalają jednak na przesyłanie dużych strumieni danych, w tym zarejestrowanych próbek sygnałów radiowych.

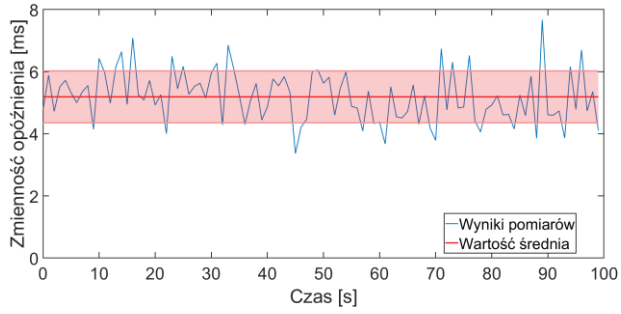


Rys. 5. Szybkość transmisji danych

Na rys. 6 i rys. 7 przedstawiono odpowiednio wykres opóźnienia w obu kierunkach pakietów oraz zmienności opóźnienia emulowanego strumienia danych o przepływności 1 Mb/s. Rozmiar każdego pakietu wynosił 1500 B.



Rys. 6. Opóźnienie transmisji



Rys. 7. Zmienność opóźnienia transmisji

Otrzymane wartości opóźnienia (wartość średnia 34,1 s) oraz jego zmienności (wartość średnia 5,2 ms) pozwalają na założenie możliwości realizacji również usług czasu rzeczywistego (np. transmisję próbek zarejestrowanych sygnałów radiowych) przez łącze zrealizowane z wykorzystaniem badanych modułów, ponieważ spełnione są podstawowe wymagania jakościowe dla tego typu usług (patrz rekomendancje ITU-T G.1010 [20] oraz Y.1541 [21]).

6. PODSUMOWANIE

W artykule przedstawiono cel, scenariusz oraz wyniki badania jakości usług transmisji danych z wykorzystaniem modułu DTC SOL8SDR na potrzeby rozwiązań dronowych. Otrzymane wyniki pokazują, że testowany moduł spełnia wymagania zdefiniowane dla różnych aplikacyjnych zastosowań (patrz Tab. 1) w systemach rozpoznania radioelektronicznego wykorzystującego BSP. Otrzymane wyniki pokazują jednocześnie zgodność z wymaganiami jakościowymi zdefiniowanymi w rekomendacjach ITU-T G.1010 oraz ITU-T Y.1541.

7. PODZIĘKOWANIA

Podziękowanie dla firmy MSP InnTech sp. z o.o. (ul. Piastów Śląskich 49, 01-494 Warszawa, mail: info@uav.com.pl) za wypożyczenie modułów DTC SOL8SDR.

Artykuł powstał w ramach projektu „Dowodzenie i kierowanie grupą bezzałogowych statków powietrznych rozpoznania radioelektronicznego COMINT na bazie nowoczesnych technologii IT" kr. UAV-COMINT, nr DOB-SZAFIR/01/B/029/03/2021, finansowanego przez Narodowe Centrum Badań i Rozwoju oraz grantu nr UGB/22-748/2024/WAT finansowanego przez Wojskową Akademię Techniczną.

LITERATURA

[1] "DJI - Official Website," DJI Official. Accessed: Jun. 01, 2024. [Online]. Available: https://www.dji.com/pl

[2] M. Specht, S. Widźgowski, A. Stateczny, C. Specht, and O. Lewicka, "Comparative analysis of unmanned aerial vehicles used in photogrammetric surveys," *TransNav, International Journal on Marine Navigation and Safety od Sea Transportation*, vol. 17, no. 2, pp. 433–443, Jun. 2023, doi: 10.12716/1001.17.02.21.

[3] H. Shakhatreh *et al.*, "Unmanned aerial vehicles (UAVs): A survey on civil applications and key research challenges," *IEEE Access*, vol. 7, pp. 48572–48634, 2019, doi: 10.1109/ACCESS.2019.2909530.

[4] Research and Markets Ltd., "Global Commercial Unmanned Aerial Vehicle Drones Market by Drone Type (Fixed Wing Drones, Hybrid Drones, Multi Rotor Drones), Application (Agriculture, Audit, Surveillance, Inspection & Monitoring, Consumer Goods & Retail) - Forecast 2023-2030." Accessed: Nov. 13, 2023. [Online]. Available: https://www.researchandmarkets.com/reports/5336496/global-commercial-unmanned-aerial-vehicle-drones

[5] S. K. Chaturvedi, R. Sekhar, S. Banerjee, and H. Kamal, "Comparative review study of military and civilian unmanned aerial vehicles (UAVs)," *INCAS BULLETIN*, vol. 11, no. 3, pp. 183–198, Sep. 2019, doi: 10.13111/2066-8201.2019.11.3.16.

[6] Research and Markets Ltd., "Military Drones Global Market Report 2023 - Research and Markets." Accessed: Nov. 13, 2023. [Online]. Available: https://www.researchandmarkets.com/reports/5735293/military-drones-global-market-report

[7] F. Rinaldi *et al.*, "Non-terrestrial networks in 5G & beyond: A survey," *IEEE Access*, vol. 8, pp. 165178–165200, 2020, doi: 10.1109/ACCESS.2020.3022981.

[8] D. Zmysłowski, P. Skokowski, K. Malon, K. Maślanka, and J. M. Kelner, "Naval use cases of 5G technology," *TransNav Int. J. Mar. Navig. Saf. Sea Transp.*, vol. 17, no. 3, pp. 595–603, Sep. 2023, doi: 10.12716/1001.17.03.11.

[9] K. Bednarz, J. Wojtuń, J. M. Kelner, and K. Różyc, "Frequency Instability Impact of Low-Cost SDRs on Doppler-Based Localization Accuracy," *Sensors*, vol. 24, no. 4, Art. no. 4, Jan. 2024, doi: 10.3390/s24041053.

[10] P. Skokowski and J. Dułowicz, "Autonomous mobile system for detecting and jamming cellular network signals using a software defined radio integrated into a UAV platform," in *2023 Signal Processing Symposium (SPSympo)*, Karpacz, Poland, Sep. 2023, pp. 149–151. doi: 10.23919/SPSympo57300.2023.10302702.

[11] P. Skokowski *et al.*, "Practical trial for low-energy effective jamming on private networks with 5G-NR and NB-IoT radio interfaces," *IEEE Access*, vol. 12, pp. 51523–51535, 2024, doi: 10.1109/ACCESS.2024.3385630.

[12] J. M. Kelner and C. Ziółkowski, "Effectiveness of mobile emitter location by cooperative swarm of unmanned aerial vehicles in various environmental conditions," *Sensors*, vol. 20, no. 9, Art. no. 9, May 2020, doi: 10.3390/s20092575.

[13] R. Szczepanik and J. M. Kelner, "Localization of modulated signal emitters using Doppler-based method implemented on single UAV," in *2023 Communication and Information Technologies (KIT)*, Vysoke Tatry, Slovakia, Oct. 2023, pp. 1–5. doi: 10.1109/KIT59097.2023.10297098.

[14] "NEO3 | MSP." Accessed: Jun. 01, 2024. [Online]. Available: http://uav.com.pl/pl/co-robimy/bsl/neo3

[15] "Domo Tactical Communications (DTC)," DTC. Accessed: Jun. 01, 2024. [Online]. Available: https://www.domotactical.com/

[16] "[Embention," Embention. Accessed: Jun. 01, 2024. [Online]. Available: https://www.embention.com/

[17] "Domo Tactical Communications. SOLO8 SDR for UxV 2x1W PCB OEM Integration Document. Revision 2.2," Domo Tactical Communications.

[18] "Domo Tactical Communications. SDR-UC Single Board UxV Transceiver," DTC. Accessed: Jun. 01, 2024. [Online]. Available: https://www.domotactical.com/products/sol8sdr-u-single-board-uxv-transceiver

[19] D. Zmysłowski and J. M. Kelner, "Mobile network operators' assessment based on drive-test campaign in urban area for iPerf scenario," *Applied Sciences*, vol. 14, no. 3, Art. no. 3, Jan. 2024, doi: 10.3390/app14031268.

[20] "ITU-T Recommendation G.1010 : End-user multimedia QoS categories," International Telecommunication Union (ITU), Geneva, Switzerland, G.1010 (11/01), Nov. 2001. Accessed: Jun. 01, 2024. [Online]. Available: https://www.itu.int/rec/T-REC-G.1010-200111-I

[21] "ITU-T Recommendation Y.1541: Network performance objectives for IP-based services," International Telecommunication Union (ITU), Geneva, Switzerland, Y.1541 (12/11), Dec. 2011. Accessed: Jun. 01, 2024. [Online]. Available: https://www.itu.int/rec/T-REC-Y.1541-201112-I/en

ANALIZA OBIEKTYWNA ALGORYTMÓW NISKIEJ LATENCJI W DASH.JS OPARTA NA QOE

QOE BASED OBJECTIVE ANALYSIS OF LOW-LATENCY ALGORITHMS IN DASH.JS

Syed Uddin[1]; Waqas ur Rahman[2]; Michal Grega[3]; Mikolaj Leszczuk[4]

[1]AGH University of Krakow, Krakow, uddin@agh.edu.pl
[2] Birmingham City University UK, Birmingham, waqas.rahman@bcu.ac.uk
[3]AGH University of Krakow, Krakow, michal.grega@agh.edu.pl
[4]AGH University of Krakow, Krakow, mikolaj.leszczuk@agh.edu.pl

DOI: 10.15199/59.2024.4.52

Streszczenie: **W ostatniej dekadzie znaczący wzrost ruchu wideo przyniósł wiele wyzwań dla dostawców transmisji wideo na żywo, którzy muszą zapewnić spójną jakość doświadczenia (QoE) dla widzów. Strumieniowanie na żywo o niskiej latencji przez HTTP z użyciem dynamicznego adaptacyjnego strumieniowania przez HTTP (LL-DASH) pojawiło się, aby dostarczać treści wideo wysokiej jakości i krótką latencję end-to-end. Zaproponowano wiele algorytmów adaptacyjnych opartych na niskiej latencji, takich jak Learn2Adapt (L2A-LL) i Low on Latency (LOL+), które podejmują decyzje na podstawie heurystycznych modeli predykcyjnych i modeli uczących się. Niniejszy artykuł analizuje, jak skutecznie algorytmy niskiej latencji poprawiają doświadczenie użytkownika. Algorytmy niskiej latencji są porównywane z tradycyjnymi algorytmami adaptacyjnego strumieniowania przez HTTP (HAS). Środowisko DASH.js jest używane do oceny eksperymentalnej. Wyniki eksperymentów pokazują, że algorytm L2A-LL osiąga najwyższą przepływność, a algorytm LOL+ zajmuje drugie miejsce pod tym względem. Algorytmy niskiej latencji osiągają wyższą przepływność wideo w porównaniu do standardowych algorytmów adaptacyjnego strumieniowania. Wyniki pomogą zidentyfikować ograniczenia istniejących algorytmów adaptacyjnych i wskazać strategie ich ulepszania.**

Abstract: **In the past decade, significant growth in video traffic has presented many challenges for live video streaming providers to deliver to ensure cosnistent quality of experience (QoE) for viewers. Low latency live streaming over HTTP using Dynamic Adaptive Streaming over HTTP (LL-DASH) has emerged to deliver high quality video content and short end-to-end latency. Multiple low-latency based adaptive algorithms have been proposed such as Learn2Adapt (L2A-LL) and Low on Latency (LOL+), which makes decisions based on heuristic predictive models and learning models. This paper analyses how effectively low-latency algorithms enhance the user experience. The low-latency algorithms are compared with traditional HTTP Adaptive Streaming (HAS) algorithms. The DASH.js environment is used for experimental evaluation. The experimental results demonstrate that the L2A-LL algorithm achieves the highest bitrate, followed by the LOL+ algorithm. The low-latency algorithm achieve a higher video rate compared to the standard adaptive streaming algorithms. The findings will help identify the limitations of existing adaptive algorithms and inform strategies for their enhancement.**

Slowa kluczowe: **adaptacyjne strumieniowanie przez HTTP, dynamiczne adaptacyjne strumieniowanie przez HTTP, jakość doświadczenia, niska latencja.**

Keywords: **adaptive streaming over HTTP, dynamic adaptive streaming over HTTP, low latency, quality of experience**

1. INTRODUCTION

The demand for video streaming has surged and it is penetrating the emerging markets in the world. There are various streaming services provided over HTTP like HTTP Adaptive Streaming (HAS) such as Dynamic Adaptive Streaming over HTTP (DASH) [1], Apple HTTP Live Streaming (HLS) [2] , Adobe HTTP Dynamic Streaming (HDS) [3] and Microsoft Smooth Streaming [4]. The MPEG-DASH is a popular standard for adaptive streaming deployed by vendors like YouTube, Facebook, Amazon and Netflix. In adaptive streaming, the video sequence is encoded at various bitrates and resolutions, then fragmented into smaller segments. These video representations are stored at a server. The client requests the video segments one by one. The server intially sends the manifest file (e.g., .mpd for DASH) containing information about the video sequence, such as video duration, segment size, quality level and codecs [5].

The ABR algorithm configured on the client-side dynamically selects the optimal video quality for the streaming session, utilizing information provided in the manifest file. The algorithm is based on throughput, buffer and hybrid approaches. The main purpose of adaptation is to ensure smooth streaming, lower stalling and switching events to enhance user experience [6][7]. The adaptation logic downloads the most feasible bitrate for each segment based on the network conditions. There are a variety of encoding parameters which possibly can impact the video quality. These parameters are bitrate, frame rate, CPU usage and Group of Pictures (GOP) [8]. The size of these video segments should be large in volume so it provides ABR algorithm sufficient time to asses network conditions [9][10].

HAS suffers from long delay between when the content is captured by the camera and when the content is playbacked on the screen. The end-to-end delay could be anywhere betwen 30 – 90 seconds which negatively impacts the user experience. In order to minimize the latency, many researchers have focused on developing

HAS algorithms to reduce latency. As mentioned earlier, the purpose of ABR algorithms is to enhance QoE by observing the network conditions and switching quality levels as requested by users [9][10].

Recently, a number of low-latency HAS algorithms were designed with the goal of minimizing the inherent latency in IP video streaming. The goal of this research is to analyze how effectively low-latency algorithms enhance user experience and understand the limitations of the exisitng ABR algorithms. This work helps in understanding the areas in which existing algorithms need improvement and paves the way for the development of algorithms that address these identified shortcomings. To this end, we have compared the low-latency algorithms with the traditional HAS algorithms. We conducted extensive experiments under various network profiles and made several observations regarding the performance of HAS algorithms. We find that no ABR algorithm produces the best QoE for all conditions.

The paper is structured as follows: Section 2 describes Literature Review, Section 3 mentions Methodology, Section 4 includes Discussion and finally Section 5 provides the Conclusion.

2. RELATED WORK

This section provides a review of exisiting work in the domain of video quality in DASH based streaming. There are two subsections, first one emphasie on ABR algorithms and second one is focusing on the QoE factor which impact the video quaity and impact the user experience.

2.1 ABR Algorithms

The ABR algorithms in streaming decide which bitrate and quality segments to download based on the network capacity. There are different ABR algorithms available [11]. There are three main categories of ABR algorithms. These are throughput-based, buffer based and hybrid.
The *throughput* based algorithm is the simplest one and it is working on the throughput estimation principle. This algorithm estimates the current network bandwidth based on the downloaded segments. The *BOLA* is a buffer based algorithm. This algorithm executes on the buffer state principle. The *Dynamic* is hybrid based algorithm. This algorithm works based on both throughput and buffer based approaches. This algorithm utilizes BOLA first, but as the buffer level is low it switches to the throughput algorithm.
There are also ABR algorithms focused on low latency [10]. The *Learn2Adapt* (L2ALL) is low latency algorithm. This algorithm ultilizes convex optimization predicting the best option for future segments deciding on the basis on previously downloaded segments and its latency. The Low on Latency (LOL+) algorithm makes decision based on heuristics predictive model and a learning model [9]. The literature demonstarte that there are some work which compare the traditional ABR alogirtm againist low-latency algorithms. In this paper some state of the art literature is provided for explanation.

2.2 QoE factors

This section describes the factors potentially impact the video quality of experience (QoE). The research work [12] reveals that stalling events and quality switching impacts users QoE. In the research [13] the different versions of stall events are investigated. The multiple short duration stall events against a single long stall event are compared and their impact on video QoE is measured. The rebuffering is also an important factor potentially impacting the overall QoE. There are research work [12] investigates the impact of rebuffering on the video QoE. The results reveal that rebuffering negatively impacts the user QoE. The research is carried out to correlate the multiple quality switching, high amplitude quality switching and stalling [12].These results reveal that these factors irritates the users and thus negatively impacts the QoE. A study [12] is carried to analyze the impact of playout buffer on the QoE in DASH. The factors of buffering delay, playback duration and interruption probability are taken into consideration. The research [15] work considers impact of buffering and interruptions on the video QoE. The results demonstrate that interruptions impacts the user QoE negatively.

3. METHODOLOGY

This section describe the methods of evaluation in detail and also includes the description of the video dataset.

3.1 Dataset

The dataset selected for experimentation, referenced as [16], has been specifically designed for evaluating adaptive streaming methods. The dataset is created based on original source (YUV format). The content is in Full HD format (FHD). The video sequences have 30 segments. Each segment is 10 seconds long duration.

3.2 Encoding Bitrate Ladder

Table 1 shows the sample resolutions and representative bitrates. There are 20 different representations of a video sequences encoded at different bitrates. A sample is shown in the following table. The encoded video streams are migrated to streaming server based on Linux configurations. The DASH.js player is used as the client.

Tab. 1. video bitrate ladder

Index	*Video resolution*	*Bitrate*	*Bitrate*
1	1280x720	1174238	1.2 Mbps
2	1280x720	1431232	1.4 Mbps
3	1920x1080	2070985	2.1 Mbps
4	1920x1080	2384387	2.4 Mbps
5	1920x1080	2884382	2.9 Mbps
6	1920x1080	3245900	3.2 Mbps
7	1920x1080	3493765	3.5 Mbps

4. RESULTS AND ANALYSIS

The testbed is shown in Fig. 1. It is implemented to evaluate the performance of the algorithms. The network emulator, based on the Ubuntu utilities, controls the maximum achieved bandwidth for the client using the Linux Traffic Control system (tc) and the Hierarchical Token Bucket, which is a classful queuing discipline (qdisc). The Big Bunny Buck video (300 seconds long) [17] was pre-encoded and stored on an Apache Web Server and the algorithms are implemented in dash.js version 4.7.4 [18]. The algorithms were evaluated under four bandwidth profiles provided in Table 2.

Figure 1 Experimental Setup

Tab. 2. Network profiles

Index	*Bandwith Profile*
1	Stream the complete session at 1mbps
2	Stream the complete session at 4mbps
3	[1 – 150 s] at 1mbps, [150 – 300 s] at 4mbps
4	[1 – 150 s] at 4mbps, [150 – 300 s] at 1mbps

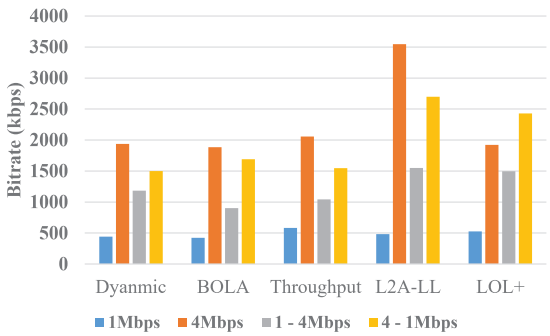


Figure 2 Average bitrates achieved by the algorithms

Fig. 2 depicts the average video rate streamed by the algorithms for each bandwidth profile. Fig. 2 shows that the L2A-LL algorithm achieves the highest bitrate, followed by the LoL+ algorithm. On average, the low-latency algorithms achieve a higher video rate compared to the standard adaptive streaming algorithms. When the bandwidth is set to 1mbps, all algorithms downloaded segments at similar bitrates. However, as the bandwidth increases, the L2A-LL algorithm downloads higher quality segments as compared to other algorithms.

Fig. 3 shows that the standard adaptive streaming algorithms experience higher video rate switches compared to the low-latency adaptive streaming. The L2A-LL algorithm experiences the lowest number of switches followed by the LoL+. However, the L2A-LL algorithm also experiences the highest number of playback interruptions. The challenge for the adaptive streaming algorithms is to select the highest feasible bitrate for the upcoming segment while minimizng the risk for playback interruption [19]. The L2A-LL is able to stream the video at higher quality; however, at the expense of playback interruption, which degrades the QoE. The dynamic algorithm experiences relatively smaller number of playback interruptions; however, it compromises on the quality of the downloaded segments. Similarly, the throughput algorithm downloaded lower quality segments and experienced a higher number of video rate switches compared to L2A-LL. However, it also experienced a lower number of interruptions compared to L2A-LL.

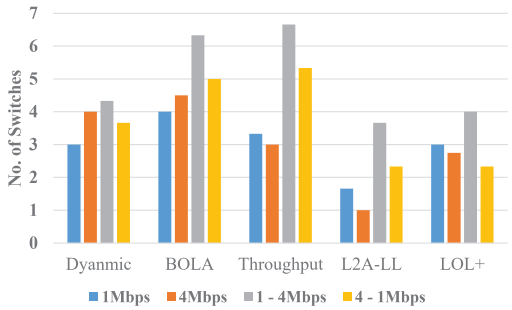


Figure 3 Average video rate switches experienced by the clients

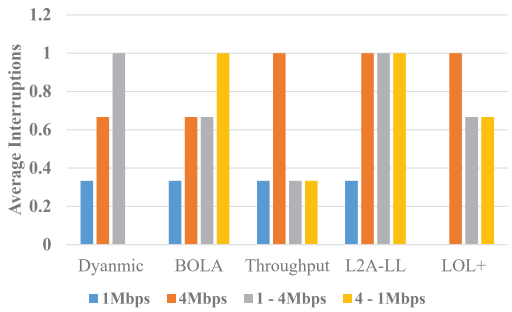


Figure 4 Average interruptions experienced by the clients

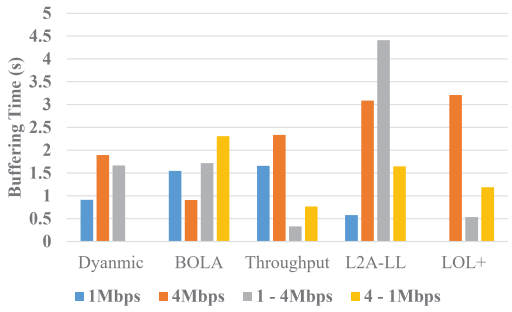


Figure 5 Duration of playback interruptions

The results reveal that no single algorithm satisfies all aspects of video quality factors. The low-latency algorithm downloads high-quality segments but cannot prevent the buffer from draining. In contrast, the standard adaptive streaming algorithms experience a lower number of playback interruptions while compromising on the video quality as shown in Fig.4 and Fig 5. respectively.

5. CONCLUSION

This research analyzes how effectively low-latency algorithms enhance the user experience. The research work compares the low-latency algorithms with traditional HTTP Adaptive Streaming (HAS) algorithms. The results reveal that low-latency algorithms achieved a higher video rate as compared to the standard HAS algorithms. There is another interesting observation when bandwidth is lowered all algorithms downloaded the segments at a similiar rate. There is a high amplitude of switching experienced in standard algorithms as compared to low-latency algorithms. The dynamic algorithm have a smaller number of interruptions but it compromised on the quality of downloaded segments. Overall, we find that no ABR algorithm produces the best QoE for all conditions.

In this study, we evaluate the performance of low-latency networks against traditional ABR algorithms by analyzing key video quality metrics, including (1) video quality, (2) rebuffering events, and (3) video quality fluctuations. In future, we plan to analyze the tradeoff between latency, bitrate effectiveness, and interrutption level to understand the gain from LL algorithms. Additionally, we plan to evaluate the QoE results when delivering legacy MPEG-DASH live content compared to chunked CMAF-based content.

REFERENCES

[1] K. Khan and W. Goodridge, "Dynamic Adaptive Streaming over HTTP (DASH) within P2P systems: a survey," *International Journal of Advanced Networking and Applications*, vol. 11, pp. 4155–4161, Apr. 2019, doi: 10.35444/IJANA.2019.11014.

[2] Apple Inc "Apple DASH ." https://developer.apple.com/streaming/

[3] Adobe Inc "ADOBE." https://helpx.adobe.com/adobe-media-server/dev/configure-dynamic-streaming-live-streaming.html

[4] Microsoft "Miscrosoft Smooth Streaming ". https://learn.microsoft.com/en-us/shows/iis-net-site-videos/smooth-streaming

[5] J. Vlaovic, S. Rimac-Drlje, D. Žagar, and L. Filipović, "Content dependent spatial resolution selection for MPEG DASH segmentation," *J Ind Inf Integr*, vol. 24, p. 100240, Sep. 2021, doi: 10.1016/j.jii.2021.100240.

[6] W. ur Rahman and K. Chung, ''Buffer-based adaptive bitrate algorithm for streaming over HTTP,'' KSII Trans. Internet Inf. Syst., vol. 9, no. 11, pp. 4585–4622, Nov. 2015, doi: 10.3837/tiis.2015.11.019 .

[7] W. Alqhtani, A. A., and M. S., "An Adaptive Quality Switch-aware Framework for Optimal Bitrate Video Streaming Delivery," *International Journal of Advanced Computer Science and Applications*, vol. Nov. 2020, doi: 10.14569/IJACSA.2020.0110871.

[8] M.Grega. Uddin Syed and M. Leszczuk, "Preliminary Study on Video Codec Optimization Using VMAF," in *Intelligent Information and Database Systems*, T. K. and T. U. and H. T.-P. and T. B. and S. E. Nguyen Ngoc Thanh and Tran, Ed., Cham: Springer International Publishing, 2022, pp. 469–480.

[9] T. Lyko, M. Broadbent, N. Race, M. Nilsson, P. Farrow, and S. Appleby, "Improving quality of experience in adaptive low latency live streaming," *Multimed Tools Appl*, Nov. 2023, doi: 10.1007/s11042-023-15895-9.

[10] P. O'Hanlon and A. Aslam, "Latency Target Based Analysis of the DASH.Js Player," in *Proceedings of the 14th Conference on ACM Multimedia Systems*, in MMSys '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 153–160. doi: 10.1145/3587819.3590971.

[11] S. S. Tadahal *et al.*, "A Survey On Adaptive Bitrate Algorithms and Their Improvisations," in *2021 International Conference on Intelligent Technologies (CONIT)*, 2021, pp. 1–7. doi: 10.1109/CONIT51480.2021.9498318.

[12] B. Taraghi, S. Haack, and C. Timmerer, "Towards Better Quality of Experience in\\HTTP Adaptive Streaming." Sep. 2023. doi: 10.48550/arXiv.2301.13523.

[13] B. Taraghi, M. Nguyen, H. Amirpour, and C. Timmerer, "Intense: In-Depth Studies on Stall Events and Quality Switches and Their Impact on the Quality of Experience in HTTP Adaptive Streaming," *IEEE Access*, vol. 9, pp. 118087–118098, 2021, doi: 10.1109/ACCESS.2021.3107619.

[14] G. Xie, H. Chen, F. Yu, and L. Xie, "Impact of playout buffer dynamics on the QoE of wireless adaptive HTTP progressive video," *ETRI Journal*, vol. 43, Apr. 2021

[15] J. Allard, A. Roskuski, and M. Claypool, 0"Measuring and modeling the impact of buffering and interrupts on streaming video quality of experience," in *Proceedings of the 18th International Conference on Advances in Mobile Computing & Multimedia*, in MoMM '20. New York, NY, USA: Association for Computing Machinery, 2021, pp. 153–160.

[16] S. Lederer, C. Müller, and C. Timmerer, "Dynamic adaptive streaming over HTTP dataset," in *Proceedings of the 3rd Multimedia Systems Conference*, in MMSys '12. New York, NY, USA: Association for Computing Machinery, 2012, pp. 89–94.

[17] T. Roosendaal, "Big Buck Bunny," in *ACM SIGGRAPH ASIA 2008 Computer Animation Festival*, in SIGGRAPH Asia '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 62.

[18] dash.js, "dash.js JAVAScript Reference Client "

[19] W. ur Rahman and E.-N. Huh, "Content-aware QoE optimization in MEC-assisted Mobile video streaming," *Multimedia Tools Appl.*, vol. 82, no. 27, pp. 42053–42085, Apr. 2023

Subiektywne postrzeganie pojedynczych zniekształceń w telewizji opartej o strumieniowanie adaptacyjne

THE SUBJECTIVE PERCEPTION OF SINGLE DISTORTIONS IN ADAPTIVE STREAMING-BASED TELEVISION

Paweł Andruloniw[1]; Piotr Zwierzykowski[2]; Karol Kowalik[3];

[1] Instytut Sieci Teleinformatycznych, Wydział Informatyki i Telekomunikacji, Politechnika Poznańska, INEA Sp. z o.o. Poznań, pawel.andruloniw@doctorate.put.poznan.pl
[2] Instytut Sieci Teleinformatycznych, Wydział Informatyki i Telekomunikacji, Politechnika Poznańska, Poznań, piotr.zwierzykowski@put.poznan.pl
[3] Fiberhost S.A., Poznań, karol.kowalik@fiberhost.com

DOI: 10.15199/59.2024.4.53

Streszczenie: **Rosnąca popularność treści opartych o strumieniowanie adaptacyjne oraz możliwości wdrożenia ich do systemów telewizyjnych wymagają, aby operatorzy podejmowali działania w obrębie analizy i polepszenia jakości usług dla swoich klientów. Niniejszy referat opisuje podejście do subiektywnej oceny jakości wideo w systemie TV jednego z polskich dostawców usług. Opisano, w jaki sposób zaimplementowano badanie oraz przedstawiono analizę ocen użytkowników dla pojedynczych błędów kodowania treści w systemie opartym o strumieniowanie adaptacyjne.**

Abstract: **The growing popularity of content based on adaptive streaming and the possibilities of implementing it in TV systems require operators to take action within the framework of analyzing and improving the quality of service for their customers. This paper describes an approach to the subjective assessment of video quality in the TV system of one Polish service provider. It describes how the study was implemented and presents an analysis of user ratings for individual content coding errors in a system based on adaptive streaming.**

Słowa kluczowe: **jakość doświadczeń, ocena subiektywna, over-the-top, strumieniowanie adaptacyjne**

Keywords: **adaptive streaming, over-the-top, subjective assessment, quality of experience,**

1. WSTĘP

Szacuje się, że w 2029 roku rynek usług OTT (ang. over-the-top) w Europie Zachodniej osiągnie przychody sięgające 48 miliardów dolarów, dając wzrost o około 55% w odniesieniu do 2023 roku. Dla polskiego rynku usług OTT oczekiwany wzrost to 700 milionów dolarów, co jest drugim wynikiem w regionie Europy Wschodniej. [1]. Migracja użytkowników analogowych usług telewizyjnych wymusza na dostawcach usług zwiększanie poziomu świadczonych usług, aby zachęcić klientów do swoich platform TV. Operatorzy muszą dostosować systemy telewizyjne, aby zapewniały klientom odpowiedni poziom strumieni audio i wideo oraz funkcji dodatkowych, takich jak treść na życzenie, odtwarzanie od początku, czy odtwarzanie do kilku dni wstecz. Usługa telewizyjna oparta o strumieniowanie adaptacyjne bardzo często dostarczana jest do klientów razem z pakietem Internetu. Aby dostarczyć wideo, wykorzystywane są między innymi standardy MPEG DASH (ang. dynamic adaptive streaming over http) [5], Apple HLS (ang. http live streaming) [8], czy też Microsoft Smooth Streaming [7]. Wspólną cechą tych standardów jest protokół http wykorzystywany do transmisji danych do klienta końcowego. Wykorzystanie powyższych standardów pozwala na dostarczenie zróżnicowanych funkcji wideo oraz optymalne wykorzystanie sieci domowej. Dzięki temu nie powinno być przerw w transmisji, a jedynie spadki jakości przy jednoczesnych spadkach wydajności sieci. Wady strumieniowania adaptacyjnego mogą obejmować przede wszystkim złożoność techniczną, opóźnienia w dostarczaniu treści, zwiększone koszty infrastruktury, a także zarządzanie zwiększoną liczbą strumieni, obejmujące kodowanie i pakowanie danych. Opóźnienia charakterystyczne dla transmisji adaptacyjnej powiązane z kodowaniem, pakowaniem, transmisją, czy też buforowaniem mogą zostać zniwelowane poprzez wdrożenie rozwiązań niskiej latencji, np. LL-HLS (ang. low latency HLS) [8], czy też LL-DASH (ang. low latency DASH) [5]. W rozdziale drugim opisano podejście do subiektywnej oceny treści wideo. W trzecim rozdziale zawarto opis badania. Rozdział czwarty i piąty opisuje odpowiednio uzyskane wyniki oraz wnioski i plan na dalsze prace.

2. SUBIEKTYWNA OCENA JAKOŚCI

Postrzeganie usługi TV przez klientów zależy także od jakości dostarczanych strumieni wideo. Wpływają na nią między innymi przepływność bitowa oraz jej zmienność w trakcie odtwarzania, zastosowane kodeki wideo, opóźnienie, czy też buforowanie przez odtwarzacz. Nie bez znaczenia pozostają także błędy transkodowania lub błędy po stronie odtwarzacza związane z prawidłowym złożeniem kawałków (ang. chunks) strumienia DASH. Niektóre problemy mogą zostać powiązane z metrykami QoS (ang. quality of service), takimi jak: przepływność, opóźnienie, zmienność opóźnienia, czy utrata pakietów, które obiektywnie określają kondycję infrastruktury. W przypadku odczuć subiektywnych należy analizować pomiary związane z QoE (ang. quality of experience), które analizują sposób postrzegania treści przez użytkownika. Uzyskując od użytkowników ocenę MOS (ang. mean opinion score), można stwierdzić jak bardzo prze-

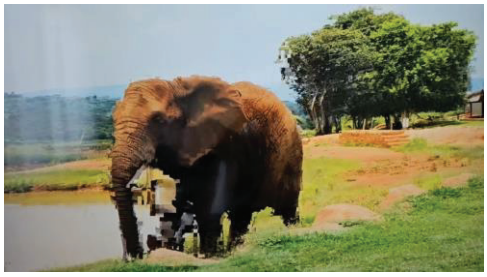
szkadzające są zniekształcenia, określony poziom przepływności bitowej, czy też buforowanie dla danego systemu telewizyjnego [6]. W prawidłowym strumieniu takie błędy nie powinny się zdarzyć. Ponadto są one praktycznie niemożliwe do wykrycia bez posiadania wyspecjalizowanego oprogramowania analizującego obraz z danego strumienia. W takim przypadku informacja o błędach może pochodzić od użytkownika. Aby ocenić poziom postrzeganej subiektywnej jakości wideo dla treści z trudno wykrywalnymi zniekształceniami, przeprowadzono badanie wykorzystując metodę DSCQS (ang. double stimulus continuous quality scale) [4]. Celem badania było uzyskanie subiektywnej opinii użytkownika w skali od 1 do 100 oraz sprawdzenie, w jakim stopniu zniekształcenia tego typu mogą przeszkadzać klientom produkcyjnego systemu telewizyjnego.

3. OPIS BADANIA

Badanie DSCQS do gromadzenia subiektywnej percepcji wideo to jedna z metod zaproponowanych przez Międzynarodowy Związek Telekomunikacyjny ITU (ang. International Telecommunication Union) w rekomendacji BT.500-15 [4]. Jest to badanie w formule podwójnej stymulacji opierające się na porównaniu obrazu referencyjnego i zniekształcenia. Oceniający ogląda parę obrazów, a następnie dokonuje oceny w skali od 1 do 100. Użytkownik może wybrać dowolną wartość całkowitą na skali ocen, która logicznie jest podzielona na 5 równych części, reprezentujących 5-stopniową skalę ACR (ang. absolute category rating) [4]. Korzystanie bezpośrednio ze skali ocen ACR operujących tylko na liczbach całkowitych, czyli od 1 do 5 powoduje spłaszczenie ocen dla małych zniekształceń oraz uwypuklenie tych bardziej znaczących. Przy skali od 1 do 100 i konwersji na ACR użytkownik może dokonać oceny zmiennoprzecinkowej odwzorowującej nieco lepiej jego nastrój i subiektywne postrzeganie danej treści. Badanie DSCQS ma dwa warianty. Pierwszy z nich jest przeznaczony dla indywidualnych użytkowników, którzy mogą dowolnie przeglądać referencje oraz zniekształcone filmy. W drugim wariancie oceny dokonuje grupa użytkowników, która ogląda sekwencję badawczą na jednym ekranie. Telewizja często oglądana jest przez wielu użytkowników jednocześnie na jednym ekranie, z tego względu zdecydowano się na drugi wariant. Dodatkową zaletą jest możliwość zebrania większej liczby ocen w określonym czasie. Badanie DSCQS zostało zaimplementowane jako treść na życzenie w katalogu VOD operatora. Użytkownik może wykonać badanie uruchamiając film z badaniem za pomocą dekodera STB podłączonego do telewizora. W badaniu wykorzystano sekwencje z obrazami kodowanymi kodekiem H.264 profil High@L4.2 [3]. Cała sekwencja trwa około 23 minuty i składa się z planszy powitalnej z instrukcją, kodu QR, za pomocą którego można uzyskać dostęp do formularza w formie online, sekwencji testowej, która obejmuje 4 materiały próbne po 30 sekund każdy oraz sekwencji właściwej składającej się z 15 par filmów. Każdy film podlegający ocenie trwa 30 sekund. Przed obejrzeniem filmu oraz po wyświetleniu pary pojawiają się plansze informacyjne mówiące użytkownikowi o kolejnych krokach badania. Czas trwania poszczególnych materiałów podlegających ocenie został dobrany w taki sposób, aby był zgodny z próbkowaniem treści na żywo. Użytkownik po zeskanowaniu kodu QR urządzeniem mobilnym uzyskuje dostęp do formularza, w którym może ocenić każdą parę, tj. film A i B w skali od 1 do 100. Formularz obejmuje 15 par filmów, a na końcu wyświetlona zostaje ankieta, która zawiera pytania związane z rozmiarem ekranu, czyli telewizora wykorzystanego w badaniu, wiekiem oraz wadą i stosowaną korekcją wzroku.

4. UZYSKANE WYNIKI

Wyniki były zbierane w dualny sposób. Pierwszy z nich obejmował badania wykonane przez użytkowników w domu, przed własnym telewizorem. Drugim sposobem badania było zebranie grupy osób i wykonanie go w kontrolowanych warunkach. Oba te sposoby są dozwolone w wariancie drugim metody DSCQS. Ze względu na to, że badanie obejmuje system telewizyjny, to wykorzystywanym ekranem były telewizory powyżej 30 cali. Wypełniający formularz określał, jaki rozmiar ma telewizor w obrębie trzech zakresów: 31-42 cale, od 43 do 55 cali lub powyżej 56 cali. Eksperyment wykonały 64 osoby, 2 wyniki zostały odrzucone ze względu na nierzetelne wykonanie badania. Pierwszy odrzucony wynik zawierał tylko oceny o wartości 50, czyli wartości domyślne dla każdej pozycji na formularzu ocen. Sugeruje to, że użytkownik wypełnił formularz bez zainteresowania faktycznym celem badania, a być może z ciekawości. Drugi wynik zawierał oceny bliskie wartości 1 niezależnie od jakości wideo. Można by stwierdzić, że oceniający użytkownik osądza jakość dostępnych treści bardzo nisko i ma do tego prawo. Jednak wypełnienie formularza nie pokrywało się z obejrzeniem całości sekwencji badawczej, co zostało sprawdzone na poziomie logów systemu. Ostatecznie stwierdzono, że użytkownik ten wypełnił formularz bez obejrzenia sekwencji testowej w całości. Po odfiltrowaniu danych odstających uzyskano 31 wyników dla ekranów w zakresie od 31 do 42 cali, 26 dla telewizorów o rozmiarze od 43 do 55 cali i 5 wyników dla ekranów powyżej 56 cali. Sekwencja testowa, objęta analizą zawierała 15 par filmów, sumarycznie dając 1860 ocen. Część treści wideo o przepływności 6, 8 oraz 10 Mbit/s zostało zaprezentowanych z pojedynczymi zniekształceniami w obrazie, czyli jeden błąd na ocenianą treść trwającą 30 sekund. Na rysunku 1 został zaprezentowany przykładowy błąd w obrazie. Jest on często przez oglądających oceniany jako pikselizacja, trwająca maksymalnie sekundę. W tabeli 1 zaprezentowano uśrednione wyniki dla wybranych poziomów strumieni. Dla strumieni o przepływności 6 i 8 Mbit/s różnica średniej oceny dla obrazów ze zniekształceniami oraz bez nich wynosi odpowiednio 2,15 i 1,63 punktu co daje wartości 0,11 i 0,08 w pięciopunktowej skali ACR. Z kolei dla strumienia 10 Mbit/s różnica wynosiła 9,55 punktu, dając w skali ACR wartość 0,48, czyli blisko połowę punktu. Zarówno dla strumieni 8 i 10 Mbit/s większa średnia została uzyskana dla strumieni bez zniekształceń. Dla strumienia 6 Mbit/s obraz ze zniekształceniami uzyskał nieznacznie wyższy wynik. Mediana ocen dla 6 i 8 Mbit/s różniła się o 2 punkty z taką samą tendencją ocen jak średnia. Dla strumienia 10 Mbit/s obraz bez zniekształceń uzyskał medianę wyższą o 8 punktów.



Rys. 1. Przykład błędów w ocenianych treściach. Opracowanie własne, na podstawie [2]

Tab. 1. Średnia ocena i mediana dla wybranych poziomów strumieni

Mbit/s	*Zniekształcenia*	*Średnia*	*Mediana*
6	NIE	69,62	72,00
6	TAK	71,77	74,00
8	NIE	67,61	72,00
8	TAK	65,98	70,00
10	NIE	81,22	83,00
10	TAK	71,67	75,00

W tabeli 2 przedstawiono średnie wyniki ocen w odniesieniu do rozmiaru ekranu. Dla strumienia 6 Mbit/s wyniki oscylują w okolicach 70 punktów. Dla pozycji ze zniekształceniami i ekranów o rozmiarze powyżej 56 cali średni wynik wyniósł 78,30 punktów, różniąc się o około 8 punktów w stosunku do innych rozmiarów. Wynik ten może wynikać z niskiej liczebności grupy o takich ekranach. Dla strumienia 8 Mbit/s średnie wyniki wahały się od 64,12 dla najmniejszego ekranu i treści ze zniekształceniami do 68,15 dla obrazu bez zniekształceń i tego samego typu ekranu. Największe różnice widoczne są dla najwyższej przepływności bitowej. Obraz ze zniekształceniami uzyskał 69,21 punktów dla najmniejszego ekranu, 4,12 punktu więcej dla ekranów od 43 do 55 cali i 9,24 punktu więcej dla największych ekranów.

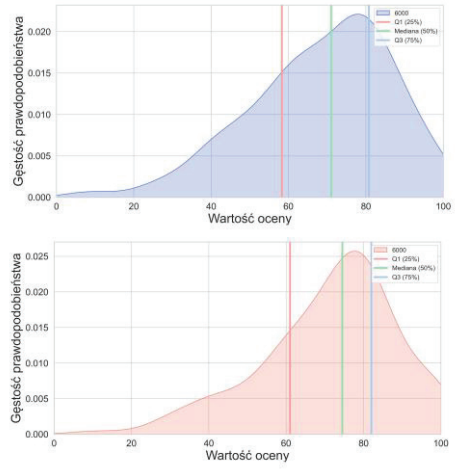
Tab. 2. Średnia ocena dla wybranych poziomów strumieni oraz rozmiaru ekranu telewizora

Mbit/s	*Zniekształcenia*	*Średnia 31-42"*	*Średnia 43-55"*	*Średnia powyżej 56"*
6	NIE	70,19	68,72	70,60
6	TAK	71,66	70,63	78,30
8	NIE	68,15	67,14	65,90
8	TAK	64,12	67,84	67,88
10	NIE	81,72	80,64	81,20
10	TAK	69,21	73,33	78,45

Uśredniona mediana ocen dla wskazanych rozmiarów ekranów została zaprezentowana w tabeli 3. Tabela zawiera średnią medianę dla zbiorów obrazów wideo o tej samej przepływności. Dla strumienia 6 Mbit/s najmniejsza uśredniona mediana dotyczyła ekranów w rozmiarze od 43 do 55 cali. Wyniosła odpowiednio 68,83 i 72,75 punktu dla obrazów bez błędów i razem z nimi. Mediana regionalna dla tych obrazów osiągała wartości od 66 do 71.5 punktu dla obrazu bez zniekształceń i od 69 do 77.5 punktu dla obrazów ze zniekształceniami. W tabeli widać, że dla strumienia 10 Mbit/s bez zniekształceń mediana ocen osiągnęła taki sam wynik dla każdego typu ekranu. Natomiast pomiędzy najmniejszym i największym ekranem i obrazem ze zniekształceniami jest 10 punktów różnicy na korzyść największych ekranów. Mediana regionalna dla strumienia 10 Mbit/s wynosiła od 77 do 85 punktów dla obrazów zniekształconych i równe 83 punkty dla obrazów bez zniekształceń. Taka rozpiętość wyników dla wszystkich strumieni sugeruje, że pojedyncze problemy w obrazie, które wystąpią raz na jakiś czas nie mają dużej wagi w stosunku do przepływności strumienia. Strumień 6 Mbit/s charakteryzuje się lepszymi ocenami dla zniekształconych obrazów. Może to być powiązane z tym, że obraz zniekształcony i prawidłowy nie zawsze był wyświetlany jeden po drugim, przez co odniesienie się do obrazu bez zniekształceń było trudniejsze w ocenie. Ponadto, pojedynczy błąd w strumieniu może zmniejszyć swoją istotność gdy dalsza treść nie wykazuje błędów. Elementy odstające w ramach zbioru ocen zostały odfiltrowane metodą kwartylową IQR (ang. ang. interquartile range).

Tab. 3. Mediana ocen dla wybranych poziomów strumieni oraz rozmiaru ekranu telewizora

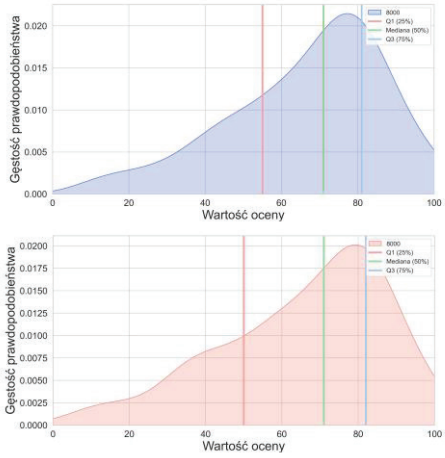
Mbit/s	*Zniekształcenia*	*Uśredniona mediana 31-42"*	*Uśredniona mediana 43-55"*	*Uśredniona mediana powyżej 56*
6	NIE	72,67	68,83	70,33
6	TAK	76,00	72,75	78,75
8	NIE	71,50	71,25	72,50
8	TAK	68,50	70,50	75,40
10	NIE	83,00	83,00	83,00
10	TAK	71,33	75,50	81,33



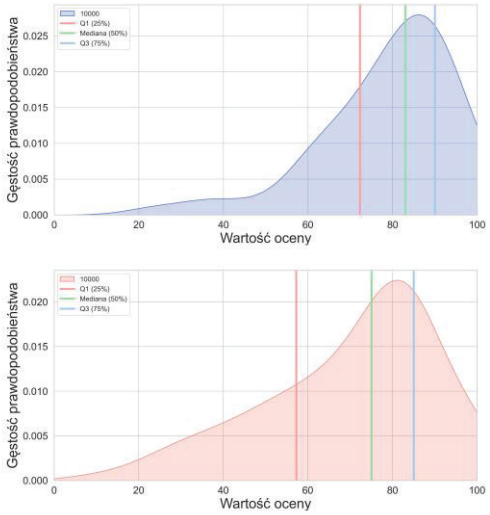
Rys. 2. Rozkład prawdopodobieństwa uzyskanych ocen dla strumienia o przepływności 6 Mbit/s.

Rozkład prawdopodobieństwa dla podanych strumieni został przedstawiony na rysunkach 2, 3 i 4, reprezentując

odpowiednio przepływności 6, 8 i 10 Mbit/s. Dla wszystkich strumieni rozkłady charakteryzują się asymetrią lewostronną. Większość ocen użytkowników skupia się na wyższych wartościach, w okolicach 61-80 punktów, co według skali ACR jest jakością dobrą.



Rys. 3. Rozkład prawdopodobieństwa uzyskanych ocen dla strumienia o przepływności 8 Mbit/s



Rys. 4. Rozkład prawdopodobieństwa uzyskanych ocen dla strumienia o przepływności 10 Mbit/s

Użytkownicy rzadko kiedy udzielają niskich i bardzo niskich ocen, podobnie jak maksymalnych, przymiotnikowo określonych odpowiednio jako jakość zła i wspaniała. Mediana wyników jest wyższa niż średnia dla każdego analizowanego strumienia. Sugeruje to wartości odstające w okolicach niższych wartości. Ze względu na subiektywność ocen, takie wyniki także powinny zostać wzięte pod uwagę dla przypadków, gdy badanie zostało wykonane poprawnie.

5. WNIOSKI I DALSZE PRACE

W przypadku systemów telewizyjnych opartych o strumieniowanie adaptacyjne pojedyncze zniekształcenia nie muszą znacząco wpływać na subiektywne postrzeganie jakości obrazu. Oceny nie różnią się znacząco dla różnych rozmiarów ekranu, na którym jest konsumowana treść. Najważniejsza jest w tym przypadku jakość powiązana z przepływnością wyświetlanego strumienia. Obserwując znaczącą liczbę ocen w okolicach 61-80 punktów, czyli jakość dobrą według skali ACR można sądzić, że użytkownicy oglądając kanały liniowe w telewizji doceniają to, że dany obraz jest dostępny i jest on w przyzwoitej jakości. Obserwatorzy nie mając solidnego powodu nie udzielają niskich ocen, podobnie niechętnie przyznają oceny maksymalne, przypuszczalnie licząc, że kolejny obraz będzie jeszcze lepszy niż poprzedni. Dalsze prace dla tego systemu transmisyjnego będą obejmować analizę niższych przepływności oraz zależności pomiędzy wiekiem oceniających a udzielanymi ocenami. Analizie zostanie także poddany fakt obecności wady wzroku i jej korekcji na postrzeganie jakości wideo w danym systemie telewizyjnym OTT.

6. PODZIĘKOWANIA

Autorzy dziękują Ministerstwu Edukacji i Nauki za wsparcie finansowe (program Doktorat Wdrożeniowy, nr DWD/4/24/2020). Publikacja została sfinansowana ze środków Ministerstwa Nauki i Szkolnictwa Wyższego (Projekt 0313/SBAD/1312).

LITERATURA

[1] Digital TV Research. 2024. “Digital TV Research’s April 2024 newsletter”. https://digitaltvresearch.com/digital-tv-researchs-april-2024-newsletter/. [data dostępu: 02.06.2024]
[2] Ehlers Magda. “Elephants and Their Habitat”. https://www.pexels.com/video/elephants-and-their-habitat-3842814/. [data dostępu: 01.12.2023]
[3] International Telecommunication Union – Telecommunication Standardization Sector. 2021. “Recommendation H.264 - Advanced video coding for generic audiovisual services”
[4] International Telecommunication Union – Radiocommunication Sector. 2023. “Recommendation ITU-R BT.500-15 (05/2023) Methodologies for the subjective assessment of the quality of television images”
[5] ISO/IEC 23009-1. 2022. “Information technology — Dynamic adaptive streaming over HTTP (DASH) — Part 1: Media presentation description and segment formats”
[6] Kowalik, Karol, Paweł Andruloniw, Bartosz Partyka, and Piotr Zwierzykowski. 2022. “Telecom Operator’s Approach to QoE”. Journal of Telecommunications and Information Technology, no. 2 (June):26-34.
[7] Microsoft. 2020. “[MS-SSTR]: Smooth Streaming Protocol”
[8] Pantos R.. Apple Inc. 2017. “RFC 8216 HTTP Live Streaming”

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

DOI: 10.15199/59.2024.4.54

REALIZACJA PRÓBNIKA SYSTEMU REJESTROWANIA I MONITOROWANIA PARAMETRÓW JAKOŚCIOWYCH STOSUJĄCEGO STANDARD IPFIX

IMPLEMENTATION OF A SAMPLE OF SYSTEM FOR RECORDING AND MONITORING QUALITY PARAMETERS USING THE IPFIX STANDARD

Sylwester Kaczmarek[1], Krzysztof Zalewski[2]

[1] Politechnika Gdańska Wydział ETI, Gdańsk, kasyl@eti.pg.edu.pl
[2] Adtran, Gdynia, krzysztof.zalewski@adtran.com

Streszczenie: **Nadzór parametrów jakościowych w sieci z protokołem IP jest jedną z ważniejszych funkcji jaką musi dysponować Operator sieci telekomunikacyjnej. W pracy przedstawiono koncepcję systemu rejestracji i monitorowania przeznaczonego do realizacji tego celu. Jest to rozwiązanie pasywne stosujące rozszerzenie standardu IPFIX, które określono mianem SIPFIX. Opisano praktyczną realizację i wyniki testów Próbnika, jednego z podstawowych elementów składowych tego systemu. Wyniki potwierdziły możliwość realizacji zaproponowanego systemu.**

Abstract**: Supervision of quality parameters in a network with the IP protocol is one of the most important functions that a telecommunications network operator must have. The paper presents the concept of a recording and monitoring system intended to achieve this goal. This is a passive solution that uses an extension of the IPFIX standard, called SIPFIX. The practical implementation and test results of the Sampler, one of the basic components of this system, are described. The results confirmed the feasibility of implementing the proposed system.**

Słowa kluczowe: **IPFIX, monitorowanie, Próbnik, QoS, YAF**

Keywords**: IPFIX, monitoring, Sampler, QoS, YAF**

1. WPROWADZENIE

Parametry jakościowe są metrykami, które są wykorzystywane do określania jakości usług zapewnianej przez sieć operatora telekomunikacyjnego. Jakość usług, czyli QoS (Quality of Service), to według definicji ITU-T pełny zbiór charakterystyk usługi telekomunikacyjnej, które decydują o jej zdolności do spełnienia określonych wymagań i potrzeb użytkownika usługi [1]. Parametry jakościowe służą zatem do określania, czy użytkownik usługi powinien być nią usatysfakcjonowany. Na ich podstawie możliwe jest określanie spodziewanej satysfakcji użytkownika z dostarczanej usługi. Już poprzez ten fakt zrozumiałym jest, że monitorowanie parametrów jakościowych jest nieodzowne do weryfikacji, czy usługi dostarczane przez operatora spełniają oczekiwania użytkowników. Dodatkowo monitorowanie tych parametrów w czasie zbliżonym do rzeczywistego daje administratorowi sieci możliwość szybkiego reagowania na wszelkie odstępstwa od przyjętych norm.

Niesatysfakcjonujące parametry jakościowe pomierzone na danym odcinku (segmencie) sieci umożliwiają również sprawniejsze namierzenie wadliwie funkcjonujących urządzeń sieciowych, dużych obciążeń na konkretnych łączach, wąskich gardeł czy potencjalnych ataków przypuszczanych na daną sieć. Jak zatem widać szereg zastosowań i korzyści wypływających z wykorzystania systemu monitorującego parametry jakościowe sprawia, że jest to niezwykle istotne narzędzie, którym powinien dysponować administrator sieci. Można tu wyróżnić dwa podejścia do realizacji takiego systemu: aktywne i pasywne. W pierwszym przypadku do sieci generowany jest ruch pomiarowy, na którym przeprowadzane są pomiary. Przykładem takiego rozwiązania jest np. opracowany specjalny protokół INT [2,3]. Wymaga to jednak dodatkowej funkcjonalności elementów sieciowych oraz powoduje wzrost ruchu w sieci i jej dodatkowe obciążenie. W przypadku pasywnego rozwiązania system znajduje się na zewnątrz sieci i ma jedynie dostęp do strumieni rejestrowanych i monitorowanych pakietów bez ingerencji w ich obsługę. Mając na względzie obiektywność oceny jakości usług powinno się dążyć do niezależności tego typu systemu w stosunku do monitorowanej sieci.

Jak ważny jest to problem świadczy fakt, że opracowano standard IPFIX dla realizacji zewnętrznego systemu rejestrowania i monitorowania danych dla określenia parametrów jakościowych, a następnie zaproponowano jego rozszerzenie do wariantu określonego mianem SIPFIX [4,5]. Na jego bazie opracowano koncepcję takiego systemu, który przedstawiono i krótko scharakteryzowano w kolejnym rozdziale a następnie w rozdziale trzecim przedstawiono propozycję i realizację podstawowego jego elementu jakim jest Próbnik. Został on poddany testom a jego wyniki omówiono w rozdziale czwartym.

2. KONCEPCJA SYSTEMU REJESTRACJI I MONITOROWANIA PARAMETRÓW

Rejestracja i monitorowanie parametrów jakościowych to bardzo złożone zagadnienie, które musi spełniać szereg wymagań, aby w skuteczny sposób dostarczać należytych informacji dotyczących realizacji usług w sieci. Poza szeroką liczbą tych wymagań, kluczowy jest również fakt, że z reguły są one niezwykle restrykcyjne, np. w przypadku usług uwarunkowanych czasowo, przy jednoczesnej potrzebie zapewnienia odpowiedniej reakcji administratora na zjawiska występujące w monitorowanej sieci. Należy mieć na uwadze, iż oba te czynniki przekładają się na znaczący wzrost trudności opracowania satysfakcjonującego rozwiązania, zarówno w sferze koncepcyjnej, jak i pod względem doboru właściwych narzędzi,

które mogą posłużyć do zaimplementowania takiego systemu.

Ogólne wymagania dotyczące realizacji systemu rejestracji i monitorowania parametrów jakościowych można przedstawić w formie następujących punktów.

1. Analiza wiadomości SIP w płaszczyźnie sygnalizacyjnej oraz pakietów RTP w płaszczyźnie danych medialnych – system monitorujący musi być dostosowany do analizy konkretnej architektury sieci IP, wykorzystującej protokoły SIP i RTP.
2. System musi być pasywny – oznacza to, że system nie może ingerować w strukturę monitorowanej sieci, np. poprzez umieszczanie w niej dodatkowych urządzeń lub stwarzanie konieczności modyfikacji konfiguracji urządzeń już funkcjonujących w ramach tejże sieci.
3. System musi monitorować połączenia na całej drodze (end-to-end) – sieć IP można podzielić na segmenty, z których każdy powinien być monitorowany niezależnie.
4. Analiza danych i prezentacja rezultatów musi odbywać się w czasie zbliżonym do rzeczywistego – jest to najbardziej restrykcyjne wymaganie z wszystkich wymienionych i zarazem niezwykle istotne.
5. System powinien dostarczać wizualną prezentację danych – wymaganie to zakłada, że administrator sieci będzie miał możliwość analizowania danych i parametrów jakościowych związanych z poszczególnymi sesjami przy pomocy graficznego interfejsu dostępnego poprzez przeglądarkę (web GUI – Graphical User Interface) w formie aplikacji klient-serwer.
6. System musi informować administratora sieci w przypadku niezapewniania odpowiedniej jakości połączeń lub nieprawidłowego działania sieci – od systemu monitorującego wymaga się, aby alarmował on administratora sieci w sytuacjach, gdy co najmniej jeden z parametrów jakościowych nie spełni zadanych kryteriów.
7. System powinien przechowywać parametry zakończonych połączeń i informacje o nich – system monitorujący powinien zawierać bazę danych, w której byłyby gromadzone wiadomości analizowanych protokołów (SIP oraz RTP), dotyczące poszczególnych połączeń oraz wyznaczone dla nich parametry jakościowe.

Opracowywany system jest pasywny o rozproszonej architekturze. Dane będą zatem napływały z monitorowanej sieci do zewnętrznie usytuowanych elementów, rozmieszczonych tak, aby każdy z nich mógł analizować ruch z innego segmentu sieci. Ponadto architektura systemu została podzielona na warstwy według funkcji pełnionych przez poszczególne moduły wchodzące w ich skład.

W najniższej warstwie znajdować się będą ***Próbniki*** odpowiedzialne za odfiltrowywanie ruchu związanego z protokołami podlegającymi analizie, czyli SIP i RTP. Dla każdego segmentu sieci przewidziane jest użycie dwóch takich modułów – jednego do odfiltrowywania wiadomości protokołu SIP, a drugiego do odfiltrowywania pakietów protokołu RTP. Dodatkowo zadaniem Próbników będzie także przyporządkowywanie wiadomości z poszczególnych sesji do wspólnych przepływów poprzez umieszczanie ich w dedykowanych do tego celu kontenerach protokołu IPFIX w formie jego rozszerzonej adaptacji – SIPFIX. Zabieg ten ma w założeniu usprawnić dalszą analizę danych odbieranych przez system ze względu na uproszczenie weryfikacji przynależności do sesji.

W skład drugiej warstwy wchodzić będą ***Mediatory*** przydzielone, tak jak Próbniki, do poszczególnych segmentów monitorowanej sieci. Mediatory mają odpowiadać za wstępną analizę odbieranych od Próbników przepływów i wyznaczanie dla nich parametrów jakościowych na odcinku połączenia należącym do segmentu, za który dany Mediator odpowiada. Dodatkowym zadaniem Mediatorów jest przechowywanie w lokalnej bazie danych wiadomości związanych z mediami (protokół RTP), umożliwiając w ten sposób zmniejszenie obciążenia generowanego w wyższych warstwach systemu.

Mediatory przydzielone do poszczególnych segmentów monitorowanej sieci przesyłają analizowane wiadomości sygnalizacyjne oraz obliczone parametry jakościowe dla poszczególnych przepływów ze swoich segmentów do ***Mediatora Centralnego***, znajdującego się w warstwie trzeciej systemu monitorującego. Mediator Centralny odpowiada za łączenie powiązanych ze sobą sesji z różnych segmentów w jeden wspólny strumień, reprezentujący połączenie na całej drodze end-to-end, a także za obliczanie parametrów jakościowych dla całej drogi end-to-end danego połączenia. W ten sposób w systemie uzyskiwane są wszystkie potrzebne parametry jakościowe – zarówno dla poszczególnych odcinków sieci, jak i całej drogi danego połączenia.

Ostatnią warstwę systemu stanowi ***Kolektor***, który odbiera od Mediatora Centralnego wiadomości sygnalizacyjne podzielone na połączenia oraz obliczone dla nich parametry jakościowe. Kolektor ma za zadanie przekazywać te dane do zewnętrznej bazy danych, która będzie je przechowywała przez określony czas oraz do aplikacji klienckiej po stronie administratora sieci, stanowiącej graficzny interfejs dostępny z poziomu przeglądarki. Kolektor odpowiadałby także za przesyłanie alarmów związanych z nieakceptowalną jakością któregoś z połączeń lub wadliwym działaniem któregoś z elementów monitorowanej sieci.

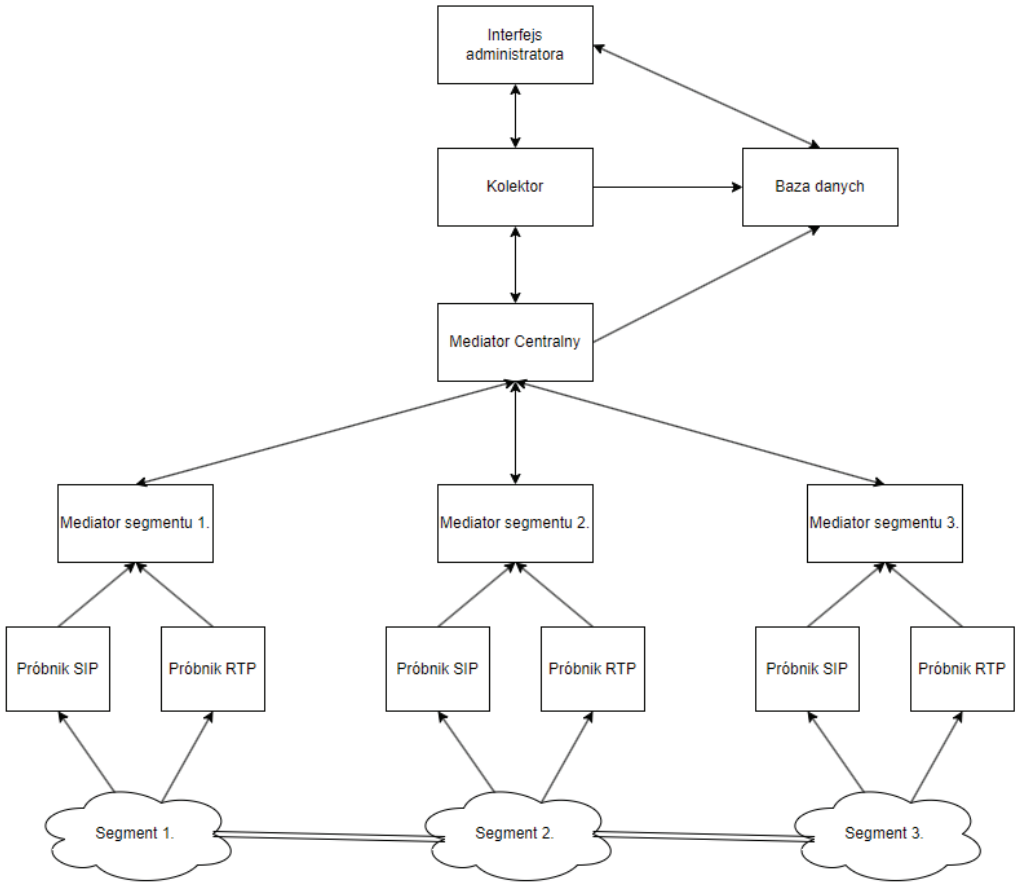
Abstrahując od warstw systemu, uwzględniono w nim wspomniane: zewnętrzną bazę danych oraz graficzny interfejs administratora sieci. Baza danych przechowuje wiadomości sygnalizacyjne, dane medialne i parametry jakościowe z rozróżnieniem na poszczególne połączenia, z którymi byłyby one związane.

Ogólny schemat warstw systemu oraz relacji między nimi dla przykładu sieci składającej się z 3. segmentów przedstawiono na rys. 1.

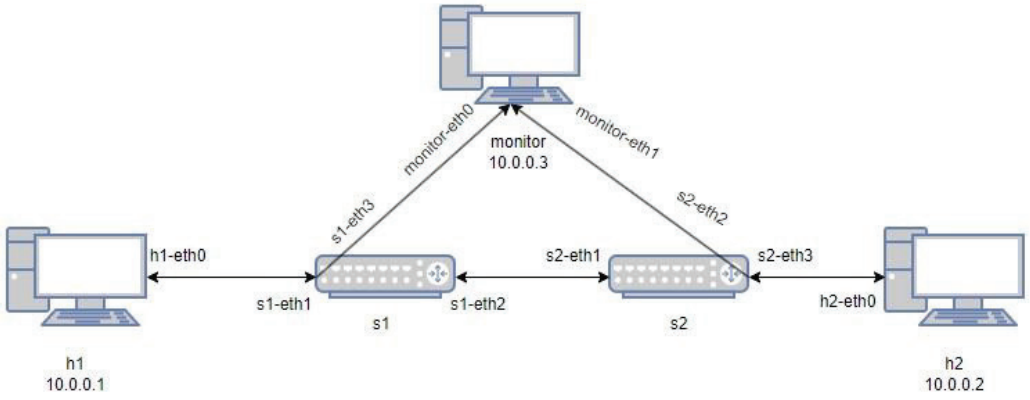
3. KONCEPCJA I REALIZACJA PRÓBNIKA SYSTEMU

Pierwszym z elementów funkcjonalnych sieci monitorującej jest Próbnik. Do jego zadań należy odfiltrowywanie ruchu i przechwytywanie tej jego części, która jest związana z analizowanymi przez system protokołami (SIP albo RTP), a także przyporządkowywanie tego ruchu do poszczególnych przepływów SIPFIX, na podstawie przynależności określonych wiadomości protokołu SIP czy też strumieni danych medialnych do konkretnej sesji połączeniowej.

Na każdy segment przypadają po dwa moduły Próbnika. Jeden z nich jest odpowiedzialny za odfiltrowywanie i przetwarzanie wiadomości protokołu SIP, a drugi natomiast przechwytuje i przetwarza dane protokołu RTP.



Rys. 1. Schemat warstw i elementów systemu rejestracji i monitorowania parametrów jakościowych



Rys. 2. Topologia sieci wykorzystanej do testów

Te dwa rodzaje Próbników różnią się jedynie protokołem, na którym będą operować, a w związku z tym wykonywane przez nie operacje będą takie same, przy czym będą się różniły informacjami pobieranymi z przetwarzanych danych. Funkcje realizowane przez Próbnik można podzielić na dwa etapy:

– odfiltrowywanie ruchu związanego z monitorowanymi protokołami (wiadomości SIP albo pakietów RTP),
– przyporządkowywanie odfiltrowanego ruchu do przepływów SIPFIX na podstawie sesji sygnalizacyjnych/strumieni danych mediów, do których należą poszczególne pakiety.

Na odbieranych przez Próbnik pakietach są realizowane podstawowe operacje związane z *Procesem obserwacji*, *Procesem pomiarowym* i *Procesem eksportującym*. W przypadku *Procesu obserwacji* realizowana jest jedna operacja, a mianowicie filtrowanie pakietów SIP albo RTP odbieranych z danego segmentu monitorowanej sieci. Jest to pierwszy etap realizowany przez Próbnik. W związku z tym na podstawie informacji dostępnych w nagłówku warstwy transportowej odczytywana jest informacja o tym, jakiego protokołu dane one transportują i na tej podstawie zachowywane do dalszego przetwarzania są tylko te, należące do protokołu SIP albo RTP. Odfiltrowane pakiety są składowane w formie plików z rozszerzeniem *.pcap* generowanych przy wykorzystaniu Packet

Capture library (*pcap*), wchodzącej w skład biblioteki *libpcap* [6,7]. Te pliki są z kolei wykorzystywane dalej przez *Proces pomiarowy*. W ramach *Procesu pomiarowego* wyróżnia się następujące operacje:
- przechwytywanie nagłówków pakietów,
- sygnowanie czasowe,
- klasyfikacja do danego przepływu,
- zarządzanie rekordami przepływów.

Przechwytywanie nagłówków pakietów jest wykorzystywane do przydzielania pakietów do odpowiednich przepływów, a także do pozyskania danych niezbędnych w procesie obliczania parametrów jakościowych w kolejnych składnikach systemu monitorującego. W przypadku Próbnika SIP, poza przechwytywaniem nagłówków wiadomości, analizowane jest także jej ciało, w celu odczytania kluczowych informacji z protokołu SDP.

Sygnowanie czasowe, podobnie jak przechwytywanie nagłówków pakietów, służy obliczaniu parametrów jakościowych, a ponadto umożliwia sortowanie przetwarzanych danych według momentu odebrania oraz wykrywa nieaktywne przepływy przeznaczone do usunięcia.

Klasyfikacja do danego przepływu odbywa się na podstawie danych przechwyconych z nagłówków wiadomości SIP oraz datagramów UDP (w przypadku pakietów RTP). W przypadku protokołu RTP, klasyfikator podejmuje decyzję na podstawie klasycznego w sieciach IP 5-elementowego zbioru, na który składają się następujące pola nagłówka datagramu UDP: adres źródłowy, adres docelowy, port źródłowy, port docelowy oraz zastosowany protokół transportowy. Natomiast w przypadku wiadomości SIP, konieczne jest odwołanie się bezpośrednio do nagłówka tego protokołu, gdyż sygnalizacja dla różnych sesji na ogół przesyłana jest tym samym portem (typowo w przypadku protokołu SIP jest to port 5060). Tak więc wiadomości SIP z różnych sesji, przesyłane między tymi samymi elementami sieciowymi, byłyby klasyfikowane do tego samego przepływu, gdyby klasyfikacja przebiegała na podstawie wspomnianego 5-elementowego zbioru. W związku z tym konieczne jest rozróżnienie przepływów sygnalizacji na podstawie identyfikatora sesji (nagłówki *To*, *From* i *Call-ID* albo nagłówek *Session-ID*).

Poza samym problemem klasyfikacji wiadomości sygnalizacyjnych do przepływów, warto w tym miejscu zaznaczyć, że w przypadku sesji sygnalizacyjnych, w każdym z segmentów, przepływ utworzony dla takiej sesji różni się względem swoich odpowiedników w pozostałych segmentach ze względu na to, że zmianie ulegają adres źródłowy i adres docelowy (co nie ma miejsca w przypadku strumieni danych mediów). Stąd również do powiązywania przepływów sygnalizacji z różnych segmentów monitorowanej sieci, należących do tej samej sesji połączeniowej, konieczne jest stosowanie nagłówków protokołu SIP, gdyż te (poza ewentualną translacją na brzegu domeny danego operatora) powinny pozostać niezmienne.

Na zarządzanie rekordami przepływów składają się następujące akcje:
- tworzenie nowych rekordów przepływu,
- aktualizowanie istniejących rekordów przepływu,
- przekazanie rekordów przepływu do *Procesu eksportującego*,
- wykrycie wygaśnięcia przepływu,
- usuwanie rekordów przepływu, które wygasły.

Ostatnia z wykonywanych przez Próbnik operacji to przesyłanie rekordów przepływów do kolejnego elementu sieci monitorującej, czyli Mediatora danego segmentu. Jest to realizowane w ramach *Procesu eksportującego*. Poza zasadniczą częścią Procesu eksportującego, Próbnik przesyła również do Mediatora segmentu niezmienione wiadomości SIP i pakiety RTP.

W ramach lokalnej bazy danych Próbnika utrzymywana jest tablica wykorzystywana na potrzeby zarządzania przepływami IPFIX. Jej struktura umożliwia sprawne zarządzanie przepływami i w czytelny sposób przedstawia ich stan. W związku z tym w tablicy można wyróżnić pola, które są kluczowe dla jednoznacznej identyfikacji przepływu (identyfikator sesji w przypadku wiadomości SIP albo 5-elementowy zbiór w przypadku danych mediów), a także dodatkowe, które uzupełniają informacje dotyczące stanu danego przepływu i warunkują jego czas życia. W tablicy przepływów IPFIX znajdującej się w Próbniku można zatem wyróżnić następujące pola:
- Identyfikator przepływu:
  - *To*, *From* i *Call-ID* albo *Session-ID* (w przypadku wiadomości SIP),
  - adres źródłowy, adres docelowy, port źródłowy, port docelowy, zastosowany protokół transportowy (w przypadku pakietów RTP),
- czas życia przepływu,
- dodatkowe informacje:
  - czas zarejestrowania przepływu,
  - czas ostatniej aktualizacji przepływu,
  - liczba pomierzonych pakietów,
  - liczba pomierzonych bajtów.

Tablica ta jest jedynym elementem wchodzącym w skład lokalnej bazy danych Próbnika, gdyż wszystkie pozostałe dane są przesyłane dalej, do Mediatora przypisanego do danego segmentu. Na tablicy zarządzającej przepływami IPFIX w Próbniku przeprowadzane są następujące operacje:
- zapis nowych rekordów do tablicy (poprzez odczytywanie danych z nagłówków wiadomości SIP/ datagramów UDP (w przypadku pakietów RTP)),
- aktualizowanie istniejących rekordów (czasu życia przepływu, czasu ostatniej aktualizacji, pomierzonych pakietów i pomierzonych bajtów), dla których zarejestrowano nowe pakiety,
- usuwanie rekordów, dla których wygasł czas życia przepływu,
- odczyt rekordów z tablicy.

Nowe rekordy są zapisywane do tablicy w momencie zarejestrowania wiadomości SIP albo pakietów RTP, dla których wyodrębniony identyfikator sesji albo 5-elementowy zbiór nie wystąpił w żadnym z istniejących już rekordów. Natomiast w przypadku znalezienia odpowiadającego zbioru nastąpi aktualizacja pasującego rekordu. Jeżeli w danym rekordzie czas życia przepływu wyzeruje się, to taki przepływ zostaje uznany za wygasły, a związany z nim rekord zostaje usunięty z tablicy. Aktywne rekordy można odczytać poprzez wyświetlenie całej tablicy.

Filtrowanie wiadomości SIP zrealizowano poprzez zastosowanie skryptu korzystającego z języka powłoki *bash* [8]. Główny mechanizm filtrowania jest realizowany

poprzez otwarte dostępne narzędzie *dumpcap* [9]. Przygotowany skrypt nazwano *sip_filter.sh* i przyjmuje on 5 opcjonalnych argumentów, przekazywanych do narzędzia *dumpcap*, Są to:

- i / interface – interfejs, na którym odbierany jest odfiltrowywany ruch (np. *eth0*),
- d / duration – czas (w sekundach), po którym odfiltrowane dane będą zapisywane do nowego pliku z rozszerzeniem pcap [7],
- fs / filesize – maksymalny rozmiar pliku pcap (w kB), po osiągnięciu którego odfiltrowywany ruch zapisywany jest do kolejnego pliku,
- p / path – ścieżka zapisu plików, w których przechowywany jest odfiltrowany ruch,
- f / filter – kryteria, na podstawie których odbywa się filtrowanie monitorowanego ruchu.

Przykładowe wywołanie skryptu filtrującego wiadomości SIP wygląda następująco: ./sip_filter.sh i=”eth0” d=10 fs=20 f=”udp port sip”

Jeżeli chodzi o moduł przeznaczony do przetwarzania pakietów RTP, to w jego przypadku filtrowanie jest ściśle związane z procesem klasyfikacji do przepływów i zrealizowane w ramach jednego narzędzia, stosowanego do obu procesów, którym jest *YAF* (Yet Another Flowmeter) [10]. Filtrowanie pakietów RTP odbywa się jako wstępny proces stosowany w ramach narzędzia *YAF*. *YAF* wykorzystuje do filtrowania analizator *tcpdump*, który podobnie, jak *dumpcap* bazuje na bibliotece *libpcap*, przeznaczonej do przechwytywania monitorowanego ruchu [6]. Zatem filtrowanie w ramach narzędzia *YAF* sprowadza się do zastosowania parametru --*filter* z odpowiednią regułą. W przypadku pakietów RTP domyślną regułą jest: udp[1] & 1 != 1 && udp[3] & 1 != 1 && udp[8] & 0x80 == 0x80 && length < 250

YAF to czujnik, który przechwytuje ruch w sieci, zarówno w czasie rzeczywistym, jak i asynchronicznie – z plików w formacie pcap. Na podstawie tego ruchu *YAF* generuje dane przepływu, zawierające informacje na temat każdego połączenia zaobserwowanego w monitorowanej sieci. *YAF*, w ramach procesu zwanego *Deep Packet Inspection*, bada zawartość pakietów, przechwytuje istotne informacje dla poszczególnych protokołów i eksportuje je w szablonie bazującym na standardzie IPFIX, z którym to narzędzie jest zgodne. *YAF* i poszczególne biblioteki wchodzące w jego skład są dostępne na licencji GNU GPL 2.0 [10,11]. Zatem *YAF* dostarcza wszystkie mechanizmy potrzebne do implementacji Próbnika. Konieczne było jednak rozszerzenie biblioteki *rtpplugin.c*, tak aby można było zapisywać także dane niezbędne do obliczania parametrów jakościowych (IPLR, IPTD, IPDV). Wymagało to wyodrębnienia jeszcze z nagłówka pakietu RTP informacji o dwóch jego polach: numerze sekwencyjnym (*sequence number*) oraz sygnaturze czasowej (*timestamp*) [12]. Nie było to trywialne zadanie lecz z uwagi na ograniczenie co do ilości tekstu nie będzie ono tu opisane.

Aby ułatwić korzystanie z narzędzia *YAF* i usprawnić proces filtrowania i klasyfikacji strumieni danych mediów, podobnie jak w przypadku filtrowania wiadomości SIP, przygotowano skrypt *flow_monitoring.sh* w języku powłoki *bash*, automatyzujący całość procesu dla pakietów RTP. Skrypt ten przyjmuje cztery opcjonalne parametry, przekazywane dalej do głównego procesu *yaf* [13], który na końcu uruchamia. Parametrami skryptu są:

- i / interface – interfejs, na którym odbierany jest odfiltrowywany ruch (np. *eth0*),
- o / out – ścieżka zapisu plików w formacie IPFIX dostosowanym do narzędzia *YAF* (rozszerzenie *.yaf*), przechowujących dane zarejestrowanych przepływów,
- r / rotate – czas (w sekundach), po którym dane zarejestrowanych przepływów będą zapisywane do nowego pliku z rozszerzeniem *.yaf*,
- pt / pcapTimer – czas (w sekundach), po którym odfiltrowane pakiety RTP będą zapisywane do kolejnego pliku w formacie pcap.

Przykładowe wywołanie skryptu *flow_monitoring.sh* filtrującego pakiety RTP wygląda następująco: ./flow_monitoring.sh i=”eth0” o=”home/admin/eth0/rtp_flows” r=10 pt=20

W ramach utworzonego skryptu *flow_monitoring.sh* wywoływany jest proces *yaf* (wraz z procesami *applabel* i *yafdpi*), który monitoruje ruch na wskazanym interfejsie, odfiltrowuje z niego pakiety RTP, klasyfikuje je do jednopakietowych przepływów i wyznacza dla nich istotne wartości wskazane przez Elementy informacyjne, co zawiera w formacie zgodnym ze standardem IPFIX, w plikach z rozszerzeniem *.yaf*. Ponadto odfiltrowane z ruchu pakiety RTP są zapisywane przez proces *yaf* do plików *pcap*. Wszystkie te wygenerowane przez proces *yaf* pliki są gotowe do dalszego wykorzystania przez kolejne składniki systemu monitorującego.

4. TESTY PRÓBNIKA SYSTEMU

Na zaimplementowanym Próbniku przeprowadzono następujące testy:

- test filtrowania wiadomości SIP (test skryptu sip_filter.sh),
- test dynamicznego przetwarzania strumieni mediów – pakietów RTP (test skryptu flow_monitoring.sh).

Jako środowisko testowe wykorzystano sieć wirtualną utworzoną przy pomocy oprogramowania Mininet w wersji 2.3.1b1, umożliwiającą dodawanie wirtualnych switchy i hostów na jednym urządzeniu [14]. Topologię tej wirtualnej sieci, przygotowano przy pomocy skryptu o nazwie *flow_monitoring_test.py* napisanego w języku Python a samą topologię przedstawiono na rys. 2. Składa się ona z 5 elementów – 3 hostów (h1, h2 oraz monitor) i 2 switchy (s1 oraz s2). Host, któremu nadano nazwę monitor pełni w tym teście rolę Próbnika, nasłuchującego na interfejsach połączonych z portami switchy, na które duplikowany jest ruch przechodzący przez te switche. Tak prosta sieć w zupełności wystarczyła do testów, gdyż istotne było zweryfikowanie, czy ruch jest poprawnie przechwytywany przez Próbnik i prawidłowo przetwarzany w ramach jego procesów.

Ruch pomiędzy hostami h1 i h2 generowano przy pomocy narzędzia SIPp [15], które można wykorzystać do symulacji połączeń zestawianych przy pomocy protokołu SIP. SIPp umożliwia realizację prostych scenariuszy pomiędzy dwoma UA w relacji UAS – UAC. Aby poprawnie odwzorować pasywność Próbnika konieczne było dostosowanie portów switchy połączonych z Próbnikiem (monitor) do rozgłaszania zduplikowanego całego ruchu przechodzącego przez pozostałe porty tych switchy

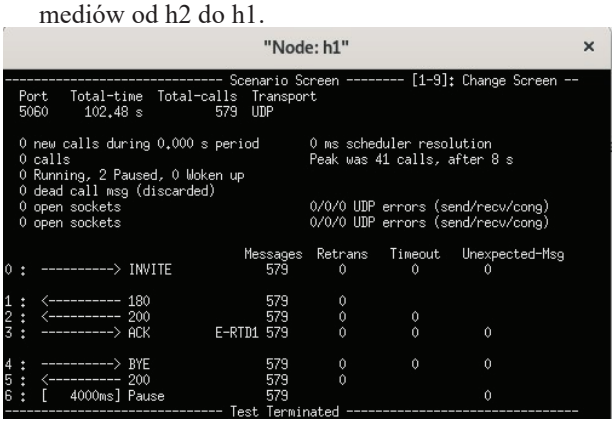
(czyli w komunikacji między hostami h1 i h2). W tym celu utworzono skrypty *s1_prepare.sh* i *s2_prepare.sh* dla switchy, odpowiednio konfigurujące porty, na które miał być rozgłaszany zduplikowany ruch. Test filtrowania wiadomości SIP przebiegał następująco:

- skonfigurowano switche s1 i s2 poprzez skrypty *s1_prepare.sh* i *s2_prepare.sh*,
- na hoście h1 uruchomiono przy pomocy narzędzia SIPp wbudowany scenariusz UAS,
- na Próbniku (monitor) uruchomiono filtrowanie wiadomości SIP poprzez skrypt *sip_filter.sh* z domyślnymi parametrami,
- na hoście h2 uruchomiono przy pomocy narzędzia SIPp wbudowany scenariusz UAC, wywołując w ten sposób zestawianie prostych sesji połączeniowych między hostami h1 i h2.

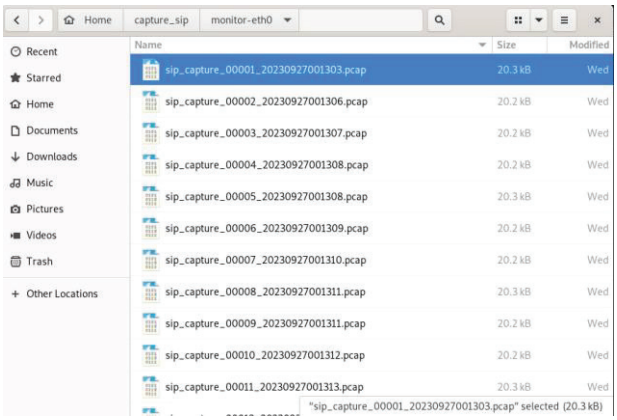
Na rys. 3 przedstawiono raport z narzędzia SIPp dla UAS (host h1). Zgodnie z ustawionymi parametrami dla skryptu filtrującego *sip_filter.sh* pliki pcap zostały tworzone po upływie 10 sekund lub po przekroczeniu rozmiaru 20kB, których katalog został pokazany na rys. 4. Z kolei na rys. 5 przedstawiono zawartość przykładowego spośród wygenerowanych plików, tzn. sip_capture_00002_20230927001306.pcap.

Drugi z przeprowadzonych testów dotyczy weryfikacji dynamicznego przetwarzania strumieni mediów przy pomocy narzędzia *YAF*, realizowanego poprzez przygotowany skrypt *flow_monitoring.sh*. Celem tego testu było zweryfikowanie poprawnego zautomatyzowania monitorowania strumieni danych mediów, sprawdzenie czy pakiety RTP są prawidłowo odfiltrowywane z monitorowanego ruchu, a także wykazanie, że narzędzie *YAF* zostało poprawnie rozszerzone o obsługę nowych Elementów informacyjnych. Test ten przebiegał następująco:

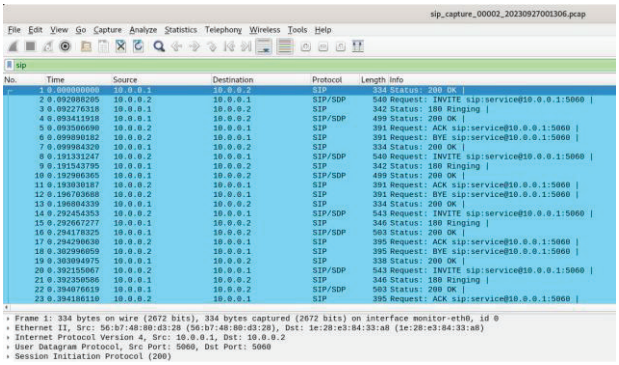
- skonfigurowano switche s1 i s2 poprzez skrypty *s1_prepare.sh* i *s2_prepare.sh*,
- na hoście h1 uruchomiono przy pomocy narzędzia SIPp wbudowany scenariusz UAS,
- na Próbniku (monitor) uruchomiono proces *yaf* poprzez skrypt *flow_monitoring.sh* z domyślnymi parametrami,
- na hoście h2 uruchomiono przy pomocy narzędzia SIPp wbudowany scenariusz UAC_PCAP, wywołując w ten sposób zestawianie prostych sesji połączeniowych między hostami h1 i h2 z przesyłaniem danych mediów od h2 do h1.



Rys. 3. Raport z narzędzia SIPp dla UAS po przeprowadzonym teście

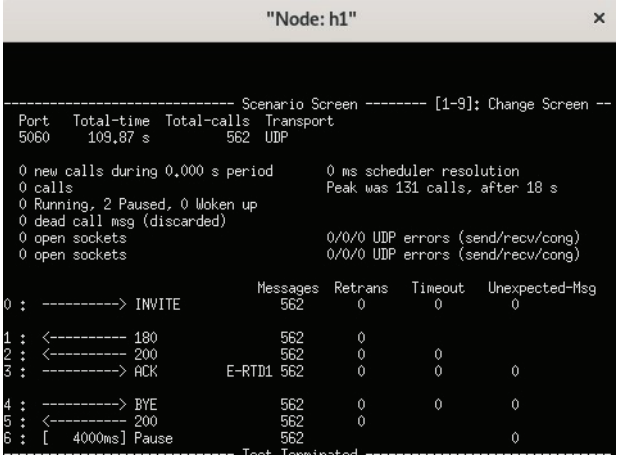


Rys. 4. Katalog utworzony przez skrypt sip_filter.sh zawierający pliki pcap z odfiltrowanymi wiadomościami SIP

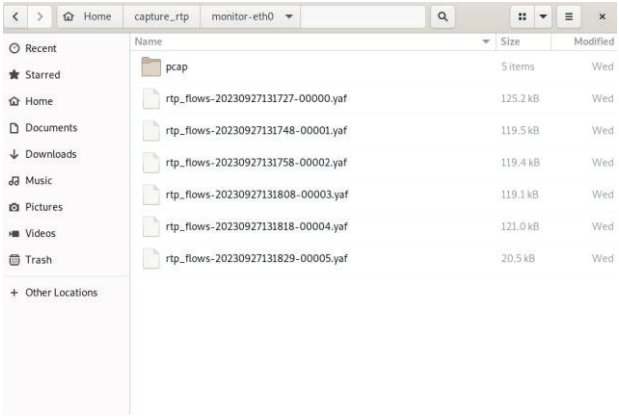


Rys. 5. Weryfikacja wystąpienia wiadomości SIP w przykładowym pliku pcap wygenerowanym przez Próbnik dla odfiltrowanego ruchu

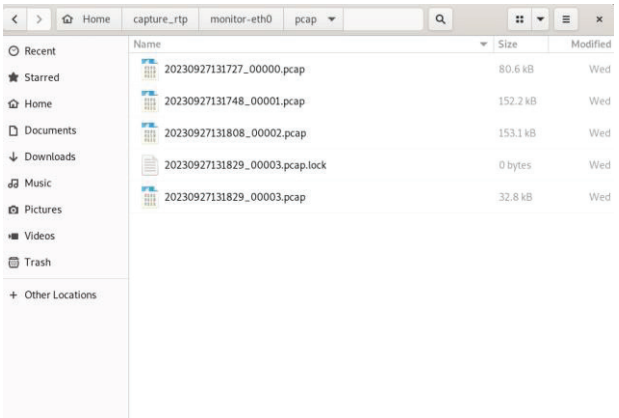
Na rys. 6 przedstawiono raport z narzędzia SIPp dla UAS (host h1). Pliki ze zgromadzonymi danymi przepływów z rozszerzeniem *.yaf* zostały utworzone pod domyślną ścieżką (home/admin/capture_rtp/), z analogicznym interfejsem, jak w przypadku pierwszego testu, czyli „monitor-eth0" (rys. 7). Na rysunku można również zauważyć, że został utworzony katalog o nazwie „pcap", przechowujący odfiltrowane z monitorowanego ruchu pakiety RTP umieszczone w plikach pcap. Zawartość katalogu „pcap" przedstawiono na rys. 8. Z kolei na rys. 9 przedstawiono zawartość przykładowego spośród wygenerowanych plików, tzn. 20230927131748_00001.pcap.



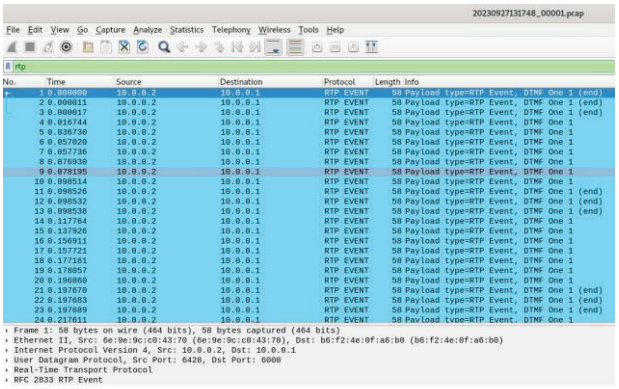
Rys. 6. Raport z narzędzia SIPp dla UAS po przeprowadzonym teście



Rys. 7. Katalog utworzony przez skrypt flow_monitoring.sh zawierający pliki .yaf z wyznaczonymi danymi przepływów RTP



Rys. 8. Katalog utworzony przez skrypt flow_monitoring.sh zawierający pliki pcap z odfiltrowanymi pakietami RTP



Rys. 9. Weryfikacja wystąpienia pakietów RTP w przykładowym pliku pcap wygenerowanym przez Próbnik dla ruchu odfiltrowanego przy pomocy procesu yaf

Przeprowadzone testy zakończyły się pozytywnymi rezultatami dowodząc, że zrealizowana implementacja Próbnika spełnia swoje zadanie założone w koncepcji systemu monitorującego.

5. PODSUMOWANIE

Nadzorowanie sieci telekomunikacyjnej jest dla jej Operatora bardzo ważną funkcją w szczególności w odniesieniu do parametrów jakościowych. Dlatego przedstawiony w pracy system rejestracji i monitorowania oraz realizacja jednego z elementów – Próbnika – jest ważnym krokiem na drodze rozwiązania problemu poruszanego w tej pracy. Uzyskane wyniki wskazują na możliwość zbudowania takiego systemu. Wymaga to realizacji pozostałych składników tego systemu według przedstawionej koncepcji i przeprowadzenia odpowiednich testów. Jednocześnie gromadzone i analizowane przez system informacje wskazują iż może on być rozbudowany dla wsparcia podejmowania decyzji przez Operatora o sztuczną inteligencję i to zarówno na poziomie centralnym jak i na niższych pośrednich poziomach tego systemu.

LITERATURA

[1] ITU-T Recommendation E.800. 2008. “Definitions of terms related to quality of service”.

[2] In-band Network Telemetry (INT) Dataplane Specification, https://p4.org/p4-spec/docs/INT_v2_1.pdf , (data dostępu 01.01.2023).

[3] Kaczmarek Sylwester, Mateusz Krasoń. 2023. „Emulacja środowiska dla zastosowania protokołu In-band Network Telemetry”. *Przegląd Telekomunikacyjny*, rocznik XCVI (4): 57-60.

[4] Claise B., B. Trammell, P. Aitken. 2013. “Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of Flow Information”. *Internet Engineering Task Force (IETF)*, RFC 7011.

[5] Anderson S., D. Hogrefe, S. Niccolini. 2009. “SIP-FIX: A Scheme For Distributed SIP Monitoring”. IFIP/IEEE International Symposium on Integrated Network Management, June: 382-389.

[6] The Tcpdump Group, tcpdump and libpcap, https://www.tcpdump.org/, (data dostępu 30.09.2023).

[7] The Tcpdump Group, pcap – Packet Capture library, https://www.tcpdump.org/manpages/pcap.3pcap.html, (data dostępu 30.09.2023).

[8] GNU Operating System, GNU Bash, https://www.gnu.org/software/bash/, (data dostępu 30.09.2023).

[9] Wireshark Foundation, dumpcap(1) Manual Page, https://www.wireshark.org/docs/man-pages/dumpcap.html, (data dostępu 30.09.2023).

[10] CERT NetSA Security Suite, Software Engineering Institute, Carnegie Mellon University, YAF, https://tools.netsa.cert.org/yaf/index.html, (data dostępu 30.09.2023).

[11] CERT NetSA Security Suite, Software Engineering Institute, Carnegie Mellon University, YAF – License, https://tools.netsa.cert.org/yaf/license.html, (data dostępu 30.09.2023).

[12] Schulzrinne H., S. Casner, R. Frederick, V. Jacobson. 2003. “RTP: A Transport Protocol for Real-Time Applications”, Network Working Group, RFC 3550.

[13] CERT NetSA Security Suite, Software Engineering Institute, Carnegie Mellon University, yaf manual page, https://tools.netsa.cert.org/yaf/yaf.html, (data dostępu 30.09.2023).

[14] Mininet Team, Mininet, http://mininet.org/, (data dostępu 30.09.2023).

[15] SIPp reference documentation, https://sipp.readthedocs.io/en/latest/index.html, (data dostępu 30.09.2023).

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

ANTENA MIKROPASKOWA NA PASMO HIGH-BAND SYSTEMU 5G

MICROSTRIP ANTENNA FOR THE HIGH-BAND BAND OF THE 5G SYSTEM

Rafał Przesmycki[1]; Marek Bugaj[2]; Bartosz Dudziński[3]; Leszek Nowosielski[4]

[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, bartosz.dudzinski@wat.edu.pl
[4] Wojskowa Akademia Techniczna, ul. Gen. Sylwestra Kaliskiego 2, 00-908 Warszawa, leszek.nowosielski@wat.edu.pl

DOI: 10.15199/59.2024.4.55

Streszczenie: **W artykule przedstawiono antenę o dużym zysku energetycznym, która wykonana została na podłożu dielektrycznym i przeznaczona jest do pracy w systemie 5G w paśmie high-band. Opracowana antena posiada zwartą konstrukcję o wymiarach (5,16 x 5,05 x 0,254) mm. Jako podłoże do budowy anteny wykorzystano laminat ROGERS RT5880 o przenikalności dielektrycznej 2,2 i grubości h=0,254 mm. W artykule przedstawiono analizę wyników symulacji i pomiarów parametrów elektrycznych i charakterystyk promieniowania zaprojektowanej anteny. Opisana w artykule antena mikropaskowa o dużym zysku energetycznym, pracująca w systemach 5G o częstotliwości środkowej 38 GHz charakteryzuje się niskim współczynnikiem odbicia -29,11 dB, bardzo dużym zyskiem energetycznym 7,61 dBi i paśmie pracy równym 1,21 GHz (3,18%).**

Abstract: **The article presents an antenna with high energy gain, made on a dielectric substrate and intended for operation in the 5G system in the high-band band. The developed antenna has a compact structure with dimensions (5.16 x 5.05 x 0.254) mm. ROGERS RT5880 laminate with a dielectric coefficient of 2.2 and a thickness of h = 0.254 mm was used as a substrate for the construction of the antenna. The article presents an analysis of the simulation results and measurements of the electrical parameters and radiation characteristics of the designed antenna. The high-gain microstrip antenna described in the article, operating in 5G systems with a center frequency of 38 GHz, is characterized by a low reflection coefficient of -29.11 dB, a very high energy gain of 7.61 dBi and an operating band of 1.21 GHz (3.18%).**

Słowa kluczowe: **antena, antena mikropaskowa, duży zysk energetyczny, system 5G.**

Keywords: **antenna, microstrip antenna, high gain, 5G wireless system.**

1. WSTĘP

System 5G jest kolejną generacją powszechnie stosowanych od wielu lat systemów radiokomunikacyjnych ruchomych lądowych, które wszyscy znają i z których korzystają na co dzień, czyli systemów telefonii komórkowej. Elementy infrastruktury w tym systemie są połączone za pomocą światłowodów tworząc sieć 5G (sieć telekomunikacyjną piątej generacji). Jest to nowy standard telekomunikacyjny, który ma umożliwić pięćdziesięcio-, a nawet stukrotne zwiększenie prędkości transmisji w porównaniu do obecnie wykorzystywanej sieci 4G. Sieć 5G ma przyspieszyć m.in. rozwój Internetu rzeczy, usług telemedycyny, autonomicznych pojazdów, czy inteligentnych miast [1]. Jednym z kluczowych warunków wdrażania 5G jest zapewnienie odpowiednich zasobów częstotliwości. Konieczne jest również wdrożenie tzw. „gęstych sieci" rozumianych jako sieci z użyciem wielu stacji bazowych o krótkich zasięgach, co pozwoli znacznie zwiększyć efektywności widmową a zatem oferowaną przepływność klientowi. W sieci 5G odchodzi się od sieci komórek dużych rozmiarów, na rzecz sieci składających się z dużej liczby małych komórek [2][3].

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) [4].

Częstotliwości pasma 700 MHz (694 – 790 MHz) umożliwiają uzyskanie zasięgu łączności nawet do kilkunastu kilometrów od stacji bazowej (więcej niż w sieci 4G). Charakteryzują się dobrą penetracją sygnału radiowego wewnątrz budynków. Z tych powodów pasmo to nazywane jest pasmem pokryciowym lub zasięgowym i bardzo dobrze nadaje się do budowy sieci na terenach pozamiejskich, gdzie konieczne jest zapewnienie zasięgu na dużych obszarach, przy jednocześnie względnie małej gęstości użytkowników sieci 5G. Z kolei ograniczenie dostępnych zasobów częstotliwości tylko do 60 MHz ogranicza przepustowość sieci, która może być porównywalna z przepustowością w sieci 4G lub nawet mniejsza [5][6].

Częstotliwości pasma 3,6 GHz (3,4 – 3,8 GHz) umożliwiają uzyskanie zasięgu łączności do kilku kilome-

trów od stacji bazowej (mniej niż ok. 5 km). Charakteryzują się słabą penetracją sygnału radiowego wewnątrz budynków. Jednak, w odróżnieniu od pasma 700 MHz, zasób częstotliwości jest zdecydowanie większy, co umożliwia uzyskanie dużych przepustowości – nawet ok. 10 krotnie większych względem sieci 4G. Z tego powodu pasmo to nazywane jest pasmem pojemnościowym i zapewnia parametry wymagane do jednoczesnej obsługi wielu podłączonych urządzeń. Ponadto pasmo (z pewnymi ograniczeniami) jest już dostępne od dawna i może być z powodzeniem wykorzystywane do budowy sieci 5G na obszarach dużych miast o gęstej zabudowie do transmisji znacznych ilości danych. Dodatkową zaletą sieci 5G pracujących w tym paśmie jest wprowadzanie bardzo małych opóźnień w transmisji danych (na poziomie pojedynczych milisekund), co ma istotne znaczenie dla obsługi urządzeń Internetu rzeczy [5][6].

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" [5][6].

W niniejszym artykule przedstawiono propozycję anteny mikropaskowej przeznaczonej do pracy w systemie 5G dla pasma high-band 38 GHz oraz wyniki przeprowadzonych symulacji i pomiarów parametrów elektrycznych tej anteny. Oprócz pasma częstotliwości innym ważnym wymaganiem stawianym dla projektowanej anteny są wymiary, które powinny być nie większe niż 10 mm x 10 mm, oraz jak największy zysk energetyczny.

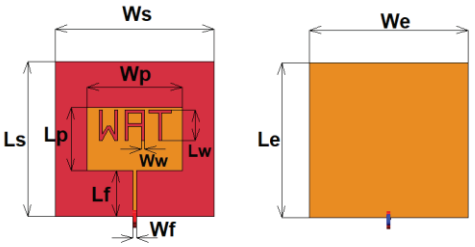
2. MODEL ANTENY

Ze względu na planowane zakresy częstotliwości pracy systemu 5G, głównym założeniem dla projektowanej anteny przeznaczonej do pracy w tym systemie, jest częstotliwość pracy w zakresie 38 GHz (high-band). Oprócz pasma częstotliwości innym ważnym wymaganiem stawianym dla projektowanej anteny są wymiary anteny, które powinny być nie większe niż 10 x 10 mm, jak największy zysk energetyczny, liniowa polaryzacja oraz rodzaj materiału, z którego wykonana ma być antena. Do projektu anteny wybrano profesjonalny laminat ROGERS RT5880. Jest to materiał o grubości h = 0,254 mm, przenikalności elektrycznej równej $\varepsilon_r = 2,2$ oraz niskich stratach $\tan(\delta) = 0,0009$. 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 na antenę opracowano model w oprogramowaniu do projektowania anten mikropaskowych CST Microwave Studio. W oprogramowaniu tym można wyznaczyć również parametry elektryczne i charakterystyki promieniowania anteny. Przy projektowaniu anteny 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 uzyskanego modelu pod kątem stawianych wymagań na antenę modyfikując jej wymiary oraz sposób zasilania. W procesie optymalizacji zmieniane były długość i szerokość podłoża anteny wraz z ekranem (W_s, W_e, L_s, L_e), długość i szerokość promiennika (L_p, W_p) oraz długość i szerokość linii zasilającej (L_f, W_f). Pozostałe elementy struktury takie jak grubość i przenikalność elektryczna podłoża, wielkość liter „WAT" zostały bez zmian. W wyniku przeprowadzonego procesu optymalizacji ostatecznie otrzymano model anteny przedstawiony na rys. 1, którego wymiary przedstawiono w Tab. 1. Dla tak uzyskanego modelu anteny wykonano model fizyczny anteny, którego wygląd pokazano na rys. 2. Dla wykonanego finalnego modelu proponowanej anteny przeprowadzono symulacje komputerowe i pomiary dotyczące wyznaczenia wartości parametrów elektrycznych anteny.

Tab. 1. Wymiary projektu anteny

Antenna Component	Symbol	Dimensions (mm)
Substrate width	W_s	5,16
Substrate length	L_s	5,05
Ground plane width	W_e	5,16
Ground plane length	L_e	5,05
Patch width	W_p	3,10
Patch length	L_p	2,06
Substrate thickness	h	0,254
Permittivity	ε_r	2,2
Feed line width	W_f	0,107
Feed line length	L_f	4,02
WAT letter height	L_w	1,00
WAT letter width	W_w	0,10



Rys. 1. Projekt anteny mikropaskowej wraz z oznaczeniami wymiarów

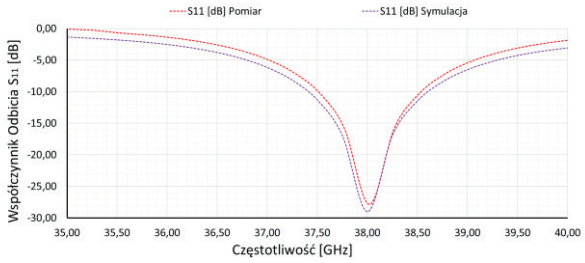
Rys. 2. Wygląd modelu fizycznego proponowanej anteny mikropaskowej

3. WYNIKI POMIARÓW DLA OPRACOWANEJ ANTENY 5G

Zaprojektowana konstrukcja anteny o wysokim zysku energetycznym składa się z trzech komponentów: ekranu, promiennika i podłoża. Dla tak zaprojektowanej anteny 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, rozkład prądów w antenie oraz charakterystyki promieniowania. Wszystkie wyżej wymienione parametry porównano ze sobą w celu analizy właściwości wykonanej anteny mikropaskowej.

3.1. Współczynnik odbicia S_{11}

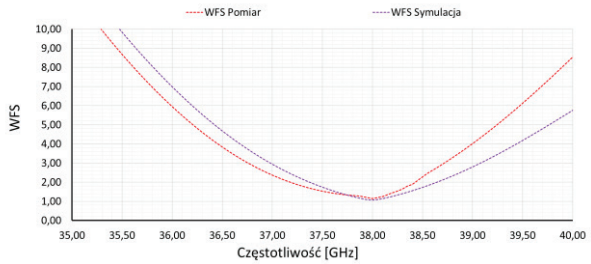
Pierwszym współczynnikiem, który pozwala stwierdzić jaki jest zakres pracy proponowanej anteny jest współczynnik odbicia S_{11}. Na rys. 3 przedstawiono jego wyniki symulacji oraz wyniki pomiarów dla proponowanej anteny mikropaskowej „WAT". Proponowana antena posiada rezonans na częstotliwości 38 GHz ze stratami odbiciowymi wynoszącymi -29,11 dB. Proponowana antena ma szerokość pasma pracy równą 1,21 GHz, co daje szerokość względną pasma pracy 3,18%. Na podstawie tego możemy powiedzieć, że jest to antena rezonansowa stosowana tylko w wybranym paśmie systemu 5G.



Rys. 3. Wartość współczynnika odbicia S_{11} w funkcji częstotliwości dla finalnego modelu proponowanej anteny

3.2. Współczynnik fali stojącej WFS

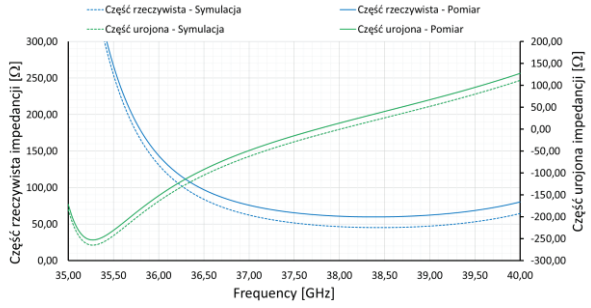
Kolejnym parametrem, który określa właściwości impedancyjne anteny jest współczynnik fali stojącej (WFS). W przypadku anteny łatowej napięciowy współczynnik fali stojącej nie powinien być większy niż 2 w całej szerokości pasma częstotliwości. Na rys. 4 przedstawiono napięciowy współczynnik fali stojącej w funkcji częstotliwości. Jak można zaobserwować wartość WFS uzyskana przy częstotliwości rezonansowej 38 GHz wynosi 1,07, wartość WFS wynosząca 2 wyznaczona została przy częstotliwościach odpowiednio 37,41 GHz i 38,62 GHz. Z powyższych wartości wynika, że proponowana antena pracuje w zakładanym wysokim paśmie częstotliwości 5G tj. 38 GHz, w którym wymagane szerokości kanału wynoszą 50 MHz, 100 MHz, 200 MHz, 400 MHz. Pasmo pracy proponowanej anteny pokrywa wszystkie szerokości kanałów stosowanych w paśmie 38 GHz, co potwierdza możliwość jej zastosowania w systemie 5G.



Rys. 4. Wartość współczynnika fali stojącej dla finalnego modelu proponowanej anteny

3.3. Impedancja wejściowa

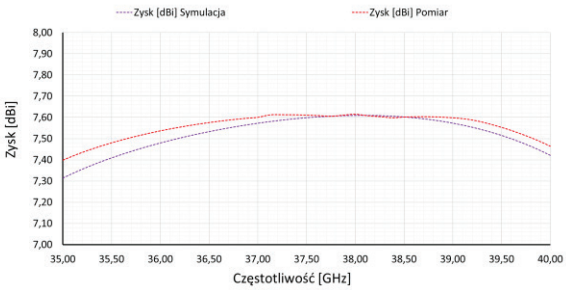
Impedancja wejściowa anteny składa się z części rzeczywistej i urojonej, których wartości zmieniają się wraz z częstotliwością. W projekcie anteny przyjęto, że impedancja linii zasilającej powinna wynosić 50 Ω. Należy tą wartość uwzględnić podczas projektowania anteny, ponieważ jest to łatwy sposób na pogorszenie właściwości tej anteny. Zaprojektowana szerokość i długość linii zasilającej w antenie pozwoliła na uzyskanie impedancji wejściowej wynoszącej Z = 46.58 – j0,24 Ω przy częstotliwości rezonansowej 38,0 GHz. Przy częstotliwościach skrajnych z pasma pracy anteny impedancja wejściowa wynosi odpowiednio dla częstotliwości 37,41 GHz Z = 53,26 – j34,84 Ω, dla częstotliwości 37,62 GHz Z = 45,44 + j31,29 Ω. Szczegółowe wartości impedancji wejściowej dla proponowanej anteny w funkcji częstotliwości przedstawiono na rys. 5.



Rys. 5. Wartość impedancji wejściowej dla finalnego modelu proponowanej anteny

3.4. Zysk energetyczny

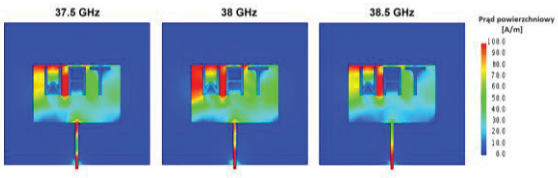
Jednym z ważniejszych parametrów branych pod uwagę podczas projektowania proponowanej anteny mikropaskowej jest zysk energetyczny, który powinien być jak największy. Zysk energetyczny anteny podawany jest w odniesieniu do anteny izotropowej i wyrażany jest w jednostkach dBi. Proponowana antena posiada zysk równy 7,61 dBi przy częstotliwości rezonansowej 38 GHz. Uzyskana wartość jest uważana za bardzo dobrą dla jednoelementowych anten mikropaskowych, co jest dużym osiągnięciem dla takiej 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 zaprojektowanej anteny „WAT" przedstawiono na rys. 6.



Rys. 6. Wartość zysku energetycznego dla finalnego modelu proponowanej anteny

3.5. Rozkład prądów w antenie

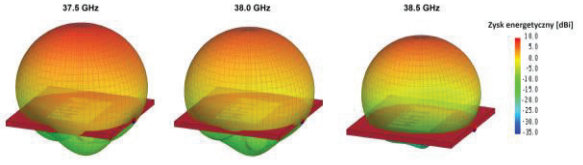
W antenie mikropaskowej na końcu elementu promieniującego (krawędzi łaty) wartość prądu powinna być minimalna. Napięcie na krawędzi łaty jest przesunięte w fazie w stosunku do prądu. W związku z tym napięcie będzie miało wartość szczytową na końcu łaty przy wartościach prądu zbliżonych do zera. Podobna sytuacja wystąpi w połowie długości fali dalej, czyli na początku łaty. W proponowanej antenie mamy potwierdzenie wyżej wymienionych zależności co pokazano na rys. 7, na którym przedstawiono rozkład prądu opracowanej anteny dla częstotliwości 37,5 GHz, 38 GHz oraz 38,5 GHz.



Rys. 7. Rozkład prądów w zaprojektowanej antenie dla częstotliwości 37,5 GHz, 38 GHz, 38,5 GHz

3.6. 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 przedstawiono trójwymiarowy wygląd charakterystyki promieniowania proponowanej anteny „WAT" dla częstotliwości 37,5 GHz, 38 GHz oraz 38,5 GHz uzyskany podczas przeprowadzonych symulacji w CST Microwave Studio.



Rys. 8. Charakterystyka promieniowania 3D anteny 5G dla częstotliwości 37.5 GHz, 38 GHz, 38.5 GHz

4. WNIOSKI

W ramach artykułu przedstawiono projekt anteny mikropaskowej „WAT" przeznaczonej do odbioru sygnałów w technologii 5G w paśmie częstotliwości „high-band". Jest ona doskonałym przykładem zastosowania innowacyjnych technologii w dziedzinie komunikacji bezprzewodowej. Antena została zaprojektowana specjalnie do odbioru sygnałów 5G, które dają możliwość transmisji bardzo dużej ilości danych w krótkim czasie. Opracowana antena to bardzo kompaktowe urządzenie, które umożliwia odbiór sygnałów 5G w sposób bardzo efektywny i niezawodny. Dzięki swojej konstrukcji 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, które w przypadku sieci 4G utrudniają przesył informacji.

Opisana w artykule antena „WAT" wykonana została na podłożu dielektrycznym i przeznaczona jest do pracy w systemie 5G w wysokim paśmie 38 GHz. Opracowana antena posiada zwartą konstrukcję o wymiarach (5,16 x 5,05 x 0,254) mm. Jako podłoże do budowy anteny wykorzystano laminat ROGERS RT5880 o współczynniku dielektrycznym 2,2 i grubości h = 0,254 mm. W artykule przedstawiono analizę wyników symulacji i pomiarów parametrów elektrycznych i charakterystyk promieniowania zaprojektowanej anteny. Opisana w artykule antena mikropaskowa o dużym zysku energetycznym, pracująca w systemach 5G o częstotliwości środkowej 38 GHz charakteryzuje się niskim współczynnikiem odbicia -29,11 dB, bardzo dużym zyskiem energetycznym 7,61 dBi i paśmie pracy równym 1,21 GHz (3,18%).

Praca została sfinansowana przez Wojskową Akademię Techniczną w ramach projektu nr UGB/749/2024 pt. "Anteny mikropaskowe o wysokim zysku energetycznym pracujące w paśmie milimetrowym"

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] Tikhomirov A., Omelyanchuk E. and Semenova A., "Recommended 5G frequency bands evaluation" 2018 Systems of Signals Generating and Processing in the Field of on Board Communications, Moscow, Russia, 2018, pp. 1-5.

[3] Marsch P. et al. (Eds), "5G System Design: Architectural and Functional Considerations and Long Term Research", Wiley, 2018

[4] Chauhan V. and Srinivasans, "A Review on 5G Network System with its limitation and different Approaches to build strong 5G Network System" 3rd International Conference on Intelligent Engineering and Management (ICIEM), London, United Kingdom, 2022, pp. 403-410.

[5] J. Senic et al., "Challenges for 5G and Beyond," 2022 16th European Conference on Antennas and Propagation (EuCAP), Madrid, Spain, 2022, pp. 1-5.

[6] 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.

ANTENA TEKSTYLNA NASOBNEGO SYSTEMU PRZESYŁANIA OBRAZU

TEXTILE ANTENNA FOR WEARABLE IMAGE REGISTRATION SYSTEM

Łukasz Januszkiewicz[1]; Iwona Nowak[2]; Karol Kropidłowski[3]

[1] Politechnika Łódzka, Instytut Elektroniki, Łódź, lukasz.januszkiewicz@p.lodz.pl
[2] Politechnika Łódzka, Instytut Materiałoznawstwa Tekstyliów i Kompozytów Polimerowych, Łódź, iwona.nowak@p.lodz.pl
[3] Politechnika Łódzka, Instytut Elektroniki, Łódź, karol.kropidlowski@p.lodz.pl

DOI: 10.15199/59.2024.4.56

Streszczenie: **W artykule przedstawiono projekt anteny tekstylnej przeznaczonej do nasobnego systemu przesyłania obrazu, wykorzystującego transmisję bezprzewodową w paśmie 2,4 GHz. Opracowana antena jest strukturą warstwową, wykorzystującą dzianinę dystansową o właściwościach elektroizolacyjnych oraz tkaninę przewodzącą. W artykule przedstawiono projekt anteny, wyniki symulacji komputerowych oraz weryfikację pomiarową dopasowania impedancyjnego anteny prototypowej.**

Abstract: **The article presents the design of a textile antenna intended for a wearable image transmission system in 2.4 GHz band. The developed antenna is a layered structure fabricated using a spacer knit with electro-insulating properties and a conductive fabric. The article presents the antenna design as well as the results of computer simulations of the antenna and the measurement verification results of the prototype structure.**

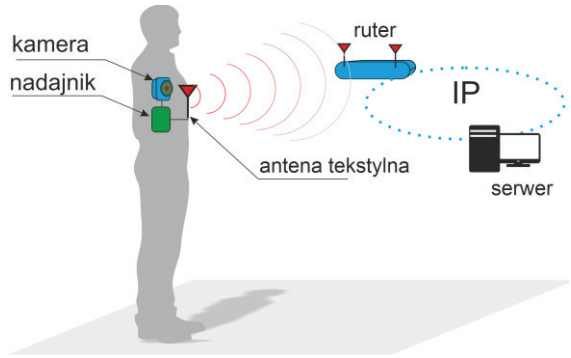
Słowa kluczowe: **antena, elektronika noszona, e-tekstylia, systemy bezprzewodowe**

Keywords: **antenna, e-textiles, wearable electronics, wireless systems.**

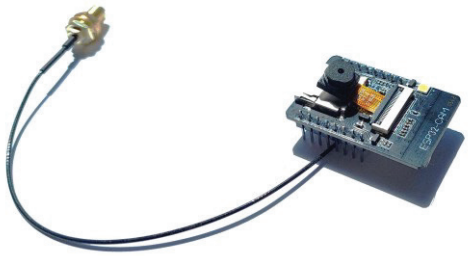
1. WSTĘP

Współczesne układy elektroniczne umożliwiają realizację złożonych zadań za pomocą pojedynczych układów scalonych, zasilanych niewielkimi źródłami energii. Umożliwiło to rozwój systemów elektroniki noszonej, która integruje urządzenia elektroniczne z odzieżą. Przykładem takiego rozwiązania jest system nasobny (przeznaczony do montażu na ciele człowieka) umożliwiający rejestrację i transmisję obrazu, który został opracowany w Instytucie Elektroniki Politechniki Łódzkiej (Rys. 1).

System składa się z miniaturowych kamer wyposażonych w układ transmisji danych wykorzystujący protokół IEEE 802.11 w paśmie 2,4 GHz oraz routera i serwera rejestrującego obraz. Dzięki zastosowaniu miniaturowego układu typu ESP32-CAM [1] przedstawionego na Rys. 2, możliwe jest wbudowanie go w element odzieży ochronnej np. ratownika lub pracownika ochrony. Układ posiada wbudowaną miniaturową antenę typu odwrócone F (z sekcją meandryczną) jednak umieszczenie urządzenia w pobliżu ciała spowoduje odstrojenie impedancyjne tej anteny, która została zaprojektowana do pracy w otwartej przestrzeni [2]. Ponadto zysk energetyczny miniaturowej anteny tego typu jest mały i nie przekracza 0 dBi [3]. Ze względu na to, że układ posiada wyjście antenowe ze złączem koncentrycznym, możliwe było opracowanie dla niego dedykowanej anteny tekstylnej o większym zysku niż antena wbudowana. W dalszej części artykułu przedstawiono jej projekt oraz wyniki symulacji i pomiarów.



Rys. 1. Schemat systemu rejestracji obrazu



Rys. 2. Moduł kamery i nadajnika ESP32-CAM

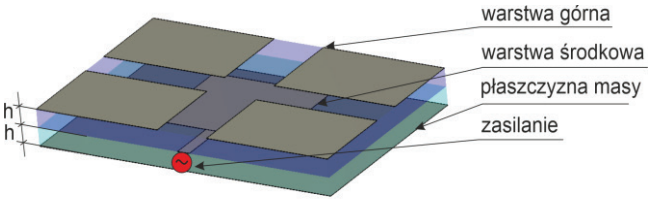
2. PROJEKT ANTENY

Projekt anteny zakładał opracowanie struktury planarnej, warstwowej, która dla pasma 2,4 GHz umożliwiałaby osiągnięcie zysku energetycznego powyżej 5 dBi. Antena powinna być zasilana niesymetrycznie, oraz powinna charakteryzować się dobrym dopasowaniem impedancyjnym do 50 Ω na ciele człowieka.

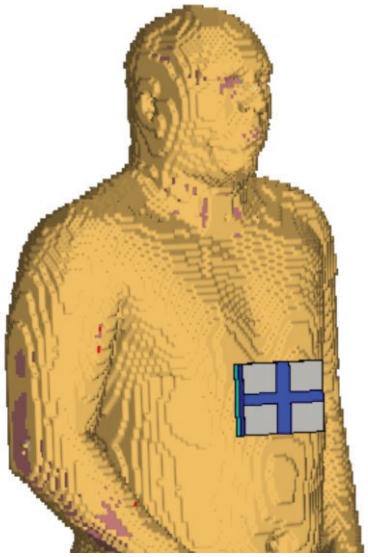
Aby spełnić założone wymagania opracowano projekt 4 elementowego szyku anten paskowych zasilanych

poprzez sprzężenie elektromagnetyczne. Zastosowanie takiej konstrukcji umożliwia poszerzenie pasma pracy anteny [4] co zwiększa jej odporność na odstrojenie impedancyjne spowodowane przysłonięciem przez materiał osłonowy oraz zmieniająca się podczas eksploatacji odległość od ciała człowieka. Ponadto, możliwe jest uzyskanie zysku energetycznego wynoszącego nawet 12 dBi [5,6].

Wstępny projekt anteny wykonano w programie Antenna Magus, w którym stosowane są typowe reguły projektowania tej anteny dla warunku wolnej przestrzeni. Następnie opracowano zmodyfikowany model anteny w programie CST Studio. Modyfikacja polegała na ograniczeniu rozmiaru anteny, co jest pożądane ze względu na przeznaczenie anteny do montażu na ciele. Ograniczono wymiary podłoża anteny jedynie do obszaru wyznaczonego przez łatki górnej warstwy. Tak zmodyfikowany model anteny przedstawiony został na Rys. 3. Wymiary anteny zostały wyznaczone w procesie optymalizacji z wykorzystaniem algorytmu ewolucyjnego. Minimalizowaną funkcją celu było odstrojenie impedancyjne anteny (S11) w paśmie 2,4 – 2,5 GHz, umieszczonej w pobliżu ciała człowieka. W tym celu do symulacji zastosowany został wielotkankowy model Hugo dostępny w programie CST Studio, odwzorowujący ciało mężczyzny za pomocą wokseli o długości boku wynoszącej 5mm. Rys. 4 przedstawia wykorzystany podczas optymalizacji model ciała wraz z umieszczoną na nim anteną. Symulacje przeprowadzono za pomocą metody FIT (ang. Finite Integration Technique) w dziedzinie czasu.



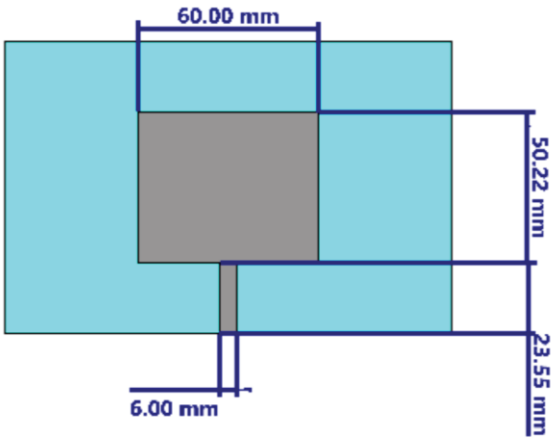
Rys. 3. Model anteny zmodyfikowanej



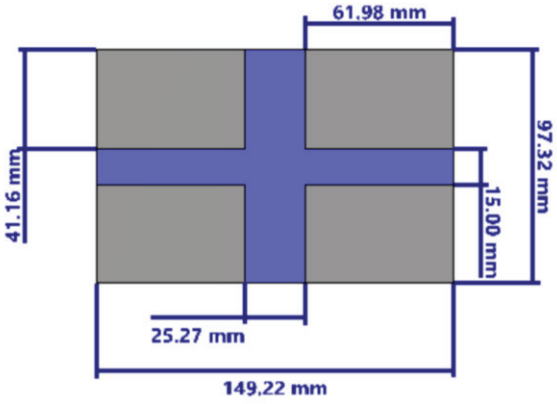
Rys. 4. Moduł anteny umieszczonej na ciele

W procesie projektowania założono, że antena wykonana zostanie z wykorzystaniem poliestrowej dzianiny dystansowej o grubości h=8mm. Przenikalność elektryczną dzianiny wynoszącą ε=2,8 wyznaczono metodą analizy odstrojenia anteny paskowej [7]. Do realizacji elementów przewodzących zastosowano tkaninę poliestrową z domieszką srebra, którą w modelu numerycznym odwzorowano materiałem o przewodności zbliżonej do miedzi.

Wymiary anteny wyznaczone w procesie optymalizacji przedstawiono na Rys. 5 (warstwa środkowa z łatką zasilającą) oraz na Rys. 6 (warstwa górna).



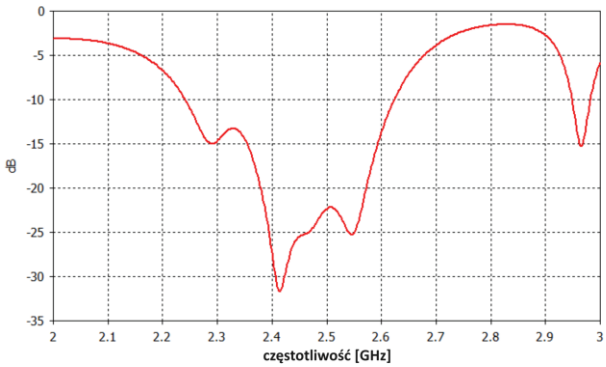
Rys. 5. Wymiary warstwy środkowej z łatką zasilającą



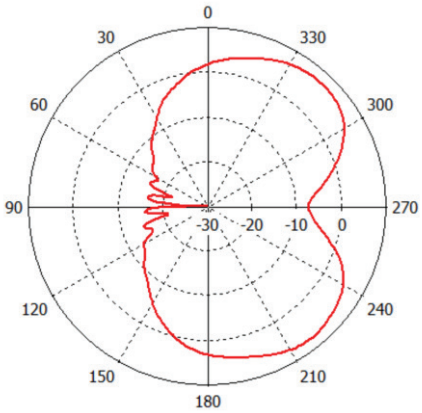
Rys. 6. Wymiary warstwy górnej

Wyniki symulacji dopasowania impedancyjnego anteny uzyskanej w procesie optymalizacji przedstawiono na Rys. 7. W paśmie 2,4 – 2,5 GHz moduł parametru S_{11} nie przekracza -20 dB. W zakresie częstotliwości od 2,24 GHz do 2,63 GHz moduł parametru S_{11} jest mniejszy od -10 dB. Wyniki symulacji dopasowania impedancyjnego anteny są zgodne z założonymi.

Wyniki symulacji charakterystyki promieniowania anteny w płaszczyźnie poziomej $G(\varphi,90°)$ przedstawiono na Rys. 8. Zysk maksymalny wyniósł 7,3 dBi co również spełnia przyjęte założenia projektowe.



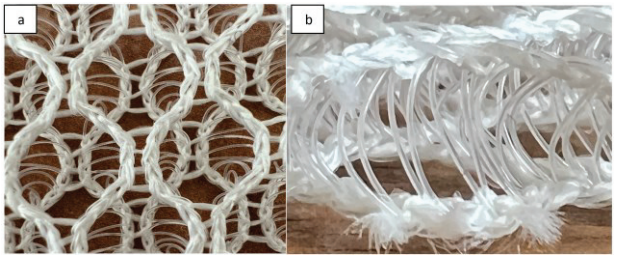
Rys. 7. Wyniki symulacji dopasowania impedancyjnego anteny uzyskanej w procesie optymalizacji



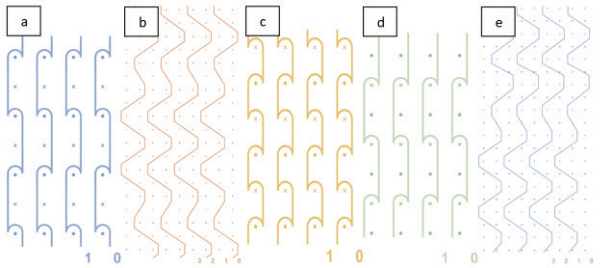
Rys. 8. Wyniki symulacji charakterystyki promieniowania anteny w płaszczyźnie poziomej G(φ,90°)

2. PROTOTYP ANTENY

Prototyp anteny wykonany został z materiałów tekstylnych połączonych za pomocą kleju polimerowego, nanoszonego za pomocą techniki rozpylania. Podłoże anteny tekstylnej stanowiła poliestrowa dzianina dystansowa o grubości 8 mm i masie powierzchniowej 449 g/m2 wytworzona na osnowarce dwugrzebieniowej firmy Karl Mayer. Technologia kolumienkowych dzianin dystansowych znana jest w literaturze i przemyśle od wielu lat. Tak wytworzone struktury dziane znajdują zastosowanie w wielu gałęziach przemysłu takich jak: produkcja materacy, obuwia, meblarstwo i tapicerstwo, przemysł samochodowy, medycyna czy odzieżownictwo. Dzianina ta, zwana również dzianiną 3D, składa się z dwóch warstw zewnętrznych zbudowanych ze splotów podstawowych oraz warstwy wewnętrznej utworzonej z odcinków nitek przebiegających pomiędzy zewnętrznymi warstwami dzianiny (tzw. dystansu). W przypadku zastosowanej dzianiny dystans został wykonany z nici monofilamentowej o średnicy równej 0,24 mm. Widok próbki pokazano na Rys. 9, wszystkie sploty składowe tworzące dzianinę dystansową przedstawiono na Rys. 10. Charakter warstwy wierzchniej definiują sploty a i b, warstwy spodniej sploty d i e natomiast splot c (splot łańcuszka dwuprawego) definiuje dystans między warstwami.



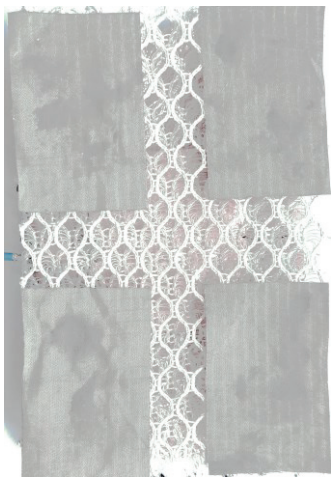
Rys. 9. Zdjęcia dzianiny 3D: a) widok z góry pokazujący splot warstwy wierzchniej (powiększenie 2,0x), b)przekrój poprzeczny wzdłuż wątków (powiększenie 3,0x).



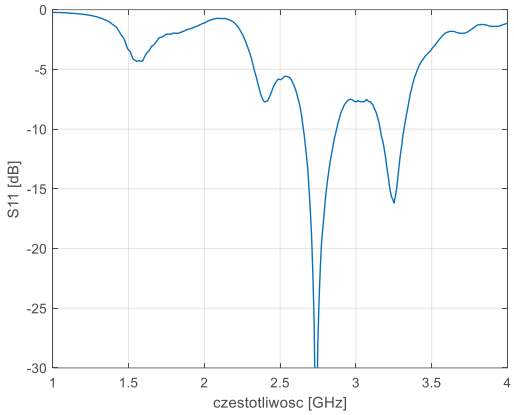
Rys. 10. Sploty składowe dzianiny dystansowej a) łańcuszek o oczkach otwartych wytwarzany na przednim grzebieniu igielnicowym (GB1), b) splot wątkowy tworzony na przednim grzebieniu igielnicowym – nawleczenie igieł: 1 pusta, 1 nawleczona (GB2), c) łańcuszek o oczkach otwartych tworzony na igłach przedniego i tylnego grzebienia (GB3), d) łańcuszek o oczkach otwartych , wytwarzany na tylnym grzebieniu igielnicowym (GB4), e) splot wątkowy produkowany na tylnym grzebieniu igielnicowym (GB5).

Do realizacji elementów przewodzących zastosowano polimerową tkaninę przewodzącą domieszkowaną srebrem. Antenę zasilono sondą koncentryczną, która została połączona z materiałem przewodzącym za pomocą kleju elektroprzewodzącego. Prototyp anteny przedstawiono na Rys. 11.

Pomiar dopasowania impedancyjnego anteny przeprowadzono dla anteny umieszczonej na ciele człowieka. Płaszczyzna kalibracji analizatora obwodów przesunięta została do punktu połączenia sondy koncentrycznej z materiałem tekstylnym. Wyniki pomiarów dopasowania impedancyjnego anteny przedstawiono na Rys. 12. W założonym paśmie częstotliwości 2,4 – 2,5 GHz, moduł parametru S11 był mniejszy od -5 dB. W tym przypadku minimum charakterystyki przesunięte jest do częstotliwości 2,72 GHz. Uzyskane wyniki pomiarów odbiegają od wyników symulacji, a poziom dopasowania impedancyjnego w paśmie 2,4 GHz jest gorszy od zakładanego.



Rys. 11. Wyniki symulacji dopasowania impedancyjnego anteny uzyskanej w procesie optymalizacji



Rys. 12. Wyniki symulacji dopasowania impedancyjnego anteny uzyskanej w procesie optymalizacji

3. PODSUMOWANIE I WNIOSKI

W artykule przedstawiono wyniki wstępnych prac nad anteną tekstylną przeznaczoną do pracy na ciele człowieka w pasmie 2,4 GHz.

Na podstawie analizy wyników symulacji komputerowych przeprowadzonych z uwzględnieniem wpływu ciała człowieka na antenę można stwierdzić, że zaprojektowana antena charakteryzuje się dobrym dopasowaniem impedancyjnym w założonym paśmie oraz posiada względnie duży zysk przekraczający 6 dBi. Płaska konstrukcja anteny umożliwia jej wykonanie w technologii tekstylnej z wykorzystaniem dzianin dystansowych i materiałów tekstylnych przewodzących prąd elektryczny.

Wyniki pomiarów dopasowania impedancyjnego opracowanego prototypu są niestety rozbieżne z wynikami symulacji komputerowych. Zaobserwowano przesunięcie minimum parametru S_{11} w kierunku wyższych częstotliwości. Może to wynikać z rozbieżności pomiędzy parametrami materiałów zastosowanych do wykonania prototypu oraz parametrami uwzględnionymi podczas symulacji komputerowych. W symulacjach nie uwzględniono kleju łączącego komponenty anteny, co również może wpływać na uzyskane wyniki. Ponadto w procesie naklejania łatek przewodzących na podłoże z dzianiny dystansowej doszło do ich przesunięcia względem założonego w projekcie położenia. Zaobserwowano również nierównomierności w warstwie kleju oraz jego częściową penetrację w głąb materiałów tekstylnych, co jest widoczne w postaci nierównomiernych zaciemnień łatek przewodzących na Rys. 11. Te zjawiska mogą uzasadniać rozbieżności pomiędzy wynikami symulacji i pomiarów.

Wyniki uzyskane w procesie symulacji komputerowych opracowanej anteny zachęcają autorów do udoskonalenia procesu wykonania prototypu. Dalsze prace w tym zakresie koncentrują się obecnie na wyeliminowaniu wymienionych wyżej niedoskonałości.

LITERATURA

[1] https://www.handsontec.com/dataspecs/module/ESP32-CAM.pdf

[2] Pradhan, Omkar, Kimberly Newman, and Frank Barnes. “Parametric Analysis of Meandered Inverted-F Antenna and Use of a High Impedance Surface Based Ground Plane for WBAN Applications.” In *2013 IEEE International Conference on Body Sensor Networks*, 1–7. Cambridge, MA, USA: IEEE, 2013.

[3] Cheung, Cheuk Yin, Joseph S. M. Yuen, and Steve W. Y. Mung. “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* 2018 (2018): 1–5.

[4] Rudyk Karol, Kubacki Roman, Lamari Salim, “Antena ultra-szerokopasmowa ze strukturą periodyczną.” *PRZEGLĄD ELEKTROTECHNICZNY* 1, no. 12 (December 5, 2016): 151–54.

[5] Lee, Kyoung-Joo, Juntaek Oh, and Sang-Hwa Yi. “Stacked Microstrip Patch Antenna Array with High-Gain and Improved Thermal-Stability for Microwave Power Transmission Applications.” In 2018 Asia-Pacific Microwave Conference (APMC), 1432–34. Kyoto: IEEE, 2018. https://doi.org/10.23919/APMC.2018.8617391.

[6] H. Legay and L. Shafai, “New Stacked Microstrip Antenna with Large Bandwidth and High Gain,” IEE Proc. Microwave, Antennas and Propagation, Pt. H, vol. 141, no. 3 (June 1994), pp. 199–204.

[7] Hausman S, Januszkiewicz Ł, Michalak M, Kacprzak T, Krucińska I. High Frequency Dielectric Permittivity of Nonwovens. Fibers & Textiles in Eastern Europe, Vol. 14, nr.5(59), 2006

CHARAKTERYSTYKA ODBICIOWA INTELIGENTNYCH MATRYC ANTENOWYCH

REFLECTION CHARACTERISTICS OF RECONFIGURABLE INTELLIGENT SURFACES

Paweł Hatka[1]; Dawid Brząkała[1]; Marcel Garczyk[1]; Paweł Płaczkiewicz[1]; Marta Sieradzka[2]; Cyryl Prentki[1]; Tomasz Jaworski[1]; Krzysztof Cichoń[3]; Adrian Kliks[3];

[1] Politechnika Poznańska, Poznań, imie.nazwisko@student.put.poznan.pl
[2] Politechnika Poznańska, Poznań, marta.sieradzka.1@student.put.poznan.pl
[3] Politechnika Poznańska, Poznań, imie.nazwisko@put.poznan.pl

DOI: 10.15199/59.2024.4.57

Streszczenie: W artykule przedstawiono wyniki pomiarów charakterystki odbiciowej dla rekonfigorowalnych matryc antenowych (RMA). Pomiary zostały przeprowadzone w środowisku nieidealnym, tj. typowym dla późniejszego praktycznego zastosowania. Podczas eksperymentów wykorzystano popularne matryce zrealizowane w projekcie otwartoźródłowym OpenSourceRIS. W badaniu skupiono się na uzyskaniu dwóch rodzajów charakterystki odbiciowej - dwu- i trójwymiarowej.

Abstract: This paper presents the results of reflection characteristics measurements for reconfigurable intelligent surfaces (RIS). The measurements were carried out in a non-ideal environment, i.e. typical for subsequent practical use of RISes. During the experiments, popular matrices implemented in the open-source project OpenSourceRIS were used. The study focused on obtaining two types of reflection characteristics - two- and three-dimensional.

Słowa kluczowe: Rekonfigurowalne Matryce Antenowe, Charakterystyka Odbiciowa, Pomiary, 6G

Keywords: Reconfigurable Intelligent Surfaces, Reflection Characteristics, Measurements, 6G

1. WPROWADZENIE

Jednym z rozwiązań szeroko rozpatrywanych w kontekście rozwoju sieci bezprzewodowych jest zastosowanie rekonfigurowalnych matryc antenowych (RMA), znanych w literaturze anglojęzycznej jako Reconfigurable Intelligent Surfaces (RIS) [1, 2]. Matryce te, zgodnie z założeniem, są kontrolowalnymi elementami umieszczonymi w środowisku propagacyjnym (tzn. pomiędzy nadajnikiem i odbiornikiem). Wspomniana możliwość zdalnej kontroli (np. przez operatora sieci bezprzewodowej) pozwala na zmianę charakterystki transmisyjnej środowiska radiowego w taki sposób, aby np. zapewnić uzyskanie oczekiwanych efektów propagacyjnych po stronie odbiornika [7]. W szczególności RMA pozwalać może na programowe definiowanie kąta odbicia padającego na matrycę sygnału (innego niż wynikałoby to z np. optyki geometrycznej) czy na częściową albo całkowitą absorpcję sygnału. Takie działanie może w konsekwencji skutkować poprawą określonych miar jakości sygnału po stronie odbiornika, np. zwiększenie mocy pożądanego sygnału odbieranego czy zminimalizowanie poziomu sygnałów interferencyjnych [6].

Macierze RMA są elementami określanymi często jako quasi-pasywne, tzn. same elementy antenowe są pasywne (w szczególności nie ma zastosowanego wzmacniacza), a zużycie mocy przez płytę RMA związane jest głównie z działaniem kontrolera [2]. Macierze posiadają dużą liczbę elementów antenowych, co bezpośrednio przekłada się na możliwości oddziaływania RMA na sygnał padający. Każdy element można bowiem - w zależności od użytego rozwiazania - ustawić w różny tryb pracy. Im większa liczba składowych elementów antenowych tym większa możliwość uzyskania różnych wzorców odbiciowych dla takiej matrycy - jeżeli każdy element antenowy mógłby pracować w jednym z M stanów (takim stanem jest np. wartość wprowawdzanego przesunięcia fazowego), to dla N-elementowej matrycy RMA można wygenerować M^N różnych wzorców. Przestrzeń możliwości jest zatem bardzo duża. W praktycznym rozwiązaniu należałoby raczej skupić się na zdefiniowaniu bogatej, ale jednocześnie istotnie ograniczonej, książki kodowej wzorców, gdzie dla każdego wzorca przypisana byłaby odpowiednia charakterystyka odbiciowa płyty RMA.

Celem niniejszej pracy jest przedstawienie wyników pomiarów laboratoryjnych w środowisku typowym (tzn. nieidealnym) charakterystyk odbiciowych płyt RMA w dwóch wariantach - dwuwymiarowym oraz trójwymiarowym. Do celów pomiarowych wykorzystano matryce opracowane przez naukowców z Kolonii, zrealizowane w ramach projektu OpenSourceRIS [5].

2. ŚRODOWISKO POMIAROWE

W pracy rozważono dwa sceniariusze pomiarowe. Celem pierwszego z nich było zmierzenie dwuwymiarowej charakterystyki odbiciowej RMA (charakterystyka horyzontalna). W drugim scenariuszu z kolei rozszerzono zakres pomiarowy o trzeci wymiar (wertykalny), uzyskując dzięki temu możliwość zbadania charakterystyki trójwymiarowej. W celu przeprowadzenia wspomnianych badań w laboratorium zestawiono następujacy układ pomiarowy. Sygnał o częstotliwości 5.5 GHz oraz mocy -10 dBm wytwarzany był za pomocą generatora Rohde & Schwarz

SMM100A i transmitowany w kierunku macierzy RMA za pomocą anteny kierunkowej METIS 5.1-5.8 GHz. Po stronie odbiorczej zastosowano natomiast analizator widma Rohde & Schwarz FSV3000, podłączony do takiej samej anteny. Poszczególne wartości nastaw analizatora są przedstawione w Tab. 1.

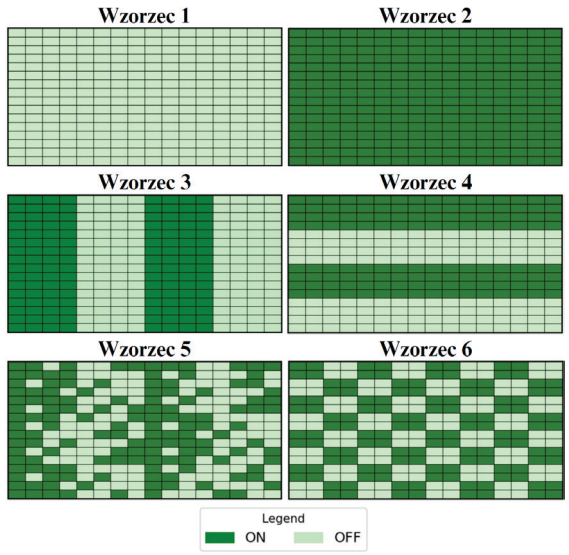
Tabela 1: Nastawy analizatora widma

Parametr	*Wartość*	*Jednostka*
Span	0	[Hz]
RBW	500	[Hz]
Sweep	50	[ms]
Reference level	-30	[dBm]
Częst. środkowa	5.5	[GHz]
Poziom szumów	-110	[dBm]
Tryb	MaxHold	

Obie anteny skierowane były na macierz RMA, dzięki czemu możliwe było zmierzenie poziomu mocu odbitej obserwowanej w odbiorniku. W celu oceny dwu- i trójwymiarowej charakterystyki odbiciowej należało zapewnić możliwość obrotu matrycy w płaszczyście poziomej i pionowej. Podczas ekpsperymenty zmiana kąta azymutu i elewacji umożliwiło specjalnie zaprojektowane i wykonane ramię wraz z silnikiem i układem sterującym. Matryca RMA posiada 16×16 elementów odbijających i sterowanych indywidualnie za pomocą jednego bitu, tzn. każdy element może odbić padający sygnał bez zmiany jego fazy, ale może także zmienić ją o 180 stopni. Należy podkreślić, że nawet tak prosta zmienność stwarza bardzo szeroki wachlarz wariantów możliwych do wygenerowania, tj. 2^{256}. Cały proces pomiaru został zautomatyzowany przy pomocy oprogramowania napisanego w języku Python, które umożliwiło zdalne sterowanie analizatorem widma, generatorem sygnałowym, głowicą ramienia przytrzymującego matrycę, oraz zmianę wzorca ustawionego na RMA [3]. Wspomniany wzorzec określa, które z 256 elementów matrycy zmieniają fazę padającego sygnału o 180 stopni (dla f $\approx$ 5.5GHz), tzn. które elementy są nieaktywne. Wspomniane 256 elementów zostało podzielone na 64 grupy każda po cztery elementy; w jednym wierszu są więc cztery podgrupy. Sterowanie matrycą polega więc na wybraniu, które z czterech elemntów w jednej z grup mają, a które nie mają odwracać fazy. Numeracja grup zaczyna się od indeksu 0, a pierwsza grupa znajduje się w lewym górnym rogu matrycy. Wybór realizowany jest poprzez wysłanie do RMA ciągu liczb w kodzie szesnastkowym, których pozycja w tym ciągu określa daną grupę. Zatem 0 - 0000 oznacza, że wszystkie elementy danego wiersza są wyłączone, F - 1111, że są włączone, a 3 - 0011, że dwie ostanie włączone itd. Przykładowy ciąg: !0X800...000 - oznacza że zostanie włączony tylko pierwszy element grupy 0 [4]. Przeprowadzono pomiary dla 27 wybranych wzorców, lecz ze względu na ograniczenia pojemności artykułu, w pracy zostaną przedstawione wyniki dla wybranych wzorców przedstawionych na Rys. 1.

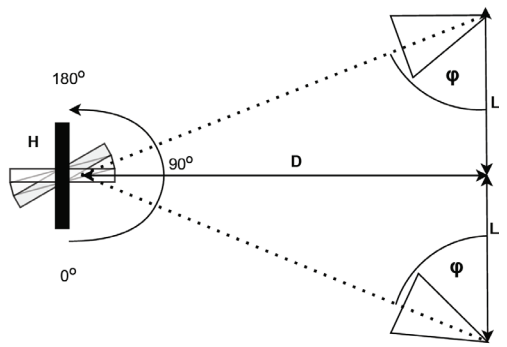
2.1 Charakterystyka w jednej płaszczyźnie

Pomiary charakterystyki w płaszczyźnie poziomej w jakiej RMA odbija fale radiowe zostały wykonane zgodnie ze schematem (Rys. 2).



Rysunek 1: Wizualizacja wzorców.

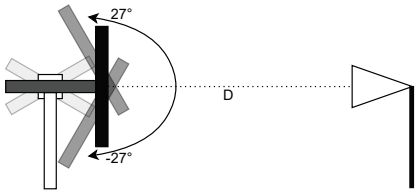
Jak wspomniano, RMA została zamontowana na obrotowej głowicy umożliwiającej ustawianie urządzenia pod zdanym kątem azymutu ϕ względem anteny nadawczej oraz odbiorczej. Odległość pomiędzy linią anten a linią matrycy RMA wynosi $D = 1.5$ m, natomiast odległość pomiędzy antenami wynosiła $2 * L = 2$ m, co w wyniku daje kąt skierowania anten na RMA $\varphi = 56.3°$. W celu otrzymania dokładnych wyników głowica wraz z zamontowaną matrycą wykonywała obrót z krokiem 1.8° w płaszczyźnie azymutu, zaczynając od kąta 0° a kończąc na kącie 180°. Osiągając kąt 90° płaszczyzna matrycy była skierowana równolegle do lini anten. Anteny i matryca były umieszczone na wysokości 1.3 m.



Rysunek 2: Schemat pomiaru charakterystyki odbiciowej w jednej płaszczyźnie.

2.2 Charakterystyka trójwymiarowa

Pomiary charakterystki przestrzennej zostały wykonane w sposób analogiczny dla charakterystyki 2D, jednak w tym przypadku wykorzystany został również obrót w płaszczyźnie elewacji zgodnie ze schematem przedstawionym na Rys. 3. Pomiary wykonane zostały z rozdzielczością 1.8° od 0° do 180° w płaszczyźnie azymutu oraz 9° od -27° do 27° w płaszczyźnie elewacji. Rozmieszczenie poszczególnych urządzeń wykorzystanych do przeprowadzenia pomiaru było identyczne jak w przypadku charakterystyki w jednej płaszczyźnie.

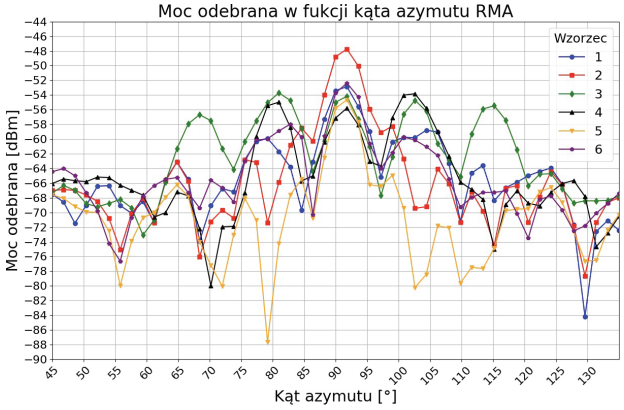


Rysunek 3: Schemat pomiaru charakterystyki przestrzennej.

3. WYNIKI POMIARÓW

3.1 Scenariusz pierwszy

Na Rys. 4 przedstawiono wyniki pomiarów charakterystyki odbiciowej w jednej płaszczyźnie dla wzorców przedstawionych na Rys. 1.

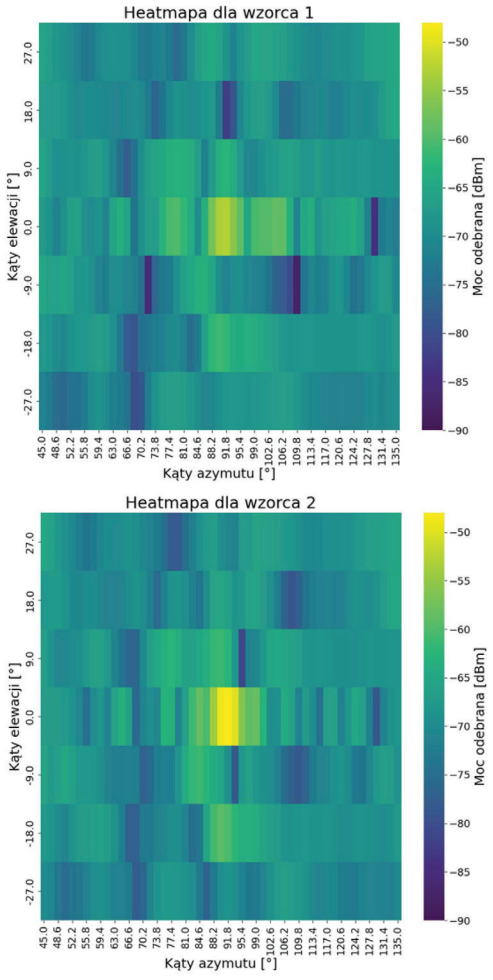


Rysunek 4: Wyniki pomiarów charakterystki odbiciowej

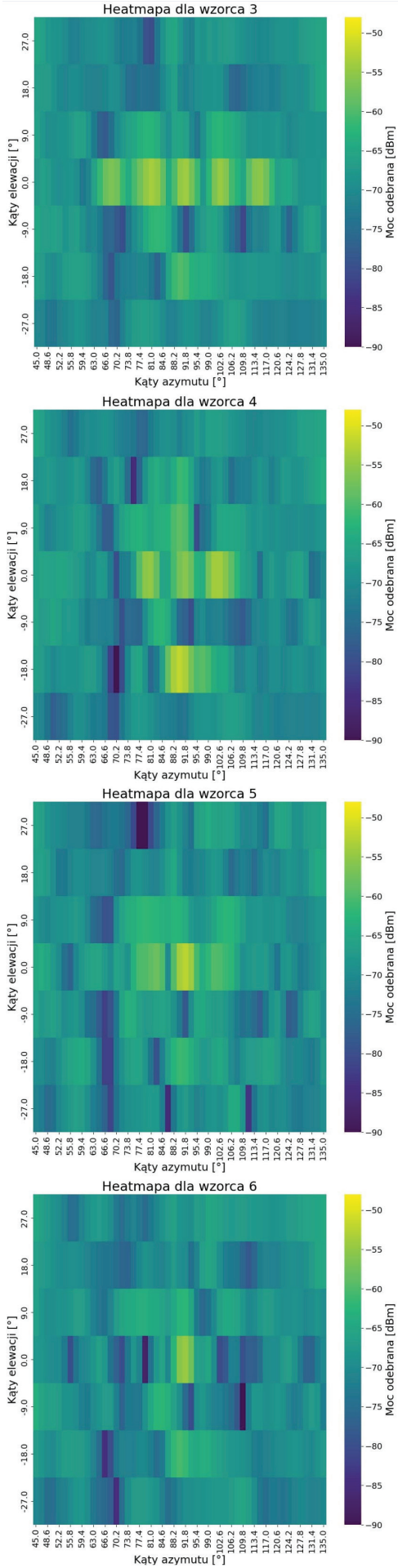
Dla wzorca pierwszego, zgodnie z oczekiwaniami, maksimum mocy odebranej występuje dla kąta 90°, zatem dla sytuacji, w której RMA wiązkę padającą odbija pod kątem równym kątowi padania. Wykorzystując wzorzec drugi widoczny jest wzrost mocy odebranej dla kąta 90° o około 5dB względem wzorca pierwszego, efekt ten wynika z aktywowania elemntów macierzy. Dla pozostałych kątów zauważalny jest brak poprawy albo wręcz spadek mocy odebranej. Dla wzorca 3 i 4 maksimum charakterystyki występuje dla kąta innego niż 90°. Wzorzec trzeci przedstawia sytuację, gdzie elementy RMA były naprzemiennie włączone/wyłączone tworząc szyk grubych pasów pionowych. Dla takiego ustawienia RMA oraz pionowej polaryzacji anten zostały uzyskane interesujące wyniki w postaci kilku maksimów odbieranej mocy sygnału, oddzielonych "zerami", które pojawiały się co 10°-12° kąta azymutu. Podobna sytuacja występuje dla wzorca czwartego czyli grubych pasów poziomych, gdzie uzyskujemy trzy maksima dla kątów azymutu 80°, 90° oraz 105°. W przypadku wzorca 5, możemy zaobserwować natomiast kilka punktów charakterystyki, dla których moc odebrana jest znacznie niższa, niż w przypadku innych wzorców. Możemy zaobserwować, że charakterystyka nie jest również symetryczna dla tego przypadku. Wyniki dla wzorca 6 ukazują występowanie przedziałów, w których charakterystyka wypłaszcza się, a poziom mocy odbieranej w całym zakresie kątowym różni się maksymalnie o ok. 2 dB. Pierwsze takie miejsce występuje miedzy kątem 59° a 66°, a drugie między 77.5° a 85°. Dla wszystkich wzorców poza 5 - losowym, kolejne ekstrema charakterystyki występują często mniej więcej symetrycznie względem położenia 90°, choć ich wartości co do mocy nie są identyczne ze względu na charakter środowiska pomiarowego.

3.2 Scenariusz drugi

Otrzymane wyniki z pomiaru charakterystyki 3D zostały przedstawione w formie wykresów typu *heatmap* (Rys. 5-6). Analiza wyników scenariusza drugiego skupia się na kątach elewacji innych niż 0 stopni ze względu na wyniki pokrywające się ze scenariuszem pierwszym. Dla każdego z analizowanych przypadków można zaobserwować spadek mocy odebranej dla kąta ok. 70° w płaszczyźnie azymutu dla każdego kąta elewacji. Wynika to z tego, że pomiary były wykonywane w nieidealnym środowisku na wyniki te mogą składać się również komponenty wielodrogowości - odbicia od ścian itp. Na wykresach otrzymanych dla wzorca 1 oraz 2 można zauważyć, że rozkład mocy odbieranej jest podobny, przy czym dla wzorca drugiego, czyli dla wszystkich elementów RMA włączonych, odbierany sygnał jest mocniejszy. W przypadku wzorca 1 dla kąta -9° elewacji występują 2 minima charakterystyki, natomiast dla drugiego wzorca pojawia się po jednym minimum dla kątów 9° i -9°. Dla wzorca 3 możemy zaobserwować większą liczbę minimów dla kąta elewacji równego -9°. Ciekawym przypadkiem wzmocnienia sygnału cechuje się wzorzec 4 dla kąta elewacji -18° i 90° kąta azymutu.



Rysunek 5: Wyniki pomiarów charakterystki odbiciowej dla różnyh kątów elewacji dla wzorców 1 i 2.



Rysunek 6: Wyniki pomiarów charakterystki odbiciowej dla różnyh kątów elewacji dla wzorców 3 - 6.

W kierunku tym występuje spore wzmocnienie sygnału względem centralnego punktu wykresu. Dla kątów elewacji -9° i 9° zauważamy takie kąty azymutu, dla którch moc sygnału odbieranego osłabiona jest o ponad 20 dB względem maksymalnego położenia. Losowo wygenerowany wzorzec stworzył sytuację, w której na kącie elewacji -27° pojawiły się dwa wyraźne osłabienia sygnału na kącie 86° i 111° o ok. o 10dB.

4. ZAKOŃCZENIE

W pracy przedstawiono wyniki pomiarów dwu- i trójwymiarowej charakterystyki odbiciowej modułów RMA w warunkach laboratoryjnych w środowisku typowym. Uzyskane wyniki pokazują dużą zmienność charakterystyk w zależności od wartosci kąta azymutu oraz elewacji. Wahania mocy mogą dochodzić nawet do 20 dB. Pozwala to stwierdzić, że istnieje możliwość wykorzystania wzorców RMA jako swoistej książki kodowej.

5. PODZIĘKOWANIA

Praca powstała w ramach projektu badawczego OPUS 2021/43/B/ST7/01365 finansowanego przez Narodowe Centrum Nauki 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] 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: June 28, 2023.

[3] GitHub. https://github.com/RISurfaces/RIS_Python_KRiT

[4] GitHub. https://github.com/mheinri/OpenSourceRIS

[5] 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. doi: 10.1109/ISAP57493.2023.10389095.

[6] Liu, R., Dou, J., Li, P., Wu, J., Cui, Y. 2022. "Simulation and Field Trial Results of Reconfigurable Intelligent Surfaces in 5G Networks". IEEE Access 10: 122786-122795. doi: 10.1109/ACCESS.2022.3223447.

[7] Renzo, M.D., 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

Rozwój systemów generacji i dystrybucji sygnałów referencyjnych w.cz. dla liniowych akceleratorów cząstek elementarnych

Development of Phase Reference Distribution Systems of Linear Particle Accelerators with Femtosecond Stability

Maciej Urbański[1], Krzysztof Czuba[2]

Institute of Electronic Systems, Warsaw University of Technology
[1] maciej.urbanski@pw.edu.pl, [2] krzysztof.czuba@pw.edu.pl

...

DOI: 10.15199/59.2024.4.58

Streszczenie:

Niniejszy artykuł stanowi streszczenie rozprawy doktorskiej pod tym samym tytułem, zrealizowanej w Instytucie Systemów Elektronicznych Politechniki Warszawskiej. Artykuł przedstawia prace dotyczące rozwoju elementów systemów generacji i dystrybucji wzorcowych sygnałów synchronizacji dla akceleratorów European XFEL, SINBAD, FLASH, działających w ośrodku DESY w Niemczech. Przedstawiono opracowania nowych i istotnie zmienionych elementów wspomnianych systemów oraz opracowaną metodykę projektowania takich elementów.

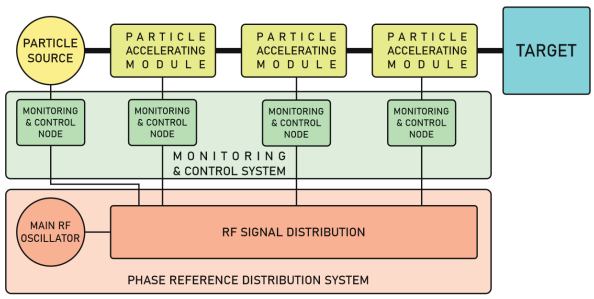
Abstract: The paper presents a PhD thesis realized at the Institute of Electronic Systems, Warsaw University of Technology, in a brief form. Topics covered include develompent of phase reference distribution systems for linear particle accelerators, like European XFEL, SINBAD, FLASH in DESY, Germany. New or significantly modified modules of these systems, and their design methodology, are presented.

Słowa kluczowe: Dryf fazy, liniowy akcelerator cząstek elementarnych, stabilizacja fazy, system generacji i dystrybucji sygnału synchronizacji.

Keywords: Phase drift, linear particle accelerator, phase stabilization, phase reference distribution system.

1. WSTĘP

Niniejszy artykuł stanowi streszczenie rozprawy doktorskiej, obronionej w 2023 roku na Politechnice Warszawskiej [9], a tematyka w nim zawarta została przedstawiona w publikacjach [8], [10],[11]. Rozprawa traktuje o rozwoju elementów i systemów generacji i dystrybucji wzorcowego sygnału synchronizacji w liniowych akceleratorach cząstek elementarnych. Uproszczony schemat blokowy typowego liniowego akceleratora cząstek przedstawiono na Rysunku 1 [8],[9].



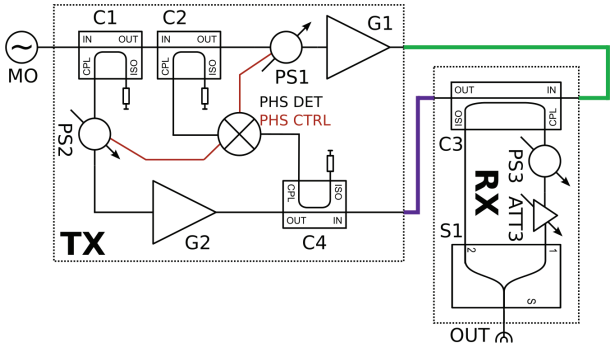
Rysunek 1: Uproszczony schemat blokowy liniowego akceleratora cząstek elementarnych [9].

Źródło cząstek dostarcza wiązkę protonów lub elektronów do kolejnych sekcji przyspieszających, sterowanych elektromagnetycznym polem w.cz wielkiej mocy. Sekcje te stopniowo zwiększają energię wiązki i kierują ją do urządzeń docelowych, zlokalizowanych zwykle w odpowiednim pomieszczeniu badawczym. Za sterowanie polem przyspieszającym w.cz odpowiadają systemy LLRF (ang. Low Level Radio Frequency) i HPRF (ang. High Power Radio Frequency) [8], oznaczone kolorem zielonym. W miarę poruszania się wzdłuż akceleratora wiązka jest naturalnie formowana w tzw. zgęstki. Aby zgęstki te mogły być przyspieszane, a nie spowalniane, położenie czoła zgęstki na wejściu każdej struktury przyspieszającej musi być w odpowiedniej relacji do fazy sterującego pola elektromagnetycznego w danej strukturze. Relacja ta musi być ściśle określona, monitorowana i stabilizowana z rozdzielczością sięgającą dziesiątek femtosekund lub mniejszą, na całej długości akceleratora, nierzadko przekraczającej kilkaset metrów, lub nawet kilometr. Aby to osiągnąć, konieczny jest wzorcowy generator, lub syntezer sygnału w.cz (tzw. sygnału synchronizacji), o doskonałej stałości fazowej - krótkoterminowej opisanej przez szum fazowy lub jitter oraz długoterminowej, opisanej przez dryft fazy medium transmisyjnego [3] użytego w systemie generacji i dystrybucji wzorcowego sygnału synchronizacji (ang. PRDS - Phase Reference Distribution System)[2], [5], [7], [15]. Opisane wyżej niestałości fazy wpływają na rozproszenie i ograniczenie energii wiązki cząstek uderzającej w cel [4], [6].

2. OPRACOWANE ELEMENTY SYSTEMÓW I METODYKA ICH PROJEKTOWANIA

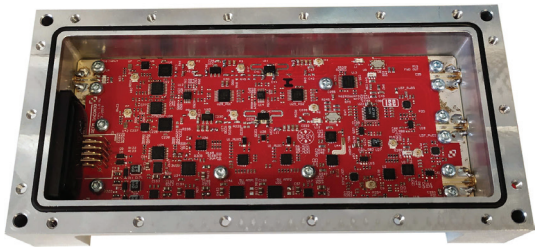
2.1 Elementy systemu PRDS dla SINBAD

W ramach prac związanych z systemem PRDS dla akceleratora SINBAD [9] opracowano m.in układ łącza z interferometryczną stabilizacją fazy, przedstawiony na Rysunku 2, bazujący na rozwiązaniach opisanych w [7].



Rysunek 2: Schemat blokowy pojedynczego łącza interferometrycznej stabilizacji fazy [9].

Łącze zostało przystosowane do pracy z sygnałem wzorcowym o częstotliwości 3 GHz. Zasada stabilizacji fazy sygnału wyjściowego jest oparta o dwa warunki. Pierwszy warunek to ciągłe monitorowanie i kompensowanie dryftów fazy w przewodach w.cz łącza (oznaczonych kolorami niebieskim i zielonym). Dzięki temu sygnał z generatora wzorcowego jest dostarczany na wejście odbiornika łącza, po przejściu drogi do odbiornika i z powrotem. Nadajnik, przedstawiony na Rysunku 3, na podstawie zmierzonego dryfu fazy w łączu wytwarza sztuczny sygnał odbity [12] i dostarcza go na drugie wejście odbiornika. W odbiorniku realizowany jest drugi warunek poprawnej pracy, czyli wyrównanie faz obu sygnałów wejściowych i ich sumowanie wektorowe. Dzięki takiemu ustawieniu układu minimalizowany jest wpływ dryfu fazy w łączu [7],[12].

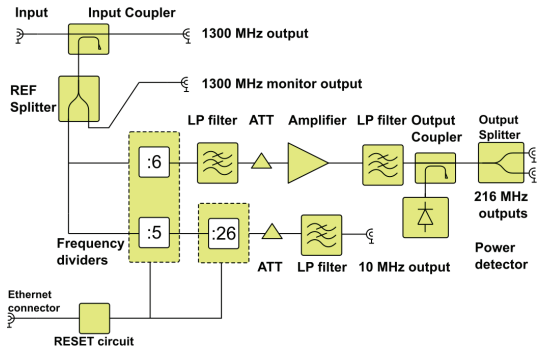


Rysunek 3: Zdjęcie modułu nadajnika łącza interferometrycznej stabilizacji fazy [9], [12].

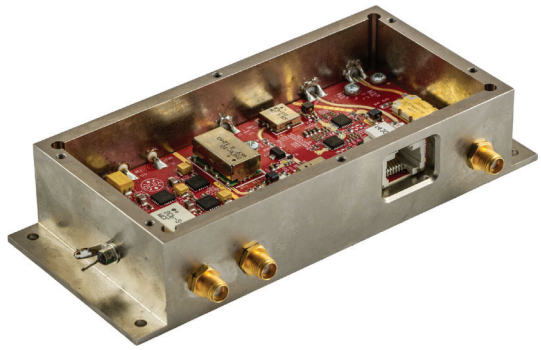
W trakcie badań modułów łącza przeprowadzono także analizę ograniczeń precyzji detekcji fazy, wprowadzanych przez elementy składowe modułów, głównie przez użyte w modułach analogowe scalone detektory fazy w.cz. Analiza ta wykazała, że obecnie dostępne analogowe metody detekcji fazy sygnałów w.cz, zarówno w postaci scalonych detektorów fazy, czy mieszaczy, dla sygnału o częstotliwości 3 GHz nie umożliwiają osiągnięcia precyzji detekcji dryftu fazy lepszej niż około 40 femtosekund, z uwagi na wprowadzany własny dryft sygnału wyjściowego.

2.2 Elementy systemu PRDS dla European XFEL

Na potrzeby systemów PRDS i LLRF akceleratora European XFEL opracowano szereg rozwiązań, takich jak łącze z interferometryczną stabilizacją fazy [7], czy układy precyzyjnych dzielników częstotliwości [13], których schemat blokowy przedstawiono na Rysunku 4, a zdjęcie modułu na Rysunku 5. Układ dzielnika jest wtrącany do toru głównego sygnału synchronizacji o częstotliwości 1.3 GHz i wytwarza zsynchronizowane z nim sygnały o częstotliwościach 216 MHz i 10 MHz, stosowane m.in. w układach timing'u [8]. Na potrzeby E-XFEL wykonano łącznie około 70 modułów, będących w ciągłym użyciu.



Rysunek 4: Schemat blokowy układu precyzyjnego dzielnika częstotliwości sygnału wzorcowego 1.3 GHz [9].

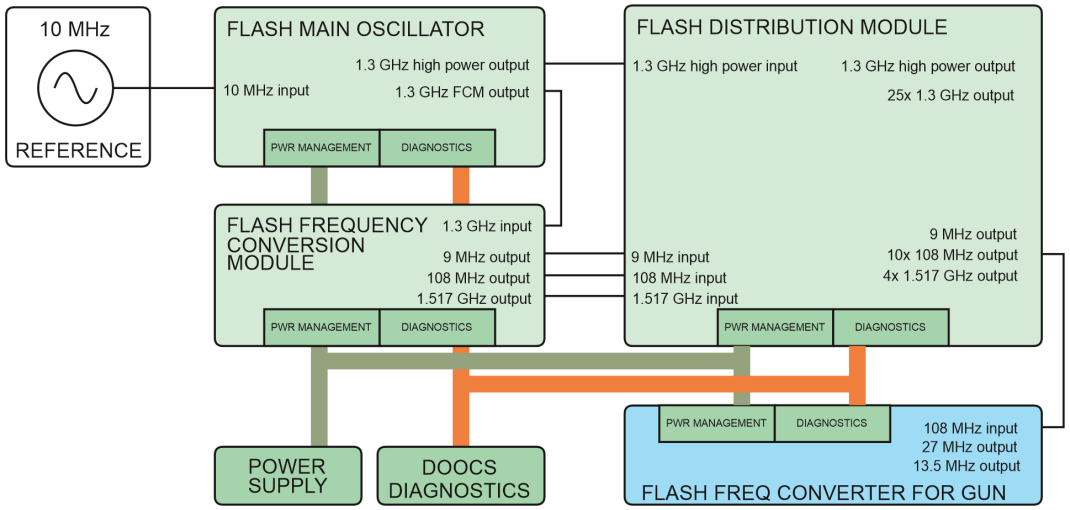


Rysunek 5: Zdjęcie opracowanego modułu dzielnika częstotliwości [9].

2.3 Elementy systemu PRDS dla FLASH2020+

Na potrzeby programu modernizacji (FLASH2020+) akceleratora FLASH w DESY zrealizowano nowy system generacji i dystrybucji wzorcowego sygnału synchronizacji. Poprzednio stosowany system z upływem lat przestał spełniać stawiane mu wymagania, głównie ze względu na starzenie elementów krytycznych, takich jak oscylatory kwarcowe. Nowy system, którego schemat blokowy przedstawiony jest na Rysunku 6, został opracowany na podstawie analogicznego rozwiązania, zrealizowanego dla European XFEL [14].

System składa się z czterech modułów składowych, z których trzy (zaznaczone na zielono) były opracowane w Instytucie Systemów Elektronicznych, we współpracy

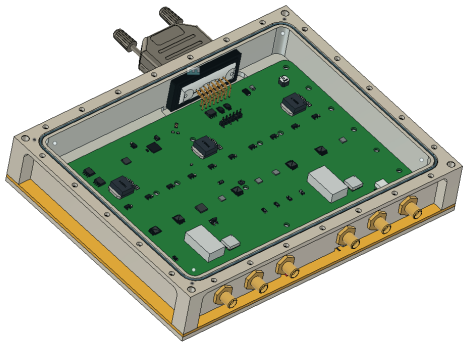


Rysunek 6: Schemat blokowy nowego systemu PRDS dla akceleratora FLASH2020+ [9], [11].

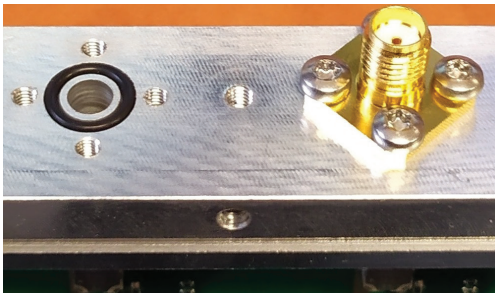
z DESY. Konstrukcja nowych modułów jest kompatybilna z modułami opracowanymi dla akceleratora European XFEL, a elementy sterujące i diagnostyczne są identyczne w obu rodzajach urządzeń. Struktura systemu jest zrealizowana w sposób możliwie uniwersalny, przez co łatwo jest go dostosować do wykorzystania w systemie sterowania innym akceleratorem. Zasada działania opiera się o bardzo precyzyjne źródło sygnału 10 MHz, które jest wykorzystane do syntezy wszystkich kolejnych częstotliwości. Moduł syntezera wzorcowego (ang. *Main Oscillator*) wytwarza sygnał o częstotliwości 1.3 GHz. Układ konwertera częstotliwości (ang. *Frequency Conversion Module*) syntezuje częstotliwości 9, 108 i 1517 MHz, zachowując synchronizację z sygnałem głównym 1.3 GHz. Osobny moduł zapewnia dystrybucję i diagnostykę wszystkich sygnałów wyjściowych. Opracowane rozwiązanie zapewnia sygnał o niezwykle małym poziomie szumów fazowych - jitter obliczony w zakresie częstotliwości od 10 Hz do 1 MHz wynosi około 10.7 fs, co było w momencie powstania systemu jednym z najlepszych na świecie wyników [9]. System został skomercjalizowany.

2.4 Metodyka projektowania elementów systemów synchronizacji

Do realizacji układów dla systemów PRDS akceleratorów cząstek opracowano ogólną metodykę projektowania modułów wrażliwych, lub wysokostabilnych fazowo [9]. Na podstawie doświadczeń zdobytych przy projektowaniu systemów dla akceleratorów European XFEL, FLASH i ESS możliwe było określenie najistotniejszych problemów, które powodują ograniczenie ostatecznych parametrów systemu PRDS. Przedstawiona metodyka została wykorzystana m.in. podczas projektowania elementów systemu PRDS dla FLASH2020+, opisanego wcześniej. Elementy syntezera częstotliwości, wrażliwe na wibracje, zostały odizolowane od źródeł wibracji mechanicznych, niekontrolowanych przepływów powietrza w urządzeniu i źródeł dużych zmian temperatury. Elementy o dużej masie, takie jak radiatory i wzmacniacze mocy, zostały umieszczone na dnie obudowy urządzenia i wykorzystane jako baza do montażu kolejnych, wrażliwych na wibracje części. Moduły stabilne fazowo muszą być umieszczone w obudowach hermetycznych, ekranujących i umożliwiających stabilizację termiczną. W tym celu opracowano ogólny model projektowania obudów, który przedstawiono na Rysunku 7. Obwód drukowany może mieć elementy tylko po jednej stronie i jest przykręcony do dna metalowej obudowy. Obudowa jest zamykana uszczelnianym wieczkiem. Ściany boczne są dodatkowo frezowane, aby zmniejszyć pojemność cieplną modułu i umożliwić jego skuteczną stabilizację termiczną, przy pomocy umieszczonych na dnie obudowy modułów Peltiera. Wszystkie zastosowane złącza muszą być uszczelniane. W przypadku złącz SMA dla układów dużej mocy konieczne było wykorzystanie dodatkowych uszczelnień, co przedstawiono na Rysunku 8.



Rysunek 7: Rysunek przedstawiający metodykę projektowania stabilnych fazowo modułów systemów PRDS [9].



Rysunek 8: Wykorzystana metoda uszczelniania złącz SMA [9].

3. PODSUMOWANIE

Artykuł przedstawia podsumowanie rozprawy doktorskiej, napisanej na podstawie ponad pięciu lat pracy związanej z systemami PRDS dla liniowych akceleratorów cząstek w ośrodku DESY w Niemczech. Opracowane w jej ramach rozwiązania są istotnie lepsze od istniejących, a część z nich została skomercjalizowana i może być z powodzeniem wykorzystana w innych systemach.

Materiał przedstawiony w rozprawie i niniejszym artykule powstał na podstawie prac wykonanych w ramach następujących projektów badawczych:

- Prace rozwojowe nad systemami generatorów wzorcowych, synchronizacji oraz sterowania akceleratorów XFEL, FLASH oraz SINBAD-Appendix 4 do umowy DESY-PW. Prace w ramach projektu międzynarodowego współfinansowanego ze środków programu Ministra Nauki i Szkolnictwa Wyższego pn. "PMW" w latach 2020-2021, umowa nr 5083/DESY/2020/0,
- Prace rozwojowe nad systemem dystrybucji sygnału synchronizacji akceleratora SINBAD, oraz modułów syntezerów częstotliwości. Prace w ramach projektu międzynarodowego współfinansowanego ze środków programu Ministra Nauki i Szkolnictwa Wyższego pn. "PMW" w latach 2019-2020, umowa nr 5028/DESY/2019/0,
- Badania i rozwój systemów RF dla akceleratorów FLASH, SINBAD i XFEL-Appendix 6 do umowy DESY-PW.

LITERATURA

[1] Cullerton Ed, Chase Brian. 2011. „Phase Averaging Reference Line". Prezentacja podczas Low-Level RF Workshop 2011.

[2] Czuba Krzysztof, et. al. 2018. „Concept of the phase reference line for the European Spallation Source". 22nd International Microwave and Radar Conference (MIKON).

[3] Czuba Krzysztof, Sikora Dominik. 2010. „Phase drift versus temperature measurements of coaxial cables". 18-th International Conference on Microwaves, Radar and Wireless Communications.

[4] H.Hanaki, Hideki Dewa, Shinsuke Suzuki, K.Yanagida. 2014. „Evaluation of Beam Energy Fluctuations Caused by Phase Noises".

[5] Lamb Thorsten, et. al. 2017. „Laser-to-RF Synchronization with Femtosecond Precision". 38th International Free Electron Laser Conference.

[6] Ludwig Frank, Hoffmann Matthias, Schlarb Holger, Simrock Stefan. 2006. „Phase Stability of the Next Generation RF Field Control for VUV- and X-ray Free Electron Laser". Proceedings of EPAC 2006.

[7] Sikora Dominik, et. al. 2020. „Phase Drift Compensating RF Link for Femtosecond Synchronization of E-XFEL". IEEE Transactions on Nuclear Science.

[8] Simrock Stefan, Geng Zheqiao. 2022. „Low-Level Radio Frequency Systems". Springer Cham.

[9] Urbański Maciej. 2023. „Development of Phase Reference Distribution Systems of Linear Particle Accelerators with Femtosecond Stability". Rozprawa doktorska. Politechnika Warszawska.

[10] Urbański Maciej, et al. 2023. „An Upgraded Universal Frequency Divider Module for the New FLASH2020+ RF Reference Generation System". 14th International Particle Accelerator Conference IPAC'23.

[11] Urbański Maciej, et al. 2022. „FLASH2020+ RF Reference Generation System Upgrade Status". Low-Level RF Workshop 2022.

[12] Urbański Maciej, et. al. 2018. „The concept of the RF phase reference distribution system for SINBAD accelerator research facility,". 22nd International Microwave and Radar Conference (MIKON).

[13] Urbański Maciej, et. al. 2018. „Frequency Divider Module for the European XFEL Phase Reference Signal Distribution System". 22nd International Microwave and Radar Conference (MIKON),

[14] Zembala Łukasz, et. al. 2014. „Master Oscillator for the European XFEL". Proceedings, 5th International Particle Accelerator Conference (IPAC 2014).

[15] Zhenyang Lin, et. al. 2020. „An active coaxial line phase reference distribution system". Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

MODELOWANIE NIESNELLOWSKICH ODBIĆ FAL RADIOWYCH OD METAPOWIERZCHNI

MODELLING OF NON-SNELLIAN REFLECTIONS OF RADIOWAVES FROM METASURFACES

Łukasz Jopek[1]; Szymon Grabarek[2]; Kinga Sikorska[3]

[1] Politechnika Łódzka, Łódź lukasz.jopek@p.lodz.pl
[2] Politechnika Łódzka, Łódź, 228992@edu.p.lodz.pl
[3] Politechnika Łódzka, Łódź, 234717@edu.p.lodz.pl

DOI: 10.15199/59.2024.4.59

Streszczenie: **W artykule przedstawiono dwie metody modelowania niesnellowskich odbić od metapowierzchni takich jak inteligentne powierzchnie odbijające. Omówiono oraz porównano możliwości modelowania za pomocą metody różnic skończonych w dziedzinie czasu oraz metody śledzenia promieni.**

Abstract**: The article discusses methods for modelling non-Snell reflections from metasurfaces such as intelligent reflecting surfaces. The possibilities of modelling using the finite-difference time-domain method and the ray tracing technique are discussed and compared.**

Słowa kluczowe: **6G, FDTD, IRS, kształtowanie wiązek, śledzenie promieni,**

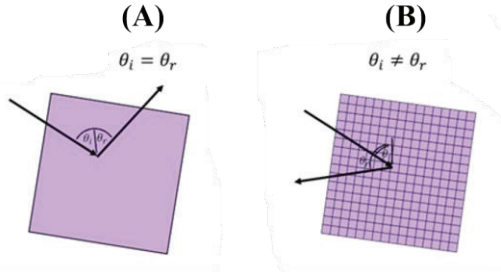
Keywords: **6G, Beamforming, FDTD, IRS, ray tracing.**

1. Metapowierzchnie

Metamateriały to materiałowe struktury, które zostały sztucznie wytworzone. Są szczególnie przydatne do tworzenia filtrów, modulatorów anten czy soczewek. Z założenia wykazują one odmienne właściwości elektromagnetyczne niż materiały, które można znaleźć w środowisku naturalnym. Zakres pracy obejmował symulacje meta-powierzchni jako dwuwymiarowej macierzy zawierające metamaatomy. Charakterystykę działania danej metapowierzchni definiuje szereg cech jak np.: geometria, kształt, wymiary oraz parametry materiałowe meta-atomów jak i materiału dielektrycznego, na którym są zazwyczaj umieszczane. Modelowanie numeryczne metamateriałów możliwe jest na wiele sposobów. Zazwyczaj jednak wymaga to znacznych nakładów obliczeniowych i pamięciowych. W niniejszym artykule zaprezentowano rozwiązanie, które, w uproszczony sposób modeluje zachowanie się metapowierzchni konkretnego typu – mianowicie inteligentnych powierzchni odbijających (Intelligent Reflecting Surfaces - IRS) [1,3-5] jednak cechuje się znacznie niższą złożonością obliczeniową niż klasyczne metody, takie jak np. metoda różnic skończonych w dziedzinie czasu.

1.1 Inteligentne Powierzchnie Odbijające

Rozwój szóstej generacji sieci mobilnej (6G), wymaga postępu technologicznego w zakresie transmisji danych, z przepływnościami aktualnie nieosiągalnymi w publicznych warunkach, używając coraz to większych częstotliwości fal - wykorzystanie fal subterahercowych wymaga pokonania problemu polegającego na bardzo dużych stratach propagacji, kiedy takie fale napotkają jakiekolwiek przeszkody. Uzyskanie bezpośredniej widzialności optycznej nie zawsze jest możliwe, ale alternatywą jest kształtowanie środowiska propagacji poprzez rozmieszczenie inteligentnych powierzchni odbijających [1,3-5] IRSs, które mogą odbijać fale w kierunkach innych niż wynikające z prawa Snella, a przy tym dynamicznie zmienianych zależnie od potrzeb, np. od zmian liczby i lokalizacji abonentów lub ich grup. IRS mogą być modelowane za pomocą macierzy układów RLC, wchodzących w skład pojedynczego metaatomu, lub po prostu - metapowierzchni. Różnica odbić, pomiędzy falą padającą oraz falą odbitą została przedstawiona na Rys. 1, gdzie kąt fali padającej został oznaczony jako θ_i, natomiast kąt fali odbitej jako θ_r. W przypadku odbicia zgodnego z generalnie przyjętym prawem Snella oba kąty mają tę samą wartość, w przypadku odbicia anomalnego (niesnellowskiego) kąty przyjmują różne wartości.



Rys. 1. A) Odbicie snellowskie B) Odbicie niesnellowskie

Przy odpowiedniej konfiguracji metaatomów można uzyskać formowanie wiązek (ang. beamforming). IRS można podzielić na dwie zasadnicze grupy; aktywne oraz pasywne. Aktywne IRS są obiektami, które w założeniu będą w stanie, zmieniać swoje parametry dynamicznie. Dla IRS realizowanych jako meta powierzchnie można to uzyskać np. poprzez zmianę charakterystyki rezonansowych poszczególnych metacząstek (rezonatorów) a przez to także fazę odbicia dla każdej z nich. Faza odbicia zmienna w funkcji współrzędnych rezonatorów (np. ułożonych w postaci dwuwymiarowej macierzy) pozwala na kształtowanie wiązki [6] lub jednocześnie kilku wiązek odbitych. Możliwe jest też uzyskiwanie odbicia rozproszonego, tj. w szerokim zakresie kątów. Aby uzyskać żądane

parametry odbicia należy zatem wyznaczyć parametry opóźnienia fazowego (oraz modułu odbicia) dla każdego metaatomu.

2. Modelowanie IRS

Celem pracy jest opracowanie metody modelowania IRS. W uproszczonym modelu każdy metaatom może być opisywany przez zaledwie trzy wartości: kąt opóźnienia fazowego, moduł odbicia oraz rozmiar samego metaatomu, który także ma wpływ na kierunek uformowanej wiązki (lub inaczej na kąt odbicia θ_r). Oba zaproponowane modele biorą po uwagę tylko te trzy parametry. Realizując rzeczywistą metapowierzchnię należy zaprojektować każdy metaatom tak aby wykazywał te trzy wartości. W pracy przestudiowano użycie dwóch modeli symulacyjnych, w celu porównania zgodności rezultatów i rozważenia efektywności użycia prostszego obliczeniowo[3], a co za tym idzie mniej obciążającego modelu w ograniczonym okresie czasowym. Oba modele, które istotnie się od siebie różnią zastosowaną metodą, dają zbliżone wyniki obliczeń. Model wykorzystujący metodę różnic skończonych w dziedzinie czasu jest w pełni falową metodą rozwiązywania równań Maxwella w postaci różniczkowej. Wykorzystano tym celu symulator Remcom™ XFdtd. Model używający śledzenia promieni, działa lepiej jako forma narzędzia ukierunkowującego osobę konfigurującą urządzenia na najbardziej prawdopodobne poprawne rezultaty.

2.1 Symulacje metodą śledzenia promieni

Ta metoda symulacyjna bazuje na środowisku stworzonym w oprogramowaniu Matlab™, które zostało rozszerzone o publicznie dostępne biblioteki upraszczające prace związane z analizą, symulacją oraz wyświetlaniem wyników.

Kształtowanie wiązek jest kluczowe dla technologii 5G/6G. Polega ono na stworzeniu pewnego układu bądź (najczęściej) macierzy anten o tym samych parametrach. Formowanie wiązki polega na odpowiednim dobieraniu opóźnień fazowych dla poszczególnych elementów macierzy metaatomów. W przypadku inteligentnych powierzchni odbijających dochodzi jeszcze problem określenia opóźnień w dotarciu sygnału źródłowego w odniesieniu do każdego elementu tej powierzchni. Jeśli przyjąć, że IRS będzie realizowana jako meta powierzchnia to dla każdego metaatomu należy obliczyć takie opóźnienie, celem synchronizacji odbieranego sygnału.

Beamfoming, jest kluczowy dla aktualnej technologii 5G i anten massive MIMO. Polega on na zastosowaniu szyku antenowego ze sterowaniem fazowym. Inteligentne powierzchnie odbijające można również zakwalifikować do układów ze sterowaniem fazowym, ale w tym przypadku brana jest pod uwagę faza odbicia, a nie przesunięcie fazy w elementach antenowych.

Przy odpowiednim zastosowaniu IRS, wspomniane wyżej kształtowanie wiązek może być użyteczne również w odniesieniu do sieci 6G. Wymagać to jednak będzie możliwości dynamicznej rekonfiguracji wartość fazy odbicia każdego meta-atomu.

2.2 Symulacje w pełni falowe

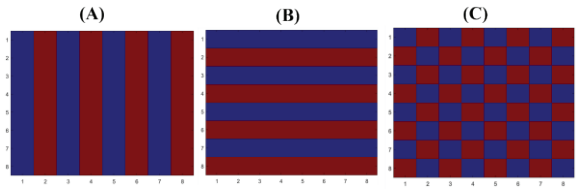
Jako symulator w pełni falowy użyto komercjalnego pakietu Remcom XFdtd ™. Oprogramowanie implementuje metodę różnic skończonych w dziedzinie czasu (Finite Discrete Time Domain - FDTD).

2.3 Materiały

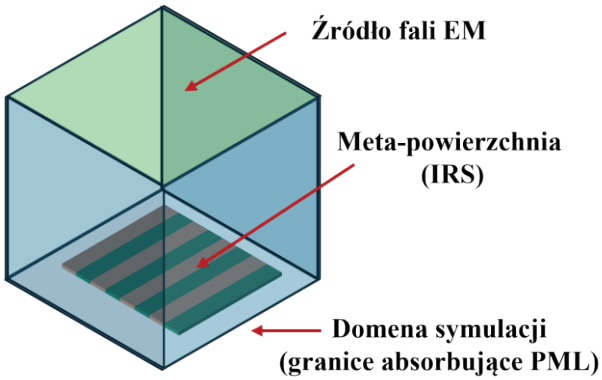
W symulacjach wykorzystujących metodę FDTD dla uzyskania różnych faz odbicia użyto materiałów: PEC – Perfect Electric Conductor, PMC – Perfect Magnetic Conductor, oraz metriały określone modelem DD – Debye-Drude[7] (model używany do opisu zachowania materiałów, dla którego można wyznaczyć efektywne wartości modułu odbicia i kąta fazowego. Teoretycznie daje to możliwość zdefiniowania metriału o dowolnym kącie odbicia fazowego).

2.4 Szczegóły symulacji

Do oceny wyników wykorzystano charakterystyki promieniowania w polu dalekim. Analizowana struktura składała się z macierzy osiem na osiem elementów (metaatomów). W przypadku symulacji metodą w pełni falową (pakiet Remcom XFdtd™) wykorzystano pobudzenie szerokopasmowe. Zaproponowano trzy podstawowe wzorce macierzy meta-powierzchni – linie poziome, pionowe oraz „szachownica". Rozmiar pojedynczego metaatomu wynosił 5 mm x 5 mm x 0.5 mm (rozmiar siatki tymczasem wynosił 0.2 mm x 0.2 mm x 0.2 mm). Dla częstotliwości f= 50 GHz wartość 5 mm stanowi ok 0.83 długości tej fali. Typowy rozmiar meta-atomu jest zazwyczaj ułamkiem długości fali (0.5 – 0.05 λ) lecz w przypadku powierzchni IRS należy także brać pod uwagę związek pomiędzy rozmiarem pojedynczej komórki a kątem odbicia fali. Wyniki były uzyskiwane z czujnika pola dalekiego. Charakterystykę promieniowania analizowano dla częstotliwości f = 50 GHz. Schemat domeny symulacji ilustruje rysunek 3. Drugą metodą symulacji była metoda bazująca na śledzeniu promieni, zrealizowana w pakiecie Matlab™, przy wykorzystaniu m.in. biblioteki phasedArray. Celem wprowadzenia tej metody symulacji była redukcja czasu obliczeń. Maksymalny czas potrzebny do wykonania niezbędnych obliczeń udało się zredukować do kilku sekund. Do analizy wybrano trzy podstawowe wzorce, znane z literatury[2] (rys. 2). Składają się one z dwóch typów elementów: o kącie przesunięcia fazowego równego 0 stopni (na rys. oznaczone kolorem niebieskim) oraz 180 stopni (na rys. oznaczone kolorem czerwonym).



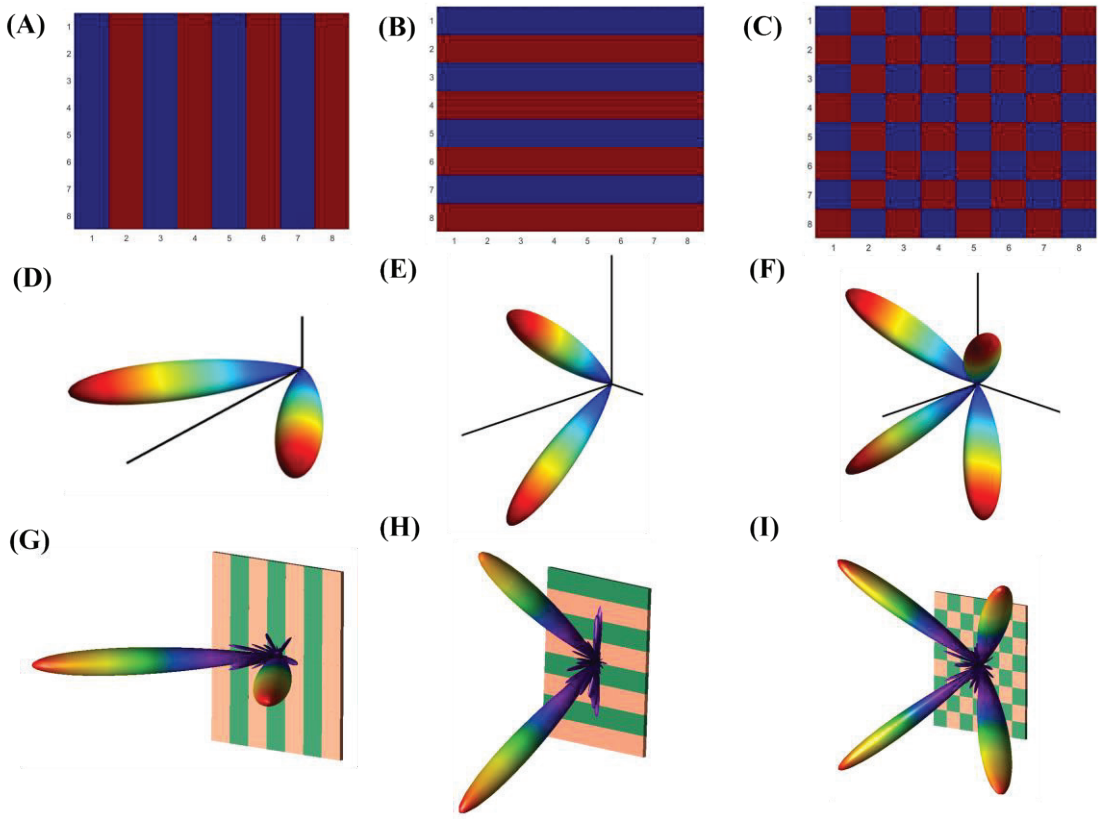
Rys. 2. Wzory przesunięcia fazowego; A) linie pionowe, B) linie poziome C) szachownica. Kolor niebieski oznacza materiał typu PMC, czerwony zaś to materiał typu PEC.



Rysunek 3: Schemat symulacji w pełni falowej

3. Wyniki modelowania

Na rysunku 4 przedstawiono wyniki symulacji przeprowadzonej dla zaproponowanych wzorców przy wykorzystaniu obu metod modelowania. Zawiera on uzyskane charakterystyki kierunkowe przewidywane dla odbić fali o częstotliwości f= 50 GHz, padającej prostopadle na badaną powierzchnie. Wartością mierzoną jest pole elektryczne mierzone w polu dalekim.

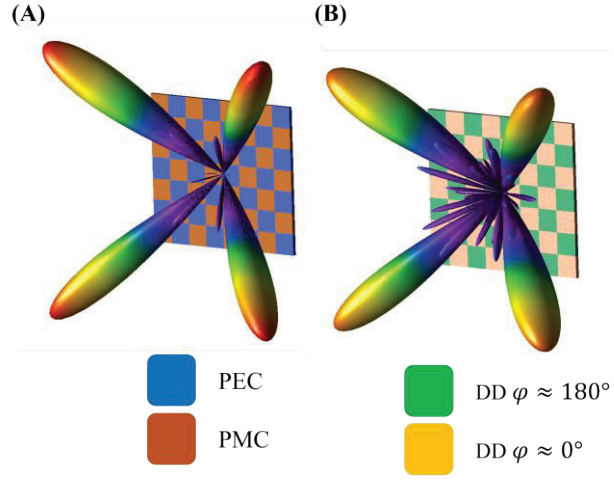


Rysunek 4: Porównanie rezultatów obu sposobów modelowania powierzchni IRS: (A,B,C) wzory przesunięć fazowych odpowiednio pionowy, poziomy i "szachownica", (D,E,F) rezultaty modelowania metodą beamformingową, (G,H,I): rezultaty modelowania metodą FDTD.

Na podstawie analizy uzyskanych wyników (rys. 4) można stwierdzić zgodność z przypadkami znanymi z literatury[2] niezależnie od czego czy zastosowano kombinację materiałów PEC/PMC czy też kombinację materiałów modelu Debye-Drude (dla 0 i 180 stopni), co potwierdza przydatność tychże do symulacji tego typu struktur w przyszłości (rysunek 5). Zaletą zastosowania materiałów opisanych modelem Debye-Drude jest niewątpliwie możliwość zdefiniowania zarówno (dowolnego) kąta opóźnienia fazowego oraz modułu odbicia ale także obniżenie złożoności obliczeniowej modelu w porównaniu do modelu zawierającego rzeczywiste metaatomy w postaci rezonatorów. Pewnym problemem może oczywiście okazać się znalezienie zestawu ośmiu parametrów opisujących żądany materiał[7]. Tutaj z pomocą mogą przyjść metody optymalizacyjne. W przypadku metody w pełni falowej (XFdtd™) wyraźne efekty uzyskano dla zakresu częstotliwości między 50-90 GHz, niezależnie od tego który meta-(dla metaatomu o rozmiarze 5 mm, co stanowi odpowiednie 0.85 do 1.5 λ). Jak już wyżej wspomniano, w przypadku IRS rozmiar czy też odległość pomiędzy sąsiadującymi ze sobą metaatomami ma także znacznie w kwestii wyznaczanie kąta odbicia θ_r.

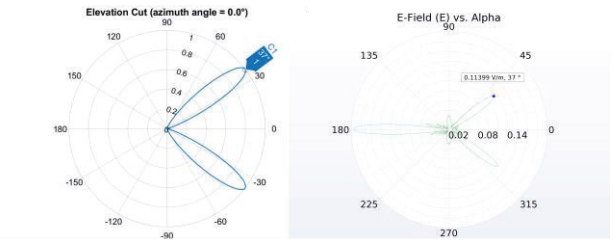
Warto również odnotować, że czas symulacji różnił się w zależności od użytej metody. Dla metody FDTD (XFdtd™) średni czas symulacji powierzchni o rozmiarach 8x8 elementów wynosił 3 godziny. Natomiast model beamformingowy potrzebował jedynie kilku sekund. Otrzymane wyniki wskazują na możliwość wykorzystania prostszej z rozważanych metod, tj. śledzenia promieni

do modelowania meta-materiałowych powierzchni odbijających.



Rysunek 5: Porównanie przykładowego wzorca promieniowania (symulacja w pełni falowa) dla meta-powierzchni zbudowanej z: (A) materiałów PEC/PMC i (B) materiałów z wykorzystaniem modelu Debye-Drude

Z kolei na rysunku 6 zaprezentowano porównanie wyników dla obu metod modelowanie pod kątem uzyskanego kąta odbicia. W eksperymencie kąt padania fali padającej na powierzchnię IRS wynosił 90° (czyli był prostopadły do powierzchni IRS), natomiast zmierzony kąt odbicia wynosił ok. 37° dla obu metod. Potwierdza to zgodność obu także zgodność uzyskanych wyników oraz możliwosc wykorzystania metody beamformingowej w miejsce metody pełni falowej.



Rysunek 6: Porównanie charakterystyki promieniowania (pole elektryczne) dla obu sposobów modelowania: (A) metoda beamformingowa, (B) metoda w pełni falowa (FDTD). Na rysunku zaznaczono kąt odbitej wiązki w płaszczyźnie poziomej. Kąt fali padającej wynosił 90°

WNIOSKI

W publikacji porównano ze sobą dwie metody, których celem jest określenie charakterystyki kierunkowej powierzchni IRS. Metoda bazująca na beamformingu jest wystarczająca do poprawnego określania tej charakterystyki. Jest to więc alternatywna metoda dla symulacji w pełni falowych (takich jak np. Wykorzystana w publikacji metoda FDTD). Natomiast jej istotną przewagą nad powszechnie stosowanym podejściem w pełni falowym jest znacznie mniejsza złożoność obliczeniowa i pamięciowa, a co za tym idzie także znacznie krótszy czas potrzebny do wykonania symulacji numerycznej konkretnej powierzchni IRS. Wykazano także możliwość zastosowanie materiałów opisanych modelem Debye-Drude jako ciekawą alternatywę dla symulacji metaatomu w postaci rezonatorów. Daje to możliwość skrócenia czasu potrzebnego do wykonania symulacji w pełni falowej (FDTD) przykładowej metapowierzchni IRS.

LITERATURA

[1] Adrian Kliks, Metamateriały, konfigurowalne matryce antenowe i komunikacja holograficzna - wstępna analiza nowej koncepcji bezprzewodowej transmisji danych. Krajowa Konferencja Radiokomunikacji i Teleinformatyki 2023

[2] Cui, T., Qi, M., Wan, X. et al. Coding metamaterials, digital metamaterials and programmable metamaterials. Light Sci Appl 3, e218 (2014).

[3] D. F. Sievenpiper, J. H. Schaffner, H. J. Song, R. Y. Loo and G. Tangonan, "Two-dimensional beam steering using an electrically tunable impedance surface," in IEEE Transactions on Antennas and Propagation, vol. 51, no. 10, pp. 2713-2722, Oct. 2003.

[4] Siqi Liu, Zhenyu Ma, Jian Pei, Qingbin Jiao, Lin Yang, Wei Zhang, Hui Li, Yuhang Li, Yubo Zou, Xin Tan; A review of anomalous refractive and reflective metasurfaces. Nanotechnol. Precis. Eng. 1 June 2022; 5 (2)

[5] Tran, Nguyen Minh, Muhammad Miftahul Amri, Je Hyeon Park, Sa Il Hwang, Dong In Kim, and Kae Won Choi. 2019. "A Novel Coding Metasurface for Wireless Power Transfer Applications" Energies 12, no. 23: 4488.

[6] Van Trees, H. Optimum Array Processing. New York: Wiley-Interscience, 2002.

[7] Wsparcie XFdtd – Remcom. 2024. 'Materiał Debye-Drude.'. Dostępne na: Debye-Drude Material – XFdtd – Remcom Support

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

Propagacja sygnału radiowego w systemach 5G wyposażonych w matryce IPR

Radio signal propagation in 5G systems equipped with RISs

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: 10.15199/59.2024.4.60

Streszczenie: W niniejszym artykule przeanalizowano charakterystykę propagacji sygnału radiowego na obszarze zawierającym się w granicach miasta Poznania (Polska). W rozważaniach przyjęto wykorzystanie sieci dostępu radiowego RAN (*ang. Radio Access Network*) systemu bezprzewodowego 5. generacji (5G NR – *ang. New Radio*), w skład którego wchodzi 8 stacji bazowych (SB) wykorzystujących technologię antenową SISO (*ang. Single Input Single Output*) lub MIMO (*ang. Multiple Input Multiple Output*) w zależności od przyjętej konfiguracji komórek sieciowych. Na badanym obszarze umieszczono również 15 matryc odbijających zwanych Inteligentnymi Powierzchniami Rekonfigurowalnymi (IPR) oraz uwzględniono ich wpływ na propagację sygnału radiowego dla różnych wysokości ich zawieszenia.

Abstract: In this paper, the characteristics of radio signal propagation within the boundaries of the city of Poznan (Poland) are analyzed. The study considers the use of a Radio Access Network (RAN) of the 5th generation wireless system (5G NR – New Radio), which includes 8 base stations (BSs) utilizing Single Input Single Output (SISO) or Multiple Input Multiple Output (MIMO) antenna technology depending on the adopted configuration of network cells. Additionally, 15 reflecting arrays known as Reconfigurable Intelligent Surfaces (RISs) were placed in the studied area, and their impact on radio signal propagation at different suspension heights was taken into account.

Słowa kluczowe: 5G, Inteligentne Powierzchnie Rekonfigurowalne, propagacja sygnału radiowego, tłumienie ścieżki, systemy bezprzewodowe.

Keywords: 5G, path loss, radio signal propagation, Reconfigurable Intelligent Surfaces, wireless systems.

1. WSTĘP

Obecnie stadium rozważań dotyczących zarówno teraźniejszych (5G), jak i przyszłych (6G) generacji systemów mobilnych wprowadza do literatury naukowej coraz to nowsze rozwiązania propagacji sygnału. W ostatnich latach duże zainteresowanie zyskały tzw. Inteligentne Powierzchnie Rekonfigurowalne (*ang. Reconfigurable Intelligent Surfaces* – RISs), które cechują się zdolnością odbijania sygnału radiowego i kierunkowania go w stronę konkretnego obszaru. Taka właściwość miałaby umożliwić mobilnym operatorom (*ang. Mobile Network Operators* – MNOs) manipulowanie pokryciem radiowym w celu *doświetlania* terenów, na których terminale odbiorcze mają utrudniony dostęp do usług mobilnych. W dodatku wykorzystanie matryc IPR może skutecznie doprowadzić do redukcji interferencji międzysygnałowych i/lub narażenia użytkowników na działanie pola elektromagnetycznego (PEM). Wykorzystanie wspomnianych matryc IPR mogłoby znaleźć swoje szczególne zastosowanie w obrębie terenów miejskich, gdzie nadawany sygnał radiowy jest załamywany i rozpraszany przez liczne przeszkody znajdujące się na jego drodze. Niemniej jednak na terenach wiejskich aktywne matryce IPR (tj. takie, które dodatkowo wzmacniają przekierowywany sygnał radiowy) mogą zostać wdrożone jako przekaźniki, redukując przy tym liczbę stacji bazowych potrzebnych do pokrycia rozważanego obszaru [1–5].

W niniejszym artykule przeanalizowano wykorzystanie wspomnianych matryc IPR na obszarze miejskim badając przy tym ich wpływ na charakterystykę tłumienia ścieżki (*ang. Path Loss* – PL) propagowanego sygnału radiowego. W rozważaniach wzięto również pod uwagę wpływ zawieszenia inteligentnych matryc na pokrycie badanego terenu.

Artykuł został podzielony na następujące części: sekcja druga opisuje rozważany scenariusz systemowy; sekcja trzecia zawiera informacje dotyczące metodologii przeprowadzanych badań oraz konfiguracji symulacyjnych; treść sekcji czwartej składa się z wyników badań zaprezentowanych w formie zarówno grafów, jak i tablic ze średnimi wartościami obserwowanych parametrów systemowych; piąta i ostatnia sekcja podsumowuje całą pracę poprzez przedstawienie wniosków wyciągniętych na podstawie zaobserwowanych rezultatów.

2. ROZWAŻANY SCENARIUSZ

Rozważana sieć została umieszczona na terenie miasta Poznania – dokładnie wzięto pod uwagę Stary Rynek oraz ulice w jego sąsiedztwie [4–6]. W skład systemu mobilnego wchodzi 8 stacji bazowych, gdzie każda z nich posiada 3 komórki transmitujące sygnał radiowy na różnych częstotliwościach – 800, 2100 i 3500 MHz. Dodatkowo każda z tych komórek wykorzystuje systemy antenowe transmitujące sygnał zgodnie z technikami SISO (800 i 2100 MHz) lub MIMO (3500 MHz), które cechują się odmienną liczbą aktywnych elementów – 1 (SISO) oraz 64 (MIMO) przypadających na pojedynczy transciver [6, 7].

Ponadto na rozważanym terenie pomiędzy budynkami rozmieszczono 15 pasywnych matryc IPR w taki sposób, że każda z nich jest w zasięgu przynajmniej jednej stacji bazowej. Matryce są umieszczone domyślnie na krawędzi każdego budynku, na którym się znajdują, jednakże ich wysokość może zostać zmanipulowana.

Do zaprojektowania systemu mobilnego (które przekładało się m.in. na ustalenie lokalizacji stacji bazowych i budynków, jak i na wytyczenie wymiarów budowli znajdujących się na terenie systemu bezprzewodowego oraz określenie planowanego obszaru pokrycia sygnałem radiowym) wykorzystano dane rzeczywiste udostępnione w bazie danych BTSearch [8] oraz w Systemie Informacji Przestrzennej Miasta Poznania [9].

3. KONFIGURACJA SYMULACYJNA

Badania zostały przeprowadzone w formie symulacyjnej przy użyciu dedykowanego oprogramowania o nazwie GRAND (*Green Radio Access Network Design*) [6, 7].

Głównym celem badań było przeanalizowanie charakterystyki tłumienia ścieżki sygnału radiowego na badanym terenie miejskim. Do ewaluacji wartości tłumienia ścieżki sygnału radiowego w danym punkcie przyjęto następujące modele propagacyjne:

- 3GPP TR 38.901 UMa (*ang. Urban Macro*) – propagacja sygnału LOS (*ang. Line-Of-Sight*) i NLOS (*ang. Non-Line-Of-Sight*) od stacji bazowej [10],
- RIS FFBC (*ang. RIS Far Field Beamforming Case*) – propagacja sygnału od stacji bazowej poprzez odbicie od matrycy IPR [2].

Co więcej, warunki propagacyjne zostały zaaranżowane (symulacyjnie) dla scenariusza systemowego, w którym początkowo żadna z dostępnych matryc nie jest wykorzystywana do pośredniej propagacji sygnału radiowego, a następnie wszystkie nich są brane pod uwagę podczas estymacji charakterystyki tłumienia jego ścieżki. W dodatku dla rozważanego scenariusza przetestowano kilka wariantów konfiguracji symulacyjnej, które różnią się między sobą, chociażby wysokością zawieszenia matryc odbijających sygnał radiowy. Poniżej wyszczególniono każdy ze wspomnianych przypadków wziętych pod uwagę podczas rozważań symulacyjnych:

I. Matryce IPR zawieszone na domyślnej wysokości h_{RIS} [m] (tj. na wysokościach prawie równych wysokościom budynków, na których się znajdują).

II. Matryce IPR zawieszone na wysokości $h_{\mathrm{RIS}} + 5$ [m].

III. Matryce IPR zawieszone na wysokości $h_{\mathrm{RIS}} + 10$ [m].

Wypunktowany powyżej zróżnicowany zbiór przypadków testowych pozwoli przeanalizować wpływ zarówno wysokości zawieszenia matryc IPR, jak i średniej wartości różnicy pomiędzy wysokościami zawieszenia układów antenowych stacji bazowych oraz matryc odbijających wiązkę radiową od tychże węzłów dostępowych (czyli de facto znajdujących się w ich zasięgu) na charakterystykę propagacyjną sygnału radiowego w obrębie rozważanego terenu pokrycia.

Konfiguracja parametrów symulacyjnych dla rozważanej sieci mobilnej została zawarta w kolejno zaprezentowanych tabelach (tj. Tab. 1 i 2). Pierwsza z nich (Tab. 1) zawiera wartości parametrów wykorzystanych do emulowania działania stacji bazowych w oprogramowaniu symulacyjnym. Wyszczególnione we wspomnianej tabeli parametry zawierają informacje dotyczące wykorzystywanego pasma, konfiguracji układów nadawczo-odbiorczych oraz modelowania warunków propagacyjnych dla każdego typu komórki sieciowej. Z kolei na Tab. 2 składają się wartości dla parametrów symulacyjnych proponujących konkretnie zdefiniowany rodzaj inteligentnych matryc IPR. Wśród tych właściwości znajdują się takie cechy jak liczba oraz wymiary aktywnych elementów, wysokość zawieszenia matryc na budynkach, czy chociażby wykorzystany model propagacyjny do oszacowania potencjalnego tłumienia sygnału.

Tabela 1: Konfiguracja parametrów (SB) [6–11]

	Parametr	Symbol	Jednostka	Wartość		
				Komórka stacji bazowej		
				1.	2.	3.
Og.	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	$N_{\mathrm{SC,u}}$	–	320		
	L. wszystkich podnośnych	$N_{\mathrm{SC,t}}$	–	512		
	Współczynnik próbkowania	SF	–	1.536		
	Współ. ponow. użycia pilota	RF	–	1		
	Czas koherencji	t_{c}	[ms]	50		
	Szerokość pasma koherencji	B_{c}	[MHz]	1		
	Przestrzenny cykl pracy	S	[%]	0	0	25
Transceivery	Wysokość anteny	h_{BS}	[m]	(32, 46)		
	L. elementów antenowych	M_{BS}	–	1	1	64
	Zysk antenowy	G_{a}	[dBi]	16	18	24
	Tłumienie kabla antenowego	L_{f}	[dBi]	2	2	3
	Maks. moc nadawcza	$P_{\mathrm{TX,\ max}}$	[dBm]	46	49	53
	Współ. szumów	NF	[dB]	8	8	7
Propagacja	Model tł. ścieżki	–	–	3GPP TR 38.901 UMa [10]		
	Margines interferncji	IM	[dB]	2		
	Margines Dopplera	DM	[dB]	3		
	Margines zaników	FM	[dB]	10		
	Margines shadowingu	SM	[dB]	12.8	15.2	10
	Strata implementacyjna	IL	[dB]	0	0	3

Tabela 2: Konfiguracja parametrów (IPR) [2]

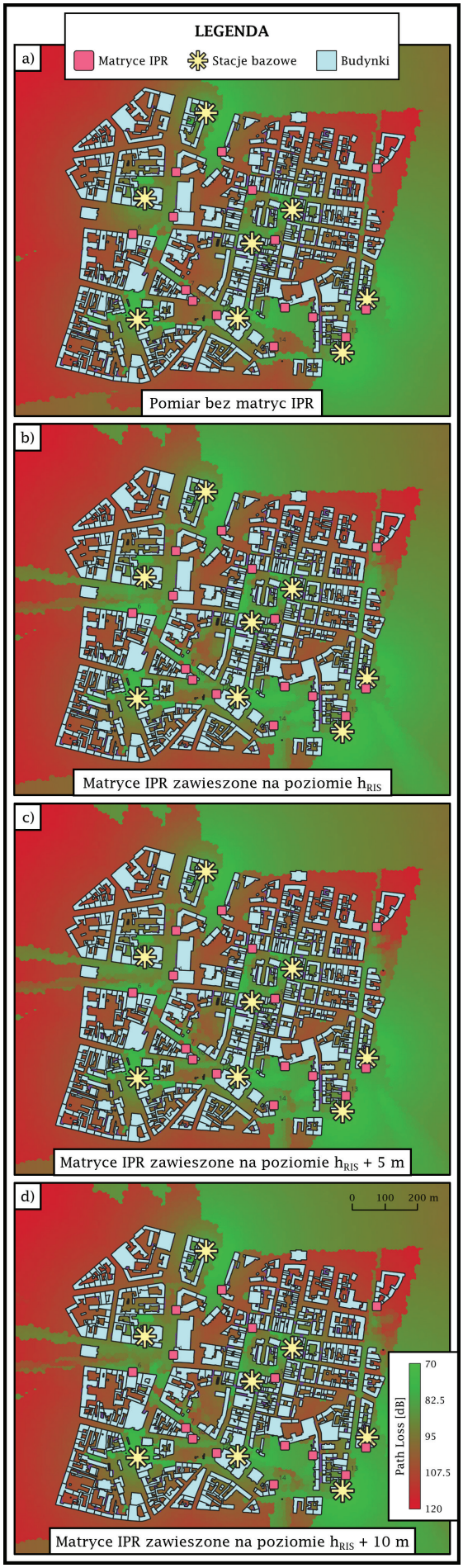
	Parametr	Symbol	Jednostka	Wartość
	Liczba	K_{RIS}	–	15
	L. elementów w kolumnie	M_{RIS}	–	102
	L. elementów w wierszu	N_{RIS}	–	100
	Szerokość kolumny	d_{m}	[m]	0.01
	Szerokość wiersza	d_{n}	[m]	0.01
	Współ. amp. syg. odbitego	A	–	0.9
	Wysokość zawieszenia	h_{RIS}	[m]	(18, 43)
	Model tł. ścieżki	–	–	RIS FFBC [2]

4. WYNIKI

Niniejsza sekcja prezentuje wyniki uzyskane podczas przeprowadzania badań symulacyjnych. Na Rys. 1.a zaprezentowano rozkład minimalnego tłumienia sygnału radiowego dla rozważanego obszaru gdy żadna z matryc nie jest brana pod uwagę. Kolejno Rys. 1.b–d prezentują te same wyniki, lecz po uwzględnieniu matryc IPR w przybliżaniu charakterystyki propagacyjnej oraz po zmanipulowaniu (lub też nie – Rys. 1.b) wysokości ich zawieszenia.

Porównując mapy zawarte w Rys. 1 zauważyć można, iż wprowadzenie matryc IPR zredukowało tłumienie sygnału w niektórych rejonach badanego obszaru. Natomiast, łatwo również dostrzec, iż sama wysokość zawieszenia inteligentnych powierzchni szczególnie oddziałuje na prezentowaną charakterystykę propagacji sygnału radiowego. Analizując każdy przypadek z osobna (Rys. 1.b–d) można zobaczyć, iż zwiększając wysokość zawieszenia matryc IPR o pierwsze 5 m w południowej części obszaru zredukowano tłumienie sygnału radiowego. Jednakże kolejne podwyższenie punktu zaczepienia matryc odbijających (o 10 m względem wysokości domyślnej) niefortunnie pogarsza ten rozkład w porównaniu do Rys. 1.b–c.

Powyższe obserwacje zostały podsumowane w Tab. 3. Zawarte we wspomnianej tabeli wartości parametrów zostały uzyskane poprzez zastosowanie poniższych wzorów oznaczonych dalej jako Równ. (1)–(3).



Rysunek 1: Mapy rozkładu tłumienia ścieżki sygnału.

Pierwszym opisywanym parametrem jest średnia różnica zawieszenia matryc IPR między stanem obecnym a przyjętą wysokością domyślną (Δh_{RIS}). Wysokość domyślna dla każdej matrycy jest równa w przybliżeniu wysokości budynku, na którym się znajduje. Wspomniana (średnia) różnica wysokości matryc odbijających może zostać opisana przez następującą formułę:

$$\Delta h_{\mathrm{RIS}} = \frac{1}{K_{\mathrm{RIS}}} \sum_{i=0}^{K_{\mathrm{RIS}}-1} \left(h_{\mathrm{RIS},j}^{\mathrm{current}} - h_{\mathrm{RIS},j}^{\mathrm{default}} \right) \tag{1}$$

gdzie K_{RIS} to liczba powierzchni umieszczonych na badanym terenie, a $h_{\mathrm{RIS},j}^{\mathrm{current}}$ i $h_{\mathrm{RIS},j}^{\mathrm{default}}$ to odpowiednio wysokość j-tej ($j = 0, \ldots, K_{\mathrm{RIS}} - 1$) matrycy IPR w chwili obecnej oraz jej wartość domyślna.

Kolejnym parametrem opisanym w Tab. 3 jest średnia różnica wysokości między stacjami bazowymi a matrycami IPR ($\Delta h_{\mathrm{BS–RIS}}$), które kierunkowały sygnał radiowy z tych stacji do poszczególnych punktów na mapie rozważanego obszaru. Wspomniany parametr opisuje następujące równanie:

$$\Delta h_{\mathrm{BS–RIS}} = \frac{\sum_{i=0}^{K_{\mathrm{BS}}-1} \sum_{j=0}^{K_{\mathrm{RIS}}-1} \chi_{i,j} \left(h_{\mathrm{BS},i} - h_{\mathrm{RIS},j}^{\mathrm{current}} \right)}{\sum_{i=0}^{K_{\mathrm{BS}}-1} \sum_{j=0}^{K_{\mathrm{RIS}}} \chi_{i,j}} \tag{2}$$

gdzie K_{BS} to liczba komórek stacji bazowych znajdujących się w obrębie sieci mobilnej. Z kolei $h_{\mathrm{BS},i}$ to wysokość zawieszenia anten i-tej komórki sieciowej ($i = 0, \ldots, K_{\mathrm{BS}} - 1$). Natomiast $\chi_{i,j}$ to parametr binarny ($\chi_{i,j} \in \{0, 1\}$) przyjmujący wartość równą 1, jeżeli j-ta matryca IPR znajduje się w zasięgu *wzroku* i-tej komórki sieciowej przynajmniej w jednej z L konfiguracji (dla niniejszego artykułu wyróżnić można 3 przypadki – $\Delta h_{\mathrm{RIS}} \in \{0, 5, 10\}$ [m]). W pozostałych przypadkach wartość $\chi_{i,j} = 0$.

Najważniejszy z parametrów ewaluacyjnych, czyli średni zysk związany z redukcją tłumienia ścieżki sygnału radiowego $\left(\Delta\overline{\mathrm{PL}}\right)$ został zaimplementowany w środowisku symulacyjnym w sposób następujący:

$$\Delta\overline{\mathrm{PL}} = \left[1 - \frac{\sum_{y=0}^{Y-1} \sum_{x=0}^{X-1} \mathrm{PL}_{\mathrm{RIS}}^{\mathrm{min}}(x, y)}{\sum_{y=0}^{Y-1} \sum_{x=0}^{X-1} \mathrm{PL}_{\mathrm{NRIS}}^{\mathrm{min}}(x, y)} \right] \cdot 100\ \% \tag{3}$$

gdzie x ($x = 0, \ldots, X - 1$) i y ($y = 0, \ldots, Y - 1$) to koordynaty danego punktu znajdującego się w obrębie badanego obszaru miejskiego. Natomiast $\mathrm{PL}_{\mathrm{RIS}}^{\mathrm{min}}(x, y)$ i $\mathrm{PL}_{\mathrm{NRIS}}^{\mathrm{min}}(x, y)$ to odpowiednio wartości strat propagacyjnych sygnału radiowego (w [dB]) w punkcie $P(x, y)$ gdy w obliczeniach uwzględnione zostanie oddziaływanie matryc IPR na transmisję tegoż sygnału, oraz gdy to oddziaływanie zostanie pominięte. Wspomniane parametry opisano wewnątrz oprogramowania symulacyjnego tak, jak to przedstawiają poniższe wzory:

$$\mathrm{PL}_{\mathrm{RIS}}^{\mathrm{min}}(x, y) = \tag{4}$$
$$\min\left\{ \mathrm{PL}_{\mathrm{LOS},i}(x, y) ; \mathrm{PL}_{\mathrm{NLOS},i}(x, y) ; \mathrm{PL}_{\mathrm{RIS},i,j}(x, y) \right\}$$

$$\mathrm{PL}_{\mathrm{NRIS}}^{\mathrm{min}}(x, y) = \min\left\{ \mathrm{PL}_{\mathrm{LOS},i}(x, y) ; \mathrm{PL}_{\mathrm{NLOS},i}(x, y) \right\} \tag{5}$$

gdzie $\mathrm{PL}_{\mathrm{LOS},i}(x, y)$, $\mathrm{PL}_{\mathrm{LOS},i}(x, y)$ oraz $\mathrm{PL}_{\mathrm{RIS},i,j}(x, y)$ to wartości dla tłumienia ścieżki sygnału radiowego odbieranego w punkcie $P(x, y)$ od i-tej komórki sieciowej bezpośrednio ($\mathrm{PL}_{\mathrm{LOS},i}$), pośrednio ($\mathrm{PL}_{\mathrm{NLOS},i}$), tj. po osłabieniu jego mocy przez obiekty fizyczne (np. budynki) lub zjawiska atmosferyczne, oraz po odbiciu wiązki radiowej

od j-tej matrycy IPR ($\mathrm{PL}_{\mathrm{RIS},i,j}$). W obu tych przypadkach (Równ. (4) i (5)) zawsze wybierana jest korzystniejsza (minimalna) wartość tłumienia ścieżki.

Odnosząc się bezpośrednio do Tab. 3 można zauważyć, iż zwiększając stopniowo wysokość zawieszenia matryc IPR (Δh_{RIS}) – o 5 m, a następnie o 10 m, pogorszeniu ulega zysk związany ze średnią mocą odbieranego sygnału radiowego $\left(\Delta\overline{\mathrm{PL}}\right)$ – dla 5 m o ok. 0.18 p.p, dla 10 m o ok. 0.4 p.p). Zależność ta prawdopodobnie jest spowodowana zmianą kątów elewacji i azymutu [2] zarówno dla wiązek sygnału radiowego padających na powierzchnie IPR (nadawanych przez stacje bazowe), jak i tych odbitych przez matryce w kierunku punktów pomiarowych $P(x, y)$. Na niekorzystną zmianę kątów padania i odbicia sygnału radiowego wskazuje również zmiana wartości parametru opisującego średnią różnicę wysokości między stacjami bazowymi a matrycami IPR ($\Delta h_{\mathrm{BS–RIS}}$).

Tabela 3: Wpływ wysokości zawieszenia matryc IPR na charakterystykę tłumienia ścieżki.

Parametr	Wartość		
Δh_{RIS}	0 m	5 m	10 m
$\Delta h_{\mathrm{BS–RIS}}$	9.57 m	4.62 m	−0.96 m
$\Delta\overline{\mathrm{PL}}$	1.39 %	1.21 %	0.99 %

5. WNIOSKI

Badania przeprowadzone w kontekście wykorzystania inteligentnych matryc IPR w sieciach mobilnych wykazały istotny potencjał tego rozwiązania w poprawie jakości odbieranego sygnału radiowego w obszarach gęsto zabudowanych. Przy uwzględnieniu różnych konfiguracji wysokości zawieszenia powierzchni odbijających oraz ich rozmieszczenia, obserwowane zmiany w tłumieniu wiązki radiowej oraz jej propagacji mogą mieć znaczący wpływ na efektywność działania całej sieci RAN.

Przeprowadzona analiza wskazuje, że optymalna wysokość montażu matryc IPR jest kluczowym czynnikiem wpływającym na poprawę jakości sygnału w sieciach bezprzewodowych. Wzrost wysokości matryc IPR prowadzi do zmian kątów elewacji i azymutu, co może negatywnie wpłynąć na efektywność odbioru sygnału. Wnioski te są zgodne z literaturą, która również sugeruje, że zmiany w geometrii ustawienia powierzchni IPR mogą wpływać na kąt padania i odbicia sygnału radiowego, a tym samym na jakość jego odbioru przez użytkownika. Dlatego konieczne jest dalsze badanie optymalnych warunków montażu IPR, aby maksymalizować korzyści płynące z ich integracji z sieciami bezprzewodowymi w rzeczywistych warunkach miejskich.

Podsumowując, wyniki badań potwierdziły, że inteligentne matryce IPR mają duży potencjał w poprawie jakości transmisji sygnału na obszarach miejskich. Jednakże ich efektywne wykorzystanie wymaga uwzględnienia wielu czynników, takich jak wysokość zawieszenia, rozmieszczenie oraz konfiguracja sieci RAN, co może być przedmiotem dalszych badań w celu optymalizacji działania (przyszłych) systemów bezprzewodowych.

PODZIĘKOWANIA

Autorzy pragną podziękować prof. Margot Deruyck z Uniwersytetu w Gandawie (IMEC) w Belgii za wsparcie niniejszej pracy poprzez udostępnienie narzędzia GRAND.

Praca została zrealizowana w ramach projektu nr 2021/43/B/ST7/01365 ufundowanego przez Narodowe Centrum Nauki (NCN) w Polsce.

Niniejszy artykuł jak i prace przedstawione w [4–6] zostały przygotowane we współpracy międzyuniwersyteckiej między Politechniką Poznańską a Uniwersytetem w Gandawie rozpoczętej w ramach programu grantowego Short-Term Scientific Mission (STSM) zorganizowanego przez akcję COST CA20120 INTERACT w roku 2023.

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.

[2] 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.

[3] A. Samorzewski, „Energy Consumption in Wireless Systems Equipped with RES, UAVs, and IRSs", *Journal of Telecommunications and Information Technology*, no. 2, pp. 35–40, 2023.

[4] A. Samorzewski and A. Kliks „5G cellular systems supported by UAVs, RESs, and RISs", in *Radiocommunication and Teleinformatics Conference 2023 (pol. Konferencja Radiokomunikacji i Teleinformatyki 2023 – KRiT 2023)*, Cracow, Poland, 2023, pp. 97–100.

[5] A. Samorzewski and A. Kliks, „5G Networks Supported by UAVs, RESs, and RISs", in *2023 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, Split, Croatia, 2023, pp. 1–6.

[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.

[7] 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.

[8] *Database and map of BTS station locations / UKE permits*. BTSearch. [Online]. Available: http://beta.btsearch.pl.

[9] *Poznan – Model 3D*. SIP (ang. Spatial Information System). [Online]. Available: http://sip.poznan.pl/model3d/#/legend.

[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.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

UŻYCIE OGRANICZONEJ KSIĄŻKI KODOWEJ DLA KONFIGUROWALNYCH MACIERZY ANTENOWYCH

APPLICATION OF REDUCED CODEBOOK FOR RECONFIGURABLE INTELLIGENT SURFACES

Marta Sieradzka[1]; Cyryl Prentki[2]; Tomasz Jaworski[2]; Paweł Hatka[2]; Dawid Brząkała[2]; Marcel Garczyk[2]; Paweł Płaczkiewicz[2]; Łukasz Kułacz[3]; Paweł Kryszkiewicz[3]; Adrian Kliks[3];

[1] Politechnika Poznańska, Poznań, marta.sieradzka.1@student.put.poznan.pl
[2] Politechnika Poznańska, Poznań, imie.nazwisko@student.put.poznan.pl
[3] Politechnika Poznańska, Poznań, imie.nazwisko@put.poznan.pl

DOI: 10.15199/59.2024.4.61

Streszczenie: w artykule przedstawiono wyniki przeprowadzonego eksperymentu pomiarowego z wykorzystaniem rekonfigurowalnych mataryc antenowych (RMA) dla polepszenia jakości detekcji sygnału. W pierwszej kolejności przedstawiono procedurę tworzenia ograniczonej książki kodowej. Następnie omówiono wybór najlepszego wzorca elementów antenowych dla polepszenia detekcji sygnału w wybranym punkcie. Zaprezentowano także porównanie wyników zastosowania prostego algorytmu detekcji energii w środowisku bez oraz z użyciem matryc antenowych.

Abstract: The paper presents the results of a measurement experiment using reconfigurable intelligent surfaces (RIS) to improve the quality of signal detection. First, the procedure for creating a reduced codebook is presented. Then, the selection of the best RIS pattern (i.e. pattern of antenna elements) to improve detection at a selected point was discussed. A comparison of the results of using a simple energy detection algorithm in an environment without and with the use of RIS is also presented.

Słowa kluczowe: Rekonfigurowlne Matryce Antenowe, Książka Kodowa, Wzorce Antenowe, Pomiary, Detekcja Sygnałów

Keywords: Reconfigurable Intelligent Antennas, Codebook, RIS Patterns, Measurements, Signal Detection

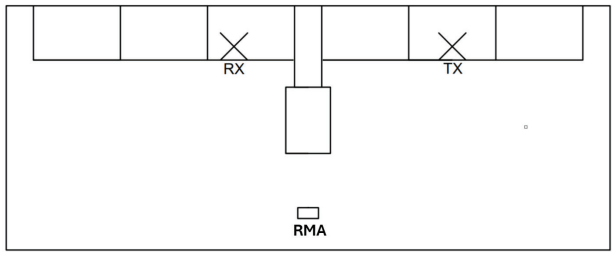
1. WPROWADZENIE

Rekonfigurowalne matryce antenowe (RMA), znane w literaturze anglojęzycznej jako Reconfigurable Intelligent Surfaces (RIS), stanowią jedno z intensywniej rozwijanych rozwiązań w kontekście poprawy efektywności przyszłych sieci bezprzewodowych [1, 2, 3]. Matryce te, zgodnie z założeniem, są kontrolowalnymi elementami umieszczonymi w środowisku propagacyjnym, pozwalając na zmianę jego charakterystyki transmisyjnej w taki sposób, aby np. zapewnić uzyskanie oczekiwanych efektów propagacyjnych po stronie odbiornika [6]. W szczególności RMA może umożliwić elastyczne zdefiniowanie kąta odbicia padającego na matrycę sygnału poprawiając w ten sposób moc odbieranego sygnału [5]
Macierze RMA są elementami określanymi często jako quasi-pasywne, gdzie same elementy antenowe nie są połączone ze wzmacniaczem mocy, a zużycie energii przez płytę RMA związane jest głównie z działaniem kontrolera [2, 3]. Macierze typowo są wyposażone w dużą liczbę elementów antenowych, które można konfigurować indywidualnie tworząc przez to wzorce odbiciowe matrycy RMA. Im większa liczba elementów antenowych tym większa jest liczba możliwych do uzyskania wzorców odbiciowych - jeżeli każdy element antenowy mógłby pracować w jednym z M stanów, to dla N-elementowej matrycy można wygenerować M^N różnych wzorców. Przestrzeń możliwości jest więc teoretycznie bardzo duża, co kontekście praktycznego jej zastosowania wymusza przyjęcie istotnych ograniczeń względem liczebności książki kodowej.
Niniejsza praca ma charakter eksperymentalny. Przedstawiono w niej wynik eksperymentu, w którym macierz RMA, dla której zaprojektowano ograniczoną książkę kodową, została wykorzystana dla polepszenia efektywności wykrywania obecności sygnału za pomocą algorytmu detekcji energii. W pracy wykorzystano matryce opracowane przez naukowców z Kolonii i zrealizowane w ramach projektu OpenSourceRIS [4].

2. SCENARIUSZ POMIAROWY

Nadrzędnym celem wykonanego eksperymentu była praktyczna weryfikacja możliwości wykorzystania rekonfigurowalnych macierzy antenowych RMA do poprawy funkcjonowania systemu bezprzewodowego. W szczególności skupiono się na ocenie potencjału użycia macierzy RMA wraz ze stworzoną dla niej ograniczoną książką kodową do poprawy algorytmu wykrywania obecności sygnału ("sensingu"). Zaplanowany eksperyment składał się z dwóch faz. W pierwszej należało stworzyć ograniczoną książkę kodową, w skład której wchodziła niewielka liczba wzorców definiujących tryb pracy poszczególnych elementów macierzy RMA. W drugiej fazie, dla zadanej książki kodowej, należało przeprowadzić próbę detekcji obecności sygnału w dwóch przypadkach - z obecnością matrycy RMA i przy jej braku. W pierwszej fazie eksperymentu rozpatrywano zbiór 27 różnych wzorców odbiciowych matrycy RMA (należy zaznaczyć, że zarówno metodologia tworzenia, jak i wybór tego zbioru wzorców wykracza poza ramy tego artykułu),o częstotliwości roboczej 5.15-

5.85 GHz oraz wymiarach 16×16 elementów odbijających, sterowanych indywidualnie za pomocą jednego bitu. W planowanym eksperymencie przyjęto położenie nadajnika, odbiornika oraz macierzy RMA tak, jak przedstawiono to na Rys. 1. Jako nadajnik wykorzystano generator sygnałowy Rohde & Schwarz SMM100A, z którego transmitowany był sygnał o częstotliwości nośnej 5.5 GHz w kierunku macierzy RMA za pomocą anteny kierunkowej METIS 5.1-5.8 GHz. Po stronie odbiorczej zastosowano natomiast analizator widma Rohde & Schwarz FSV3000, podłączony do takiej samej anteny. Aby odizolować od siebie nadajnik i odbiornik, pomiędzy nimi ustawiono ścianę składającą się z dwóch wysokich skrzyń, z czego jedna dodatkowo pokryta była farbą ekranującą pochłaniającą promieniowanie elektromagnetyczne. Anteny nadawcza i odbiorcza zostały umieszczone na tej samej wysokości.

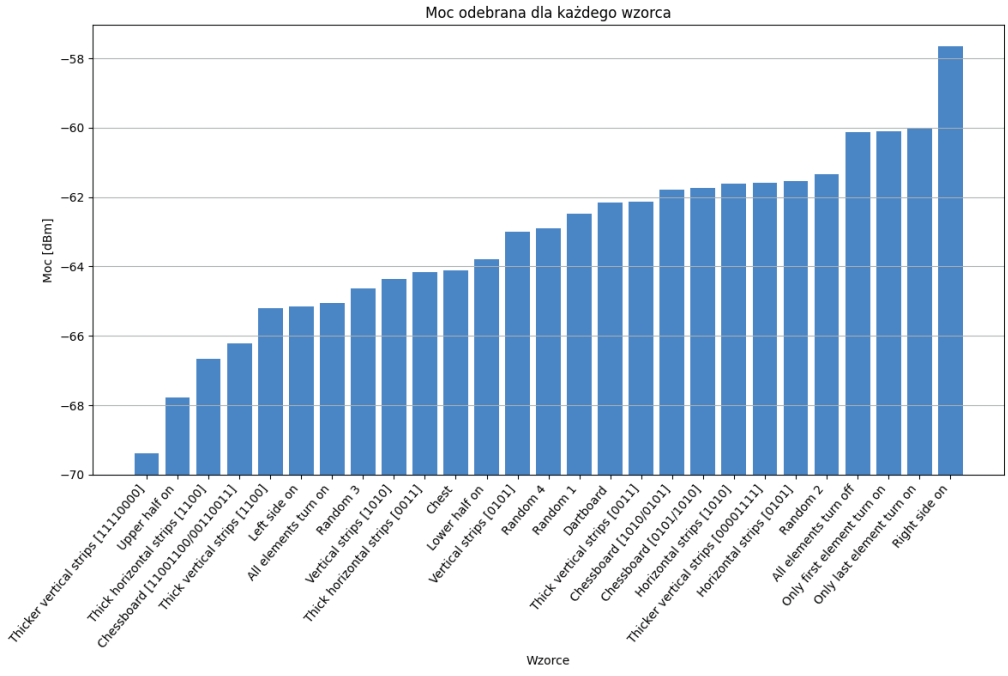


Rysunek 1: Rozmieszczenie elementów pomiarowych: nadajnika (TX), odbiornika (RX) oraz macierzy RMA

Zakładając, że chcemy maksymalizować prawdopodobieństwo wykrywania obecności sygnału użytkownika w punkcie za ścianą (oznaczonym na Rys. 1 jako RX), zmierzono wartości mocy odbieranych w tym punkcie dla każdego z dostępnych 27 wzorców. Proces ten został zautomatyzowany poprzez napisanie odpowiedniego oprogramowania w języku Python, dzięki któremu można było zmieniać ustawienia generatora i analizatora i badać charakterystykę odbiociową różnych wzorców. W konsekwencji udało stworzyć się wspomniany zbiór 27 wzorców testowych, z których tworzona była później ograniczona książka kodowa.

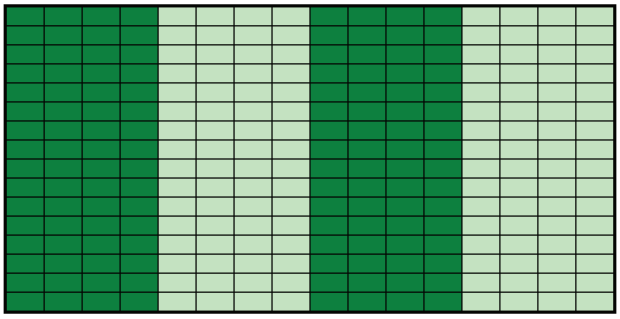
3. EKSPERYMENT POMIAROWY - WYBÓR WZORCA

Jak już wspomniano, pierwsza faza eksperymentu została zautomatyzowana. Badanie właściwości odbiciowych poszczególnych wzorców odbywało się poprzez programową zmianę badanego w danej chwili wzorca ustawionego na module RMA i zmierzenie poziomu mocy odbieranej przez antenę odbiorczą w badanym punkcie RX. Nadawany był stały w częstotliwości sygnał sinusoidalny o mocy -10 dBm.Antena nadawcza (TX), jak i odbiorcza (RX) mały polaryzację pionową względem RMA. Wynik pomiaru dla wszystkich badanych wzorców przedstawiono na Rys. 2. Ponieważ calem drugiej fazy eksperymentu było pokazanie możliwości poprawy jakości działania systemu poprzez dobór odpowiedniego wzorca, przy tworzeniu ograniczonej książki kodowej kierowano się zasadą wyboru dwóch wzorców: takiego, który maksymalnie polepsza jakość odbieranego sygnału w badanym punkcie, i takiego, który nie wnosi do działania systemu istotnej poprawy. Stosując przyjęte kryterium wyboru, analiza przedstawionego wykresu (Rys. 2) jednoznacznie wskazuje na dwa wzorce, które powinny być uwzględnione w książce kodowej. W szczególności są to: wzorzec opisany na rysunku jako "Thicker vertical strips [11110000]" (dla którego mierzona moc sygnału w badanym punkcie wynosiła -69,39 dBm) oraz wzorzec oznaczony jako "Right side on" (ze zmierzoną mocą sygnału -57,66 dBm). Różnica poziomów mocy odbieranej w badanym punkcie różni się pomiędzy dwoma wybranymi wzorcami o około 12 dB.

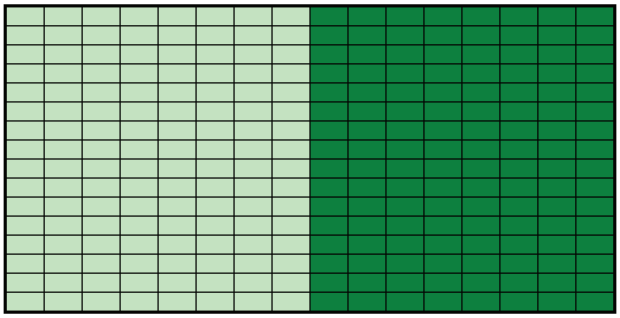


Rysunek 2: Wyniki pomiarów mocy odbieranej w punkcie RX dla wszystkich 27 wzorców testowych

Pierwszy wybrany wzorzec (Rys. 3) składał się z dwóch pionowych grusów, pasów pionowych. Pasy te naprzemiennie albo odbijały sygnał bez zmiany fazy, albo powodowały jej przesunięcie o 180 stopni. Drugi wzorzec natomiast (Rys. 4) dzielił płaszczyznę matrycy RMA na dwie duże równe części, gdzie jedna połowa odbijała sygnał bez zmiany fazy, druga zaś powodowała obrócenie fazy sygnału padającego o 180 stopni.



Rysunek 3: Pierwszy z wybranych wzorców odbijania (jasnozielone elementy są wyłączone, natomiast ciemnozielone włączone)



Rysunek 4: Drugi z wybranych wzorców odbijania (jasnozielone elementy są wyłączone, natomiast ciemnozielone włączone)

4. DETEKCJA SYGNAŁU

Po wyselekcjonowaniu dwóch wzorców odbiciowych matrycy RMA przeprowadzono drugą fazę eksperymentu, w której podczas pomiarów zbierano 1000 próbek poziomów mocy przez 20 sekund. Rozważano trzy przypadki eksperymentalne: pomiar samego szumu (traktowany jako pomiar odniesienia), detekcja sygnału dla obu wybranych wzorców RMA. Dodatkowo te trzy przypadki rozpatrzono w dwóch wariantach: A - kiedy moc sygnałów odebranych jest wyraźnie większa od mocy szumu (dzięki czemu można osiągać bardzo niskie wartości prawdopodobieństwa fałszywego alarmu), B (kiedy odebrana moc jest zbliżona do mocy szumu, przez co prawdopodobieństwo alarmu może być trudne do zapewnienia nawet dla długiego czasu obserwacji). Moc nadawaną przez generator ustawiono na -10 dBm w pierwszym przypadku (A) i -45 dBm w drugim przypadku (B). Wyniki pomiarów przedstawiono na rysunkach 5 oraz 6.

W celu wykrycia obecności sygnału zastosowano najprostszy algorytm detekcji energii, którego główna zasada polega na porównaniu zmierzonej średniej mocy z K próbek z progiem decyzyjnym. W pierwszym kroku konieczne jest wyznaczenie średniej mocy szumu σ^2, a następnie dla określonego prawdopodobieństwa fałszywego alarmu P_{fa} obliczenie progu decyzyjnego Γ zgodnie z wzorem [7]:

$$\Gamma = \sigma * \left(Q(P_{fa}) \sqrt{\frac{2}{K}} + 1 \right), \tag{1}$$

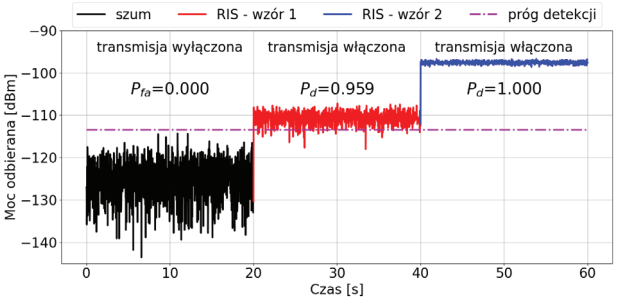
gdzie Q jest funkcją dopełnienia dystrybuanty rozkładu normalnego, a K jest liczbą próbek. Ostatnim krokiem jest podjęcie decyzji Ψ o obecności sygnału poprzez porównanie z obliczonym wcześniej progiem decyzyjnym:

$$\Psi = \begin{cases} 1 & \text{dla } P_{rx} \geq \Gamma \\ 0 & \text{w pozostałych przypadkach} \end{cases} \tag{2}$$

gdzie P_{rx} jest mocą odbieranego sygnału.

Jeżeli wartość zmierzonej mocy jest powyżej progu uznaje się, że hipoteza testowa o obecności sygnału jest prawdziwa. W przeciwnym razie przyjmuje się, że poprawna jest hipoteza przeciwna. Próg decyzyjny dla algorytmu detekcji energii dobierać można na podstawie różnych założeń, w niniejszym eksperymencie przyjęto dość popularne podejście, w którym próg decyzyjny wyznacza się dla przyjętej stałej wartości prawdopodobieństwa fałszywego alarmu P_{fa}. W szczególności:

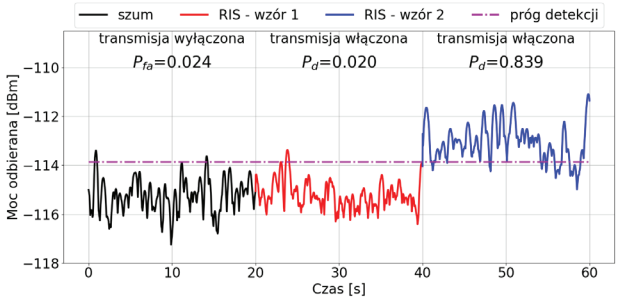
W scenariuszu A gdzie moc odbieranego sygnału jest dość wysoka (ponad 10 dB powyżej mocy szumu), wykorzystanie odpowiedniej konfiguracji (wybór odpowiedniego wzorca), zapewnia poprawę prawdopodobieństwa detekcji sygnału z 95.9% na 100.0% (zakładając próg detekcji energii dla prawdopodobieństwa fałszywego alarmu na bardzo niskim poziomie 10^{-10}).



Rysunek 5: Wyniki pomiaru dla scenariusza A

W scenariuszu B, typowym dla detekcji zajętości pasma, gdzie moc sygnału jest bardzo zbliżona do mocy szumu, wykorzystanie odpowiedniej konfiguracji macierzy RMA zapewnia poprawę prawdopodobieństwa detekcji sygnału z 2.0% na 83.9%. Oznacza to w praktyce, że jesteśmy w stanie dokonać poprawnej detekcji sygnału, którego bez matrycy RMA nie

bylibyśmy w stanie wykryć. Przedstawione wyniki pokazują także, że wybór niewłaściwego wzorca nie będzie skutkował poprawą jakości działania algorytmu detekcji.



Rysunek 6: Wyniki pomiaru dla scenariusza B

5. PODSUMOWANIE

W pracy przedstawiono wyniki eksperymentu pokazującego na możliwość praktycznego wykorzystania ograniczonej książki kodowej (składającej się z różnych wzorców odbiciowych matryc) do polepszenia wybranych parametrów systemu bezprzewodowego. W szczególności skupiono się na zwiększeniu wydajności metod detekcji sygnału. W pracy udało się wykazać, że dzięki zastosowaniu odpowiedniego wzorca sygnał padający można skierować w kierunku wskazanego celu w taki sposób, że moc sygnału obserwowanego w tej lokalizacji zostaje zwiększona. Z kolei dobranie złego wzorca odbiciowego nie powoduje polepszenia działania systemu. W takim ujęciu matryca RMA może być wykorzystana np. w fabrykach czy hala produkcyjnych. Autonomiczne urządzenia poruszające się w takim środowisku zwykle wymagają stałego połączenia z siecią. Odpowiednie rozmieszczenie urządzeń typu inteligentnych matryc mogło by pozwolić na podążanie sygnałem radiowym za robotem.

6. PODZIĘKOWANIA

Praca powstała w ramach projektu badawczego OPUS 2021/43/B/ST7/01365 finansowanego przez Narodowe Centrum Nauki 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] 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, dostęp online: 28 czerwca 2023 r.

[3] ETSI. 2023. "Reconfigurable Intelligent Surfaces (RIS); Communication Models, Channel Models, Channel Estimation and Evaluation Methodology". ETSI GR RIS 003 V1.1.1 (2023-06). https://www.etsi.org/committee/1966-ris, dostęp online: 28 czerwca 2023 r.

[4] 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. doi: 10.1109/ISAP57493.2023.10389095.

[5] Liu, R., Dou, J., Li, P., Wu, J., Cui, Y. 2022. "Simulation and Field Trial Results of Reconfigurable Intelligent Surfaces in 5G Networks". IEEE Access 10: 122786-122795. doi: 10.1109/ACCESS.2022.3223447.

[6] Renzo, M.D., 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

[7] Yucek, T., Arslan, H. 2009. "A survey of spectrum sensing algorithms for cognitive radio applications". IEEE Communications Surveys & Tutorials 11 (1): 116-130. doi: 10.1109/SURV.2009.090109.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

ANALIZA POSTACI SYGNAŁU SYNCHRONIZACYJNEGO NPSS W NB-IOT

NPSS NB-IOT SYNCHRONIZATION SIGNAL FORM ANALYSIS

Olga Błaszkiewicz[1]; Jarosław Sadowski[2]; Piotr Rajchowski[3]

[1] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, olga.blaszkiewicz@pg.edu.pl
[2] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, jaroslaw.sadowski1@pg.edu.pl
[3] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, piotr.rajchowski@pg.edu.pl

DOI: 10.15199/59.2024.4.62

Streszczenie**: W artykule zaprezentowane zostały możliwości zmiany postaci sekwencji Zadoff-Chu używanej do wygenerowania sygnału synchronizacyjnego NPSS w interfejsie radiowym NB-IoT. Modyfikacji poddano elementy charakterystyczne sekwencji z uwzględnieniem: *root index*, ciągu binarnego czy składnika funkcji wykładniczej w celu poprawy właściwości korelacyjnych na potrzeby synchronizacji pracy terminali użytkowników.**

Abstract**: In the article Zadoff-Chu sequence, which is used to generate NPSS synchronization signal in NB-IoT interface, modified forms are presented. Modification of sequences included characteristic elements of it such as: index root, binary sequence and a part of exponential function. The main purpose of this work is to improve correlation properties for synchronization of users' terminals.**

Słowa kluczowe**: NB-IoT, NPSS, sekwencja binarna, Zadoff-Chu.**

Keywords**: NB-IoT, NPSS, binary sequence, Zadoff-Chu.**

1. WSTĘP

Transmitowane cyklicznie przez stacje bazowe sygnały synchronizacyjne są potrzebne do przeprowadzenia procesu synchronizacji, który jest niezbędny dla zapewniania komunikacji terminali użytkowników ze stacjami bazowymi zgodnie ze strukturą czasową obowiązującą w danym standardzie.

W interfejsie radiowym NB-IoT (ang. *Narrowband Internet of Things*) wyróżniane są dwa typy sygnałów synchronizacyjnych, tj. NPSS (ang. *Narrowband Primary Synchronization Signal*) oraz NSSS (ang. *Narrowband Secondary Synchronization Signal*). W niniejszym artykule analizie poddano sygnał synchronizacyjny NPSS, który transmitowany jest cyklicznie co 10 ms, w #5 podramce struktury ramkowej w łączu w dół [1].

Niniejszy artykuł został podzielony na trzy główne rozdziały: rozdział drugi prezentujący sygnały synchronizacyjne typu PSS (ang. *Primary Synchronization Signal*) w różnych interfejsach radiowych z uwzględnieniem elementów charakterystycznych dla analizowanego sygnału synchronizacyjnego, rozdział trzeci traktujący o wprowadzonych modyfikacjach do sekwencji wzorcowej służącej do generowania sygnału synchronizacyjnego dla interfejsu NB-IoT oraz rozdział zawierający podsumowanie i opis przyszłych prac.

2. SYGNAŁY SYNCHORNIZACYJNE TYPU PSS W INTERFEJSACH RADIOWYCH: LTE, NB-IoT i 5G NR

Na potrzeby niniejszych analiz zapoznano się z postaciami sekwencji wzorcowych używanych do generowania sygnałów synchronizacyjnych typu PSS w standardzie LTE (ang. *Long-Term Evolution*), NB-IoT oraz 5G NR (ang. *5th Generation New Radio*). Decyzja ta umotywowana jest bezpośrednią bliskością poszczególnych wydań standardów (Rel. 13 dla NB-IoT i Rel. 15 dla 5G NR) przez organizację 3GPP (ang. *Third Generation Partnership Project*), a także faktem bazowania na rozwiązaniach znanych ze standardu LTE w celu opracowania rozwiązań dla interfejsu NB-IoT.

Sygnał synchronizacyjny typu PSS zarówno w interfejsie LTE (zależność (1)) jak i w NB-IoT (zależność (2)) generowany jest przy użyciu sekwencji Zadoff-Chu [1].

$$d_{u_{LT}}(n_{LT}) = \begin{cases} e^{-j\frac{\pi u_{LT} n_{LT}(n_{LT}+1)}{63}}, n_{LT} = 0, 1, \dots, 30, \\ e^{-j\frac{\pi u_{LT} n_{LT}(n_{LT}+1)(n_{LT}+2)}{63}}, n_{LT} = 31, 32, \dots, 61, \end{cases} \quad (1)$$

gdzie:

$$u_{LT} = \begin{cases} 25, \ N_{ID_LT}^{(2)} = 0, \\ 29, N_{ID_LT}^{(2)} = 1, \\ 34, N_{ID_LT}^{(2)} = 2. \end{cases}$$

$$d_{l_{NB}}(n_{NB}) = S(l_{NB}) \cdot e^{-j\frac{\pi u_{NB} n_{NB}(n_{NB}+1)}{11}}, \quad (2)$$

gdzie:

$n_{NB} = 0, 1, \dots, 10,$

$u_{NB} = 5,$

$l_{NB} = 0, 1, \dots, 10,$

$S(l_{NB}) = [1\ 1\ 1\ 1\ -1\ -1\ 1\ 1\ 1\ -1\ 1].$

Sygnał synchronizacyjny w interfejsie 5G NR [2], w odróżnieniu od sygnałów PSS z LTE i NB-IoT, generowany jest już na podstawie sekwencji typu m (ang. *m-sequence*) zgodnie z zależnością (3).

$$d_{PSS}(n_{NR}) = 1 - 2x_{NR}(m_{NR}), \quad (3)$$

gdzie:

$$m_{NR} = \left(n_{NR} + 43N_{ID_NR}^{(2)}\right) mod\ 127,$$

$$0 \le n_{NR} < 127,$$

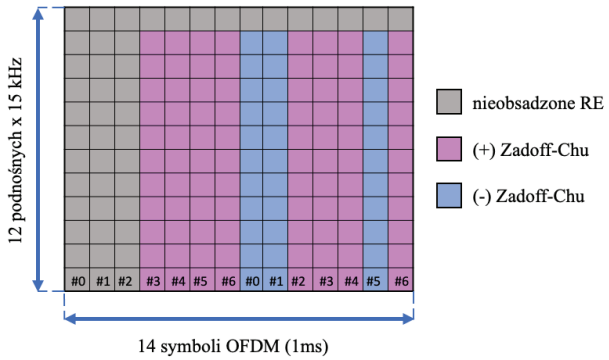
$$x_{NR}(i_{NR} + 7) = \left(x_{NR}(i_{NR} + 4) + x_{NR}(i_{NR})\right) mod\ 2,$$

$$[x_{NR}(6)\ x_{NR}(5)\ x_{NR}(4)\ x_{NR}(3)\ x_{NR}(2)\ x_{NR}(1)\ x_{NR}(0)] = [1\ 1\ 1\ 0\ 1\ 1\ 0].$$

Analizując przytoczone zależności można zauważyć, że w LTE występują trzy postaci sygnałów synchronizacyjnych PSS, których wartość zdeterminowana jest wartością identyfikatora komórki, natomiast w NB-IoT występuje tylko jedna postać tego sygnału niezależna od innych zmiennych.

Dodatkowo, cechami charakterystycznymi sekwencji Zadoff-Chu jest występowanie składnika *root index* u (w analizowanych przypadkach to u_{LT} oraz u_{NB}), którego wartości mogą zmieniać się w zakresie od 1 do $N_{ZC} - 1$, gdzie N_{ZC} odpowiada długości sekwencji i stanowi mianownik funkcji wykładniczej. Wartość ta musi być liczbą nieparzystą (bywa też liczbą pierwszą) [3].

Na podstawie zależności (2) generowanych jest 11 symboli Zadoff-Chu w dziedzinie częstotliwości, powielanych 11 razy w dziedzinie czasu z uwzględnieniem mnożenia przez sekwencję binarną, co przekłada się na 121 elementów zasobów RE (ang. *Resource Element*) które zostają przypisane do siatki zasobów RG (ang. *Resource Grid*) w sposób zaprezentowany na rys. 1 zgodny z wytycznymi pochodzącymi ze standardu [1] a dotyczącymi nieobsadzania RE z powodu innych elementów występujących na pierwszych trzech symbolach OFDM (ang. *Orthogonal Frequency Division Multiplexing*) [7].



Rys. 1. Sposób ulokowania elementów sekwencji Zadoff-Chu w siatce zasobów.

Pomimo podobieństw w strukturze ramkowej wymienionych interfejsów, w 5G NR wprowadzono zmienną konfigurację interfejsu radiowego (rozdział 4.2 [2]), która wpływa na liczbę szczelin czasowych w ramach podramki co przekłada się na czas trwania symbolu OFDM. Dodatkowo, sygnał synchronizacyjny typu PSS został ulokowany zgodnie ze sposobem zaprezentowanym w rozdziale 7.4.3 [2] na 127 podnośnych. Jako że przedmiotem niniejszego opracowania nie jest porównywanie sposobu ulokowania, a tylko rodzaj sekwencji synchronizacyjnej to przedstawiony na rys. 1 sposób ulokowania elementów sekwencji synchronizacyjnej typu PSS jest aktualny tylko dla interfejsu NB-IoT.

3. MODYFIKACJE WPROWADZONE W POSTACI SEKWENCJI WZORCOWEJ

Mając na uwadze, że właściwości sygnału synchronizacyjnego zależą od sekwencji, na podstawie której zostały wygenerowane, przed rozpoczęciem wprowadzania modyfikacji do jej postaci zdefiniowano miarę, która pozwoliła na jakościowe porównanie uzyskiwanych wyników. Szczególnie interesująca, z punktu widzenia zagadnienia synchronizacji, jest właściwość korelacyjna wygenerowanego sygnału, stąd przyjętą miarą jakości jest stosunek procentowy (SP) prążka o maksymalnej wartości lokalnej funkcji korelacji sygnału w odniesieniu do prążka o maksymalnej wartości globalnej tejże funkcji. Przy czym maksimum lokalne należy rozumieć jako wartość maksymalną ze zbioru wartości funkcji korelacji z wyzerowaną wartością prążka o maksymalnej wartości globalnej.

Przyjęta miara jakościowego porównania swoje uzasadnienie znajduje w stosowanych metodach korelacyjnych służących do określenia momentu wystąpienia sygnału synchronizacyjnego i bazujących na detekcji wartości maksymalnej lub wartości prążka powyżej jakiegoś zdefiniowanego progu.

3.1. Zmiana wartości *root index*

Pierwsza z wprowadzonych w sekwencji wzorcowej modyfikacji dotyczyła zmiany wyłącznie wartości *root index*, która zgodnie z postacią zależności (2) równa się $u_{NB} = 5$. Zmian dokonano w zakresie od 1 do 10 z krokiem co 1. Wyniki przeprowadzonych analiz zamieszczono w tab. 1.

Tab. 1. Wpływ zmiany wartości root index na SP

u_{NB}	*SP [%]*
1	18,56
2	23,76
3	25,54
4	31,88
5	31,41
6	31,13
7	32,07
8	25,33
9	23,51
10	18,67

3.2. Zmiana wartości dodajnika

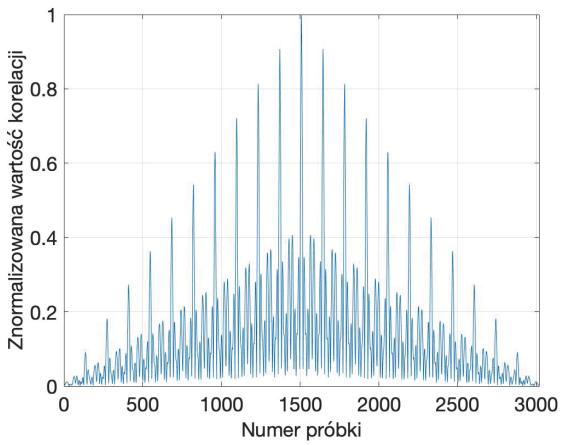
Druga z wprowadzonych w sekwencji wzorcowej modyfikacji dotyczyła zmiany wartości dodajnika wyrażenia ujętego w zależności (2). Tym razem zakres zmian wartości tej części sumy mieścił się w przedziale od 1 do 22 z krokiem co 1. Wyniki przeprowadzonych obliczeń zamieszczono w tab. 2.

Tab. 2. Wpływ zmiany wartości dodajnika na SP

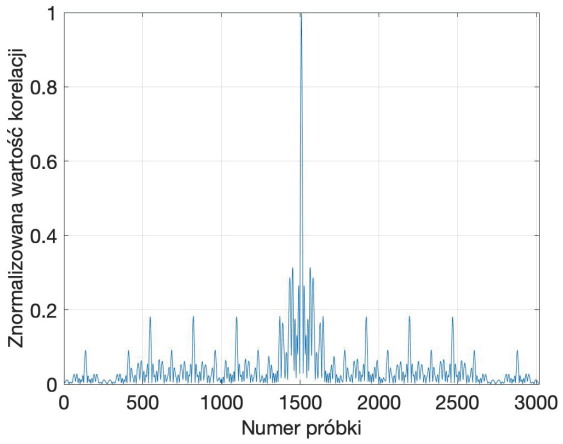
dodajnik	*SP [%]*	*dodajnik*	*SP [%]*
1	31,41	12	37,32
2	41,89	13	42,91
3	40,65	14	32,1
4	41,22	15	42,39
5	30,98	16	35,97
6	38,12	17	38,6
7	30,06	18	28,38
8	30,28	19	31,68
9	35,04	20	33,43
10	30,27	21	30,38
11	39,96	22	40,29

3.3. Zmiana postaci ciągu binarnego

Trzecia z wprowadzonych w sekwencji wzorcowej modyfikacji wymagała wytypowania ograniczonego zbioru ciągów binarnych o dobrych właściwościach korelacyjnych spośród 2^{11} możliwych jej postaci, a także sprawdzenia zasadności stosowania jakiegokolwiek ciągu binarnego. Stąd na rys. 2 i 3 zaprezentowano dwie wynikowe korelacje dla sekwencji wzorcowej Zadoff-Chu z ciągiem binarnym o postaci zaprezentowanej w standardzie [1] oraz bez ciągu binarnego.



Rys. 2. Funkcja korelacji dla sekwencji wzorcowej Zadoff-Chu bez ciągu binarnego



Rys. 3. Funkcja korelacji dla sekwencji wzorcowej Zadoff-Chu z ciągiem binarnym

Wyraźnie widać tu przewagę rozwiązania sekwencji wzorcowej z ciągiem binarnym, gdyż różnica pomiędzy maksymalnym prążkiem funkcji autokorelacji a prążkiem o wartości maksimum lokalnego jest znacząca. Występowanie wielu prążków o tak wysokiej energii jak te zaprezentowane na rys. 2 mogłyby doprowadzić do błędnie przeprowadzonego procesu synchronizacji w warunkach propagacyjnych dalekich od idealnych.

Wstępnie, na podstawie źródeł literaturowych [4] i [6] wytypowano pięć typów ciągów binarnych (tab. 3) i określono dla nich wartości wprowadzonej miary jakościowej SP (tab. 4).

Tab. 3. Postaci wybranych 11-elementowych ciągów binarnych

Nazwa ciągu	*Postać ciągu*
Barkera	1 1 1 -1 -1 -1 1 -1 -1 1 -1
Legrange'a	-1 1 -1 1 1 1 -1 -1 -1 1 -1
Williarda	1 1 1 -1 1 1 -1 1 -1 -1 -1
Naumanna-Hoffmana	-1 -1 -1 -1 1 -1 1 1 -1 -1 1
Weila 1	-1 -1 -1 1 1 -1 1 1 -1 -1 1
Weila 2	1 1 -1 1 -1 -1 1 -1 1 -1 -1
Weila 3	-1 1 -1 -1 -1 -1 -1 1 1 1 1
Weila 4	-1 1 1 -1 -1 1 1 1 -1 -1 -1
Weila 5	-1 -1 1 -1 1 -1 1 -1 1 1 -1
ciąg ze standardu NB-IoT	1 1 1 1 -1 -1 1 1 1 -1 1

Tab. 4. Wpływ zmiany ciągów binarnych na SP

Nazwa ciągu	*SP [%]*
Barkera	29,27
Legrange'a	36,02
Williarda	29,59
Naumanna-Hoffmana	29,59
Weila 1	45,15
Weila 2	45,46
Weila 3	36,17
Weila 4	45,15
Weila 5	54,47
ciąg ze standardu NB-IoT	31,41

4. PODSUMOWANIE I PRZYSZŁE PRACE

W niniejszym artykule poruszono zagadnienie związane z procesem synchronizacji w interfejsie NB-IoT.

W rozdziale drugim zaprezentowano ogólną postać sekwencji synchronizacyjnej używanej do generowania sygnału synchronizacyjnego NPSS zgodnie ze standardem NB-IoT, a także wymieniono jej charakterystyczne cechy.

W rozdziale trzecim przedstawiono wpływ zmian wartości *root index*, dodajnika oraz postaci ciągu binarnego na uzyskiwane wartości wprowadzonej miary jakościowej. Każda z wprowadzonych modyfikacji wpłynęła na poprawę uzyskiwanych wyników (im niższa wartość SP, tym lepiej), co stało się podstawą do twierdzenia, że zasadne jest wprowadzanie dalszych modyfikacji sekwencji wzorcowej, gdyż jej ustandaryzowana postać nie

jest najlepsza biorąc pod uwagę wprowadzoną miarę jakościową.

Dotychczas sprawdzane osobno zmiany wartości parametrów charakterystycznych można zatem rozszerzyć o wprowadzenie modyfikacji związanej z testowaniem zmian dokonywanych w wartościach kilku parametrów jednocześnie.

W celu ograniczenia przeszukiwanego zbioru rozwiązań przyjęto dodatkowe kryterium związane z wyborem postaci ciągu binarnego. W celu jakościowego porównania tych ciągów przyjęto następujące miary: PSL (ang. *Peak Side Lobe*), MPS (ang. *Minimal Peak Side*) oraz MF (ang. *Merit Factor*).

Pierwsza z miar określa maksymalną wartość prążka bocznego funkcji autokorelacji. Minimalną wartość tego prążka określa MPS, a MF to stosunek liczby elementów ciągu binarnego do energii prążków bocznych funkcji autokorelacji [5]. Wyniki zbiorcze dla trzech miar ujęto w tab. 4.

Tab. 4. Wartości PSL, MPS i MF dla 11-elementowych ciągów binarnych

Nazwa ciągu	*Wartość PSL*	*Wartość MPS*	*Wartość MF*
Barkera	1	1	12,1
Legrange'a	4	2	1,34
Williarda	3	1	2,88
Naumanna-Hoffmana	2	1	4,65
Weila 1	5	2	1,14
Weila 2	5	2	0,88
Weila 3	4	2	1,06
Weila 4	5	2	1,06
Weila 5	3	3	0,93
ciąg ze standardu NB-IoT	2	2	2,88

Analiza wartości wynikowych parametrów PSL, MPS oraz MF (tab. 4) pozwala na stwierdzenie, że najlepszymi właściwościami korelacyjnymi charakteryzuje się ciąg Barkera, gdyż wartości parametrów PSL i MPS są najmniejsze, a wartość MF największa, stąd do dalszych analiz wytypowano ciąg Barkera.

Oddzielna modyfikacja wybranych wartości parametrów charakterystycznych sekwencji synchronizacyjnej NPSS wykazała, że można znaleźć sekwencje o lepszych właściwościach korelacyjnych mogących wpływać na wyższą jakość synchronizacji terminali użytkowników. Trudność w znalezieniu jawnych zależności łączących strukturę sekwencji synchronizacyjnej z przyjętą miarą jakościową przekłada się na brak użyteczności podejścia analitycznego w celu znalezienia optymalnej postaci sekwencji synchronizacyjnej dla interfejsu NB-IoT. Z tego powodu dalsze prace będą nakierowane na opracowanie metody oraz narzędzia do poszukiwania korzystnej postaci sekwencji synchronizacyjnej z użyciem algorytmów genetycznych.

LITERATURA

[1] 3GPP TS 36.211 version 14.14.0 Release 14. 2020
[2] 3GPP TS 38.211 version 15.2.0 Release 15. 2018
[3] Andrews G. Jeffrey. 2022. A Primer on Zadoff Chu Sequence
[4] Holmes K. Jack. 2007. Spread Spectrum Systems for GNSS and Wireless Communications. Artech House, INC.
[5] Leukhin N.A., Parsaev V.N., Bezrodnyi I.V., Kokovihina A.N. 2017. The Exhaustive Search for Optimum Minimum Peak Sidelobe Binary Sequences. Bulletin of the Russian Academy of Sciences: Physics, Vol. 81, No. 5, pp. 575–578.
[6] Rushanan J. Joseph. 2006. Weil Sequences: A Family of Binary Sequences with Good Correlation Properties. 2006 IEEE International Symposium on Information Theory
[7] Schlienz J., Raddino D. Whitepaper Narrow-Band_IoT ─ 1MA266_0e.

FRAMEWORK INTERFEJSU RADIOWEGO NB-IOT

NB-IOT RADIO INTERFACE FRAMEWORK

Olga Błaszkiewicz; Jarosław Magiera; Alicja Olejniczak; Piotr Rajchowski;
Jarosław Sadowski; Jacek Stefański; Krzysztof K. Cwalina

Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk,
{olga.blaszkiewicz, jaroslaw.magiera, aliolejn, piorajch,
jarsadow1, jacek.stefanski1, kkcwalina}@pg.edu.pl

DOI: 10.15199/59.2024.4.63

Streszczenie: W niniejszym artykule opisano rezultaty uzyskane w ramach zrealizowanego projektu definiowanego programowo frameworka wąskopasmowego interfejsu radiowego dla urządzeń Internetu Rzeczy. Przedstawiono elementy składowe konfigurowalnego interfejsu radiowego oraz jego charakterystykę eksploatacyjną.

Abstract: In the article the results obtained in the design of a software-defined framework for a narrowband radio interface for Internet of Things devices are described. The components of the configurable radio interface and its operational characteristics are presented.

Słowa kluczowe: framework, Internet Rzeczy, IoT, NB-IoT

Keywords: framework, Internet of Things, IoT, NB-IoT

1. WSTĘP

Rozwijający się rynek urządzeń Internetu Rzeczy IoT (Internet of Things) oraz rosnące zapotrzebowanie na urządzenia, które mają wspierać działanie nowoczesnych systemów telemetrycznych, wymuszają także rozwój interfejsów radiowych. Interfejsy te mają być używane do komunikacji i przesyłania danych pomiarowych, także w ramach krajowej infrastruktury krytycznej i w zastosowaniach specjalnych, jak np. systemach alarmowych, licznikach gazu, energii elektrycznej, wody czy też do sygnalizacji szeroko pojętych służb specjalnych. Jeden z takich interfejsów dla inteligentnych urządzeń IoT stanowi wąskopasmowa komunikacja NB-IoT (Narrowband IoT) pomiędzy urządzeniami D2D (Device-to-Device). Wykorzystuje ona istniejącą infrastrukturę stałą sieci LTE (Long Term Evolution) oznaczoną jako LTE kat. NB1/NB2, która jest i będzie wykorzystywana szczególnie w przypadkach gdy urządzenia IoT będą wymagały energooszczędnych trybów pracy i tym samym także efektywnej energetycznie komunikacji radiowej i przesyłania telemetrycznych danych użytecznych z niskimi przepływnościami rzędu kilkuset kb/s.

Odpowiedzią na tak zdefiniowane zapotrzebowanie rynku urządzeń IoT jest zrealizowany framework interfejsu radiowego NB-IoT w ramach konsorcjum firmy DGT Sp. z o.o. (DGT) jako lidera oraz Politechniki Gdańskiej (PG) jako konsorcjanta opracowującego i realizującego część programową interfejsu radiowego. W niniejszym artykule opisano założenia, wytyczne oraz architekturę zrealizowanej części programowej definiowanego programowo, uniwersalnego interfejsu radiowego do inteligentnych urządzeń IoT o akronimie: DUCH IoT, które należy rozumieć jako podsumowanie prac opisywanych wcześniej w [1, 6].

2. ZAŁOŻENIA PROJEKTOWE

Założenia dotyczące postaci frameworka interfejsu radiowego determinowane są głównie przez: wymagania normatywne i techniczne stosu protokołów NB- IoT kat. 13 i kat. 14, zakładane innowacyjne funkcjonalności produktu końcowego oraz koncepcję i metodykę realizacji projektu (produktu końcowego) z uwzględnieniem spełnienia wymagań VIII poziomu gotowości technologicznej (PGT) jako rezultatu projektu na koniec prac badawczo-rozwojowych.

Funkcje torów nadawczo-odbiorczych miały zostać zrealizowane jako uniwersalne moduły programowo-sprzętowe, o konfigurowalnym podziale funkcjonalności pomiędzy realizacje programową (realizowaną przez PG) i sprzętową (realizowaną przez firmę DGT) za pomocą interfejsów komunikacyjnych (IK) w zależności od możliwości obliczeniowych i zasobów sprzętowych platformy uruchomieniowej. Umożliwia to wymienność funkcji NB1/NB2 zgodnie ze specyfikacją 3GPP (Third Generation Partnership Project) pomiędzy modułami zaimplementowanymi w części programowej (w języku C/C++, jako oprogramowanie proceduralne), a zdublowanymi modułami sprzętowymi typu IP (Intellectual Property) Core (w języku VHDL/Verilog) w części sprzętowej. Dzięki temu architektura może być zdekomponowana na niezależne bloki realizowane na różnych urządzeniach, a nawet w różnych architekturach. Określono także liczbę możliwych do obsłużenia platform uruchomieniowych (platform sprzętowych urządzeń IoT) i założono obsługę m.in. systemu SoC (System-on-Chip) w technologii FPGA (Field-Programmable Gate Array), systemu SoC opartego na mikroprocesorze w architekturze ARM (Advanced RISC Machine) oraz architekturę rozproszoną (CPU x86-64 + FPGA) z zastosowaniem platformy radia programowalnego SDR (Software Defined Radio).

Osiągnięcie wymienionych funkcjonalności z oczywi-

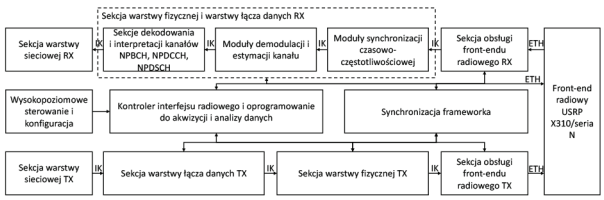
stych względów wymagało opracowania architektury frameworka wspierającej współbieżną, równoczesną analizę sygnałów radiowych. Dodatkowo, zgodnie z przeznaczeniem pracy urządzeń wykorzystujących interfejs radiowy NB-IoT w różnych środowiskach charakteryzujących się zmiennymi i trudnymi warunkami propagacyjnymi, istotnym elementem była realizacja adaptacyjnych algorytmów do odbioru sygnałów radiowych i estymacji stanu kanału radiowego [4, 7, 9]. Szczególnie uwzględniając docelowe warunki pracy m.in. przy niskiej mocy sygnału odbieranego, w obecności zakłóceń przemysłowych oraz interferencji wewnątrzsystemowych. Przy czym, w dokumentach standaryzacyjnych 3GPP opisywane są wymagania dotyczące toru odbiorczego urządzeń radiowych, a nie metody przetwarzania sygnałów, stąd zarówno koncepcja jak i algorytmy przetwarzania sygnałów w torze odbiorczym w kontekście zastosowania konfigurowalności frameworka stanowiły element innowacyjny i badawczy w projekcie.

Framework powinien umożliwiać dwustronną (tj. dla łącza w górę oraz w dół) komunikację z pozostałymi elementami urządzenia IoT, w którym będzie zainstalowany. Ze względu na sposób komunikacji innych elementów frameworka (w tym z częścią sprzętową z użyciem gniazd sieciowych) oraz budowy urządzeń, w których planowane jest instalowanie wąskopasmowego interfejsu radiowego NB-IoT zasadnym biznesowo oraz ekonomicznie jest stosowanie połączenia z frameworkiem także z użyciem komunikacji IP.

W procesie realizacji całego projektu, zgodność ze standardem i weryfikację poprawności działania interfejsu radiowego (zarówno w zakresie poprawności formowania/odbioru sygnałów w warstwie fizycznej, jak i na poziomie protokolarnym w warstwach wyższych) zrealizowano z użyciem testera radiowego symulującego rzeczywiste warunki operacyjne stacjonarnego urządzenia IoT. Biorąc pod uwagę charakterystykę kanału radiowego, założono iż weryfikowane będą profile środowiskowe EPA (Extended Pedestrian A) oraz ETU (Extended Typical Urban) dla odchyłki dopplerowskiej maks. 5 Hz i kanał typu AWGN (Additive White Gaussian Noise) ze zmiennymi wartościami mocy sygnału odbieranego i wynikową wartością stosunku sygnału do szumu SNR (Signal-to-Noise Ratio).

3. ARCHITEKTURA INTERFEJSU RADIOWEGO ORAZ JEGO CHARAKTERYSTYKA EKSPLOATACYJNA

Na rys. 1 przedstawiono schemat blokowy zrealizowanej w ramach projektu architektury części programowej interfejsu radiowego w formie wielowątkowego oprogramowania.



Rys.1. Schemat blokowy wielowątkowego oprogramowania architektury części programowej interfejsu radiowego.

Warto zaznaczyć, iż przedstawiona w niniejszym artykule końcowa architektura stanowi rozwinięcie względem prezentowanej w [1, 6]. Na schemacie oznaczono interfejsy komunikacyjne (IK), które należy rozumieć jako interfejsy wewnętrzne (IW) lub zewnętrzne (IZ) do komunikacji wewnątrz części programowej (IW) lub pomiędzy częścią programową a sprzętową (IZ) architektury interfejsu radiowego. IK umożliwiają przesłanie wektorów danych wejścia/wyjścia dla poszczególnych modułów czy sekcji (tj. zbiorów połączonych modułów) frameworka wraz z danymi sterującymi (np. wektorami parametrów konfiguracyjnych dla poprawnej pracy interfejsu radiowego zgodnie z wymaganiami standardu 3GPP).

3.1 Kontroler i obsługa front-endu

Wydzielony moduł kontrolera interfejsu radiowego i oprogramowania do akwizycji i analizy danych jest elementem nadrzędnym całego frameworka pełniąc funkcję kontrolno-sterującą, w tym, m.in. inicjalizację i zarządzanie środowiskiem wielowątkowym, identyfikację poprawności działania frameworka a w szczególności integralności i poprawności działania poszczególnych elementów architektury wraz z front-endem radiowym oraz umożliwia konfigurację i dekompozycję frameworka dla danego urządzenia IoT i wybranego front-endu radiowego.

Mając na względzie sposób realizacji frameworka i jego uniwersalność, oczywistym jest fakt, iż niezbędna jest obsługa front-endu radiowego w części programowej architektury interfejsu radiowego. Biorąc pod uwagę asynchroniczny charakter przetwarzania sygnałów i niezależność stosu NB-IoT w torze nadawczym oraz odbiorczym, zdecydowano się zrealizować dwie sekcje do obsługi front-endu radiowego, tj. sekcję dla łącza w dół do odbioru sygnałów radiowych pochodzących ze stacji bazowej oraz sekcję dla łącza w górę do emisji sygnałów radiowych z terminala użytkownika (urządzenia IoT). W torze odbiorczym zrealizowano ciągły odbiór sygnałów radiowych w wybranym kanale częstotliwościowym w paśmie nominalnym 200 kHz, który jest niezbędny do spełnienia wymagań normatywnych transmisji NB-IoT kat. NB1/NB2 typu półdupleks. Umożliwia to na zapewnienie poprawnego działania interfejsu we wszystkich stanach jego pracy, tj. zarówno podczas początkowej synchronizacji terminala i detekcji sygnałów stacji bazowych, w trakcie procedury przyłączania do stacji bazowej, wymiany danych użytecznych czy trybie uśpienia i utrzymywania synchronizacji.

Współbieżne przetwarzanie sygnałów w torze nadawczym oraz odbiorczym determinuje podział architektury na rozdzielne sekcje dla toru nadawczego oraz toru odbiorczego. Mając na względzie strukturę stosu NB-IoT zgodną z międzynarodowym standardem 3GPP oraz przedstawiony sposób realizacji frameworka, sekcje przetwarzania sygnałów w warstwie fizycznej, warstwie łącza danych oraz warstwie sieciowej stanowią fundamentalną część interfejsu radiowego i zawierają bloki funkcjonalne umożliwiające formowanie oraz interpretację sygnałów radiowych.

3.2 Tor nadawczy TX

W ogólności sekcja warstwy fizycznej TX umożliwia formowanie segmentów próbek sygnału radiowego, które są przekazywane do emisji w sekcji obsługi front-endu radiowego na podstawie otrzymywanych asynchronicznie pakietów z warstwy łącza danych. Formowane sygnały są w pełni zgodne ze standardem NB-IoT dla łącza w górę, co

potwierdzono licznymi pomiarami kanału NPUSCH (Narrowband Physical Uplink Shared Channel) z uwzględnieniem sygnałów pilotowych DMRS (Demodulation Reference Signal) i kanału NPRACH (Narrowband Physical Random Access Channel) z użyciem testera radiowego R&S CMW500, zarówno na poziomie parametrów bazowych, tj. przydziału odpowiedniej liczby podnośnych, rastra (3,75 kHz lub 15 kHz), nominalnej szerokości zajmowanego pasma (200 kHz, tj. 180 kHz bloku zasobów oraz dwóch 10 kHz przedziałów ochronnych), rodzaju modulacji fazy, emisji pozapasmowej zgodnie z metryką ACRL (Adjacent Channel Leakage Ratio), emisji wewnątrzpasmowej IE (Inband Emmissions), wyznaczenia metryki EVM (Error Vector Magnitude), jak i zgodności parametrów wytwarzanego sygnału binarnego a w następstwie także formowania wyjściowego sygnału OFDM (Orthogonal Frequency-Division Multiplexing), tj. formatu transmisji, kodowania kanałowego, scramblingu, dopasowania szybkości transmisji, odwzorowania elementów zasobów [2, 11].

Ze względu na sposób generowania pakietów w łączu w górę (emisja sygnałów o różnej liczbie podnośnych i różnym rastrze) rozdzielono logikę przetwarzania kanałów logicznych od warstwy fizycznej. W ramach sekcji łącza danych zaimplementowano funkcje przetwarzania kanałów logicznych i formowania kanałów transportowych rozumianych jako implementacja tzw. schedulera warstwy łącza danych. Umożliwia on połączenie ze stacją bazową i przesyłanie niezbędnych wiadomości wynikających z procedury dostępu do medium transmisyjnego oraz wymiany danych użytkowych. Przy czym, dane użytkowe rozumiane są nie tylko jako dane użytkownika z określonym priorytetem, ale także dane wymagane do realizacji dalszych procedur komunikacji ze stacją bazową, informacji sterujących wynikających z funkcjonowania urządzenia IoT jako elementu infrastruktury krytycznej, czy też danych diagnostycznych na potrzeby monitorowania stanu pracy sieci w danym miejscu instalacji urządzenia IoT.

W takim przypadku sekcja ta, oprócz odpowiedniego formowania pakietów oraz przekazywania do niższych warstw parametrów danej transmisji przydzielonych do danego pakietu, realizuje także funkcję fragmentacji i translacji danych pozyskiwanych z warstwy sieciowej w kanałach logicznych zgodnie z ustaloną (na podstawie wymienianych ze stacją bazową informacji zgodnych ze standardem [12]) wielkością bloku transportowego TBS (Transport Block Size). Warto też nadmienić, iż w tej sekcji realizowana jest także procedura pozyskiwania danych z priorytetyzacją ruchu zgodnie z wymaganiami standaryzacyjnymi 3GPP opisanymi w [13].

Szczegółowe informacje dotyczące sposobu programowej realizacji toru nadawczego wraz z uzyskanymi wynikami i analizą wydajności poszczególnych modułów przetwarzania danych przedstawiono w [2].

3.3 Tor odbiorczy RX

Sekcja odbiorcza warstwy fizycznej przetwarza otrzymywane synchronicznie (ze względu na synchroniczny charakter odbieranych pakietów próbek z sekcji obsługi frontendu radiowego) sygnały radiowe, które następnie przekazuje do wyższych warstw stosu NB-IoT w kanałach transportowych w sposób asynchroniczny (ze względu na możliwy różny czas przetwarzania sygnałów wewnątrz sekcji warstwy fizycznej). Należy podkreślić, iż ze względu na niezwykle wysoki stopień skomplikowania danych przetwarzanych w torze odbiorczym oraz zależności pomiędzy parametrami warstwy fizycznej i warstwy łącza danych w zrealizowanym projekcie zdecydowano się na wewnętrzną integrację części przetwarzania danych pomiędzy tymi warstwami (zgodnie z rys. 1). Takie podejście umożliwiło realizację części odbiorczej nie tylko w formie w pełni modularnej, ale także przetwarzanie równoległe danych w poszczególnych kanałach fizycznych i logicznych celem optymalizacji zajęcia dostępnych zasobów na różnych platformach uruchomieniowych.

Podobnie jak w przypadku sekcji warstwy fizycznej toru nadawczego, możliwy jest odbiór w wybranym kanale częstotliwościowym pasma UHF (Ultra High Frequency) przy założeniach, że:

- segmenty sygnału formowane są w technice OFDM w pojedynczym bloku zasobów LTE, tj. 180 kHz, gdzie odstęp między 12. podnośnymi wynosi 15 kHz;
- segmenty próbek sygnałów odpowiadają czasowi trwania pojedynczej ramki, na który składa się 20 szczelin czasowych o czasie trwania 0,5 ms zawierających w każdej szczelinie: 7 symboli OFDM wraz ze standardowym przedrostkiem cyklicznym stosowanym w NB-IoT;
- w ramach techniki OFDM zastosowano modulację kluczowania fazy QPSK dla poszczególnych podnośnych;
- dokonywana jest korekta częstotliwościowa sygnału wejściowego równa połowie odległości pomiędzy podnośnymi, tj. dla rastra 15 kHz korekta wynosi 7,5 kHz.

Szczegółowe informacje dotyczące sposobu programowej realizacji synchronizacji interfejsu radiowego, pozyskania informacji o numerze identyfikatora komórki, numerze ramki radiowej w formie modułów synchronizacji czasu oraz synchronizacji częstotliwości przedstawiono w [5, 7]. W tym miejscu warto zaznaczyć, iż moduły te testowano zarówno z użyciem sygnałów testowych stacji bazowej eNodeB (Evolved Node B) generowanych przez tester radiowy [8], ale też z rzeczywistej, komercyjnej sieci operatorskiej. Zweryfikowano także możliwość realizacji lokalizacji urządzenia IoT na podstawie analizy sygnałów synchronizacyjnych, co może stanowić integralne uzupełnienie całego frameworka [3].

Następnie, realizowana jest procedura estymacji i korekcji wpływu kanału radiowego przed dekodowaniem i przetwarzaniem kanałów roboczych i sterujących NB-IoT stosując interpolację liniową w dziedzinie czasu z uśrednianiem w dziedzinie częstotliwości współczynników kanału, która zgodnie z analizowanymi metodami estymacji parametrów kanału charakteryzującymi się niską złożonością obliczeniową, umożliwia jej zastosowanie w urządzeniach IoT. Szczegółowe informacje oraz uzyskane wyniki badań można znaleźć w [4, 9].

3.4 Dwustronna komunikacja ze stacją bazową eNodeB

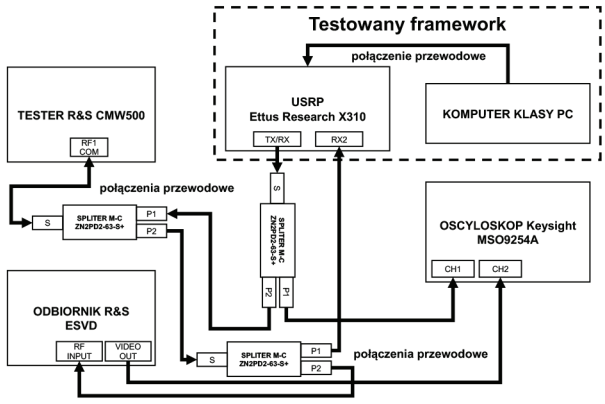
W pierwszej kolejności określono wynikową dokładność synchronizacji całego interfejsu radiowego, która jest jednym z parametrów jakościowych definiowanych przez standard 3GPP, a który potwierdza także poprawność realizacji toru nadawczego, odbiorczego i ich integracji. W celu określenia wartości dokładności synchronizacji frameworka przygotowano oprogramowanie, w którym

na podstawie przeprowadzonej korelacji pomiędzy sygnałem wzorcowym NPSS generowanym przez tester radiowy a sygnałem odebranym przez dodatkowy odbiornik wyznaczono moment rozpoczęcia ramki zgodnie ze sposobem czasowej organizacji transmisji w NB-IoT. Następnie określono różnicę pomiędzy momentem jej rozpoczęcia a momentem rozpoczęcia nadawania sygnałów w łączu w górę przez framework, który synchronizowany był także odebranymi sygnałami z testera i transmitował pakiety NPRACH tak by momenty emisji sygnałów pokrywały się z granicami szczelin i symboli w strukturze czasowej ramki interfejsu NB-IoT.

Wartość średnia różnicy czasu pomiędzy momentem rozpoczęcia nadawania a początkiem ramki wyniosła 1,35 μs, wartość minimalna wyniosła 0,85 μs, maksymalna 1,73 μs, a odchylenie standardowe 0,18 μs co oznacza spełnienie kompatybilności wyjściowego sygnału radiowego i jego zgodności ze standardem 3GPP w zakresie tolerancji parametrów czasowych sygnału radiowego nie gorszej niż 20 μs [11].

W celu weryfikacji poprawności działania interfejsu radiowego na poziomie warstwy fizycznej oraz warstwy protokolarnej, zgodnie ze standaryzacją 3GPP, zrealizowane zostały testy procedury losowego dostępu do medium transmisyjnego, będącego pierwszym etapem komunikacji urządzenia NB-IoT ze stacją bazową. Jako że do nawiązania połączenia z eNodeB niezbędna jest komunikacja dwukierunkowa, przeprowadzone badania wymagały nie tylko prawidłowego formowania sygnału w części nadawczej, prawidłowego odbioru i przetwarzania sygnałów w części odbiorczej, ale również pełnej integracji funkcjonalnej i synchronizacji toru nadawczego i odbiorczego, tak aby dane przesyłane przez stację bazową (w tym dane konfiguracyjne) były interpretowane i przekazywane do części nadawczej.

Na rys. 2 przedstawiono schemat blokowy stanowiska laboratoryjnego na potrzeby testów weryfikacji dwustronnej komunikacji z testerem radiowym.



Rys.2. Schemat blokowy stanowiska laboratoryjnego na potrzeby testów weryfikacji dwustronnej komunikacji z testerem radiowym.

Analizując zagadnienia zgodności protokolarnej oraz poprawności integracji torów TX RX, kluczowe są przede wszystkim trzy pierwsze wiadomości na etapie inicjowania komunikacji dwukierunkowej. Wiadomość inicjalizacyjna, tj. MSG1 wysyłana przez terminal, formowana jest w postaci preambuły NPRACH, podczas gdy wiadomość trzecia (MSG3) formowana jest w ramach kanału NPUSCH na podstawie uzyskanych danych konfiguracyjnych, ustalonych przez eNodeB i wysłanych do terminala w wiadomości drugiej (MSG2). Zestaw wymienionych wiadomości wraz z potwierdzeniem odbioru zasobów NPUSCH (MSG4) stanowi pełną analizę zgodności protokolarnej dla wszystkich bloków toru nadawczego, co udało się uzyskać i tym samym potwierdzić poprawność zrealizowanego interfejsu radiowego w części programowej działającego w czasie rzeczywistym i jego zgodność ze standardem 3GPP, co było jednym z głównych celów projektu.

4. PODSUMOWANIE

W ramach prac badawczo-rozwojowych, w których uczestniczyła Politechnika Gdańska w zakresie projektowania i programowej realizacji interfejsu radiowego wraz z konfigurowalną architekturą frameworka, zrealizowano prototyp demonstracyjny, spełniający założone funkcje interfejsu radiowego NB-IoT z konfigurowalną architekturą sprzętowo-programową, który został przetestowany w kilku środowiskach uruchomieniowych oraz pracuje w ramach trzech najniższych warstw modelu ISO/OSI z możliwością nawiązania z nim dwustronnej komunikacji z elementów zewnętrznych urządzenia NB-IoT. Framework odpowiada VIII poziomowi gotowości technologicznej PGT oraz na podstawie testów w licznych środowiskach uruchomieniowych spełnia wymagania i parametry techniczne, które umożliwiły prowadzenie dalszych prac przedwdrożeniowych przez firmę DGT Sp. z o.o.

Rezultatem prac jest uzyskanie poprawnej (co do zgodności ze standardem 3GPP) dwustronnej i działającej w czasie rzeczywistym komunikację z emulowaną przez tester radiowy R&S CMW500 stacją bazową eNodeB.

Uzyskane wyniki projektu stanowić mogą przyczynek do realizacji dalszych prac w zakresie realizacji komunikacji radiowej i zmniejszenia jej podatności na celowe zakłócenia [10], w szczególności infrastruktury krytycznej, co w dalszej perspektywie mogłoby oznaczać wypracowanie krajowej technologii realizacji chipsetu, a następnie samodzielnego modułu radiowego NB-IoT dla sieci 5G w Polsce.

5. FINANSOWANIE

Framework zrealizowano w ramach projektu w latach 2020 - 2023 pt. „Definiowany programowo, uniwersalny interfejs radiowy inteligentnych urządzeń Internetu Rzeczy", współfinansowanego ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego, umowa nr POIR.01.01.01-00-1025/19-00 z dnia 28.05.2020 r.

Ze względu na zakończenie projektu w 2023 roku, niniejszy artykuł został sfinansowany w ramach działalności statutowej Wydziału Elektroniki, Telekomunikacji i Informatyki Politechniki Gdańskiej.

LITERATURA

[1] Błaszkiewicz O., Magiera J., Olejniczak A., Rajchowski P., Sadowski J., Ambroziak S., Cwalina K. 2021. "Definiowany programowo uniwersalny interfejs radiowy inteligentnych urządzeń internetu rzeczy", Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne, DOI: 10.15199/59.2021.5.2

[2] Olejniczak A., Błaszkiewicz O., Cwalina K.K., Rajchowski P., Sadowski J. 2021. "Software-Defined NB-IoT Uplink Framework—The Design, Implementation and Use Cases", Sensors, wol. 21, wyd. 24, DOI: 10.3390/s21248234

[3] Rajchowski P., Cwalina K., Sadowski J. 2022. "Lokalizowanie terminala użytkownika z użyciem interfejsu NB-IoT", Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne, DOI: 10.15199/59.2022.4.76

[4] Magiera J., Rajchowski, P., Cwalina, K. 2022. "Estymacja parametrów kanału NB-IoT w łączu w dół", Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne, DOI: 10.15199/59.2022.4.74

[5] Błaszkiewicz O., Rajchowski P., Sadowski J., 2022. "Analiza metod synchronizacji w sieci NB-IoT", Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne, DOI: 10.15199/59.2022.4.75

[6] Błaszkiewicz O., Magiera J., Olejniczak A., Rajchowski P., Sadowski J., Stefański J., Cwalina K.K. 2022. "Realizacja definiowanego programowo interfejsu radiowego NB-IoT", Materiały konferencyjne XVIII Konferencji Automatyzacji i Eksploatacji Systemów Sterowania i Łączności ASMOR

[7] Rajchowski P., Błaszkiewicz O., Cwalina K.K., Magiera J., Olejniczak A., Sadowski J. 2022. "Programowa realizacja modułów synchronizacji czasu i częstotliości terminala NB-IoT", Materiały konferencyjne XVIII Konferencji Automatyzacji i Eksploatacji Systemów Sterowania i Łączności ASMOR

[8] Rajchowski P., Cwalina K., Sadowski J., 2021. "Signals of the NB-IoT network generated by using the radiocommunication tester R&S CMW500", dataset, Politechnika Gdańska, DOI: 10.34808/12me-xb40

[9] Magiera, J. 2023. "A Simplistic Downlink Channel Estimation Method for NB-IoT", Applied Sciences, wol. 13, wyd. 23, DOI: 10.3390/app132312615

[10] Magiera, J. 2023. "Selektywne zakłócanie transmisji OFDM w interfejsie radiowym NB-IoT", Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne, DOI: 10.15199/59.2023.4.78

[11] 3GPP, TS 36.101, Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) radio transmission and reception, wersja: 14.3.0

[12] 3GPP, TS 36.213, Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures, wersja: 14.6.0

[13] 3GPP, TS 23.203, Policy and charging control architecture, wersja 14.6.0

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

DOI: 10.15199/59.2024.4.64

ANALITYCZNY MODEL PROFILU GĘSTOŚCI ELEKTRONÓW JAKO WYNIK PRZETWARZANIA OBRAZOWEGO DANYCH POMIAROWYCH Z JONOGRAMÓW

ANALYTICAL MODEL OF THE ELECTRON DENSITY PROFILE AS A RESULT OF IMAGE PROCESSING MEASUREMENT DATA FROM IONOGRAMS

Tomasz Walczyna[1]; Jarosław Wojtuń[2]; Cezary Ziółkowski[3]; Bogdan Uljasz[4]
[1,2,3,4] Wojskowa Akademia Techniczna im. Jarosława Dąbrowskiego, Wydział Elektroniki,
Instytut Systemów Łączności, ul. Gen. Sylwestra Kaliskiego 2, 00-908 Warszawa,
{[1]tomasz.walczyna, [2]jaroslaw.wojtun, [3]cezary.ziolkowski, [4]bogdan.uljasz}@wat.edu.pl

Streszczenie: **Zmienność profilu gęstości ładunków ma istotne znaczenie dla modelowania i analizy możliwości transmisji fal radiowych w jonosferze. Aproksymacja profilu, której celem jest jego predykcja, stanowi temat tego opracowania. Podstawą prezentowanej metodyki są dane pomiarowe w postaci jonogramów, które można pozyskiwać ze stacji sondujących rozmieszczonych w różnych miejscach Ziemi. W artykule opisano podstawowe systemy stacji sondujących, które zapewniają dostęp do baz danych z wynikami testowania jonosfery. Opis procedury wyznaczania profilu gęstości elektronów na podstawie ekstrakcji danych z jonogramów jest głównym celem tego artykułu.**

Abstract: **The variability of the charge density profile is important for modeling and analyzing the transmission properties of radio waves in the ionosphere. Profile approximation, the purpose of which is its prediction, is the topic of this study. The basis of the presented analysis are measurement data in the form of ionograms, which can be obtained from sounding stations located in various places on Earth. The article describes the basic systems of sounding stations that provide access to databases with ionospheric testing results. The procedure for determining the electron density profile based on data extraction from ionograms is the main goal of the issues presented in this study.**

Słowa kluczowe: **propagacja fal radiowych w jonosferze, sondowanie jonosfery, jonogramy, gęstość elektronów, przetwarzanie obrazowe.**

Keywords: **radio waves propagation in ionosphere, ionosphere sounding, ionograms, electron density, image processing.**

1. WSTĘP

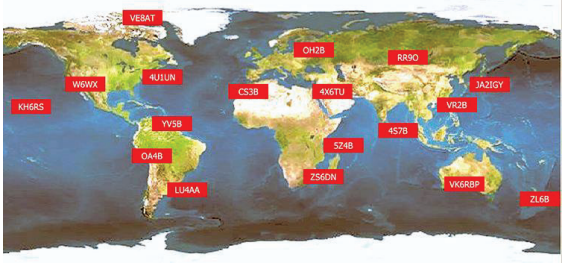
Właściwości transmisyjne ośrodka jakim jest jonosfera istotnie zależą od wielu parametrów transmitowanego sygnału oraz od terminu realizacji łącza radiowego. Zatem, zapewnienie poprawnej estymacji zarówno postaci odbieranego sygnału jak i jego poziomu wymaga zdefiniowania zbioru wartości danych wejściowych m.in. takich jak: częstotliwość fali nośnej, pasmo transmitowanego sygnału, terminu realizacji łącza, moc nadajnika, poziom sygnałów zakłócających, profil gęstości ładunków jonosfery oraz współrzędne położenia nadajnika i odbiornika.

Celem artykułu jest przedstawienie metody wyznaczania bieżącego profilu gęstości elektronów jonosfery na podstawie danych pomiarowych w postaci jonogramów.

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 i predykcji rzeczywistych warunków propagacji fal radiowych w jonosferze jest odwzorowanie statystycznych zmian parametrów opisujących profil gęstości ładunków. Zmiany parametrów, które definiują profil zarówno w dziedzinie wartości jak i czasu determinują jego właściwości statystyczne.

2. CHARAKTERYSTYKA BAZ DANYCH POMIAROWYCH

Bazy danych pomiarowych powstają w wyniku sondowania jonosfery zarówno w kierunku pionowym jak i poziomym względem powierzchni Ziemi. Przykład wykorzystania pomiarów poziomych stanowią stacje sondujące Becons, które są elementami największego międzynarodowego systemu sondującego o nazwie NCDXF (*Northern California DX Foundation*).

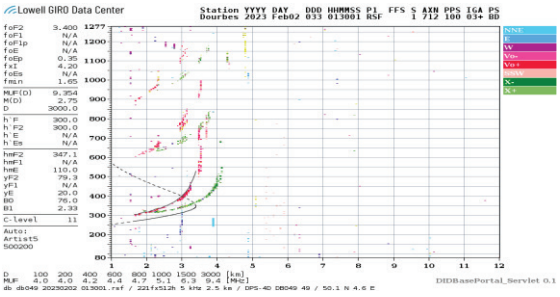


Rys. 1. Rozmieszczenie stacji Becon NCDXF [1]

Inny przykład zbioru wyników pomiarów można znaleźć na serwerze ITU. Dostępne są tutaj dwie wieloletnie bazy danych pomiarowych Databank D1 i D2 [2] z zarejestrowanym poziomem sygnału. W tym przypadku, do pomiarów parametrów opisujących właściwości transmisyjne wykorzystywane są stacje sondujące, które promieniują fale elektromagnetyczne w kierunku pionowym. Dużą liczbę danych pomiarowych dostarczają stacje sondujące stowarzyszone w grupie GIRO (ang. *Global Ionospheric Radio Observatory*). Na stronie GIRO można znaleźć ponad 17 milionów jonogramów wraz z opisem aktualnych i historycznych.

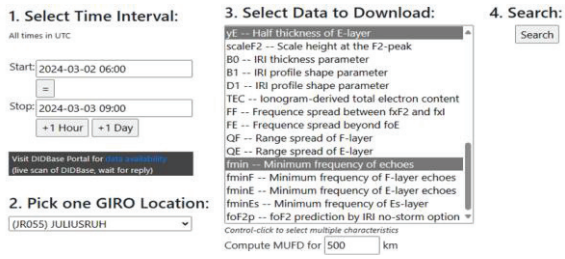
Rys. 2. Rozmieszczenie sond jonosferycznych zrzeszonych w GIRO [3]

GIRO we współpracy z Grupą Doradczą Sieci Jonosond URSI (INAG, ang. *Ionosonde Network Advisory Group*), LGDC (*ang. Lowell GIRO Data Center)* zawiera umowy o współpracy z obserwatoriami rozmieszczonych na całym świecie. Celem jest pozyskiwanie i przetwarzania danych w czasie rzeczywistym z monitorowania radiowego jonosfery w zakresie HF. Dzięki temu możliwe jest prowadzenie różnorodnych badań w zakresie właściwości transmisyjnych, które korzystają z szybkich, szczegółowych i dokładnych opisów zmienności jonosfery w skali globalnej. GIRO zapewnia jednopunktowy dostęp do globalnej sieci cyfrowych sond jonowych, z których wiele przesyła swoje dane w czasie rzeczywistym do baz danych GIRO.



Rys. 3. Jonogram poddany analizie – stacja DB049 – 2 lutego 2023 r. godz. 01:30 [3]

Na rysunku 3 przedstawiony jest przykładowy jonogram. Oprócz kolorowych wyników pomiarów jonosondą, znajduje się wyznacza numerycznie krzywa h_r [km] = f(f[MHz]), która reprezentuje zmiany wysokości pozornego promienia zwyczajnego odbicia w funkcji częstotliwości sygnału sondującego. Dodatkowo podstawowe parametry przedstawione są w postaci numerycznej [4] m. in. takie jak: częstotliwości krytyczne dla poszczególnych warstw (foF2, foF1, foE, foEs), maksymalna częstotliwość odbijana przez jonosferę (fxI), minimalne częstotliwości odbijane dla poszczególnych warstw (fmin, fminF, fminE, fminEs), maksymalna częstotliwość użytkowa dla łączności na odległość 3000 km (MUFD), rozpiętość częstotliwości pomiędzy fxF2 i fxI (FF), minimalne wirtualne wysokości dla poszczególnych warstw (hF2, hF, hE, hEs), maksymalne wysokości warstw (hmF2, hmF1, hmE), wartości połowy grubości warstw (yF2, yF1, yE) oraz rzeczywista wysokość przy połowie maksymalnej gęstości w warstwie F2 (zhafNm). Parametry te wyznaczane są przez oprogramowanie ARTIST-5. Dane numeryczne jak i jonogramy są dostarczane do bazy danych co 5 minut a na głównej stronie GIRO można obserwować zmieniający się jonogram co 30 sekund. Przykład strony GIRO do pobierania danych pomiarowych przedstawia rysunek 4



Rys. 4. Strona GIRO do pobierania danych numerycznych [3]

W Polsce znajdują się dwie stacje sondujące: CBK (Centrum Badań Kosmicznych), które nie jest członkiem GIRO i nie udostępnia swoich danych. Druga stacja stowarzyszona z GIRO od 2014 do 15.05.2016 roku znajduje się w Olsztynie (OL246). Obie stacje wykorzystują jonosondy typu VISRC2.

3. ANALIZA DANYCH POMIAROWYCH

Analiza danych pomiarowych z jonogramów wymaga szczegółowego podejścia do ekstrakcji zawartych w nich informacji, co można osiągnąć na dwa główne sposoby wykorzystując bądź bezpośredni dostęp do baz danych bądź też stosując przetwarzanie obrazowe w odniesieniu do graficznej reprezentacji danych pomiarowych.

Początkowo, analizy danych pomiarowych przeprowadzano wykorzystując jonogramy w formacie PNG. Ten sposób wymagał analizy pikseli na obrazie, detekcji kolorów i wzorców oraz dopasowania skali. Mimo że metoda ta była innowacyjna, to jednak jej niedokładność stanowiła znaczne ograniczenie. Problematyczna była precyzyjna ekstrakcja interesujących parametrów obrazu z pominięciem tych mniej istotnych często nachodzących na siebie.

Znaczącą zmianą było uzyskanie dostępu do wyników z bazy danych Lowell DIDBase. W tej bazie danych znajdują się dane numeryczne dotyczące wysokości, na której nastąpiło odbicie sygnału sondującego zarówno dla promienia zwyczajnego jak i nadzwyczajnego. Dzięki temu, do automatyzacji procesu ekstrakcji danych możliwe stało się bezpośrednie wykorzystanie oprogramowania takiego jak Microsoft Power Automate.

Mimo wprowadzenia licznych usprawnień w procesie analizy danych pomiarowych, napotkano na nowe wyzwanie związane z wielokrotnymi odbiciami sygnałów radiowych. W tym przypadku fala odbita od warstwy jonosfery powraca na Ziemię, odbija się od jej powierzchni i ponownie odbija się od jonosfery. Zjawisko to może powtarzać się wielokrotnie [5] [6].

Aby rozwiązać ten problem, zastosowano algorytm klasteryzacji danych pomiarowych DBSCAN (ang. *Density-Based Spatial Clustering of Applications with Noise*). Algorytm ten wykorzystuje metodykę bazującą na gęstości występowania danych pomiarowych na jonogramie. Algorytm DBSCAN wykorzystuje gęstość danych pomiarowych, aby identyfikować obszary o dużej

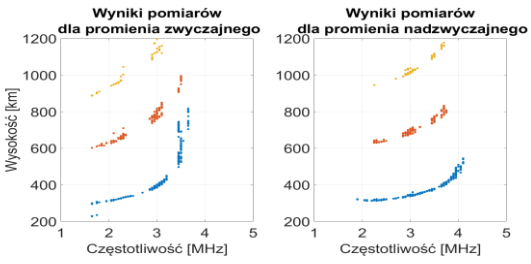
koncentracji punktów pomiarowych, co odpowiada rzeczywistym wartościom. Dzięki temu możliwe jest efektywne oddzielenie od siebie występujących odbić, eliminując jednocześnie szumy i mniej istotne dane. Ta procedura umożliwia skuteczne grupowanie danych pomiarowych, które są wynikiem odbić propagującej w jonosferze fali elektromagnetycznej.

Procedura przetwarzania i analizy danych pomiarowych składa się z następujących etapów:

1. Grupowanie: Zastosowanie algorytmu DBSCAN w celu identyfikacji i segregacji danych pomiarowych na podstawie obserwowanych odbić dla promienia zwyczajnego i nadzwyczajnego.
2. Ekstrakcja: Usunięcie mniej istotnych danych i skupienie się na wyodrębnionych klastrach, które są wynikiem odbić fali radiowej w rzeczywistych, permanentnie zmieniających się warunkach propagacyjnych.
3. Analiza: Przeprowadzenie dokładnej analizy wyodrębnionych danych, normalizacja wartości wysokości oraz dalsza interpretacja wyników.

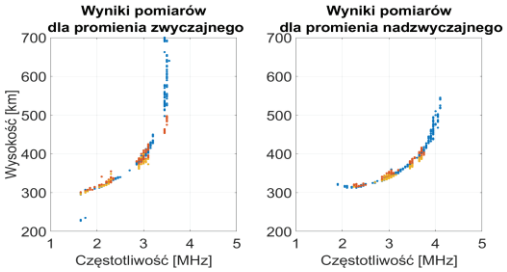
Procedura przetwarzania i analizy danych pomiarowych została zilustrowana na przykładzie jonogramu prezentowanego na rysunku 3.

W procesie grupowania utworzono klastry na podstawie dostępnych danych pomiarowych, co umożliwiło wskazanie grup odpowiadających kolejnym odbiciom, a także wyodrębnienie pojedynczych punktów (anomalii). W procesie ekstrakcji danych z jonogramów usunięto dane istotnie różniące się położeniem na płaszczyźnie jonogramu. Dzięki temu pozostawiono dane, które w istotny sposób decydują o poziomie odbieranego sygnału. . Efektywność tego podejścia została zilustrowana na rysunku poniżej, gdzie prezentowane są wyniki grupowania, które umożliwiają dalszą analizę charakterystycznych zjawisk występujących w jonosferze.



Rys. 5. Odbicia promienia zwyczajnego i nadzwyczajnego

W kolejnym etapie przetwarzania danych, obserwowane odbicia były poddane procesowi normalizacji. Operacja ta polegała na podzieleniu zarejestrowanych wartości wysokości przez odpowiednią liczbę całkowitą, zależną od typu zidentyfikowanego odbicia (pierwsze, drugie lub trzecie). Ostateczny wynik przetwarzania przedstawiony jest na rys. 7.



Rys. 6. Uśrednione wyniki pomiarów promienia zwyczajnego i nadzwyczajnego

4. PROFIL GĘSTOŚCI ELEKTRONÓW W JONOSFERZE

Parametry transmisyjne ośrodka propagacji jakim jest jonosfera cechuje duża zmienność, która występuje zarówno w krótkich jak i długich przedziałach czasu. Jest to wynikiem zmian gęstości ładunku, zależnego głównie od zmian gęstości atmosfery i promieniowania słonecznego. Estymacja rzeczywistych warunków propagacji występujących w jonosferze jest zatem ściśle związana z odwzorowaniem statystycznych właściwości parametrów opisujących profil gęstości ładunku $N(z)$. Charakterystyka ta stanowi podstawę do definiowania charakterystycznej dla jonosfery częstotliwości $f_N(z)$ zwanej częstotliwością plazmową [11]

$$f_N^2(z) = \frac{e^2}{4\pi^2 \varepsilon_0 m} N(z) \qquad (1)$$

gdzie: e oraz m to odpowiednio ładunek elektronu i jego masa, ε_0 oznacza przenikalność elektryczną próżni.

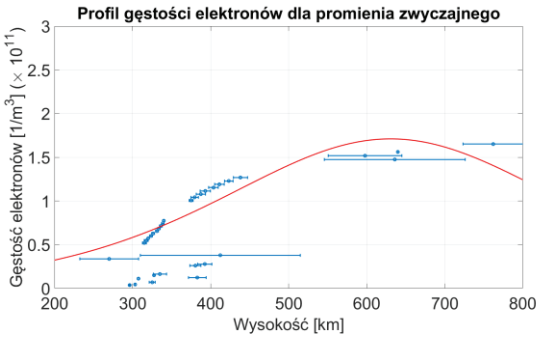
Z zależności (1) wynika, że na podstawie charakterystyki $f_N(z)$ można wyznaczyć profil gęstości ładunku jonosfery. Oddziaływanie pola magnetycznego Ziemi na propagującą w jonosferze falę powoduje powstanie tzw. fali zwyczajnej i nadzwyczajnej. W interpretacji optyki geometrycznej mówimy o promieniu zwyczajnym i nadzwyczajnym. W [11] pokazano, że dla promienia zwyczajnego współczynnik refrakcji przyjmuje wartość równą zero dla częstotliwości propagującej fali odpowiadającej częstotliwości plazmowej. To oznacza, że dla tej częstotliwości mamy do czynienia ze zjawiskiem odbicia. Zatem, na podstawie danych z jonogramów dotyczących promienia zwyczajnego możemy wyznaczyć $f_N(z)$ a w konsekwencji zależności (1) $N(z)$.

Wśród znanych modeli opisujących profil gęstości ładunku $N(z)$ [7] [8] możemy wyróżnić modele: liniowy, wykładniczy, paraboliczny, Chapmana oraz sech^2. W [9] [10] przedstawiono analizę porównawczą wskazującą, że najmniejszy błąd aproksymacji danych pomiarowych uzyskuje się z wykorzystaniem modelu $sech^2$.

$$N(z) = N_m sech^2\left(\frac{z - z_m}{a}\right) \qquad (2)$$

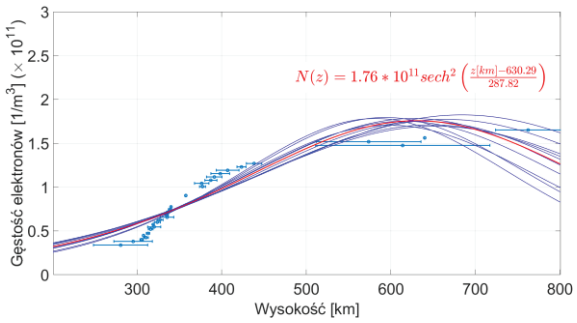
gdzie: N_m maksymalna gęstość elektronów, z_m wysokość, na której występuje maksimum gęstości elektronów, a to rozproszenie jonosfery w kierunku prostopadłym do powierzchni Ziemi.

Wartości średnie danych pomiarowe są wykorzystywane jako kluczowe punkty kalibracyjne dla funkcji aproksymującej. Rozrzut wartości pomiarowych wokół wartości średniej jest analizowany w celu oceny wiarygodności próbki, co ma zasadnicze znaczenie dla stosowanej metody najmniejszych kwadratów, służącej do iteracyjnego dopasowania modelu. Na rys. 7 przedstawiono przykładowy model profilu gęstości elektronów wyznaczony na podstawie danych z jonogramów.



Rys. 7. Profile gęstości elektronów dla promienia zwyczajnego

Dla określonej wartości częstotliwości otrzymujemy szereg punktów pomiarowych, które odpowiadają różnym wysokościom występowania odbić. To zróżnicowanie wykorzystywane jest w procesie tworzenia funkcji aproksymującej. Na rys. 8 przedstawiono 10 przykładowych wyników aproksymacji na bazie modelu sech^2. Krzywa w kolorze czerwonym wyznaczana jest jako wartość średnia uzyskana na podstawie poszczególnych funkcji aproksymujących $N(z)$. To podejście umożliwia bardziej precyzyjne modelowanie i analizę profilu gęstości ładunku w jonosferze, gdyż realizowane jest na podstawie bieżących danych, które odwzorowują rzeczywisty stan jonosfery.



Rys. 8. Dopasowanie modelu sech^2 do danych pomiarowych dla promienia zwyczajnego

5. PODSUMOWANIE

Prezentowana w artykule metodyka aproksymacji profilu gęstości ładunku w jonosferze daje możliwość oceny bieżących właściwości transmisyjnych jonosfery. Wyznaczanie parametrów modelu $sech^2$ jest podstawą dla oceny współczynnika pozornego odbicia propagującej fali elektromagnetycznej od warstw jonosfery. Daje to możliwość wyznaczania struktur czasowo-częstotliwościowych odbieranych sygnałów co warunkuje ocenę jakości transmisji informacji w łączach KF.

LITERATURA

[1] Northern California DX Foundation, „International Beacon Project," [Online]. Available: https://www.ncdxf.org/pages/beacons.html. [Data uzyskania dostępu: 09 Maj 2024].

[2] International Telecommunication Union, „Software, Data and Validation examples for ionospheric and tropospheric radio wave propagation and radio noise," [Online]. Available: https://www.itu.int/en/ITU-R/study-groups/rsg3/Pages/iono-tropo-spheric.aspx. [Data uzyskania dostępu: 09 Maj 2024].

[3] GLOBAL IONOSPHERE RADIO OBSERVATORY, „GLOBAL IONOSPHERE RADIO OBSERVATORY," [Online]. Available: https://giro.uml.edu/. [Data uzyskania dostępu: 09 Maj 2024].

[4] GIRO, „FastChar - Digital Ionogram Data Base (DIDBase)," [Online]. Available: https://giro.uml.edu/didbase/scaled.php. [Data uzyskania dostępu: 27 05 2024].

[5] British Antarctic Survey, „Interpretation of Ionograms, a real example," 09 Maj 2024. [Online]. Available: http://www.nerc-bas.ac.uk/public/uasd/instrums/dyna/interp2.html.

[6] A. HIYADUTUJE, „THE EFFECTS OF TRAVELING IONOSPHERIC DISTURBANCES ON SUPERDARN NEAR RANGE ECHOES," University of KwaZulu Natal, Westville, 2022.

[7] S. M. Stankov, N. Jakowski, S. Heise, P. Muhtarov, I. Kutiev i R. Warnant, „A new method for reconstruction of the vertical electron density distribution in the upper ionosphere and plasmasphere," *Journal of Geophysical Research,* tom 108, nr A5, 2003.

[8] Università degli Studi di Firenze., „The Ionospheric Propagation Channel," [Online]. Available: https://flore.unifi.it/retrieve/e398c378-eae6-179a-e053-3705fe0a4cff/SermiPhDthesisLowRes06.pdf.pdf. [Data uzyskania dostępu: 10 Maj 2024].

[9] C. Ziółkowski, J. M. Kelner i B. Uljasz, „Modele profili gęstości elektronów w analizie łączy jonosferycznych," w *XVIII Konferencja Automatyzacji i Eksploatacji Systemów Sterowania i Łączności (ASMOR)*, Władysławowo, 2022.

[10] B. Uljasz i C. Ziółkowski, „Analiza porównawcza modeli profilu gęstosci ładunków jonosfery wykorzystywanych do predykcji sygnałów w kanałach KF," *Przegląd Telekomunikacyjny - Wiadomosci Telekomunikacyjne,* tom Zeszyt 4, 2023.

MAPY TŁUMIENIA PROPAGACYJNEGO Z UWZGLĘDNIENIEM WARSTWY BUDYNKÓW

PROPAGATION ATTENUATION MAPS INCLUDING BUILDING LAYER

Marcin Moskal[1]; Michał Kryk[2]; Krzysztof Malon[3]; Jan M. Kelner[4]

[1] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, marcin.moskal@student.wat.edu.pl
[2] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, michal.kryk@wat.edu.pl
[3] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, krzysztof.malon@wat.edu.pl
[4] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, jan.kelner@wat.edu.pl

DOI: 10.15199/59.2024.4.65

Streszczenie: **Mapy środowiska radiowego stanowią pomocne narzędzie wykorzystywane do planowania sieci radiowych i zarządzania nimi. Znajdują one także zastosowanie np. w sieciach radia kognitywnego lub mobilnych sieciach doraźnych. Istotnym aspektem wyznaczania takich map jest dokładność odwzorowania rzeczywistych warunków propagacyjnych. W artykule przedstawiono rozszerzenie algorytmu generowania map tłumienia o uwzględnienie warstwy budynków wraz z uzyskanymi rezultatami badań.**

Abstract: **Radio environment maps are helpful tools for radio network planning and management. They are also used, for example, in cognitive radio networks or mobile ad hoc networks. An essential aspect of determining such maps is the accuracy of the representation of the actual propagation conditions. The article presents an extension of the algorithm for generating attenuation maps, including a building layer and test results.**

Słowa kluczowe: **mapa środowiska radiowego, mapa tłumienia propagacyjnego, propagacja fal radiowych, warstwa budynków.**

Keywords: **radio environment map, propagation attenuation map, radio wave propagation, building layer.**

1. WSTĘP

Mapa środowiska radiowego REM (ang. *Radio Environment Map*) to sposób reprezentowania złożonych i wielodziedzinowych informacji [1][7], które są wykorzystywane do zarządzania sieciami radiowymi, w tym także mobilnymi sieciami ad-hoc MANET (ang. *Mobile Ad-Hoc Network*) [10]. Obecnie można również dostrzec wiele nowoczesnych technologii, które mogą współpracować lub być oparte na REM. Wśród nich wyróżnić można dynamiczne zarządzanie widmem czy sieci radia kognitywnego [9] związane z algorytmami sztucznej inteligencji stosowanymi w 5G [3]. Analiza środowiska propagacji jest jedną z domen informacyjnych REM. Informacje na temat lokalizacji węzłów sieci MANET, stopnia zurbanizowania terenu oraz innych czynników charakteryzujących środowisko propagacji pozwalają na określenie tzw. map tłumienia propagacji PAM (ang. *Propagation Attenuation Map*). PAM zawiera wartości tłumienia propagacyjnego sygnału w pobliżu nadajnika co pozwala oszacować jego zasięg na danym obszarze oraz może wpływać na topologię i routing w sieci MANET. W przypadku generacji PAM na zróżnicowanym obszarze wykorzystanie statystycznych modeli propagacyjnych jest często niewystarczające [4]. W związku z tym, aby uwzględnić topografię terenu, roślinność, budynki i parametry gruntu, stosowane są deterministyczne modele propagacji. Uzupełnienie metod obliczeniowych stanowią również kampanie pomiarowe wykonywane w rzeczywistym środowisku [2] [8].

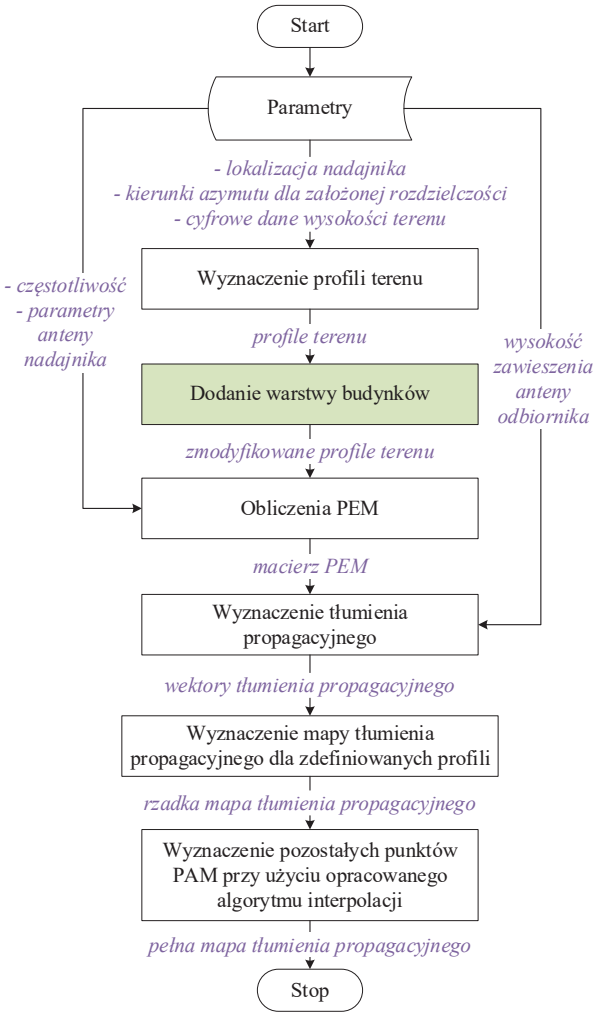
W celu dokładniejszego odwzorowania rzeczywistych warunków propagacyjnych, autorzy zaproponowali uwzględnienie, w algorytmie wyznaczania map tłumienia, warstwy budynków.

2. ALGORYTM GENEROWANIA MAP TŁUMIENIA PROPAGACYJNEGO

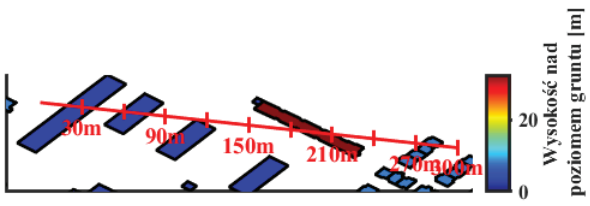
Niniejszy artykuł stanowi rozszerzenie wcześniejszych prac dotyczących map tłumienia propagacyjnego [6]. W opracowanym algorytmie, w celu dokładniejszego odwzorowania rzeczywistych warunków propagacyjnych, uwzględniono dodatkową warstwę zawierającą informacje o wysokościach budynków znajdujących się na określonym obszarze. Na Rys. 1 przedstawiono algorytm generowania PAM z oznaczeniem wprowadzonej modyfikacji. W pierwszym kroku, bazując na lokalizacji nadajnika, cyfrowych danych wysokości terenu oraz założonej rozdzielczości kątowej wyznaczane są wysokościowe profile terenu. Następnie po uwzględnieniu warstwy budynków, algorytm modyfikuje wysokości w poszczególnych profilach. Na podstawie zmodyfikowanych profili, dodatkowych parametrów anteny nadawczej oraz częstotliwości, przeprowadzane są obliczenia PEM. Proponowane rozwiązanie opiera się na numerycznym rozwiązaniu równania fali parabolicznej dla natężenia pola elektrycznego wzdłuż założonego promienia (profilu terenu) i dla dowolnej wysokości. Zastosowany model propagacyjny uwzględnia następujące zjawiska fizyczne: refrakcję powietrza, dyfrakcję terenu i przewodnictwo gruntu. Bazując na macierzy PEM i wysokości zawieszenia anteny odbiornika możliwe jest utworzenie tzw. rzadkiej macierzy z wartościami tłumienia drogi propagacyjnej na zdefiniowanych azymutach. Aby uzyskać pełną macierz tłumień wokół analizowanej pozycji nadajnika, należy wyznaczyć tłumienie również w obszarach poza punktami odpowiadającymi zdefiniowanym profilom terenu.

W tym celu zaproponowano dedykowany algorytm interpolacji opisany w [5].

Do wprowadzenia modyfikacji w algorytmie generowania PAM wykorzystano pliki SHP (Shapefile) zawierające m.in. informacje o położeniu oraz liczbie kondygnacji budynków. Na Rys. 2 zaprezentowano przykładowy fragment mapy z budynkami z uwzględnieniem wybranego profilu oznaczonego linią czerwoną. Profile wysokościowe z uwzględnieniem warstwy budynków oraz bez tych danych przedstawiono na Rys. 3. Zgodnie z zaprezentowanym algorytmem w kolejnych krokach wykonywane są obliczenia PEM oraz wyznaczenie wektorów tłumienia propagacyjnego dla poszczególnych profili.

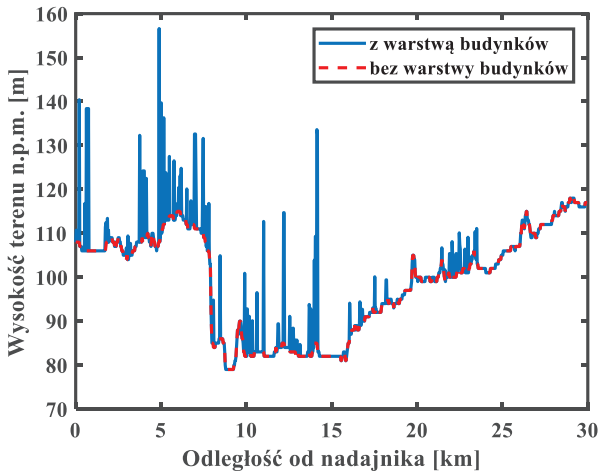


Rys. 1. Algorytm generowania PAM z uwzględnieniem warstwy budynków

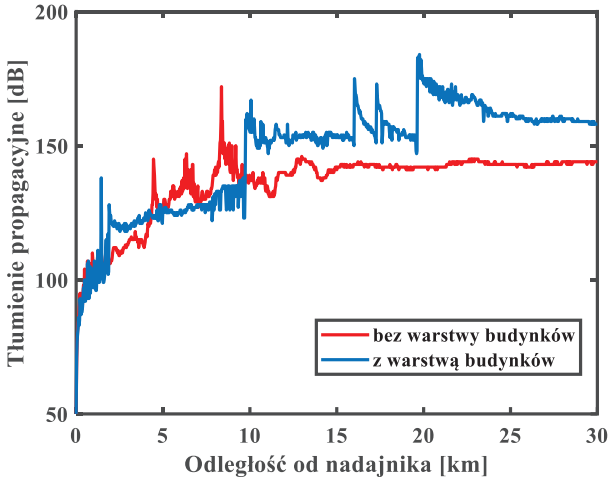


Rys. 2. Mapa z wysokościami budynków dla wybranego profilu

Porównanie wartości tłumień propagacyjnych dla dwóch profili (bazowego – bez warstwy budynków i zmodyfikowanego – z warstwą budynków), przedstawiono na Rys. 4. Wykres prezentuje zależność tłumień propagacyjnych sygnału od odległości pomiędzy nadajnikiem a odbiornikiem. Przebiegi wartości tłumień propagacyjnych wykazują różny charakter w zależności od analizowanego profilu: bazowego lub zmodyfikowanego. W obu analizowanych przypadkach zastosowano takie same wysokości zwieszenia anten nadajnika i odbiornika wynoszące 25 m. Dla profilu bez warstwy budynków można zaobserwować znacznie mniej skoków tłumienia w porównaniu do profilu z budynkami. W obu analizowanych przypadkach w miejscach, gdzie występują budynki lub przeszkody ukształtowania terenu, takie jak wzgórza i pagórki, obserwuje się spadek tłumienia propagacyjnego. Wynika to z faktu iż odczyt wartości tłumienia realizowany jest w punkcie zawieszenia anteny odbiornika, która w analizowanym algorytmie umieszczana jest na stałej, względnej wysokości nad terenem i budynkami. W związku z tym w przypadku wystąpienia przeszkody, bezwzględna wysokość anteny odbiorczej jest również większa powodując zmniejszenie tłumienia na drodze pomiędzy nadajnika a odbiornikiem. Tłumienie propagacyjne wzrasta natomiast, gdy pomiar tłumienia odbywa się po przejściu za badane przeszkody (terenowe lub budynki).



Rys. 3. Wysokość punktów na wybranym profilu z oraz bez uwzględnienia warstwy budynków



Rys. 4. Wykres wartości tłumienia propagacyjnego dla profilu z oraz bez uwzględnienia warstwy budynków

3. BADANIA I ANALIZA WYNIKÓW

Przeprowadzone badania miały na celu określenie wpływu warstwy budynków oraz zmiennej rozdzielczości kątowej Δα na dokładność wyznaczania map tłumienia propagacyjnego. Rozdzielczość kątowa definiuje liczbę profili, dla których wykonywane są obliczenia PEM. Wraz ze zwiększaniem Δα zmniejsza się liczba punktów obliczanych według PEM natomiast zwiększa się udział elementów pełnej macierzy PAM podlegających interpolacji.

3.1. Miary

Pierwszą miarą wykorzystywaną podczas badań jest mapa błędów dla zadanej rozdzielczości kątowej obliczana względem mapy referencyjnej o rozdzielczości kątowej 1˚. Dla każdego punktu na mapie **PAM**(j, k) błąd wyznaczenia tłumienia obliczany jest według następującego wzoru:

$$\Delta L(j, k) = \mathbf{PAM}_{\Delta\alpha}(j, k) - \mathbf{PAM}_{\Delta\alpha=1^\circ}(j, k), \qquad (1)$$

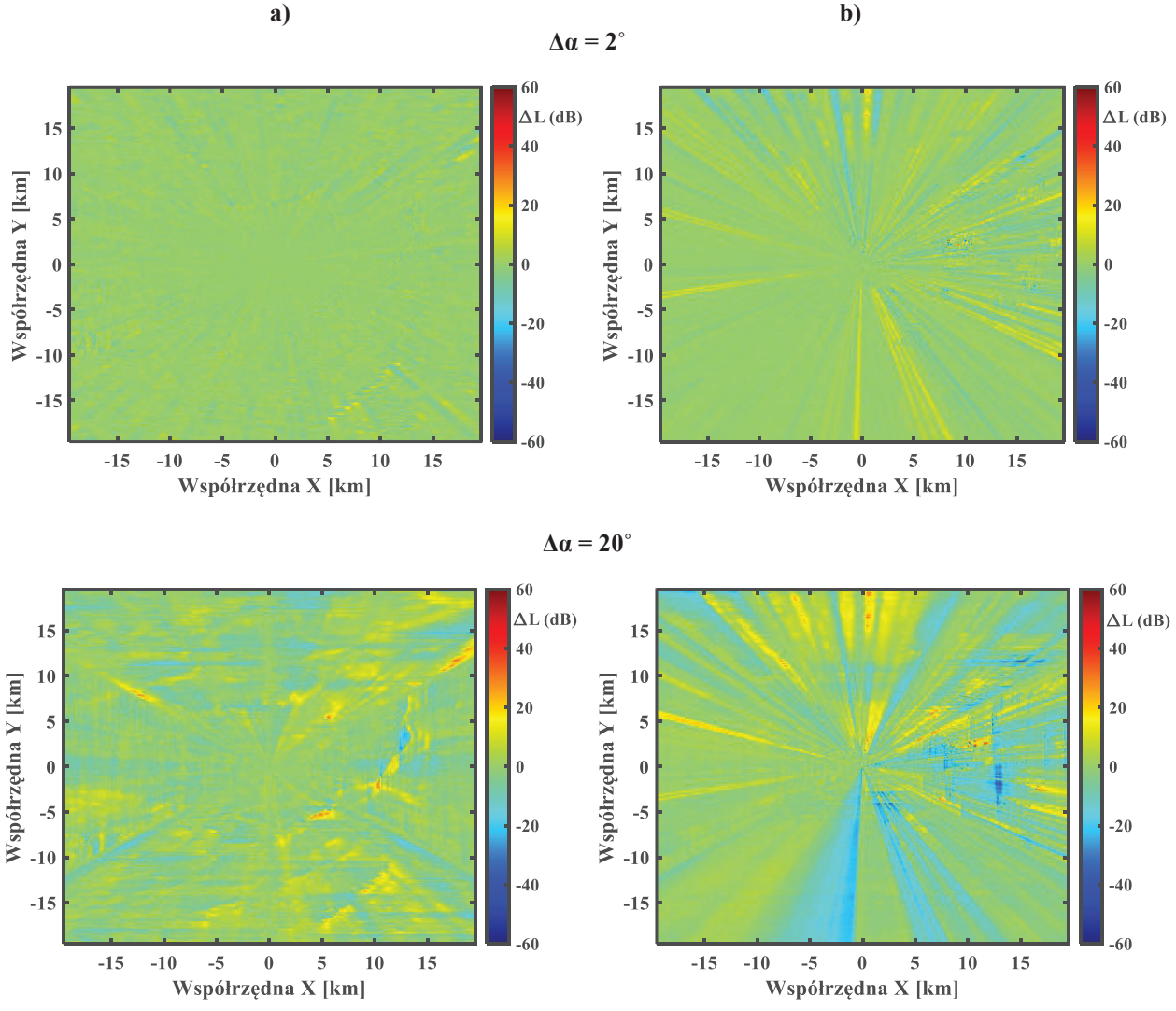
gdzie $\mathbf{PAM}_{\Delta\alpha}(j, k)$ oraz $\mathbf{PAM}_{\Delta\alpha=1^\circ}(j, k)$ są wartościami tłumienia uzyskanymi dla analizowanych rozdzielczości Δα i referencyjnej rozdzielczości Δα = 1˚. Dzięki zobrazowaniu map błędów (Rys. 5) można dostrzec obszary, w których występują największe błędy oszacowania oraz określić ich rodzaj: niedoszacowanie lub przeszacowanie. Na podstawie mapy referencyjnej oraz mapy wygenerowanej dla określonej rozdzielczości kątowej zaproponowano również miarę RMSE (ang. *Root Mean Square Error*). Jest to powszechnie stosowana miara błędu prognozowania, wyznaczana jako pierwiastek średniej kwadratowej różnic między wartościami przewidywanymi, a rzeczywistymi:

$$\mathrm{RMSE}(\Delta\alpha) = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(\mathbf{PAM}_{\Delta\alpha}(j, k) - \mathbf{PAM}_{\Delta\alpha=1^\circ}(j, k))^2} \qquad (2)$$

gdzie n oznacza całkowitą liczbę elementów macierzy.

3.2. Wyniki badań

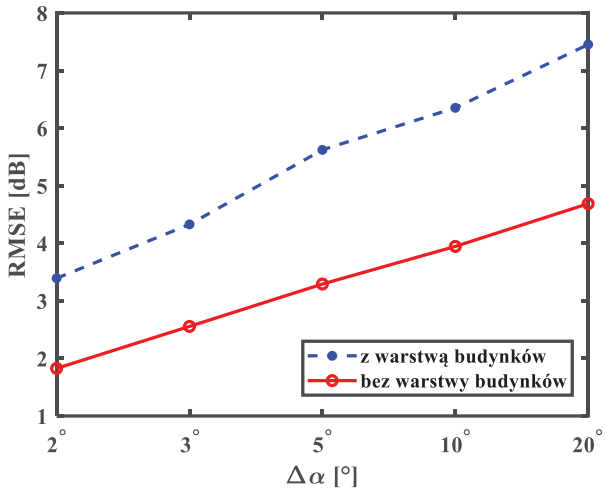
Na Rys. 5 zaprezentowano przykładowe mapy błędów dla dwóch rozdzielczości kątowych profili terenu. Wartości zgodne z wartościami referencyjnych PAM, wynoszą ΔL (dB) = 0 (kolor zielony).



Rys. 5. Mapy błędów dla rozdzielczości kątowej profili terenu równej 2° oraz 20° dla (a) obszaru nieuwzględniającego warstwy budynków oraz (b) obszaru uwzględniającego warstwę budynków

Analiza map błędów prezentowanych na Rys. 5 (b) ukazuje znaczące różnice w stosunku do PAM referencyjnych dla obu analizowanych rozdzielczości kątowych profili terenu. W przypadku obszarów bez uwzględnienia warstwy budynków (Rys. 5 (a)), znaczące różnice stają się widoczne dopiero przy rozdzielczości $\Delta\alpha = 20^{\circ}$. Obserwowane błędy wynikają ze zwiększonej liczby interpolowanych komórek w macierzy (wyznaczenie pełnej macierzy na podstawie macierzy rzadkiej w algorytmie na Rys. 1). Analiza wygenerowanych map błędów pozwala wywnioskować, że uwzględnienie warstwy budynków powoduje zwiększenie zróżnicowania wysokościowego terenu co w rezultacie prowadzi do znaczących błędów w przypadku zwiększania wartości $\Delta\alpha$.

Analiza wartości RMSE, ilustrowanej na Rys. 6, dla wybranych rozdzielczości kątowych wskazuje, że błąd średniokwadratowy dla mapy uwzględniającej wysokości budynków jest znacząco większy w porównaniu z mapą, która nie bierze pod uwagę tego parametru, niezależnie od analizowanej rozdzielczości kątowej. Wyniki interpolacji dla obszarów z budynkami wykazują niższą dokładność w odniesieniu do danych referencyjnych. Różnice te są wynikiem znaczących wahań wysokościowych, które nie występują w terenie niezabudowanym. W obszarach bez budynków, interpolacja danych jest bardziej jednorodna, co skutkuje lepszą korelacją z wartościami referencyjnymi. Niemniej jednak, w obszarach uwzględniających wysokości budynków, wyraźna zmienność wysokościowa wprowadza znaczące wyzwanie dla metody interpolacji liniowej, prowadząc do zwiększenia błędu średniokwadratowego.



Rys. 6. RMSE dla wybranych rozdzielczości kątowych profili terenu dla PAM z oraz bez uwzględnienia warstwy budynków

4. PODSUMOWANIE

W artykule przedstawiono rozszerzenie algorytmu wyznaczania map tłumienia propagacyjnego, polegające na uwzględnieniu dodatkowej warstwy danych. Zaproponowane rozwiązanie ma na celu zwiększenie dokładności oszacowania tłumienia propagacyjnego wokół nadajnika znajdującego się w określonej lokalizacji. W ramach pracy przeprowadzono analizę błędów wyznaczania map tłumienia dla zmiennej rozdzielczości kątowej profili, co w efekcie prowadzi do różnej liczby punktów docelowej macierzy tłumienia obliczanych według PEM oraz interpolowanych. Wraz ze zwiększaniem wartości $\Delta\alpha$ następuje ograniczenie udziału obliczeń PEM skutkujące również skróceniem czasu wyznaczania PAM. W takim przypadku natomiast zwiększa się wykorzystanie interpolacji co powoduje pojawianie się większych błędów określania PAM. Ponadto z przeprowadzonej analizy wynika iż mapy tłumienia wygenerowane z uwzględnieniem warstwy budynków w porównaniu do map uzyskanych dla takiej samej rozdzielczości kątowej bez warstwy budynków charakteryzuje się większymi błędami. Jest to spowodowane większym zróżnicowaniem wysokościowym terenu. Podsumowując, należy podkreślić że dodatkowa warstwa budynków przyczynia się do dokładniejszego odwzorowania rzeczywistych warunków propagacyjnych lecz w algorytmie wyznaczania PAM należy ograniczyć zwiększanie wartości rozdzielczości kątowej obliczeń.

LITERATURA

[1] Dagres I. et al. 2011. "Radio environmental maps: Information models and reference model". ICT-248351 FARAMIR, Document no. D4.1.

[2] Kaniewski P., J. Romanik, E. Golan, K. Zubel. 2021. „Spectrum awareness for cognitive radios supported by radio environment maps: Zonal approach". *Applied Sciences*, 11 (7).

[3] Kliks A., L. Goratti, T. Chen. 2016. "REM: Revisiting a cognitive tool for virtualized 5G networks". *23rd International Conference on Telecommunications (ICT)*, 1–5.

[4] Kryk M. 2021. „Creating radio environment maps (REMs) based on various path loss models". *37th International Business Information Management Conference (IBIMA)*, 9254–9264.

[5] Kryk M., K. Malon, J.M. Kelner. 2022. "Propagation attenuation maps based on parabolic equation method". *Sensors*, 22 (11) : 4063.

[6] Kryk, M., K. Malon, J.M. Kelner. 2021. „Algorithm of Creating Propagation Attenuation Maps Based on Parabolic Equation Method". *Proceedings of the 2021 Signal Processing Symposium (SPSympo)*, 38–142.

[7] Pesko M., T. Javornik, A. Košir, M. Štular, M. Mohorčič. 2014. „Radio environment maps: The survey of construction methods". *KSII Transactions on Internet and Information System*, 8 (11) : 3789–3809.

[8] Suchański M., P. Kaniewski, J. Romanik, E. Golan, K. Zubel. 2020. „Radio environment maps for military cognitive networks: Density of small-scale sensor network vs. map quality". *EURASIP Journal on Wireless Communications and Networking*, 189.

[9] Yilmaz H. B., T. Tugcu, F. Alagöz, S. Bayhan. 2013. „Radio environment map as enabler for practical cognitive radio networks". *IEEE Communications Magazine*, 51 (12) : 162–169.

[10] Younis O., L. Kant, K. Chang, K. Young, C. Graff. 2009. „Cognitive MANET design for mission-critical networks". *IEEE Communications Magazine*, 47 (10) : 64–71.

MODELOWANIE I POMIARY TŁUMIENIA SYGNAŁU RADIOWEGO WEWNĄTRZ SALI WYKŁADOWEJ NA CZĘSTOTLIWOŚCI 20 GHZ

MODELING AND MEASUREMENTS OF RADIO SIGNAL ATTENUATION INSIDE A LECTURE HALL AT 20 GHZ FREQUENCY

DOI: 10.15199/59.2024.4.66

Jakub Kostuj, Miłosz Maćkowski-Nowak, Kamil Dalil, Dawid Karnafel, Krzysztof Cichoń

[1] Politechnika Poznańska, Poznań, {jakub.kostuj, milosz.mackowski-nowak, kamil.dalil, dawid.karnafel}@student.put.poznan.pl, krzysztof.cichon@put.poznan.pl

Streszczenie: **Celem artykułu jest analiza tłumienia sygnału radiowego w sali wykładowej dla częstotliwości 20 GHz. Przeprowadzono badanie rzetelności wyników symulacji propagacji fali elektromagnetycznej w modelu sali wykładowej stworzonej w programie Altair Feko, a następnie porównano otrzymane dane z wynikami badań empirycznych. Pomiary w sali wykładowej wykonano za pomocą dwóch anten dookólnych, umieszczonych na statywach na wysokości 60 cm.**

Abstract: **The purpose of this article is to analyze the attenuation of radio signals in a lecture hall at frequency of 20 GHz. A study of the reliability of the results of simulations of electromagnetic wave propagation in the lecture hall model created in Altair Feko software was carried out, and then the obtained data were compared with the results of empirical studies. Measurements in the lecture hall were made using two omnidirectional antennas, placed on tripods at a height of 60 cm.**

Słowa kluczowe: **pomiary tłumienia propagacyjnego, modele wewnątrzbudynkowe, propagacja fali.**

Keywords: **path loss measurements, indoor-to-indoor models, wave propagation.**

1. WSTĘP

Coraz większa ilość informacji przesyłana w dzisiejszych czasach pomiędzy urządzeniami stwarza zapotrzebowanie na wykorzystywanie wyższych częstotliwości, które z uwagi na większą ilość dostępnych zasobów pozwalają na zwiększenie przepustowości łącza. W przypadku fal elektromagnetycznych dla zakresu częstotliwości FR2 [6], obiekty takie jak biurka, krzesła, elementy naścienne oraz przedmioty podobnej wielkości, mogą wprowadzać istotne tłumienie w budżecie łącza.

W przypadku systemów telekomunikacyjnych kluczowym aspektem ich działania oraz niezawodności jest rzetelność wyników symulacji programów do propagacji fal elektromagnetycznych użytych do ich projektowania. Odchyły względem faktycznych wyników pomiarów mogą mieć krytyczny wpływ na działanie systemu.

Spośród wielu prac traktujących o modelowaniu kanału radiowego wewnątrz sal wykładowych lub konferencyjnych na uwagę zasługują prace poświęcone propagacji w paśmie tzw. sub-THz [3, 4]. W pracy [2] wyznaczono współczynniki modelu α, β, γ dla sali konferencyjnej dla częstotliwości 7,1 GHz.

Interesujące porównanie dla trzech częstotliwości 10, 60, 300 GHz przedstawiono w artykule [8] w formie tzw. wykresów PAS (Power Angular Spread). Przedstawiają one moc względem kąta nadejścia sygnału dla pustej sali konferencyjnej. W artykule [5] przedstawiono wyniki tłumienia propagacyjnego dla pasma fal centymetrowych dla typowej sali konferencyjnej. Pomiary wykonano za pomocą macierzy anten. Scenariusz typowej sali wykładowej rozważono w pracy [7], gdzie dla częstotliwości 13-17 GHz wykonano serię pomiarów z anteną o dookólnej charakterystyce promieniowania i anteną odbiorczą o kącie połowy mocy wynoszącym 10 lub 30°. Istotnym aspektem pracy była ekstrakcja komponentów wielodrogowych i pogrupowanie ich w domenach opóźnienia i kąta nadejścia.

Celem artykułu jest analiza tłumienia propagacyjnego w sali wykładowej dla częstotliwości w paśmie fal centymetrowych. W ramach pracy wykonano weryfikację zgodności wyników symulacji i wykonanych pomiarów oraz wyznaczono współczynniki modelu propagacyjnego dla sali wykładowej, gdy anteny były umieszczone poniżej stolików wykładowych, ponieważ ich wpływ na tłumienie propagacyjne jest najbardziej interesujący. Zestawiono wyniki symulacji oraz pomiarów dla fali o częstotliwości 20 GHz. Wybór takiego pasma został umotywowany brakiem podobnych pomiarów dla zadanej częstotliwości.

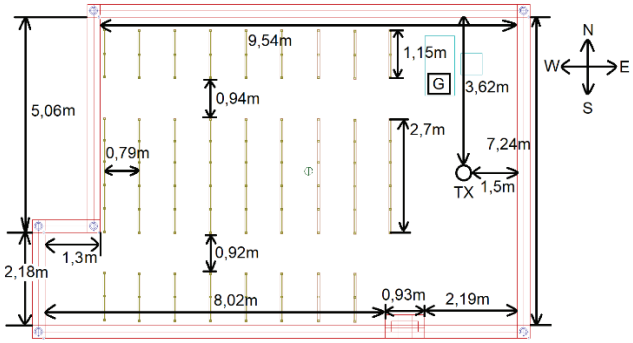
Plan artykułu jest następujący. W rozdziale drugim przedstawiono założenia dotyczące tworzenie cyfrowego bliźniaka sali wykładowej, natomiast w rozdziale trzecim opisano metodologię pomiarową. Wyniki pomiarów przedstawiono w rozdziale czwartym a w rozdziale piątym wykonano analizę porównawczą wykonanych symulacji i pomiarów. Całość podsumowano w rozdziale szóstym.

2. CYFROWY BLIŹNIAK SALI WYKŁADOWEJ

W celu przeprowadzenia symulacji, najpierw wykonano pomiary geometryczne sali oraz elementów takich jak krzesła, biurko oraz elementy naścienne, aby stworzyć model 3D pomieszczenia w programie Altair Wallman (rys. 1.) z uwzględnieniem metalowych podstaw krzeseł oraz stołów. Dobrano elementy konstrukcyjne, których parametry zestawiono w tabeli 1.

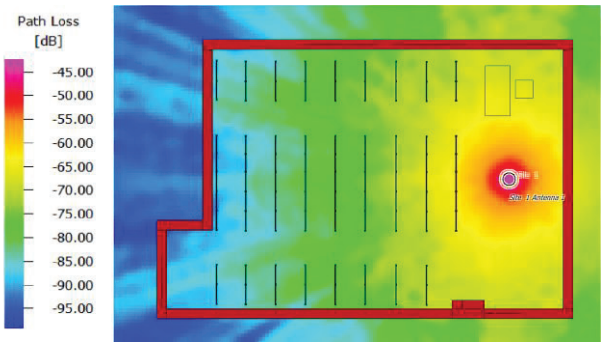
Tab. 1. Parametry elementów konstrukcyjnych w modelu sali.

Element	*Grubość [mm]*	*Przenikalność elektryczna*
Ceglane ściany	300	4.0
Szklane drzwi	5	6.0
Metalowe drzwi	5	1.0
Szklane okna	5	6.0
Drewniane stoliki	50	1.8
Drewniane biurko	50	1.8



Rys. 1. Rzut sali wykładowej z naniesionymi wymiarami oraz zaznaczonym umiejscowieniem generatora i anteny nadawczej

Następnie za pomocą programu Altair Proman przeprowadzono symulację propagacji fali. Zostały ustawione następujące parametry symulacji: maksymalna liczba odbić: 2, maksymalna liczba dyfrakcji: 1. Wyznaczenie tłumienia propagacyjnego zostało zrealizowane z wykorzystaniem współczynników względnej przenikalności elektrycznej materiałów znajdujących się w sali. Antena nadawcza została ustawiona 1,5 m od ściany wschodniej (E) oraz w równej odległości od ścian północnej (N) i południowej (S), na wysokości 60 cm od ziemi (rys. 1). Częstotliwość generowanej fali wynosiła 20 GHz, a moc 0 dBm. Wynik poziomu tłumienia uzyskany za pomocą symulacji został przedstawiony na rysunku 2.

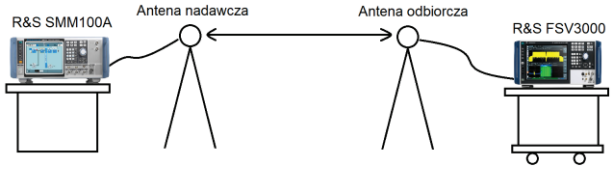


Rys. 2. Wyniki poziomu tłumienia propagacyjnego wewnątrz sali wykładowej

3. METODOLOGIA POMIARU

Pomiary przeprowadzono według konfiguracji sprzętowej przedstawionej na rys. 3. W części nadawczej wektorowy generator sygnałowy Rohde & Schwarz SMM100A został połączony z anteną o charakterystyce zbliżonej do dookólnej przewodem w. cz. o długości 5 m. Miejsce ustawienia anteny nadawczej było tożsame z przeprowadzonymi eksperymentami symulacyjnymi.

Szczegółowa lokalizacja anteny nadawczej (TX) została ujęta na rys. 1. W części odbiorczej sygnał z anteny odbiorczej (o charakterystyce promieniowania, jak antena nadawcza) był odbierany przez analizator widma Rohde & Schwarz FSV3000. W tab. 2. przedstawiono nastawy analizatora widma. W sali wykładowej wyznaczono 38 punktów pomiarowych oznaczających położenie anteny odbiorczej (rys. 4). Wysokość anteny odbiorczej, jak i nadawczej była taka sama i wynosiła 0,6 m podczas, gdy wysokość siedzeń wynosiła 0,95 m.

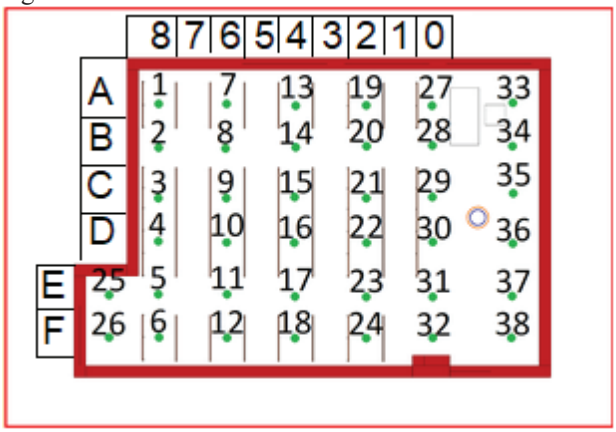


Rys. 3. Schemat przyrządów pomiarowych

Tab. 2. Nastawy analizatora widma

Parametr	*Wartość*	*Jednostka*
Zakres (Span)	15	[kHz]
Minimalne pasmo rozdzielcze (RBW)	100	[Hz]
Czas przemiatania (Sweep)	53	[ms]
Poziom odniesienia	-50	[dBm]
Częstotliwość środkowa	20	[GHz]
Poziom szumów	-130	[dBm]

Antenę odbiorczą ustawiono w pierwszym punkcie pomiarowym i zapisano wartość mocy odebranej odczytanej z analizatora widma, a następnie powtórzono procedurę dla kolejnych punktów pomiarowych (rys. 4). Cyframi 0-8 oznaczono rzędy krzeseł, natomiast litery A-F reprezentują kolejne wiersze punktów pomiarowych.. Zbadano również tłumienie kabli koncentrycznych łączących generator z anteną nadawczą oraz antenę odbiorczą z analizatorem widma, a następnie skorygowano wyniki pomiarów empirycznych o wartość tego tłumienia.

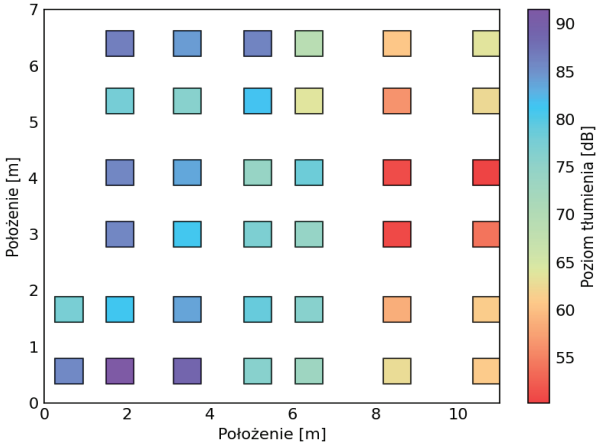


Rys. 4. Umiejscowienie punktów pomiarowych w sali wykładowej

4. WYNIKI POMIARÓW

Wyniki wartości mocy odebranej dla 38 punktów pomiarowych przedstawiono na rys. 5. Wypełnienie kwadratu dla każdego punktu pomiarowego zgodnie z legendą oznacza wartość mocy odebranej. Na krawędziach sali wykładowej zapisano skalę w metrach.

Przedstawione wyniki mocy odebranej to mediana wartości zmierzonych dla trzech serii pomiarowych.



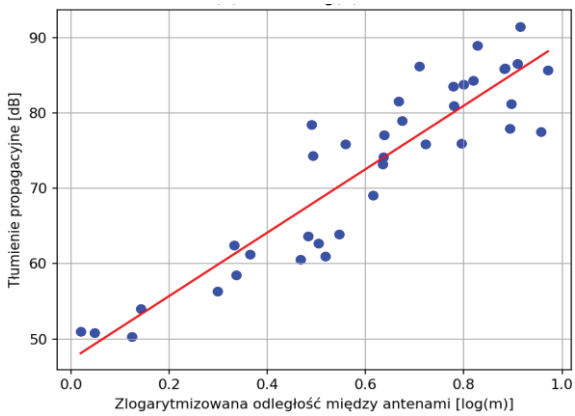
Rys. 5. Wykres poziomu tłumienia propagacyjnego w punktach pomiarowych

Na podstawie wykonanych pomiarów określono wzór (1) dotyczący tłumienia propagacyjnego w sali wykładowej. Jako, że pomiary zostały wykonane dla tylko jednej częstotliwości, zdecydowano, aby wyznaczyć model FI [1].

$$PL(d) = 42{,}1 \log\left(\frac{d}{1m}\right) + 47{,}28 + \chi_\sigma \qquad (1)$$

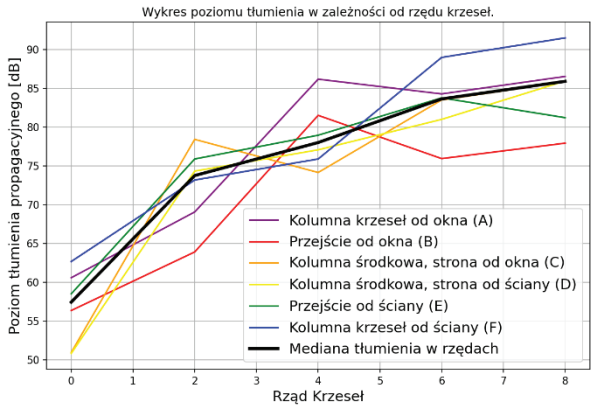
gdzie d jest odległością euklidesową liczoną między położeniem nadajnika i odbiornika. χ_σ jest gaussowską zmienną losową dla log-normalnego modelu przesłaniania, gdzie σ równa jest 4,8 dB.

Krzywa modelu wyznaczona w równaniu (1) została przedstawiona na rys. 6 wraz z wszystkimi punktami pomiarowymi w funkcji logarytmu odległości *d*.



Rys. 6. Wykres poziomu tłumienia propagacyjnego w zależności od odległości d wraz z krzywą modelu

Jednakże tłumienie propagacyjne przedstawione we wzorze (1) oraz na rys. 6 nie uwzględniają obecności i liczby składanych stolików. Dlatego na rys. 7 przedstawiono tłumienie propagacyjne w zależności od rzędu jak i liczby stolików w sali. Oznaczenia rzędów dla poszczególnych punktów pomiarowych zostały przedstawione na rys. 4. Należy jednak zwrócić uwagę, że liczba przeszkód jest ściśle skorelowana z odległością euklidesową między nadajnikiem a odbiornikiem. Z uwagi na ten fakt i wysoki współczynnik korelacji modelu PL(d) uznano, że nie jest konieczna dalsza rozbudowa modelu o kolejne parametry.



Rys. 7.Wykres zależności poziomu tłumienia od rzędu krzeseł dla poszczególnych kolumn

5. PORÓWNANIE WYNIKÓW POMIARÓW Z SYMULACJĄ ORAZ OPRACOWANYM MODELEM

W celu oceny skuteczności predykcji modelu wyznaczono współczynnik korelacji Pearsona, gdyż dostrzeżono silną tendencję liniową dla wyników tłumienia propagacyjnego w funkcji logarytmu odległości pomiędzy nadajnikiem, a odbiornikiem.

Wzór na korelację Pearsona:

$$r_{xy} = \frac{\sum_{i=0}^{n}(x_i - \underline{x})(y_i - \underline{y})}{\sqrt{\sum_{i=0}^{n}(x_i - \underline{x})^2}\sqrt{\sum_{i=0}^{n}(y_i - \underline{y})^2}} \qquad (2)$$

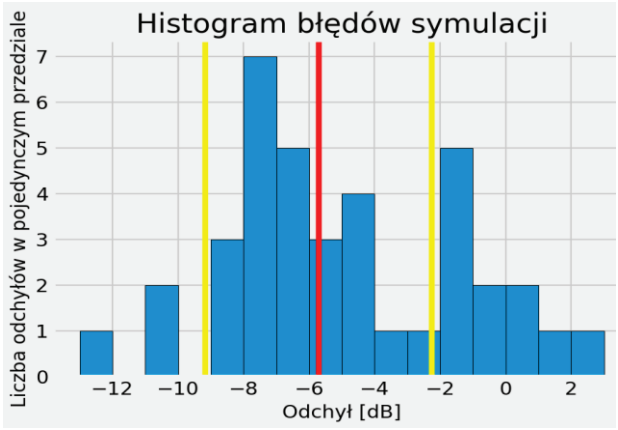
Współczynnik korelacji jest liczony, aby ocenić siłę i kierunek związku liniowego między dwiema zmiennymi ilościowymi X i Y (zbiorami). Zbiory X i Y muszą mieć równą liczbę elementów. Zbiorami w badanym przypadku są wyniki symulacji oraz pomiarów empirycznych.

X={ x_0, x_1...x_n}; Y={y_0, y_1...y_n} gdzie:
- $\underline{x}$, $\underline{y}$ są wartościami średnimi wyników pomiarów;
- x_i, y_i są i-tymi elementami zbiorów X i Y o rozmiarach N;
- n to indeks ostatniego elementu w zbiorach X i Y (n=N-1)

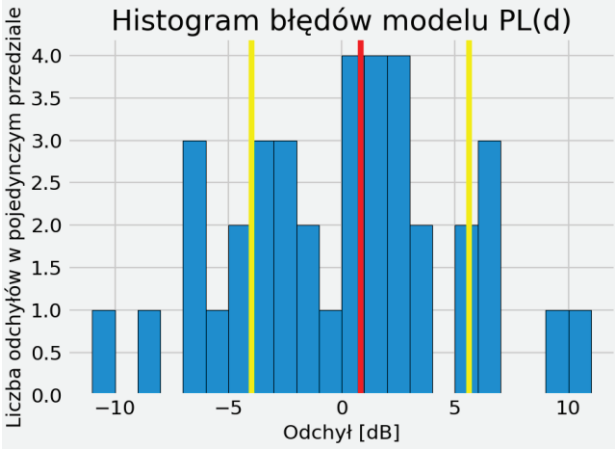
Dla utworzonego modelu estymator korelacji Pearsona pomiędzy predykcją modelu a empirycznymi wartościami tłumienia, osiągnął wartość na poziomie 91,4%.

Współczynnik Pearsona jest użytecznym, lecz nie wystarczającym kryterium do stwierdzenia skuteczności predykcji modelu tłumienia propagacyjnego. Współczynnik korelacji Pearsona mierzy jedynie liniową zależność między dwiema zmiennymi ilościowymi. W takim wypadku wysoka korelacja nie jest elementem determinującym poprawne przewidywane przez model wartości tłumienia. Korelacja jest silnie wrażliwa na wartości odstające które mogą znacznie wpływać na rzeczywistą zależność między zmiennymi. Wynikiem tego są problemy z tworzeniem modeli predykcyjnych. Korelacja nie uwzględnia również wielkości oraz częstości występowania błędów predykcji.

W celu oceny różnicy estymacji tłumienia propagacyjnego wyznaczonego na podstawie pomiarów i pakietu symulacyjnego wyznaczono histogramy przedstawione na rys. 8. Odchył zaznaczony na osi odciętych wyrażony jest jako różnica wartości poziomu tłumienia w pomiarach od wartości poziomu tłumienia w symulacji. Podobnie po opracowaniu modelu tłumienia w funkcji odległości nadajnika od odbiornika również utworzono histogram błędów pomiędzy modelem tłumienia, a empirycznymi danymi (rys. 9). W przypadku tego histogramu odchył wyraża różnicę poziomu tłumienia propagacyjnego uzyskanego w pomiarach od poziomu tłumienia propagacyjnego wskazywanego przez model w tych samych punktach. Na histogramach zaznaczono czerwoną, pionową linią wartość mediany błędów, zaś żółtymi, pionowymi liniami wyróżniono obszar w granicach do pierwszego odchylenia standardowego błędów. Szerokość pojedynczego przedziału w histogramie ustalono na 1 dB.



Rys. 8. Histogram błędów symulacji względem pomiarów poziomu tłumienia



Rys. 9. Histogram błędów modelu PL(d) względem pomiarów poziomu tłumienia

6. WNIOSKI

W pracy wykonano analizę tłumienia propagacyjnego w sali wykładowej na podstawie wykonanej symulacji oraz serii pomiarów. Należy zwrócić uwagę, że mimo tego, że współczynnik korelacji Pearsona jest na poziomie 95,95 %, to jednak rozkład odchyłów wskazuje na to, że symulacja ma tendencje do zawyżania poziomu tłumienia o więcej niż 6 dB w ok. 50% przypadków pomiarów i o więcej niż 0 dB w 95% przypadków pomiarów.

Histogram błędów modelu empirycznego cechuje się dobrze usytuowaną medianą w okolicy zera, lecz jak można zauważyć wyróżnia go całkiem duża wariancja błędów. W najgorszym przypadku model nie doszacowuje tłumienia o ok. 10 dB. Pomimo faktu, że współczynnik korelacji między modelem a pomiarami jest nieco mniejszy niż współczynnik korelacji pomiędzy rezultatami symulacji a pomiarami, to jednak porównanie histogramów błędów jasno pokazuje, że predykcja wynikająca z modelu jest bardziej trafna, gdyż rozkład błędów modelu cechuje się zdecydowanie lepiej usytuowaną medianą, choć wariancja rozkładu błędów modelu jest nieco większa.

PODZIĘKOWANIA

Praca została sfinansowana przy wsparciu projektu nr: 0312/SBAD/8163.

LITERATURA

[1] Cheng C.-L., Kim S., and Zajić., 2017. “Comparison of path loss models for indoor 30 GHz, 140 GHz, and 300 GHz channels”. 11th European Conference on Antennas and Propagation (EUCAP), Paris, France, Mar. pp. 716–720.

[2] Chong Y. -J., Hong J. -Y., Kim J., Lim J. -S. 2022. "Propagation Characteristics of Indoor Channels at 7.1 GHz Frequency Band". 13th International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Korea, str. 1762-1765.

[3] De Beelde B., Joseph W., Plets D., Tanghe E. 2022. "Indoor Radio Channel Modeling at D-Band Frequencies". 16th European Conference on Antennas and Propagation (EuCAP), Madrid, str.1-4.

[4] Dupleich D., Galdo G. D., Landmann M., Müller R., Skoblikov S., Thomä R. 2020. "Characterization of the Propagation Channel in Conference Room Scenario at 190 GHz". 14th European Conference on Antennas and Propagation (EuCAP), Copenhagen, Denmark, str. 1-5.

[5] Lin L., Liu Y., Zhang R. 2016. "MIMO channel sounder and millimeter wave measurements in a conference room". 11th International Symposium on Antennas, Propagation and EM Theory (ISAPE), Guilin, China, str. 812-814.

[6] Lin X. 2022. "An Overview of 5G Advanced Evolution in 3GPP Release 18". *IEEE Communications Standards Magazine*, tom. 6, nr. 3, str.. 77-83.

[7] Ling Cen, Wang Haowen, Yin Xuefeng, Zhang Xiaomei. 2016. „Comparison of Characteristics of 13-17 GHz Propagation Channels in Indoor Environments with Different Measurement Configurations”. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications.

[8] Vitucci E. M. et al. 2018. "Tri-Band Mm-wave Directional Channel Measurements in Indoor Environment". IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Bologna, str. 205-209.

PROJEKTOWANIE I WERYFIKACJA POMIAROWA LINII TEM

DESIGN AND MEASUREMENT VERIFICATION OF TEM CELL

Nikoloz Glonti[1]; Marta Sieradzka[1]; Cyryl Prentki[1]; Krzysztof Cichoń[1]; Jarosław Szóstka[1]

[1] KN Spektrum, Wydział Informatyki i Telekomunikacji, Politechnika Poznańska, Poznań, {nikoloz.glonti; marta.sieradzka.1; cyryl.prentki}@student.put.poznan.pl, {krzysztof.cichon; jaroslaw.szostka}@put.poznan.pl

DOI: 10.15199/59.2024.4.67

Streszczenie**: W pracy poruszono tematykę projektowania linii TEM w celu wytworzenia jednorodnego pola elektromagnetycznego (EM). Założenia projektowe poddano symulacji, aby dla wymaganych rozmiarów linii uzyskać najlepsze możliwe parametry S11/S12 i najbardziej jednorodne pole EM. Zbudowaną linię zweryfikowano podczas pomiarów, gdzie uzyskano wysoką zgodność z wartościami wynikającymi z symulacji.**

Abstract**: This paper addresses the topic of designing TEM cell to generate a homogeneous electromagnetic (EM) field. The design assumptions were simulated to achieve the best possible S11/S12 parameters and the most homogeneous EM field for the required line dimensions. The constructed cell was verified through measurements, which showed high agreement with the simulation results.**

Słowa kluczowe**: Linia TEM, natężenie pola EM, S11/S21, pomiar mocy odbieranej.**

Keywords**: TEM cell, EM field intensity, S11/S21, received power measurement.**

1. WSTĘP

Nieustanny gwałtowny rozwój systemów łączności bezprzewodowej powoduje, że otacza nas coraz więcej urządzeń przesyłających sygnały w sposób bezprzewodowy. Tak jak u początku sieci komórkowych możliwość realizacji połączenia bezprzewodowego była możliwa tylko dla nielicznych, obecnie według prognozy firmy Ericsson globalne natężenie ruchu w sieciach mobilnych ma wynieść 562 EB na miesiąc w roku 2029 [3]. Poza imponującym rozwojem sieci komórkowych, coraz więcej urządzeń ma moduł bezprzewodowy pozwalający przesyłać rzadkie i niewielkie komunikaty [5], aż do sieci bezprzewodowych mogących funkcjonować nawet wewnątrz ludzkiego organizmu [4]. Każde urządzenie bezprzewodowe podczas transmisji jest źródłem pola elektromagnetycznego (EM), co przyczynia się do niejednorodności pola w przestrzeni. Co ciekawe, wiele kierunków badań dotyczy wpływu pola EM na organizmy żywe. Wzbudziło to chęć przeprowadzenia badania wpływu pola EM na wzrost roślin. W tym celu przygotowano specjalne komory, w których wysiane zostaną rośliny poddane wpływowi pola elektromagnetycznego. Jednak samo zagadnienie wytworzenia jednorodnego pola elektromagnetycznego o danym natężeniu jest wyzwaniem. Badanie wpływu pola wymaga zmiany wartości natężenia jednorodnego pola na badane próbki.

Celem artykułu jest przedstawienie procesu projektowania linii TEM, która pozwala na wytworzenie jednorodnego pola EM dla określenia wpływu pola na wzrost roślin. Projekt linii wymagał wykonania symulacji licznych parametrów linii. Po jej zbudowaniu nastąpiła weryfikacja pomiarowa założonych parametrów.

Artykuł jest uporządkowany następująco. W rozdziale drugim przedstawiono czym jest linia TEM i przywołano główne wzory istotne dla jej projektowania. W rozdziale trzecim opisano proces projektowania linii TEM, założenia i wyniki badań symulacyjnych, dodatkowo zestawiając wartości macierzy rozproszenia S11, S12 otrzymane z symulacji z wartościami zmierzonymi. W rozdziale czwartym przedstawiono wyniki pomiarów pola elektrycznego wykonanych linii TEM. Całość podsumowano w rozdziale piątym.

2. LINIA TEM

Linie TEM (ang. *Transverse Electromagnetic*) to linie, które zazwyczaj są realizowane jako linie współosiowe z centralnym przewodnikiem otoczonym cylindrycznym ekranem lub jako linie mikropaskowe i paskowe [2]. Szczególnym przypadkiem jest linia płaska TEM. Jedno z jej zastosowań to badania kompatybilności elektromagnetycznej, gdy linia albo sprzęga się z energią o częstotliwości radiowej (w.cz.) emitowaną przez urządzenia poddawane testom w celu przeprowadzenia badań emisji, albo symuluje promieniowane pole w przestrzeni dla testów podatności urządzeń. Linia może zastępować antenę odbiorczą lub nadawczą stosowaną w konwencjonalnych pomiarach w ekranowanych pomieszczeniach. Charakteryzuje się niezwykle szerokim pasmem z liniową odpowiedzią fazową i amplitudową od częstotliwości zerowej do częstotliwości odcięcia komory. Ta cecha umożliwia przeprowadzanie testów sygnałów ciągłych (CW), z modulacją częstotliwości, a także sygnałów impulsowych i modulowanych [1]. Linia TEM ma jednak swoje ograniczenia. Przede wszystkim jej górna użyteczna częstotliwość jest ograniczona przez jej fizyczny rozmiar, który z kolei wpływa na wielkość

testowanego obiektu. Górne ograniczenie częstotliwości wynika z rezonansów wysokiej częstotliwości i wielomodowości występujących w komorze, które zniekształcają jednolite właściwości trybu TEM komory [1]. Długość zwężenia komory również wpływa na wielomodowość wewnątrz komory.

2.1. **Zasada działania**

Po wprowadzeniu do linii sygnału w. cz., między środkową płytą (septum) a zewnętrzną osłoną linii zaczyna wytwarzać się pole EM. Przy odpowiednim dobraniu wymiarów linii i częstotliwości sygnału pobudzającego, pole to w pewnej przestrzeni wewnątrz linii stanie się jednorodne. Przestrzeń ta najbardziej nadaje się do pomiarów wpływu pola EM na rośliny. Rozmiar użytecznej przestrzeni wewnątrz linii, graniczna częstotliwość, impedancja charakterystyczna oraz WFS linii są bezpośrednio zależne od wymiarów linii. Impedancja charakterystyczna linii jest dana wzorem [1]:

$$Z_0 \approx \frac{\eta_0}{4\left[\frac{a}{b} - \frac{2}{\pi}\ln\left(sinh\left(\frac{\pi g}{2b}\right)\right)\right] - \frac{\Delta C}{\varepsilon_0}}, \quad (1)$$

gdzie *b, a, g* to wymiary linii (oznaczone na Rys. 1), ΔC to pojemność brzegowa między krawędziami septum a ścianami bocznymi linii (jeśli stosunek a do b jest większy od 1, to wartość ta jest pomijalnie mała), η_0 to impedancja falowa wolnej przestrzeni, ε_0 to przenikalność elektryczna wolnej przestrzeni. Wzór na częstotliwość rezonansową [1]:

$$f_{res} = \sqrt{{f_c}^2 + \left(\frac{c}{2d}\right)^2}, \quad (2)$$

gdzie *d* to długość rezonansowa linii, c to prędkość światła a f_c to częstotliwość odcięcia modu TE_{10} [1]:

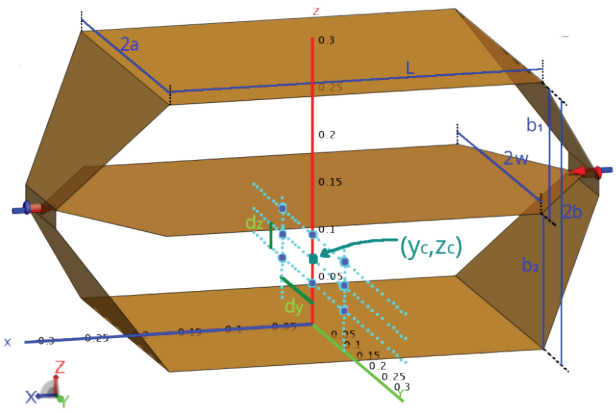
$$f_c = \frac{75}{a}\sqrt{1 + \frac{4ab}{\pi b_1 b_2 \ln\left(\frac{8a}{\pi g}\right)}} \quad (3)$$

2.2. **Pasmo pracy**

Mody TE to jedne z pierwszych modów pojawiających się w linii TEM. Pierwszym jest mod TEM, który powinien propagować się od jednego końca linii do drugiego. Mody TE pojawiają się powyżej ich częstotliwości granicznej. Mody wyższego rzędu ograniczają maksymalne pasmo pracy linii. Stłumienie modów powyżej TE_{10} jest kluczowe, aby zwiększyć pasmo pracy linii.

3. WYNIKI SYMULACJI, PORÓWNANIE PARAMETRÓW S11/S12

Linia TEM została zaprojektowana w oprogramowaniu Altair FEKO [6], szczególnie użytecznym do wykonania zaawansowanych symulacji elektromagnetycznych struktur trójwymiarowych. Schemat zaprojektowanej linii przedstawiono na rysunku 1. Model Linii TEM został sparametryzowany, zatem istnieje możliwość - poprzez zmianę wartości jednej zmiennej - modyfikacji długości linii TEM, czy odległości między górną oraz dolną częścią, a septum. Linia TEM w oprogramowaniu FEKO została zamodelowana z użyciem idealnego przewodnika, a parametry geometryczne obu portów zostały dobrane tak, aby ich impedancja charakterystyczna wynosiła 50 omów.

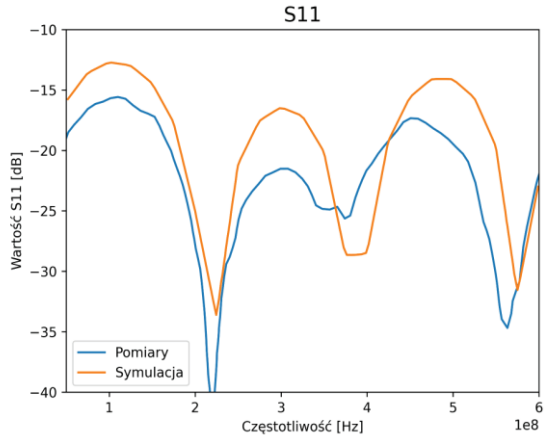


Rys. 1. Zaprojektowany model linii TEM. Na rysunku zaznaczono rozmiary linii $a, b, b1, b2, L, w$ oraz trzy osie: X (kolorem niebieskim), Y (w kolorze zielonym) oraz Z (w kolorze czerwonym). Zaznaczono także odległości używane przy określeniu obszaru roboczego.

Początkowe rozmiary linii TEM zostały wybrane z użyciem wzorów z rozdziału drugiego. Po pierwszych symulacjach, dalsza optymalizacja rozmiarów linii TEM okazała się konieczna z powodu zbyt dużej wartości parametru S11. Optymalizacja wymiarów została przeprowadzona z użyciem wbudowanej funkcjonalności pozwalającej na wykonanie wielu symulacji, w których modyfikacji podlega jedna zmienna. Celem optymalizacji w naszym przypadku było uzyskanie jak najniższej wartości współczynnika S11 (rys. 2). Jego niska wartość przekłada się na niską moc odbitą od wejścia linii. Pozwala to na użycie wzmacniacza wielkiej częstotliwości dużej mocy bez ryzyka jego uszkodzenia.

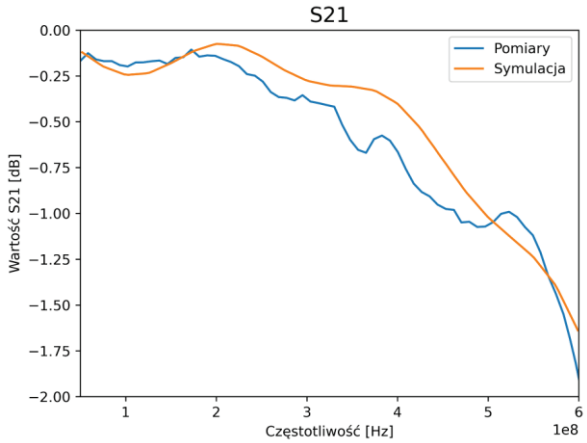
Po optymalizacji wymiarów linii TEM przeprowadzone zostały symulacje zarówno parametrów S, jak i rozkładu pola elektrycznego wewnątrz linii. Pomiary parametrów S zostały wykonane z użyciem analizatora sieci wektorowych ZVL firmy Rohde Schwarz.

Pierwszym istotnym parametrem charakteryzującym linię TEM jest S21, który informuje o tłumieniu linii w zależności od częstotliwości (rys. 3). Straty mogą wynikać zarówno z promieniowania energii



Rys. 2. Porównanie S11 otrzymanego podczas symulacji z wynikiem pomiarów

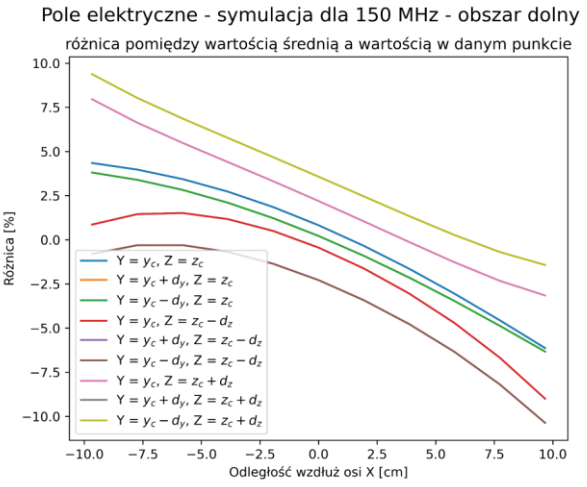
na zewnątrz linii TEM(ma to miejsce na wyższych częstotliwościach, powyżej oraz w pobliżu częstotliwości odcięcia), jak i zastosowanego materiału.



Rys. 3. Porównanie S21 otrzymanego podczas symulacji z wynikiem pomiarów

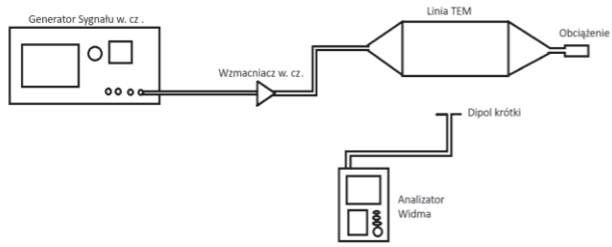
Różnice między wartościami zmierzonymi i obliczonymi wynikają ze strat materiału wykonania linii TEM, oraz niedokładności wykonania mechanicznego. Wyniki symulacji z symulacji oraz pomiarów są zbieżne zarówno dla parametru S11, jak i S12. Kluczowym elementem zbieżności dwóch współczynników są minima na zbliżonych częstotliwościach.

Najważniejszym jednak parametrem, na podstawie którego możemy określić przydatność linii TEM do badań, jest jednorodność wytwarzanego pola elektrycznego. Linie TEM przeznaczone do badań posiadają obszar roboczy, który jest ograniczony.



Rys. 4. Względne natężenie pola elektrycznego dla obszaru dolnego linii TEM

Rysunek 4 przedstawia wyniki symulacji względnego natężenia pola elektrycznego Powyższy wykres otrzymano w następujący sposób. Wartości pola elektrycznego wzdłuż osi X (od -100 do 100 cm) - dla dziewięciu par Y, Z (zaznaczonych punktami na rysunku 1) podzielono przez ich łączną wartość średnia. Następnie otrzymane wartości zamieniono na postać procentową. Punkt (y_c, z_c) to (0, 65.86 cm), natomiast d_y to 80 cm, a d_z 21.46 cm.



Rys 5. Schemat układu pomiarowego

Wyniki natężenia pola przedstawione na rys. 4 jednoznacznie wskazują, że pole E jest jednorodne wewnątrz linii na obszarze o rozmiarach 160 x 200 x 42 mm (odpowiednio szerokość, długość oraz wysokość). W obrębie tego obszaru wartość E waha się w maksymalnie granicach +-10%.

Na rys. 6 przedstawiono znormalizowane wartości pola E na płaszczyznach odpowiednio XY, XZ oraz ZY.

4. WERYFIKACJA POMIAROWA

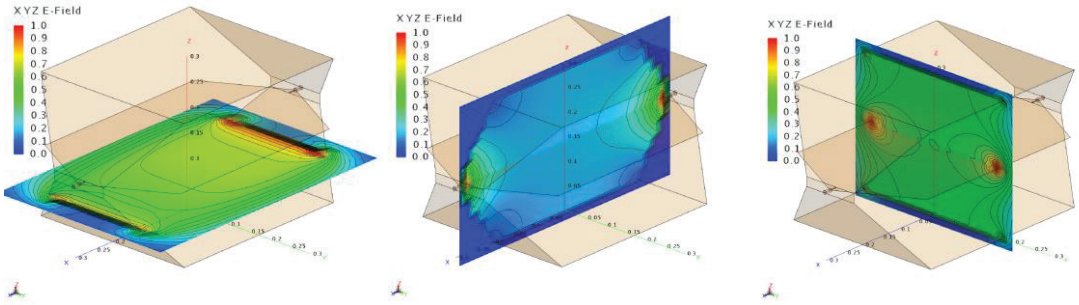
4.1 Wyniki pomiarów mocy odebranej

Po wykonaniu licznych symulacji i zbudowaniu linii wykonano serię pomiarów weryfikujących rzeczywiste parametry. Pomiary przeprowadzono na częstotliwości $f = 150$ MHz za pomocą zestawu pomiarowego przedstawionego na rysunku 5. Pomiaru mocy odebranej wewnątrz linii wykonano za pomocą krótkiego dipola wykonanego z kabla koncentrycznego. W tab. 1 przedstawiono rozdzielczość pomiarową i liczbę wykonanych pomiarów wzdłuż jednej osi.

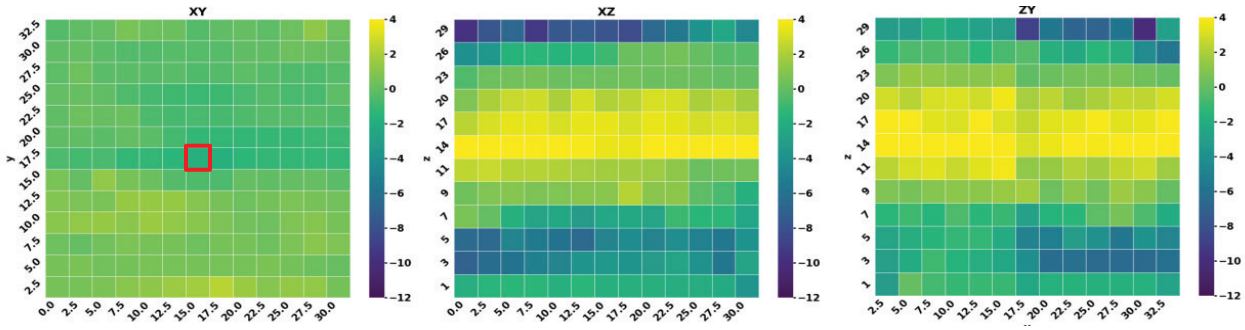
Tab. 1. Parametry wykonanych pomiarów

Oś	*Parametr (rys. 1)*	*Rozmiar [cm]*	*Rozdzielczość pomiarowa [cm]*	*Liczba pomiarów w osi*
X	L	30	2,5	14
Y	a	32,5	2,5	13
Z	b	28	2	12

Pomiary wykonano dla każdej z trzech osi linii. W efekcie, przykładowo w płaszczyźnie poziomej (XY) znalazło się 13·14=182 punkty pomiarowe a takich płaszczyzn pomiarowych w sumie było 12. Razem oznacza to wykonanie ponad dwóch tysięcy pomiarów. Na rys. 7 przedstawiono jedynie część wykonanych pomiarów w formie znormalizowanych wyników mocy odebranej. Wyniki pomiarów przedstawiono w formie tzw. mapy cieplnej, gdzie kolor zgodnie z legendą oznacza względną wartość mocy odniesioną do wartości mocy w środku obszaru roboczego, oznaczonej czerwonym kwadratem na rys. 7a.



Rys.6. Znormalizowane pole E a) przekrój XY, Z = zc, b) *przekrój XZ, Y = 0, c) przekrój ZY. X = 0*



Rys. 7. Wyniki pomiarów względnej mocy w trzech płaszczyznach (por. opisy osi przedstawione na rys. 1): a) w płaszczyźnie XY, b) w płaszczyźnie XZ oraz c) w płaszczyźnie YZ. Na osiach zapisano rozmiary linii w cm. Czerwonym kwadratem oznaczono środek przestrzeni roboczej linii TEM, użytej jako wartość mocy odniesienia.

4.2 Wyniki pomiarów mocy odebranej

Kluczowym wyzwaniem była kalibracja wartości mocy odebranej. W tym celu wykonano pomiar natężenia pola elektromagnetycznego za pomocą miernika natężenia pola MEH-25.

Przy mocy przyłożonej do wejścia linii TEM 23.5 dBm, moc odebrana na dipolu krótkim wyniosła -32.4 dBm a wartość pola, zmierzona w miejscu dipola krótkiego, wyniosła 30 V/m.

Przeprowadzone symulacje pola elektromagnetycznego w płaszczyznach XY, XZ oraz ZY, przedstawione na Rys. 6, wykazują zgodność z wynikami pomiarów mocy odebranej przedstawionymi na mapach ciepła na Rys. 7. Analiza porównawcza symulacji z wynikami pomiarów pokazuje, że rozkład mocy w płaszczyznach jest spójny. W symulacjach można zaobserwować maksymalne wartości pola elektromagnetycznego w tych samych obszarach, w których na mapach ciepła pojawiają się najwyższe wartości mocy w dB. Na przykład, w płaszczyźnie XY symulacje jak i pomiary, wskazują na wyraźne maksimum w centralnym obszarze, co potwierdza poprawność modelu symulacyjnego. W płaszczyznach XZ i ZY wyniki również są zgodne - obszary o najwyższym natężeniu pola elektromagnetycznego w symulacjach odpowiadają obszarom z najwyższymi wartościami mocy odebranej w wynikach pomiarów.

5. WNIOSKI

W pracy przedstawiono problematykę projektowania, symulacji i weryfikacji pomiarowej linii TEM. Wykazano dużą zgodność wyników prowadzonych badań symulacyjnych i pomiarowych. Zbudowana linia TEM może z powodzeniem być stosowana do planowanych eksperymentów, bowiem pozwala na wytworzenie jednorodnego pola EM.

6. PODZIĘKOWANIA

Autorzy składają podziękowania dla: firmy Endego za udostępnione oprogramowanie symulacyjne; dla Pawła Cały z firmy Endego za jego wskazówki merytoryczne; członków Koła Naukowego Spektrum za pomoc w realizacji wykonania linii TEM.

Prace nad linią TEM były finansowane ze środków przeznaczonych na działalność studenckiego ruchu naukowego na Politechnice Poznańskiej – w ramach grantu „Badanie wpływu pola elektromagnetycznego na rozwój roślin” przyznanego dla Koła Naukowego Spektrum.

LITERATURA

[1] M. L. Crawford, J. L. Workman and C. L. Thomas, "Expanding the Bandwidth of TEM Cells for EMC Measurements", *IEEE Transactions on Electromagnetic Compatibility*, EMC-20(3), str. 368-375, 1978.

[2] J. Szóstka, *Fale i anteny*, Wydawnictwa Komunikacji i Łączności, Warszawa 2000.

[3] Ericsson Mobility Report, Business Review 2024, https://ericsson.com/mobility-report.

[4] R. Asorey-Cacheda, L. Correia, C. Garcia-Pardo, K. Wójcik, K. Turbic, P. Kułakowski., “Bridging Nano-and Body Area Networks: A Full Architecture for Cardiovascular Health Applications”, *IEEE Internet of Things Journa*l, 2022.

[5] L. Chettri, R. Bera, "A Comprehensive Survey on Internet of Things (IoT) Toward 5G Wireless Systems", *IEEE Internet of Things Journal*, 7(1), str. 16-32, 2020.

[6] Altair FEKO, https://altair.com/feko.

ZWIĘKSZENIE EFEKTYWNOŚCI METODY ELEMENTÓW SKOŃCZONYCH Z WYKORZYSTANIEM WŁASNOŚCI GEOMETRII STRUKTUR

INCREASING THE EFFICIENCY OF THE FINITE ELEMENT METHOD USING THE PROPERTIES OF STRUCTURE GEOMETRY

Małgorzata Warecka[1]

[1] Politechnika Gdańska, Gdańsk, malgorzata.warecka@pg.edu.pl

DOI: 10.15199/59.2024.4.68

Streszczenie: Niniejsza praca jest omówieniem fragmentów rozprawy doktorskiej autorki. Przedstawia ideę połączenia opisu analitycznego z metodą elementów skończonych poprzez uogólnioną macierz impedancji. Przedstawione zostały analizy dwóch struktur z wykorzystaniem metody hybrydowej, które wraz z porównaniem względem innych metod stanowią potwierdzenie zasadności prezentowanego podejścia.

Abstract: This work is a discussion of fragments of the author's doctoral dissertation. It presents the idea of combining analytical description with the finite element method through a generalized impedance matrix. Analyzes of two structures using a hybrid method are presented, which, together with a comparison to other methods, confirms the validity of the presented approach.

Słowa kluczowe: metoda elementów skończonych, metody hybrydowe, struktury rezonansowe, uogólniona macierz impedancji

Keywords: finite element method, generalized impedance matrix, hybrid methods, resonant structures

1. WSTĘP

Współczesna elektronika i telekomunikacja w coraz większym stopniu polegają na efektywnych i dokładnych metodach symulacji komputerowych, które są niezbędne przy projektowaniu i optymalizacji urządzeń. Wzrost poziomu złożoności układów jest jednak na tyle szybki, że sam rozwój infrastruktury obliczeniowej okazuje się niewystarczający, aby sprostać aktualnym wymaganiom środowiska naukowego czy przemysłu. Z tego względu od dziesięcioleci równolegle rozwijane są algorytmy pozwalające zwiększyć efektywność symulacji tak, aby możliwe było osiągnięcie zadanej dokładności w akceptowalnym czasie.

Jednym z kluczowych problemów dzisiejszej elektroniki jest skuteczne modelowanie rozkładu pól elektromagnetycznych w różnego typu strukturach czy układach. W tym celu stosuje się metody elektrodynamiki obliczeniowej, które podzielić można na kilka podgrup. Metodom analitycznym przypisuje się największą dokładność jak również krótki czas obliczeń. Ich zastosowanie wiąże się jednak z ograniczeniem do struktur o prostych geometriach takich jak cylindry czy sfery. Kolejną, bardziej uniwersalną, grupą metod są metody równań całkowych. Ich większa uniwersalność związana jest jednak z całkowaniem osobliwości w obrębie funkcji Greena czy koniecznością odpowiedniego doboru funkcji bazowych. Za najbardziej uniwersalne metody uznaje się metody dyskretne, do których zaliczyć można metodę różnic skończonych czy metodę elementów skończonych. Cieszą się one dużą popularnością wśród komercyjnego oprogramowania ze względu na brak ograniczeń związanych z geometrią struktury. Należy zaznaczyć, że układy skomplikowane pod względem geometrii mogą wymagać zastosowania bardzo gęstej siatki co wydłuża czas obliczeń. Co więcej, problem stanowić może również symulacja przestrzeni otwartej. Do bardziej znanych rozwiązań należy zakończenie dziedziny warstwą PML (perfectly matched layer). Ma ona na celu zniwelowanie odbić od brzegu dziedziny. Niestety jej zastosowanie wymaga poszerzenia dziedziny obliczeniowej oraz doboru zestawu parametrów, które mogą mieć wpływ na dokładność symulacji.

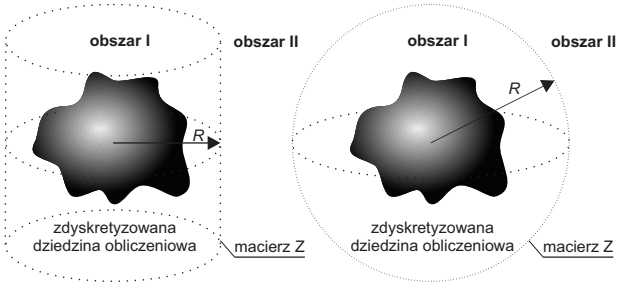
Odpowiedzią na powyższe problemy wydają się być metody hybrydowe. Pozwalają one między innymi na zachowanie elastyczności metod dyskretnych, zwiększenie ich dokładności przy jednoczesnej redukcji złożoności obliczeniowej. Większość tego typu podejść bazuje na dekompozycji dziedziny na mniejsze obszary w zależności od materiału czy geometrii, gdzie w różnych podobszarach stosowane są różne metody odpowiadające zadanym problemom.

Niniejsza praca stanowi omówienie fragmentów rozprawy doktorskiej pt. "Techniki zwiększania efektywności metody elementów skończonych poprzez redukcję dziedziny obliczeniowej z wykorzystaniem własności geometrii struktur" [1]. Zaproponowane rozwiązania hybrydowe są połączeniem metody elementów skończonych i podejścia analitycznego. Metoda elementów skończonych pozwala na dużą dowolność dotyczącą budowy analizowanej struktury, natomiast zastosowanie podejścia analitycznego redukuje rozmiar, zwiększa dokładność przy wykorzystaniu mniejszych zasobów pamięciowych i obliczeniowych.

2. ZARYS METODY

Redukcja dziedziny obliczeniowej z wykorzystaniem własności geometrii struktury może odbywać się poprzez dekompozycję dziedziny obliczeniowej. Wówczas dziedzinę

obliczeniową dzielimy na dwa lub więcej obszarów. Przy czym podobszary te w zależności od swoich własności mogą być opisywane w sposób analityczny lub też z wykorzystaniem metody elementów skończonych. Następnie na granicy podobszarów wprowadzona zostaje uogólniona macierz impedancji - patrz Rys. 1. Wrota opisywane są na powierzchni prostej bryły, którą stanowić mogą cylinder [2] czy sfera [3].



Rysunek 1: Schemat podziału dziedziny obliczeniowej

2.1 Opis analityczny

Prezentowane w niniejszym artykule struktury lub też ich elementy charakteryzuje symetria osiowa co pozwoliło na dalszą redukcję dziedziny obliczeniowej z przypadku 3D do 2.5D. W zależności od kształtu wrót oraz prezentowanego problemu opis analityczny może się różnić. W niniejszym opracowaniu ograniczę się do opisu zagadnienia w układzie współrzędnych cylindrycznych oraz sferycznych.

Dla współrzędnych cylindrycznych pole elektryczne oraz magnetyczne zapisać można następująco:

$$E(\rho,\varphi,z) = E(\rho,\varphi)e^{-\gamma z} \tag{1}$$

$$H(\rho,\varphi,z) = H(\rho,\varphi)e^{-\gamma z} \tag{2}$$

Pole w obszarze drugim w ogólnym przypadku przyjmuje formę:

$$\vec{E}_z^{II}(\rho,\varphi) = \sum_{m=-M}^{M} \left(a_m^E J_m(\kappa\rho) + b_m^E H_m^{(2)}(\kappa\rho)\right) e^{\mathrm{j}m\varphi}\vec{i}_z \tag{3}$$

$$\vec{H}_z^{II}(\rho,\varphi) = \sum_{m=-M}^{M} \left(a_m^H J_m(\kappa\rho) + b_m^H H_m^{(2)}(\kappa\rho)\right) e^{\mathrm{j}m\varphi}\vec{i}_z \tag{4}$$

gdzie m jest liczbą całkowitą oznaczającą zmienność po φ, $\kappa = \sqrt{(k_0^2 - \gamma^2)}$, γ współczynnikiem propagacji a k_0 liczbą falową. Dla współrzędnych sferycznych:

$$\vec{E}^{II}(r,\theta,\varphi) = \sum_{n=1}^{N}\sum_{m=-n}^{n} \left[A_{nm}^{(1,TE)}\vec{M}_{nm}^{(1)}(kr,\theta,\phi) + A_{nm}^{(1,TM)}\vec{N}_{nm}^{(1)}(kr,\theta,\phi) + A_{nm}^{(2,TE)}\vec{M}_{nm}^{(2)}(kr,\theta,\phi) + A_{nm}^{(2,TM)}\vec{N}_{nm}^{(2)}(kr,\theta,\phi)\right] \tag{5}$$

oraz

$$\vec{H}^{II}(r,\theta,\varphi) = -\frac{k}{j\omega\mu}\sum_{n=1}^{N}\sum_{m=-n}^{n} \left[A_{nm}^{(1,TE)}\vec{N}_{nm}^{(1)}(kr,\theta,\phi) + A_{nm}^{(1,TM)}\vec{M}_{nm}^{(1)}(kr,\theta,\phi) + A_{nm}^{(2,TE)}\vec{N}_{nm}^{(2)}(kr,\theta,\phi) + A_{nm}^{(2,TM)}\vec{M}_{nm}^{(2)}(kr,\theta,\phi)\right] \tag{6}$$

gdzie

$$\begin{aligned}\vec{M}_{nm}^{(i)}(kr,\theta,\varphi) &= z_n^{(i)}(kr)e^{jm\varphi} \times \\ &\times \left[\frac{jm}{\sin\theta}P_n^m(\cos\theta)\vec{i}_\theta - \frac{\partial P_n^m(\cos\theta)}{\partial\theta}\vec{i}_\varphi\right] \\ \vec{N}_{nm}^{(i)}(kr,\theta,\varphi) &= \frac{1}{kr}(n(n+1))z_n^{(i)}(kr)e^{jm\varphi} \times \\ &\times P_n^m(\cos\theta)\vec{i}_r + \frac{1}{kr}\frac{\partial(rz_n^i(kr))}{\partial r}e^{jm\varphi} \times \\ &\times \left(\frac{\partial P_n^m(\cos\theta)}{\partial\theta}\vec{i}_\theta + \frac{jm}{\sin\theta}P_n^m(\cos\theta)\vec{i}_\varphi\right)\end{aligned} \tag{7}$$

$P_n^m(\cdot)$ są uogólnionymi wielomianami Legendre'a natomiast, $z_n^{(1)}, z_n^{(2)}$ to odpowiednio sferyczna funkcja Bessla oraz Hankela drugiego rodzaju.

2.2 Uogólniona macierz impedancji

Uogólniona macierz impedancji opisuje relację między składowymi stycznymi pola elektrycznego i magnetycznego, a ściślej między współczynnikami rozwinięcia odpowiednio pola elektrycznego oraz magnetycznego w ustalonej bazie. Funkcje te dla powierzchni sferycznej mogą przyjmować postać:

$$\begin{aligned}\vec{e}_{nm}^{(TE)}(\theta,\varphi) &= \vec{h}_{nm}^{(TM)}(\theta,\varphi) = \\ &= e^{jm\varphi}\left(\frac{jm}{\sin\theta}P_n^m(\cos\theta)\vec{i}_\theta - \frac{\partial P_n^m(\cos\theta)}{\partial\theta}\vec{i}_\varphi\right)\end{aligned} \tag{8}$$

oraz

$$\begin{aligned}\vec{e}_{nm}^{(TM)}(\theta,\varphi) &= \vec{h}_{nm}^{(TE)}(\theta,\varphi) = \\ &= e^{jm\varphi}\left(\frac{\partial P_n^m(\cos\theta)}{\partial\theta}\vec{i}_\theta + \frac{jm}{\sin\theta}P_n^m(\cos\theta)\vec{i}_\varphi\right)\end{aligned} \tag{9}$$

natomiast dla powierzchni walcowej:

$$\begin{aligned}\vec{e}_m^{(TE)}(\varphi,z) = \vec{h}_m^{(TM)}(\varphi,z) &= \vec{e}_m^{(TM)}(\varphi,z) = \\ &= \vec{h}_m^{(TE)}(\varphi,z) = e^{-\gamma z}e^{jm\varphi}\end{aligned} \tag{10}$$

Zapis wspomnianych współczynników w postaci wektorów $\mathbf{V}$ i $\mathbf{I}$ (w pewnym sensie uogólnione napięcia i prądy) pozwala na zdefiniowanie macierzy impedancji $\mathbf{Z}$ jako

$$\mathbf{V} = \mathbf{ZI} \tag{11}$$

W zależności od zagadnienia wektory $\mathbf{V}$ i $\mathbf{I}$ będą przyjmować różną postać oraz konfigurację [3, 4]. Macierz $\mathbf{Z}$ dla prostych struktur może być wyznaczona analitycznie, jednak w bardziej skomplikowanych przypadkach niezbędne jest zastosowanie metod dyskretnych np. metody elementów skończonych, dla której

$$\mathbf{Z} = j\omega\mu_0\mathbf{\Delta}^{-1}\mathbf{B}^H\mathbf{G}^{-1}\mathbf{B}. \tag{12}$$

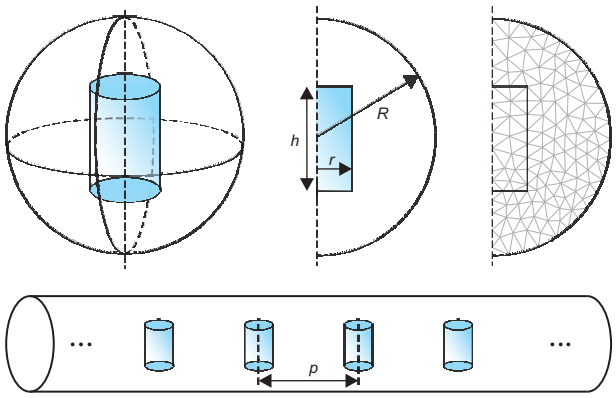
gdzie poszczególne macierze opisane zostały w artykułach [3, 4]. Jak można zauważyć powstała macierz jest niezależna od pobudzenia, co umożliwia jej wielokrotne wykorzystywanie.

3. WYNIKI NUMERYCZNE

W ramach wyników numerycznych przedstawione zostaną dwie struktury: filtrująca linia periodyczna oraz otwarty rezonator Fabry-Perot. W obu przypadkach zastosowane zostały dwa sposoby redukcji dziedziny obliczeniowej tj. dekompozycja oraz symetria osiowa.

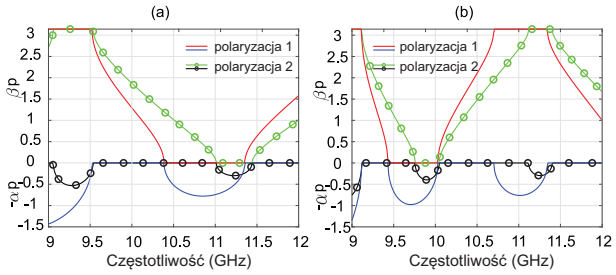
3.1 Struktura periodyczna

Struktura periodyczna zaprezentowana w [3] zbudowana jest z dielektrycznych cylindrów o promieniu $r = 2.3$ mm, wysokości $h = 10$ mm i względnej przenikalności elektrycznej $\varepsilon_r = 12$, które umieszczone są w odległości p równej odpowiednio 40 mm i 60 mm - patrz Rys. 2.



Rysunek 2: Schemat analizowanej struktury periodycznej

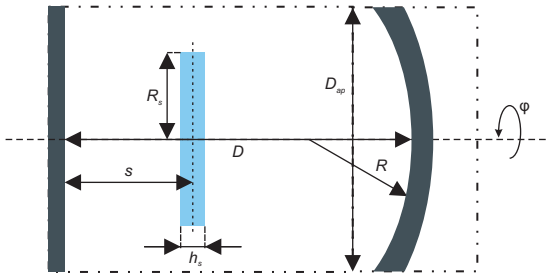
Wyniki zaprezentowane zostały na Rys. 3. Zgodnie z przewidywaniami pasmo zabronione ulega przemieszczeniu wraz ze zmianą okresu p (odległości między cylindrami).



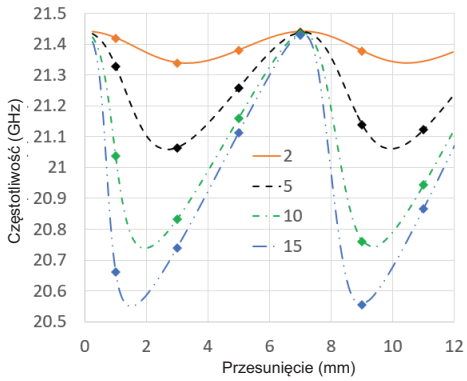
Rysunek 3: Znormalizowany współczynnik propagacji w funkcji częstotliwości dla cylindrów umieszczonych z okresem a) $p = 40$ mm oraz b) $p = 60$ mm

3.2 Otwarty rezonator Fabry-Perot

Drugim prezentowanym przykładem jest rezonator Fabry-Perot. Ze względu na swoje duże wymiary w stosunku do długości fali oraz szeroki zakres częstotliwości, w którym jest analizowany, jest on strukturą wymagającą w kontekście obliczeń numerycznych metodami dyskretnymi. Rezonator podzielony został na następujące części: obszar pomiędzy zwierciadłami wraz z próbką i zwierciadłem płaskim, który opisany został analitycznie, oraz zwierciadło, które zamodelowane zostało z wykorzystaniem metody elementów skończonych. Wymiary analizowanego rezonatora są następujące: $D = 100$ mm, $R = 150$ mm i $D_{ap} = 180$ mm - patrz Rys. 4. Wewnątrz umieszczona została próbka o grubości $h_s = 0.5$ mm i promieniu $R_s = D_{ap}/2$. Wyniki zaprezentowane zostały w [5] i na Rys. 5 oraz porównane z wynikami otrzymanymi metodą SMM [6].



Rysunek 4: Schemat otwartego rezonatora Fabry-Perot obciążonego dyskiem dielektrycznym



Rysunek 5: Częstotliwość rezonansowa rezonatora Fabry-Perot obciążonego dielektrycznym dyskiem w funkcji przesunięcia próbki względem zwierciadła płaskiego

4. PODSUMOWANIE

W niniejszym artykule przedstawiona została idea połączenia opisu analitycznego i metody elementów skończonych poprzez uogólnioną macierz impedancji. Zaprezentowane zostały analizy dla dwóch typów struktur: filtrującej linii periodycznej oraz otwartego rezonatora Fabry-Perot. Wyniki porównane zostały z innymi metodami, a otrzymana zgodność potwierdza poprawność prezentowanego podejścia.

LITERATURA

[1] Warecka Małgorzata. „Techniki zwiększania efektywności metody elementów skończonych poprzez redukcję dziedziny obliczeniowej z wykorzystaniem własności geometrii struktur". (rozprawa doktorska. Politechnika Gdańska. 2023)

[2] Piotr Kowalczyk, Rafal Lech, Malgorzata Warecka, Adam Kusiek. 2019. „Electromagnetic plane wave scattering from a cylindrical object with an arbitrary cross section using a hybrid technique, Journal of Electromagnetic Waves and Applications". 33(2) : 178-192

[3] Warecka Małgorzata, Lech Rafał, Kowalczyk Piotr. 2020. „Hybrid Analysis of Structures Composed of Axially Symmetric Objects". IEEE Transactions on Microwave Theory and Techniques, 68(11) : 4528-4535.

[4] Warecka Małgorzata, Lech Rafał, Kowalczyk Piotr. 2019. „Efficient Finite Element Analysis of Axially

Symmetrical Waveguides and Waveguide Discontinuities". IEEE Transactions on Microwave Theory and Techniques, 67(11) : 4291-4297.

[5] Salski Barłomiej, Karpisz Tomasz, Warecka Małgorzata, Kowalczyk Piotr, Czekała Piotr, Kopyt Paweł. 2022. „Microwave Characterization of Dielectric Sheets in a Plano-Concave Fabry-Perot Open Resonator". IEEE Transactions on Microwave Theory and Techniques, 70(5) : 2732-2742.

[6] Salski Bartłomiej, Karpisz Tomasz, Kopyt Paweł, Krupka Jerzy. 2020. „Rigorous Scattering Matrix Analysis of a Fabry–Perot Open Resonator". IEEE Transactions on Microwave Theory and Techniques, 68(12) : 5093-5102

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

SPÓJNA CZASOWO OBIEKTYWNA MIARA JAKOŚCI DLA WIZJI WSZECHOGARNIAJĄCEJ

TEMPORALLY-CONSISTENT OBJECTIVE QUALITY METRIC FOR IMMERSIVE VIDEO

Jakub Stankowski, Bartosz Sojka, Tomasz Grajek, Adrian Dziembowski

Instytut Telekomunikacji Multimedialnej, Politechnika Poznańska, Poznań
{jakub.stankowski, tomasz.grajek, adrian.dziembowski}@put.poznan.pl

DOI: 10.15199/59.2024.4.69

Streszczenie: **W niniejszym artykule przedstawiono efektywne rozwinięcie metryki IV-PSNR, służącej do obiektywnego pomiaru jakości dla wizji wszechogarniającej. Zaproponowany algorytm – IV-PSNR$_{YUVM}$ – uwzględnia analizę ruchu w dziedzinie czasu. Analiza ta możliwa jest dzięki dodaniu wektorów ruchu jako czwartej składowej dla każdego piksela obrazu. Przeprowadzone eksperymenty pokazują, że zaproponowana modyfikacja znacząco zwiększa korelację między obiektywną i subiektywną jakością wizji wszechogarniającej.**

Abstract: **In this paper, an efficient enhancement to the IV-PSNR metric is introduced. The proposal – IV-PSNR$_{YUVM}$ – incorporates a motion analysis in the temporal domain by adding motion vectors as a fourth component to video (Y, CB, and CR). Performed experiments demonstrate that this modification significantly improves the correlation between objective and subjective quality assessments in immersive video.**

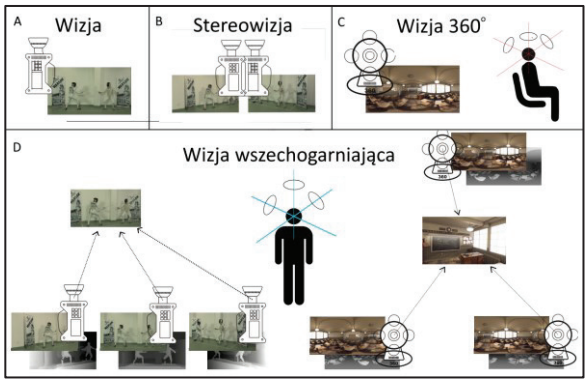
Słowa kluczowe: **wizja wszechogarniająca, ocena jakości wizji, obiektywny pomiar jakości.**

Keywords: **immersive video, video quality evaluation, objective quality assessment.**

1. WSTĘP

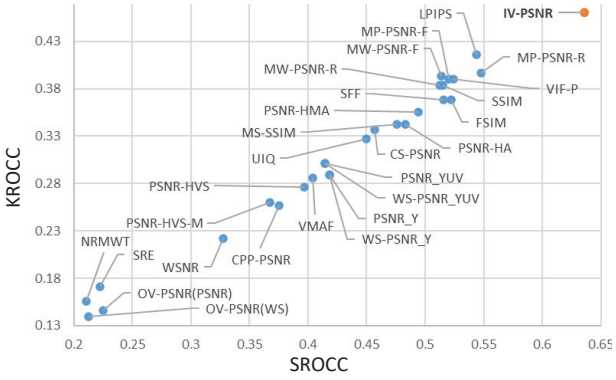
Wizja wszechogarniająca, zwana także immersyjną, to taka, która pozwala „zanurzyć się" widzowi w prezentowanym świecie. To zanurzenie może dotyczyć treści zarówno naturalnych (zarejestrowanych odpowiednimi zestawami kamer), jak i wygenerowanych komputerowo [3], [5]. Ciągle prowadzone są prace nad wieloma technologiami związanymi z wizją wszechogarniającą, z których wszystkie prezentują różne poziomy/stopnie immersji (rys. 1). Niektóre są bardzo ograniczone, np. pozwalają użytkownikowi jedynie zmieniać kierunek widzenia [9], podczas gdy inne są dość zaawansowane, np. pozwalają użytkownikowi swobodnie się poruszać [12].

Pomimo iż systemy wizji wszechogarniającej (ang. *immersive video*) [1], [15] są o wiele bardziej zaawansowane niż tradycyjne systemy wizji dwuwymiarowej, ich wspólny cel pozostaje niezmieniony: zapewnienie jak najlepszej jakości wrażeń (ang. QoE, *Quality of Experience*) poprzez maksymalizację subiektywnej satysfakcji użytkowników. Niestety, przeprowadzenie kompleksowych testów subiektywnych dla oceny jakości jest pracochłonnym i czasochłonnym przedsięwzięciem [4], co czyni je wysoce niepraktycznymi. Dużo praktyczniejszym rozwiązaniem jest ocena jakości w sposób obiektywny.



Rys. 1. Rodzaje systemów wizyjnych. Rysunek z [5].

Dziedzina obiektywnego pomiaru jakości jest obszerna, szczególnie w obrębie przetwarzania obrazów i wizji. W wielu scenariuszach, takich jak tradycyjna wizja dwuwymiarowa, najnowsze metryki skutecznie symulują subiektywne postrzeganie jakości wizji, np. PSNR, SSIM [14] lub VMAF [11]. Jednakże, nietypowa charakterystyka wizji wszechogarniającej, obejmująca przerzutowanie danych zarejestrowanych przez kilka kamer [12] czyni ją znacznie trudniejszą w ocenie, a tradycyjne metryki oceny jakości obiektywnej nie radzą sobie w jej kontekście wystarczająco skutecznie.

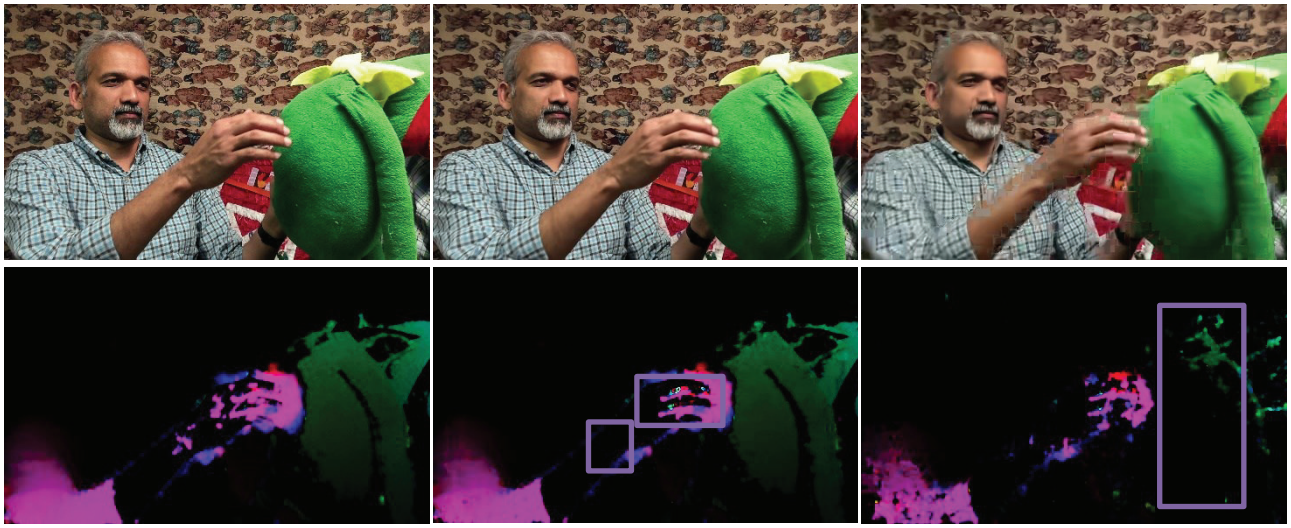


Rys. 2. Korelacja (mierzona z użyciem metryk SROCC i KROCC [10]) pomiędzy subiektywną oceną jakości a jakością mierzoną obiektywnie, z użyciem metryk opisanych w literaturze. Rysunek z [4].

(A) Widok wejściowy

(B) Widok stworzony (wyrenderowany) na podstawie nieskompresowanych widoków wejściowych

(C) Widok stworzony (wyrenderowany) na podstawie silnie skompresowanych widoków wejściowych



Rys. 3. Przykładowe pole ruchu wyznaczone dla sekwencji oryginalnej (A) oraz zniekształconej (B i C).

Aby uwzględnić zniekształcenia typowe dla wizji immersyjnej, wprowadzono metrykę IV-PSNR [4]. Metryka IV-PSNR jest koncepcyjnie oparta na PSNR (*Peak signal-to-noise ratio*) ale zawiera znaczne modyfikacje pozwalające na uwzględnienie zniekształceń występujących w obrazie powstałym w wyniku przerzutowania danych z kamer źródłowych w pozycję kamery wirtualnej. Najistotniejsze modyfikacje to dodanie narzędzia uwzględniającego globalną różnicę między komponentami (GCR – Global Component Difference) oraz mechanizmu kompensującego przesunięcie odpowiadających sobie pikseli (CPS – Corresponding Pixel Shift).

Jak przedstawiono na rys. 2, podstawowa wersja metryki IV-PSNR wykazuje wyższą korelację z oceną jakości subiektywnej niż inne metryki. Jednakże nie wykorzystuje ona analizy wizji w dziedzinie czasu. W niniejszym artykule zaproponowaliśmy uwzględnienie w obliczeniach metryki IV-PSNR analizy podobieństwa między charakterystykami czasowymi dwóch porównywanych sekwencji wizyjnych.

2. ZAPROPONOWANA METRYKA JAKOŚCI

Zaproponowane podejście opiera się na rozszerzeniu metryki IV-PSNR o informacje dotyczące zmian w charakterystyce czasowej mierzonej sekwencji. Najbardziej powszechne metryki jakości stosowane w aplikacjach wizji wszechogarniającej (PSNR, SSIM, IV-PSNR oraz WS-PSNR [13]) działają na pojedynczym obrazie, a jakość sekwencji obliczana jest przez proste uśrednienie jakości obliczonej dla poszczególnych klatek sekwencji wizyjnej. Niestety, takie podejście sprawia, że wspomniane metryki są niewrażliwe na pogorszenie spójności ruchu. Niektóre artefakty związane z czasem, takie jak migotanie krawędzi obiektów, są bardzo irytujące dla widza i powinny być uwzględnione w mierzonej jakości obiektywnej.

Zaproponowane podejście obejmuje wykorzystanie analizy pola ruchu [2] do oceny spójności ruchu sekwencji. Pole ruchu jest określane za pomocą gęstego przepływu optycznego (ang. DOF, *dense optical flow*) z wykorzystaniem opisanego w [6] algorytmu. Obliczenia przepływu optycznego są wykonywane dla obu sekwencji (referencyjnej i ocenianej), a ich wynik (w postaci wektorów ruchu) traktowany jest jako czwarty składnik obrazu. Wizualizacja różnicy pomiędzy polami ruchu obliczonymi dla obrazu oryginalnego i zniekształconego jest przedstawiona na rysunku 3.

Wszystkie rozważane metryki opierają się na PSNR, dlatego najważniejszym etapem wyznaczania każdej z nich jest obliczenie sumy kwadratów różnic (ang. SSD, *sum of squared differences*).

$$PSNR_c = 10\log_{10}\frac{W\cdot H\cdot\left(2^b-1\right)^2}{SSD_c}, \qquad (1)$$

gdzie: $PSNR_c$ – wartość miary PSNR dla składowej c; W, H – wysokość i szerokość obrazu; b – głębia bitowa; SSD_c – wartość SSD dla składowej c.

Obliczenie wspomnianych kwadratów różnic (SD) dla składników obrazu jest proste (ponieważ wartości składników koloru są skalarami) i dla pojedynczego punktu danej składowej barwnej wygląda następująco:

$$SD_c(y,x) = \left(I_c(y,x) - J_c(y,x)\right)^2, \qquad (2)$$

gdzie: $SD_c(y,x)$ – wartość SD dla punktu o współrzędnych (y,x) należącego do składowej c; $I_c(y,x)$ – próbka o współrzędnych (y,x) należącego do składowej c obrazu referencyjnego (I); $J_c(y,x)$ – próbka o współrzędnych (y,x) należącego do składowej c obrazu ocenianego (J).

W przypadku miary PSNR suma kwadratów różnic (SSD_c) jest obliczana przez wyznaczenie wartości $SD_c(y,x)$ dla całej składowej barwnej a następnie ich zsumowanie.

Natomiast dla wektorów ruchu SD jest obliczane jako kwadrat długości wektora powstałego przez obliczenie różnicy pomiędzy wektorem ruchu w obrazie referencyjnym ($\vec{V_l}$) a wektorem ruchu w obrazie ocenianym ($\vec{V_l}$):

$$SD_M(y,x) = \left|\overrightarrow{V_I(y,x)} - \overrightarrow{V_J(y,x)}\right|^2. \qquad (3)$$

Obrazy wejściowe mogą cechować się różnymi wartościami głębi bitowej. Obecnie najczęściej spotykanymi wartościami są 8 lub 10 bitów na próbkę. W celu zachowania spójności wyznaczanych metryk i uniezależnienia ich od głębi bitowej zastosowano korekcję wartości $SD_M(y,x)$ zależną od głębi bitowej. Jako wartość bazową przyjęto 10 bitów na próbkę, zaś sama operacja korekty została wyrażona w następujący sposób:

$$\grave{SD_M}(y,x) = SD_M(y,x) \cdot 2^{(b-10)}. \qquad (4)$$

gdzie: $\grave{SD_M}(y,x)$ – skorygowana wartość SD dla pola ruchu; b – głębia bitowa.

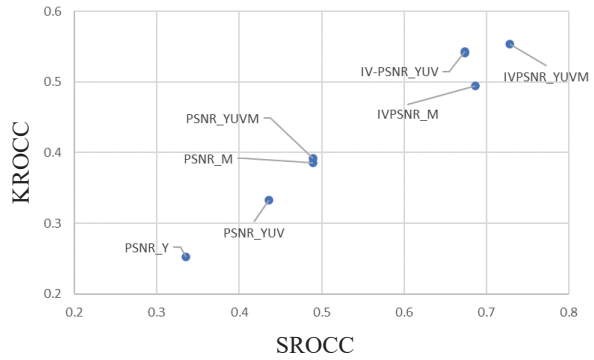
3. EKSPERYMENT

Zaproponowane podejście zostało ocenione na podstawie wyników „MPEG Call for Proposals on 3DoF+ Visual" [8], zawierających sekwencje wielowidokowe zakodowane przy użyciu siedmiu różnych technik kodowania wizji wszechogarniającej. Szczegółowa metodologia eksperymentu jest opisana w rozdziale VII artykułu [4], ponieważ została wcześniej wykorzystana do oceny podstawowej metryki IV-PSNR. Ze względu na dużą złożoność obliczeniową związaną z wyznaczaniem gęstego przepływu optycznego oraz dostępność sekwencji testowych, ocena została przeprowadzona na jednej grupie obrazów (ang. GOP, *group of pictures*). Propozycja została porównana z niezmienionym IV-PSNR, ponieważ zapewnia on najwyższą korelację z jakością subiektywną w aplikacjach wizji wszechogarniającej (rys. 2). Ponadto zaprezentowaliśmy wyniki dla zmodyfikowanego i niezmodyfikowanego PSNR, pokazując wszechstronność proponowanej modyfikacji (wykorzystanie informacji o ruchu). W eksperymencie metryki porównywano za pomocą dwóch powszechnie znanych i stosowanych współczynników korelacji: SROCC (Spearman rank correlation coefficient) i KROCC (Kendall rank correlation coefficient) [10]. Łącznie porównano siedem wariantów metryk:

- trzy metryki znane z literatury: **PSNR$_Y$** (PSNR wyłącznie dla składowej luminancji), **PSNR$_{YUV}$** (średnia ważona PSNR dla trzech składowych barwnych) i **IV-PSNR$_{YUV}$** (podstawowa wersja IV-PSNR, opisana w [4], uwzględniająca trzy składowe barwne),
- PSNR i IV-PSNR obliczone na sekwencjach pola ruchu, nieuwzględniające tekstury (ostatni rząd rys. 3): **PSNR$_M$** i **IV-PSNR$_M$**,
- Czterokomponentowe warianty miar PSNR i IV-PSNR, gdzie wektory ruchu traktowane są jako czwarty komponent (M), wraz z luminancją (Y) i obiema chrominancjami (Cb, Cr): **PSNR$_{YUVM}$** i **IV-PSNR$_{YUVM}$**.

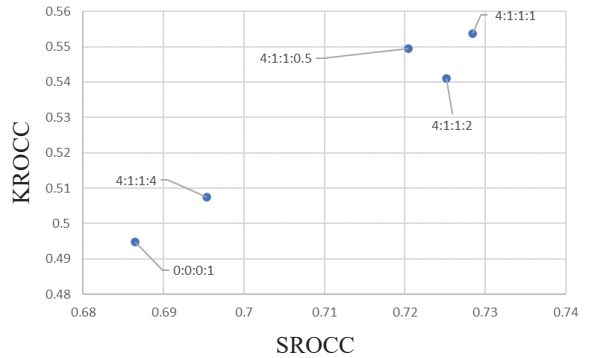
Wartość PNSR$_{YUV}$ wyznaczana była poprzez uśrednienie wartości PSNR dla luminancji i obu chrominancji, z wagami – odpowiednio – 6:1:1 (zgodnie z [7]). W przypadku IV-PSNR$_{YUV}$ wagi te zostały ustalone na 4:1:1 (zgodnie z [4]).

Korelacja między jakością subiektywną (obliczaną jako MOS, ang. *mean opinion score*) a wszystkimi testowanymi wariantami metryk jest przedstawiona na rys. 4. Jak pokazano, czteroskładnikowy IV-PSNR$_{YUVM}$ osiągnął najwyższą korelację z jakością subiektywną, zarówno pod względem SROCC, jak i KROCC. Użycie tylko wektorów ruchu (IV-PSNR$_M$) jest mniej skuteczne i charakteryzuje się podobną efektywnością do IV-PSNR obliczonego dla trzech składowych koloru (IV-PSNR$_{YUV}$).



Rys. 4. Korelacja pomiędzy subiektywną oceną jakości a jakością mierzoną obiektywnie; porównanie testowanych wariantów metryk PSNR i IV-PSNR.

Jak przedstawiono, dodanie wektorów ruchu jako czwartej składowej jest efektywne również dla metryki PSNR. W przypadku PSNR podobna korelacja z MOS może być osiągnięta poprzez użycie tylko wektorów ruchu. Przyczyną takiego zjawiska jest obliczenie samego pola ruchu – ruch jest szacowany poprzez wyszukiwanie podobnych bloków o rozmiarze 11×11 w kolejnych klatkach sekwencji, a dla wizji wszechogarniającej analiza oparta na blokach jest bardziej efektywna niż oparta na pikselach (por. PSNR vs. IV-PSNR).



Rys. 5. Korelacja pomiędzy subiektywną oceną jakości a jakością mierzoną obiektywnie; porównanie pięciu wariantów metryki IV-PSNR$_{YUVM}$.

Na rys. 5 przedstawiono porównanie pięciu wariantów IV-PSNR$_{YUVM}$. Warianty różnią się wagami przyporządkowanymi do poszczególnych składowych (Y:Cb:Cr:M). Cztery z nich (4:1:1:4, 4:1:1:2, 4:1:1:1 i 4:1:1:0.5) używają domyślnych wag dla luminancji i obu

chrominancji i różnią się w zestawie wag dla wektorów ruchu. Wariant 0:0:0:1 używa tylko wektorów ruchu, co odpowiada wartości IV-PSNR$_M$. Jak przedstawiono, najbardziej efektywny schemat ważenia równoważy wagi między wektorami ruchu a obiema chrominancjami. Podsumowując wyniki przedstawionych eksperymentów, stwierdzić można, iż wagi przyporządkowane do poszczególnych składowych Y:Cb:Cr:M równe 4:1:1:1 dają najwyższą korelację z ocenami subiektywnymi.

4. PODSUMOWANIE

W artykule przedstawiono efektywną modyfikację metryki IV-PSNR [4], która obejmuje analizę porównywanych sekwencji w dziedzinie czasu poprzez pomiar spójności ruchu, tj. podobieństwo gęstego pola ruchu w sekwencji referencyjnej i testowanej.

Propozycja opiera się na dodaniu wektorów ruchu jako czwartej składowej, uzupełniając trzy składowe barwne: Y, C_B i C_R. Wyniki eksperymentów pokazały, że proponowana modyfikacja IV-PSNR pozwala na zwiększenie korelacji między jakością obiektywną a subiektywną w ocenie jakości wizji wszechogarniającej. Biorąc pod uwagę, że propozycja poprawia metrykę IV-PSNR, a podstawowa metryka IV-PSNR przewyższa inne najnowsze metryki oceny jakości obiektywnej, można stwierdzić, że zaproponowana metryka jest wartościowa i bardzo praktyczna do zastosowania w rzeczywistych systemach wizji wszechogarniającej.

PODZIĘKOWANIA

Praca finansowana ze środków przyznanych przez Ministerstwo Nauki i Szkolnictwa Wyższego.

LITERATURA

[1] Boyce Jill, Dore Renaud, Dziembowski Adrian, Fleureau Julien, Jung Joel, Kroon Bart, Salahieh Basel, Kumar Malamal Vadakital Vinod, Yu Lu. 2021. "MPEG Immersive Video coding standard". *Proceedings of the IEEE*, 119 (9): 1521-1536.

[2] Brox Thomas, Malik Jitendra. 2011. Large Displacement Optical Flow: "Descriptor Matching in Variational Motion Estimation". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33 (3): 500-513.

[3] Domański Marek, Stankiewicz Olgierd, Wegner Krzysztof, Grajek Tomasz. 2017. "Immersive visual media — MPEG-I: 360 video, virtual navigation and beyond". *24th International Conference on Systems, Signals and Image Processing*, Poznań, Polska.

[4] Dziembowski Adrian, Mieloch Dawid, Stankowski Jakub, Grzelka Adam. 2022. "IV-PSNR – The objective quality metric for immersive video applications". *IEEE Transactions on Circuits and Systems for Video Technology*, 32 (11): 7575-7591.

[5] Dziembowski Adrian, Mieloch Dawid. 2022. "Nowe techniki kompresji wizji dla rzeczywistości wirtualnej – MPEG Immersive Video". *Przegląd Telekomunikacyjny - Wiadomości Telekomunikacyjne.* 4: 116-123.

[6] Farnebäck Gunnar. 2003. "Two-frame motion estimation based on polynomial expansion". *Scandinavian conference on Image analysis*, Berlin, Niemcy.

[7] Huang Yue, Qi Honggang, Li Bin, Xu Jizheng. 2014. "Adaptive weighted distortion optimization for video coding in RGB color space". *IEEE International Conference on Image Processing, ICIP*, Paryż, Francja.

[8] ISO/IEC. 2019. "Call for Proposals on 3DoF+ Visual". *ISO/IEC JTC1/SC29/WG11 MPEG N18145*, Marakesz, Maroko.

[9] Liu Tsu-Ming, Ju Chi-Cheng, Huang Yu-Hao, Chang Tsui-Shan, Yang Kai-Min, Lin Yi-Ting. 2016. "A 360-degree 4K×2K panoramic video recording over smart-phones". *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, Seattle, USA.

[10] Myers Jerome, Well Arnold. 2003. "Research design and statistical analysis". *Lawrence Erlbaum Associates*, Londyn, UK.

[11] The Netflix Tech Blog. 2016. Dostęp online: https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652

[12] Stankiewicz Olgierd, Domański Marek, Dziembowski Adrian, Grzelka Adam, Mieloch Dawid, Samelak Jarosław. 2018. "A free-viewpoint television system for horizontal virtual navigation". *IEEE Transactions on Multimedia*, 20 (8): 2182-2195.

[13] Sun Yule, Lu Ang, Yu Lu. 2017. "Weighted-to-Spherically-Uniform Quality Evaluation for Omnidirectional Video". *IEEE Signal Processing Letters*, 24 (9): 1408-1412.

[14] Wang Zhou, Bovik Alan, Sheikh Hamid, Simoncelli Eero. 2004. "Image quality assessment: From error measurement to structural similarity". *IEEE Transactions on Image Processing*, 13 (4): 600-612.

[15] Wien Mathias, Boyce Jill, Stockhammer Thomas, Peng Wen-Hsiao. 2019. "Standardization status of immersive video coding". *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 9 (9): 5-17.

CZY W BADANIACH QoE POWINNO SIĘ OGRANICZAĆ DOSTĘP TESTEROM W ZŁYM STANIE PSYCHOFIZYCZNYM?

SHOULD WE RESTRICT AN ACCESS TO QoE EXPERIMENTS FOR TESTERS IN BAD PSYCHOPHYSICAL CONDITION?

Tomasz Konaszyński[1]; Avrajyoti Dutta[2]; Dawid Juszka[3]; Mikołaj Leszczuk[4]

[1] AGH Akademia Górniczo-Hutnicza, Kraków, Tomasz.Konaszynski@agh.edu.pl
[2] AGH Akademia Górniczo-Hutnicza, Kraków, Avrajyoti.Dutta@agh.edu.pl
[3] AGH Akademia Górniczo-Hutnicza, Kraków, Dawid.Juszka@agh.edu.pl
[4] AGH Akademia Górniczo-Hutnicza, Kraków, Mikolaj.Leszczuk@agh.edu.pl

DOI: 10.15199/59.2024.4.70

Streszczenie: **W przeprowadzonym eksperymencie badawczym w obszarze Quality of Experience dla wideo 2D został przeanalizowany aspekt wpływu stanu psychofizycznego (zmęczenia, złego nastroju) uczestników eksperymentu na uzyskane wyniki. Wykazano związek między negatywnym stanem psychofizycznym testera a jakością i adekwatnością uzyskanych wyników badań, co stawia pod znakiem zapytania zasadność udziału testerów w złym stanie psychofizycznym w eksperymentach QoE o ograniczonym budżecie i dostępności testerów.**

Abstract: **In the conducted experiment, in Quality of Experience for 2D videos, an aspect of the influence of psychophysical conditions (i.e. fatigue, bad mood) of experiment participants on the obtained subjective results was analyzed. A relationship was demonstrated between the tester's negative psychophysical condition and the quality and adequacy of the obtained test results, raising the question of why the participation of testers in bad psychophysical conditions in QoE experiments with a limited budget and availability of testers.**

Słowa kluczowe: **QoE, ocena subiektywna, forma psychofizyczna testera, nastrój i poziom zmęczenia, mikro-struktura treści.**

Keywords: **QoE, multimedia subjective assessment psychophysical tester's condition, mood and tiredness, content micro-structure.**

1. WSTĘP

Przedmiot badań wpisuje się w dziedzinę Quality of Experience [4][9], tj. badanie stopnia zadowolenia lub irytacji użytkownika aplikacji lub usługi, wynikającego ze spełnienia jego oczekiwań co do użyteczności i/lub przyjemności korzystania w świetle jego osobowości i aktualnego stanu. Ocena subiektywna wynika z szeregu czynników wpływających na percepcję oglądającego, jak: aspekt technologiczny, społeczny, środowiskowy, ludzki czy charakterystyka prezentowanej treści [11].

W publikacji zaprezentowano rezultaty eksperymentu, dotyczącego wpływu na ocenę subiektywną jakości wideo 2D **czynnika stanu psychofizycznego testerów, deklarowanego przy rozpoczęciu procesu badawczego**. Literatura opisuje aspekt zmęczenia testerów podczas eksperymentu w wyniku np. niskiej jakości mediów [3], czy opóźnień transmisyjnych i zakłóceń w sieci [12]. Wartym analizy zagadnieniem jest jednak stan psychofizyczny w jakim tester podchodzi do eksperymentu. Zagadnienie to jest szczególnie istotne w przypadku badań przy ograniczonym budżecie i niskiej dostępności testerów, gdzie pozyskanie ocen wysokiej jakości jest kluczowe dla wartości poznawczej eksperymentu.

2. ZAGADNIENIE BADAWCZE

Postawiono pytanie badawcze: **Czy na ocenę subiektywną wideo 2D może mieć wpływ zadeklarowany zły stan psychofizyczny testera?** Określono również dwie powiązane hipotezy badawcze:

- **H1:** Deklarowany zły stan psychofizyczny testera ma wpływ na kwalifikowalność uzyskanych wyników do dalszej analizy
- **H2:** Deklarowany zły stan psychofizyczny testera ma wpływ na ocenę subiektywną wyników zakwalifikowanych do dalszej analizy

Odpowiedź na postawione pytanie badawcze jest ważna z punktu widzenia jakości badań w wielu dziedzinach nauki, wykraczając poza obszar QoE, gdyż odchylenia wyników wynikające ze złego stanu psychofizycznego testerów skutkują odrzuceniem ich istotnej części i „marnotrawstwem" ograniczonych zasobów oraz pogorszeniem jakości wyników.

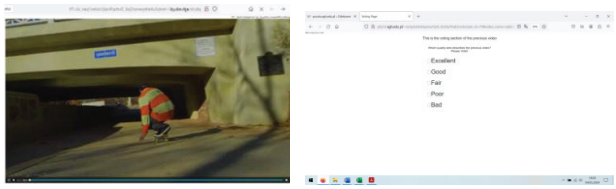
3. OPIS EKSPERYMENTU

Wstępne założenie badawcze. W eksperymencie przyjęto wstępne założenie badawcze, w którym stan psychofizyczny testera jest definiowany przez dwa komponenty, tj. deklarowany poziom zmęczenia oraz subiektywnie odczuwany nastrój. Zgodnie z Tab. 1 uzyskane wyniki zostały przypisane do jednego z dwóch obszarów analizy, odpowiadających dobremu (kolor zielony) lub złemu (kolor czerwony) stanowi psychofizycznemu testera.

Tab. 1. Obszary stanu psychofizycznego testerów

		Poziom zmęczenia		
		Niski	Średni	Wysoki
Nastrój	Pozytywny	dobry	dobry	zły
	Neutralny	dobry	dobry	zły
	Negatywny	zły	zły	zły

Interfejs do zbierania wyników. Eksperyment został przeprowadzony w laboratorium komputerowym Instytutu Telekomunikacji AGH, z użyciem autorskiej platformy badawczej w postaci aplikacji internetowej udostępnionej na serwerze uczelnianym. Po uruchomieniu aplikacji każdy tester został poproszony o wypełnienie formularza, zawierającego m.in. deklarację poziomu zmęczenia (wysoki, średni, niski) oraz nastroju (pozytywny, neutralny, negatywny). Po wypełnieniu formularza naprzemiennie wyświetlano film o czasie trwania 10 s, po czym pojawiał się ekran oceny, jak na rys. 1, aż do wyczerpania zbioru dostępnych filmów.



Rys. 1. Zrzut ekranu aplikacji do zbierania wyników subiektywnych.

Projekt eksperymentu. Eksperyment odwołuje się do wcześniej przeprowadzonych badań [7], przez użycie infrastruktury badawczej oraz identycznego zestawu danych. Eksperyment został przeprowadzony zgodnie z zaleceniami Rekomendacji ITU-T P.913 [6] oraz ITU-T P. 800 [5] dla badań QoE, w celu pozyskania ocen w 5-stopniowej skali Absolute Category Rating (ACR), tj. (5 – excellent, 4 – good, 3 – fair, 2 – poor, 1 – bad), dla zbioru sekwencji PVS (ang. *Processed Video Sequences),* wyświetlanych w odpowiednim porządku i krotności.

Testerów podzielono na dwie grupy: Grupę 1, w której każdy tester oceniał 170 sekwencji PVS jeden raz oraz Grupę 2 w której każdy tester oceniał 60 sekwencji PVS po trzy razy, tj. łącznie 180 filmów. W rezultacie otrzymano zbiór ocen subiektywnych z dodatkowymi danymi, jak m.in. poziom jakości obiektywnej filmu, deklarowany nastrój oraz poziom zmęczenia testera.

Dane badawcze. W eksperymencie użyto zestawu 34 unikalnych sekwencji wideo SRC (ang. *Source Reference Circuit*) FullHD pozyskanych z baz Netflix, CableLabs, SJTU Media Lab oraz Xiph.org Video Test Media. Zbiór sekwencji PVS powstał poprzez przetworzenie każdego SRC i zdegradowanie do 5 poziomów jakości obiektywnej, mierzonej przy użyciu metryki jakości wideo VMAF (ang. *Video Multimethod Assessment Fusion*) [10]. Dla każdej sekwencji SRC wygenerowano pliki w standardzie YUV4MPEG2, przy poziomach przepływności w przedziale pomiędzy 100000 b/s a 21000000 b/s.

Testerzy. W eksperymencie udział wzięło 47 testerów, w tym 5 kobiet oraz 42 mężczyzn, podzielonych na dwie grupy w taki sposób, że Grupa 1 liczyła 24 osoby, zaś Grupa 2 – 23 testerów.

4. ANALIZA REZULTATÓW

Po przeprowadzeniu eksperymentu dobrano metody weryfikujące uzyskane wyniki i przeprowadzono analizę uzyskanych rezultatów w odniesieniu do pytania badawczego i postawionych hipotez H1 i H2.

4.1. Weryfikacja hipotezy H1.

Weryfikacja kwalifikowalności uzyskanych wyników została przeprowadzona względem poszczególnych testerów, przy użyciu metody walidacji obserwacji odstających opisanej w [6].

Metoda weryfikacji kwalifikowalności wyników.

Wybrana metoda bada korelację średnich wyników MOS testera z odpowiednimi średnimi wynikami MOS wszystkich testerów przy użyciu współczynnika korelacji liniowej Pearsona (LCC), obliczanego jako:

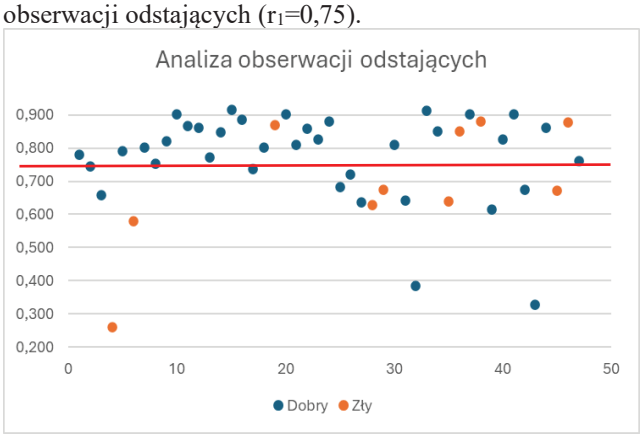
$$\mathrm{LCC}(x,y)=\frac{\sum_{i=1}^{n} x_i y_i-\frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2-\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)\left(\sum_{i=1}^{n} y_i^2-\frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right)}} \quad (1)$$

, gdzie x i y to tablice danych, a n to liczba punktów danych, x_i: MOS wszystkich testerów dla PVS, y_i: pojedynczy wynik oceny testera dla PVS, n: całkowita liczba PVS, i: Numer kolejny PVS

Aby obliczyć LCC dla poszczególnych bodźców badawczych (PVS), oblicza się wartość $r_l(x,y)=\mathrm{LCC}(x,y)$. Testerzy są odrzucani, jeśli wskaźnik r_l spadnie poniżej ustalonego progu. W przypadku testów ACR dla materiałów wideo zalecany próg odrzucenia wynosi ($r_l < 0{,}75$).

Otrzymane wyniki.

W pierwszym przebiegu procedury uzyskano wartości r_l przedstawione na rys. 2. Kolorem niebieskim zaznaczono testerów w deklarowanym dobrym, zaś kolorem pomarańczowym - złym stanie psychofizycznym. Czerwona linia oznacza próg odcięcia obserwacji odstających (r_l=0,75).



Rys. 2. Wartości współczynnika r_l w pierwszym przebiegu procedury.

W dalszych przebiegach procedury ostatecznie odrzucono oceny 6 testerów w Grupie 1 oraz 8 testerów w Grupie 2. Do dalszej analizy dopuszczono oceny 18 testerów w Grupie 1 oraz 15 testerów w Grupie 2. Tab. 2 przedstawia rozkład obserwacji odstających w zależności od deklarowanego stanu psychofizycznego testerów.

Tab. 2. Rozkład kwalifikacji testerów dla obszarów analitycznych dobrego i złego stanu psychofizycznego

	Zakwalifikowani	Odrzuceni
Dobry stan PF	29	8
Zły stan PF	4	6

W Tab. 3 porównano oceny MOS testerów zakwalifikowanych do dalszej analizy oraz testerów w złym stanie psychofizycznym, których oceny zostały odrzucone.

Tab. 3. Zależność ocen MOS dla grupy testerów zakwalifikowanych, w zestawieniu z wynikami odrzuconych testerów w złym stanie psychofizycznym.

Grupa 1	Zakwalifikowane	Outliers - zły stan PF
QG A	4,30	3,95
QG B	3,52	3,42
QG C	2,68	3,19
QG D	2,02	2,69
QG E	1,39	2,41
Grupa 2	Zakwalifikowane	Outliers - zły stan PF
QG A	4,24	3,95
QG B	3,62	3,72
QG C	2,91	3,11
QG D	2,19	2,60
QG E	1,43	2,03

Przeprowadzona analiza statystyczna z użyciem testu U Manna-Whitneya nie potwierdziła jednak istotności statystycznej otrzymanych wyników w żadnej z grup jakościowych QG A/QG B/QG C/QG D/QG E ($0{,}22 \leq p \leq .0{,}97$).

Interpretacja wyników.

Przedstawione wyniki wskazują na istotny wpływ złego stanu psychofizycznego testerów na kwalifikowalność wyników subiektywnych. Wśród testerów o deklarowanej dobrej formie psychofizycznej odsetek wyników odrzuconych wynosi 0,22, podczas gdy w grupie o złej formie psychofizycznej wynosi 0,6. Oznacza to, że w tej grupie, w sytuacji ograniczonej dostępności testerów lub ograniczonego budżetu, 60% zaangażowanych i finansowanych zasobów nie przyniosło kwalifikowalnego wyniku badawczego.

Pomimo wyraźnie zarysowanej w grupie testerów o deklarowanym złym stanie psychofizycznym tendencji „uśredniania" wyników MOS w porównaniu z wynikami testerów zakwalifikowanych do dalszej analizy, nie uzyskano istotności statystycznej wyników. Oznacza to, że jakość danych w grupie odrzuconych testerów w złym stanie psychofizycznym jest na niskim poziomie. Uzyskane wyniki i zaangażowane zasoby w tej grupie z perspektywy całego eksperymentu są zatem zmarnowane.

Ponadto, wszystkie obserwacje odrzucone w badaniu dla grupy testerów z deklarowanym złym stanem psychofizycznym dotyczą testerów, którzy zadeklarowali wysoki poziom zmęczenia, podczas gdy w przypadku testerów w nastroju negatywnym wyniki odrzucone nie pojawiły się gdy poziom zmęczenia testera był inny niż wysoki. Obserwacja ta wskazuje na dominujący wpływ czynnika wysokiego zmęczenia a nie złego nastroju na wysoki odsetek obserwacji odstających.

4.2. Weryfikacja hipotezy H2.

Weryfikacja hipotezy H2 została przeprowadzona poprzez analizę statystyczną otrzymanych wyników z użyciem testu nieparametrycznego U Manna-Whitneya dla prób niezależnych, porównując różnice między ocenami MOS uzyskanymi dla testerów o deklarowanym dobrym oraz złym stanie psychofizycznym.

Otrzymane wyniki.

Po odrzuceniu części uzyskanych wyników w rezultacie analizy obserwacji odstających porównano wartości MOS w każdej z grup jakości obiektywnej dla testerów w deklarowanym dobrym i złym stanie psychofizycznym. Wyniki przedstawiono w tab. 4.

Tab.4. Porównanie MOS dla testerów o deklarowanym dobrym i złym stanie psychofizycznym

Grupa 1	Dobry stan PF	Zły stan PF
QG A	4,31	4,47
QG B	3,52	3,89
QG C	2,68	3,31
QG D	2,02	2,47
QG E	1,39	1,70
Grupa 2	Dobry stan PF	Zły stan PF
QG A	4,24	4,31
QG B	3,62	3,76
QG C	2,91	3,03
QG D	2,19	2,44
QG E	1,43	1,53

Na poziomie istotności statystycznej α=0,05 dla wszystkich poziomów grup jakościowych w obydwu grupach 1 i 2, pomimo widocznych wyraźnych różnic w ocenach MOS nie wykazano istotności statystycznej różnic średniej oceny subiektywnej sekwencji MOS w obydwu grupach ($0{,}067 \leq p \leq .1$).

Interpretacja wyników.

Nieuzyskanie istotności statystycznej różnic pomiędzy grupami o dobrym i złym stanie psychofizycznym pomimo znaczących różnic wartości MOS wynika w znacznym stopniu z bardzo dużej różnoliczności tych grup. W wyniku analizy obserwacji odstających odrzucono 60% wyników uzyskanych w grupie testerów w złym deklarowanym stanie psychofizycznym, co obniżyło jej liczebność do poziomu nieumożliwiającego wyciągania rzetelnych wniosków.

Należy jednak podkreślić, że w każdej z grup jakościowych wartość MOS otrzymana dla grupy testerów w złej deklarowanej formie psychofizycznej jest

wyższa niż w grupie testerów w dobrej formie psychofizycznej, co może stanowić punkt wyjścia do dalszych badań, z użyciem większej kwalifikowalnej grupy testerów.

5. DYSKUSJA I DALSZE KIERUNKI BADAWCZE

Uzyskane wyniki wskazują na niepomijalny wpływ złej formy psychofizycznej testera na uzyskiwane wyniki badań subiektywnych, ich jakość i stabilność. Studia literaturowe szeroko opisują zjawisko wpływu przebiegu eksperymentu na stan psychofizyczny testera. Również w innych obszarach badawczych, jak np. badania haptyczne [1] obserwuje się prawidłowości wpływu przebiegu badania na uzyskane wyniki. Stan rozwoju nauki koncentruje się jednak na aspekcie zmiany stanu psychofizycznego testerów w trakcie badania, nie zaś na stanie w jakim tester do badania podchodzi. W kontekście efektywności badań QoE kluczowe jest jednak zagadnienie odwrotne, tj. aspekty stanu psychofizycznego, zmniejszające prawdopodobieństwo uzyskania wysokiej jakości wyników. W szczególności istotny staje się czynnik wysokiego poziomu zmęczenia, który w badaniach okazał się dominujący względem złego nastroju. Identyfikuje się szereg czynników o charakterze psychofizycznym, wpływających na QoE [13], co wskazuje, że wartościowym rozszerzeniem badań byłoby pogłębienie rozumienia aspektu zmęczenia testera.

Należy także podkreślić, że stan psychofizyczny testerów został określony w sposób deklaratywny, bez zobiektywizowanego badania poziomu zmęczenia testerów w eksperymencie. Kwalifikacja testerów do grup o dobrym i złym stanie psychofizycznym w niniejszym badaniu bazowała na przyjętym założeniu badawczym, nie wynika zaś z parametrów zobiektywizowanych, jak np. badań neurologicznych mózgu, korzystających z metody analizy elektrofizjologicznej oraz elektroencefalografii (EEG), czy badań drżenia kończyn, w obszarze badawczym dotyczącym jakości doświadczenia (QoE) [2].

Dalsze kierunki badawcze mogą zatem obejmować:

- Zaangażowanie zobiektywizowanych narzędzi weryfikacji stanu psychofizycznego testera
- Powtórzenie logiki badań, przy rozszerzonej liczebności grup badawczych
- Analizę zmian w czasie ocen, w zależności od parametrów prezentowanego wideo
- Badanie korelacji czynnika stanu psychofizycznego z innymi czynnikami, jak np. porządek prezentacji bodźców badawczych, czy pamięć testera [7]

Odpowiadając na pytanie postawione w tytule, z perspektywy badań z ograniczeniami co do dostępności zasobów i budżetu, zasadne jest pogłębienie badań głębokości i kierunków wpływu złego stanu psychofizycznego. W razie potwierdzenia obserwacji, odpowiedź na pytanie tytułowe wydaje się twierdząca.

Finansowanie: Praca została sfinansowana przez Ministerstwo Nauki i Szkolnictwa Wyższego w ramach subwencji dla Wydziału Informatyki, Elektroniki i Telekomunikacji AGH.

LITERATURA

[1] Alja'afreh, M., et al (2020, February). Towards a comprehensive study of fatigue deducing techniques for evaluating the quality of experience of haptic-visual applications. In *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)* (pp. 339-344). IEEE.

[2] Antons, J. N., Arndt, S., Schleicher, R., & Möller, S. (2014). Brain activity correlates of quality of experience. *Quality of Experience: Advanced Concepts, Applications and Methods*, 109-119.

[3] Arndt, S., Antons, J. N., & Möller, S. (2014, September). Is low quality media affecting the level of fatigue? In *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)* (pp. 47-48). IEEE.

[4] Brunnström, et al. 2013. Qualinet white paper on definitions of quality of experience.

[5] ITU-T. Recommendation ITU-T P.800: Methods for subjective determination of transmission quality, International Telecommunication Union, 1996

[6] 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.

[7] Konaszyński, T., Juszka, D., & Leszczuk, M. (2023). Impact of the Stimulus Presentation Structure on Subjective Video Quality Assessment. Electronics, 12(22), 4593.

[8] Konaszyński, T., Juszka, D., & Leszczuk, M. (2023). Analiza wpływu mikro-struktury prezentacji treści na subiektywną ocenę jakości wideo i jej stabilność. Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne.

[9] Raake, A., Egger, S. 2014. Quality and Quality of Experience. In: Möller, S., Raake, A. (eds) Quality of Experience. T-Labs Series in Telecommunication Services. Springer, Cham.

[10] Rassool R. 2017. "VMAF reproducibility: Validating a perceptual practical video quality metric," IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), Cagliari, pp. 1-2.

[11] 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.

[12] Taraghi, B., Nguyen, M., Amirpour, H., & Timmerer, C. (2021). Intense: In-depth studies on stall events and quality switches and their impact on the quality of experience in http adaptive streaming. *IEEE Access*, *9*, 118087-118098.

[13] Urvoy, M., Barkowsky, M., & Le Callet, P. (2013). How visual fatigue and discomfort impact 3D-TV quality of experience: a comprehensive review of technological, psychophysical, and psychological factors. *annals of telecommunications-annales des télécommunications*, *68*(11), 641-655.

BADANIE CROWDSOURCINGOWE JAKOŚCI SEKWENCJI WIZYJNYCH

A CROWDSOURCING STUDY OF VIDEO QUALITY

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
[4] AGH University of Krakow, dawid.juszka@agh.edu.pl
[3] AGH University of Krakow, mikolaj.leszczuk@agh.edu.pl

DOI: 10.15199/59.2024.4.71

Streszczenie: Niniejsze badanie bada ocenę subiektywnej jakości wizyjnej za pomocą krótkich sekwencji wizyjnych na platformie crowdsourcingowej. Łącznie 47 uczestników oceniło 170 krótkich sekwencji wizyjnych, korzystając z skali Absolute Category Rating (ACR). Mimo że wszystkie krótkie sekwencje wizjne były w pełnej rozdzielczości HD, sekwencje wizyjne o jakości „dostatecznej" skutkowały dłuższym czasem odpowiedzi (1867,19 ms), przekraczającym inne kategorie przy limicie 5000 ms. Eksperyment przeprowadzono wyłącznie w przeglądarce Google Chrome w trybie incognito, aby wyeliminować zewnętrzne zakłócenia. Użycie automatycznego sekwencyjnego odtwarzania umożliwiło efektywne zbieranie danych, co pozwolało użytkownikowi zakończyć eksperyment w około 55 minut. Wyniki wskazują, że sekwencje wizyjne o jakości „dostatecznej" regularnie otrzymywały oceny po upływie dłuższego czasu, co sugeruje obecność potencjalnych stronniczości w subiektywnych ocenach. Badanie podkreśla znaczenie uwzględniania zmiennych kontekstowych, takich jak rozdzielczość monitora i projekt eksperymentu, podczas oceny jakości sekwencji wizyjnych. Wyniki poprawiają nasze zrozumienie sposobu, w jaki użytkownicy postrzegają jakość sekwencji wizyjnych, i dostarczają cennych wniosków na przyszłe badania nad technologią multimedialną.

Abstract: This study investigates the evaluation of subjective video quality utilizing short video clips on a crowdsourcing platform. A total of 47 participants provided ratings for 170 short videos using the Absolute Category Rating (ACR) scale. Although all short videos are in full HD, clips of fair quality resulted in lengthier response times (1867.19 ms), exceeding other categories within a 5000 ms limit. The experiment was only conducted on Google Chrome in incognito mode to eliminate extraneous interference. The use of automated sequential playing allowed for efficient data collection, resulting in the completion of the experiment in around 55 minutes. The results indicate that videos of Fair quality regularly received higher timings, suggesting the presence of potential biases in subjective evaluations. The study emphasizes the significance of taking into account contextual variables, such as monitor resolution and experimental design, when evaluating video quality. The findings improve our understanding of how users perceive video quality and provide valuable insights for future research on multimedia technology.

Słowa kluczowe: Crowdsourcing, Processed Video Sequences (PVS), Quality of experience (QoE), ocena subiektywna jakości wideo, Absolute Category Rating (ACR) Level, czas odpowiedzi.

Keywords: Crowdsourcing, Processed Video Sequences (PVS), Quality of experience (QoE), Video Quality Subjective Assessment, Absolute Category Rating (ACR) Level, Response Time.

1. INTRODUCTION

The rise of online video material in the digital era has increased the demand for efficient quality assessment methods [1]. The use of crowd-sourcing evaluation has emerged as a potential method of assessing user opinions on video quality [2]. This strategy leverages the collective wisdom of on-line communities to acquire a wide range of feedback. Processed video sequences (PVS) and user-generated content (UGC) are important parts of the current multimedia environment. Both types of content present distinct difficulties when it comes to assessing their quality.

Ensuring an exceptional experience for viewers is of utmost importance, since it directly influences user satisfaction and engagement. Quality of experience (QoE) metrics offer significant insights into user preferences and habits, helping content creators and platform developers refine their offers. Subjective experiments are crucial instruments in this effort, allowing researchers to record subjective impressions and preferences of viewers through controlled evaluations [3].

Video quality assessment (VQA) approaches, such as Absolute Category Rating (ACR) levels, allow participants to give subjective scores based on pre-established quality categories. Response time, which refers to the time it takes to make high-quality judgments, provides vital information on the cognitive processing and decision-making involved in subjective evaluations [4].

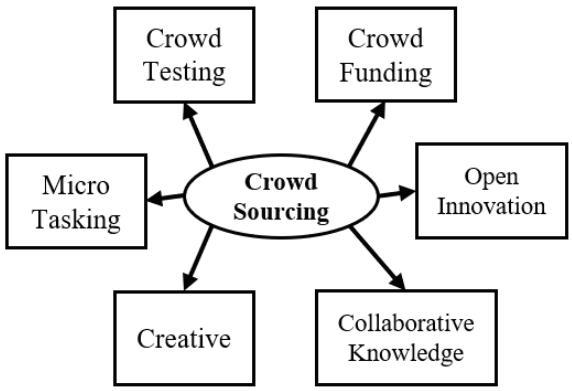
This study aims to examine subjective video quality evaluation by using short video clips on a specialised crowd-sourcing platform. Our objective is to clarify viewer impressions of video quality in various genres of content and quality levels by using Absolute Category Rating (ACR) levels and monitoring response times [5]. By conducting a thorough experimentation and analysis, our aim is to enhance the existing knowledge on video quality as-

sessment and its impact on multimedia technology and content developers.

2. CROWDSOURCING EXPERIMENT

Crowdsourcing is a technique that involves dividing a task into smaller parts and assigning them to multiple individuals for completion. The collective intelligence of a group can be harnessed to perform tasks that are excessively difficult or beyond the capabilities of machines operating independently [6].

This involves assigning tasks to multiple individuals, typically using a digital platform. This approach involves disseminating a concise overview to a wide-ranging audience through an internet-based platform, and subsequently gathering feedback from the individuals comprising that audience [7]. Although this approach has the potential to produce a substantial amount of information quickly, the effectiveness of feedback may vary depending on its execution. Each person is responsible for independently doing the task, and the combined results are used to determine the final result.



Rysunek 1: Schematic representation of crowdsourcing and its different aspects

Crowdsourcing has proven to be highly effective in various applications such as image labeling, data annotation, and speech recognition [8]. Furthermore, the crowdsourcing process, as seen in Figure 1, can have extensive implications for both content creation and content curation. Human-centered tests on perceptual quality are widely acknowledged as the most reliable method for evaluating the quality of an experience [9]. UGC, a powerful means to foster participation and develop communities, can be achieved by employing crowd-sourcing [10]. In addition, crowd-sourcing can be employed to curate material specifically tailored for particular groups, such as educational clips for young people or condensed news updates for busy professionals. This is a further use of crowd-sourcing. Crowdsourcing is increasingly being used as an alternative to traditional techniques for assessing video quality.

This technique has multiple advantages, including the ability to gather summaries from a wide range of individuals quickly, at a low cost, and with no effort. As part of our strategic plan, we enlisted individuals to view short video clips and evaluate synopses of their observations. Individuals are required to determine the short video clips according to the guidelines provided in the recommendations of ITU-T P.910 [11], which were given to them along with the instructions. The popularity of generating summaries from short video clips has increased in recent years due to the efficiency of crowd-sourcing as a strategy.

The use of crowd-sourcing for the assessment of video quality has significant potential for future advantages [12]. Crowdsourcing provides a comprehensive and diverse perspective on video material, resulting in summaries that more accurately reflect the preferences and needs of a wide range of customers. Furthermore, the utilisation of crowd-sourcing can significantly reduce the time and financial resources needed for video quality assessment, hence enhancing its efficacy and accessibility.

3. METHODOLOGY

The experiment was conducted using a customised website that was created utilising programming languages such as PHP, HTML5, and Java Script. This website allowed for the assessment of subjective video quality. Participants utilized the platform by accessing it through LAN connected laboratory computers.

The experiment includes a selection of 170 short video clips taken from various databases such as Netflix, CableLabs, SJTU Media Lab, and Xiph.org Video Test Media. All videos were of full HD resolution, specifically 1080x1920 pixels. The experimental procedure involved the random presentation of video clips to participants, with durations ranging from 9 to 21 seconds.

After seeing each clip, the participants evaluated its quality using the Absolute Category Rating (ACR) scale, which includes categories such as Excellent, Good, Fair, Poor and Bad. Participant assessments were recorded in conjunction with response times, which were quantified in milliseconds. The experiment was carried out solely on Google Chrome browsers in incognito mode to guarantee uniformity.

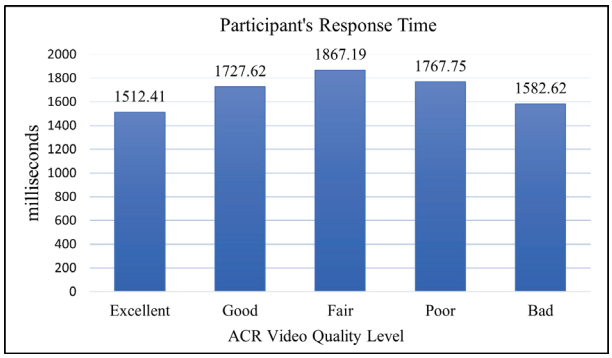
Participants were explicitly advised to refrain from engaging with the videos during playing, such as pausing or right clicking. Following completion of each video clip, a voting page would be displayed for evaluation purposes. Statistical analysis was conducted to compare decision times in various video quality categories. Response times were calculated and compared for each quality level. The experiment was completed in a total duration of approximately 55 minutes. After completion, the participants received a message confirming their successful participation [13].

4. RESULTS

The study included a total of 47 participants (including 5 women and 42 men). The sample consisted of individuals of various age groups, educational levels, and video watching patterns, ensuring a representative sample for the experiment.

Participants assigned subjective quality scores to 170 brief video clips using the Absolute Category Rating (ACR) scale. Figure 2 shows the distribution of the evaluations between various quality levels, highlighting the variability in the perceptions of the quality of the video by participants.

The duration it took participants to rate the video clips was assessed in milliseconds. Table 1 displays descriptive statistics, such as the average values and the variability represented by standard deviations, for reaction times in various quality categories.



Rysunek 2: The histogram of participant's response time of votes for different ACR video quality level mentioned in Table 1. It is visible that Fair quality has taken much more time from the others.

An analysis of the data revealed trends in quality ratings across various degrees of video quality. Significantly, clips of Fair quality regularly garnered higher scores in comparison to other categories, as participants dedicated more time to carefully consider their evaluations.

A comparison was made between the mean response times of videos classified as Excellent, Good, Fair, Poor, and Bad quality. Table1 illustrates the average response times for each quality level, highlighting the much longer response times reported for videos of Fair quality (1867.19 ms).

Tabela 1: Statistical Analysis Table

ACR Video Quality Level	*Mean Response Time (ms)*	*SD (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

A threshold of 5000 ms was set to determine the maximum acceptable reaction times. Most of the participants finished their quality assessments within this range, suggesting effective decision-making processes.

The participants viewed the clips on laboratory monitors with a resolution of 1440x900, even though the videos were presented in full HD resolution (1080x1920). Further investigation is necessary to determine whether this difference in resolution has an impact on the quality judgments and response times of the participants.

The average duration of the experiment, which includes setup instructions, video viewing, and quality rating tasks, was approximately 55 minutes per participant. This time allowed for thorough data gathering while minimising participant fatigue and attrition rates.

The validity and reliability of the acquired data was evaluated by statistical studies. The consistency and agreement among participants' quality assessments were assessed using measures such as internal consistency and inter-rater agreement.

The study's findings offer a useful understanding of how users perceive video quality and the elements that impact their subjective ratings of quality. These findings have significant implications for multimedia technologies, content creators and platform developers, providing valuable information to design more captivating and immersive video experiences for global audiences.

5. DISCUSSION

Interpretation of the Findings: The study findings uncover fascinating trends in how participants evaluate the subjective quality of short video clips. Significantly, videos of Fair quality consistently obtained superior timings in comparison to those of other quality levels, even though all clips were in full HD resolution. This discovery implies that elements other than the quality of the visuals, such as the relevancy of the material, the level of engagement, and the perceived level of novelty, might affect how viewers perceive the quality of a video.

Impact of Response Times: The considerably longer response durations found for Fair Quality videos (1867.19 ms) suggest that participants took more time to carefully consider their assessments, potentially indicating a greater level of cognitive processing and evaluative effort. This phenomenon necessitates additional exploration of the cognitive mechanisms that underlie subjective quality evaluations and its implications for the design and evaluation of multimedia material.

Analysis of the threshold: Most participants successfully completed their quality assessments within the designated threshold of 5000 ms, suggesting effective decision-making processes and active participation in the activity. This discovery indicates that the participants were capable of making quick assessments of video quality, even when brief video clips were shown.

Influence of Monitor Resolution: The disparity in resolution between the video clips (1080x1920) and the laboratory monitors (1440x900) may have impacted participants' quality evaluations and their response times. Future research should investigate the impact of monitor resolution on subjective quality assessments and take steps to control for this variable to ensure the accuracy and consistency of experimental findings.

Experimental Duration: The experimental period was meticulously planned to strike a compromise between collecting thorough data and ensuring participant participation and avoiding fatigue. On average, each participant spent roughly 55 minutes in the experiment. The brevity of the experiment likely increased participant retention rates and reduced dropout rates, hence improving the dependability of the obtained data.

Practical implications The study's findings have practical consequences for anyone involved in multimedia technologies, content creation, and platform development. Stakeholders can improve video content delivery strategies, increase user engagement, and enhance overall viewing experiences by understanding the aspects that affect subjective quality assessments.

Future directions Subsequent investigations should investigate additional variables that could impact subjective quality assessments, such as audio fidelity, material relevancy, and user inclinations. In addition, longitudinal studies could examine the progression of viewers' evaluations of video quality over time and in various viewing situations, offering useful insights into audience dynamics and trends in content consumption.

Limitations: It is crucial to recognise the constraints of the study, such as the relatively limited sample size, potential biases arising from the experimental arrangement,

and the reliance on a solitary crowd-sourcing platform for data collection. In order to overcome these constraints, future research should aim to use larger and more diverse samples, apply various data collection methods, and include experimental controls to reduce the impact of confounding variables.

The study's findings enhance our comprehension of subjective video quality assessment and its impact on multimedia technologies and content development. This research aims to clarify the factors that affect how viewers perceive video quality. In doing so, it provides valuable information to improve video compression algorithms, content delivery strategies, and quality enhancement techniques. Ultimately, this research seeks to improve the overall viewing experience for audiences around the world.

6. CONCLUSION

Ultimately, this study provides information on subjective video quality evaluation through the use of short clips on a crowd-sourcing platform. The findings emphasise the intricate nature of user perceptions and the multifaceted character of video quality assessment. Although all videos were in full HD definition, participants consistently scored videos of fair quality higher, suggesting that variables other than visual fidelity had a major impact on quality assessments.

The extended response times found for clips of Fair quality indicate that participants used greater cognitive effort in assessing these clips, presumably indicating a higher level of participation or increased inspection. This event highlights the importance of taking into account user participation and cognitive processing when evaluating subjective quality.

Setting a specific limit for answer times offers vital information on the effectiveness of participants' decision-making abilities. Most of the participants successfully completed their quality assessments within the stipulated threshold, demonstrating swift and effective judgments even when exposed to brief video segments.

Although the difference in monitor resolution may have affected the way participants evaluated quality, it is important for future studies to investigate other aspects that could also influence subjective quality evaluations, such as audio quality, relevance of content, and user preferences. Longitudinal studies have the potential to offer more understanding of how user perceptions change over time and in various viewing situations.

Although the study has some limitations, such as restrictions on sample size and potential biases caused by the experimental methodology, the results have practical consequences for multimedia technologies, content providers, and platform developers. An inherent limitation of this experimental procedure is the lack of ability to remotely assess the psycho-physiological state of the participant, hence impacting the subjective selection of the ACR scale evaluation.

In summary, this research improves our understanding of subjective video quality assessment and establishes a basis for future investigations in this domain. To enhance video experiences for global audiences, researchers and practitioners could focus on understanding and catering to user perceptions and preferences.

Funding: This work was funded by Ministry of Science and Higher Education as part of a subsidy for the Faculty of Computer Science, Electronics and Telecommunications, AGH University of Krakow.

LITERATURA

[1] M. Grega, L. Janowski, M. Leszczuk, P. Romaniak, and Z. Papir, 2008, "Quality of experience evaluation for multimedia services," Przeglad Telekomunikacyjny I Wiadomosci Telekomunikacyjne, vol. 81, pp. 142–153, 01 .

[2] T. Hoßfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, 2014, "Best practices for qoe crowdtesting: Qoe assessment with crowdsourcing," IEEE Transactions on Multimedia, vol. 16, no. 2, pp. 541–558.

[3] A. Dutta, M. Leszczuk, D. Juska, M. Grega, 2023, "Subjective Quality Assessment of Video Summarization Algorithms: A Crowdsourcing Approach" vol. 96, no. 4, pp. 333-336, Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne.

[4] P. Romaniak, M. Muy, A. Mauthe, S. D'Antonio, and M. Leszczuk, 2011, "Framework for the integrated video quality assessment," Multimedia Tools and Applications, vol. 61, 12.

[5] M. Leszczuk, M. Grega, A. Kozbiał, J. Gliwski, K. Wasieczko, and K. Smaıli, 2017, "Video summarization framework for newscasts and reports – work in progress," in Multimedia Communications, Services and Security, A. Dziech and A. Czyzewski, Eds. Cham: Springer International Publishing, pp. 86–97.

[6] W.-T. Tsai, L. Zhang, S. Hu, Z. Fan, and Q. Wang, 2023, "Crowdtesting practices and models: An empirical approach," Information and Software Technology, vol. 154, p. 107103.

[7] M. Shahid, J. Søgaard, J. Pokhrel, K. Brunnstrom, K. Wang, S. Tavakoli, and N. Gracia, 2014, "Crowdsourcing based subjective quality assessment of adaptive video streaming," in 2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX), pp. 53–54.

[8] S.-Y. Wu, R. Thawonmas, and K.-T. Chen, 2011, "Video summarization via crowdsourcing," in CHI '11 Extended Abstracts on Human Factors in Computing Systems, ser. CHI EA '11. New York, NY, USA: Association for Computing Machinery, p. 1531–1536.

[9] M. Leszczuk, L. Janowski, J. Nawała, and M. Grega, 2022, "User-generated content (ugc)/in-the-wild video content recognition," in Intelligent Information and Database Systems, N. T. Nguyen, T. K. Tran, U. Tukayev, T.-P. Hong, B. Trawinski, and E. Szczerbicki, Eds. Cham: Springer Nature Switzerland, pp. 356–368.

[10] P. Romaniak, L. Janowski, M. Leszczuk, and Z. Papir, 2012, "Perceptual quality assessment for h.264/avc compression," in IEEE Consumer Communications and Networking Conference (CCNC), pp. 597–602.

[11] ITU-T. Recommendations: ITU-T P.910: 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.

[12] K. Borchert, A. Seufert, E. Gamboa, M. Hirth, and T. Hossfeld, 2021, "In vitro vs in vivo: Does the study's interface design influence crowdsourced video qoe?" Quality and User Experience, vol. 6, 12.

MIARA PODOBIEŃSTWA STRUKTURALNEGO DLA WIZJI WSZECHOGARNIAJĄCEJ

STRUCTURAL SIMILARITY METRIC FOR IMMERSIVE VIDEO

Jakub Stankowski, Weronika Nowak, Tomasz Grajek, Adrian Dziembowski

Instytut Telekomunikacji Multimedialnej, Politechnika Poznańska, Poznań
{jakub.stankowski, tomasz.grajek, adrian.dziembowski}@put.poznan.pl

DOI: 10.15199/59.2024.4.72

Streszczenie: **W pracy przedstawiono nową miarę obiektywnego pomiaru jakości dla wizji wszechogarniającej. Zaproponowana metryka powstała w sposób analogiczny do tego jak opracowano miarę IV-PSNR na podstawie miary PSNR. Tym razem zamiast PSNR wykorzystano SSIM. Przeprowadzone eksperymenty pokazały, że zaproponowana miara SSIV jest jeszcze lepiej skorelowana z wynikami oceny subiektywnej (MOS) niż inne znane metryki, w tym IV-PSNR.**

Abstract: **In this paper, a new objective quality metric for immersive video is presented. The proposed metric was created in a manner analogous to how the IV-PSNR metric was based on the PSNR one. This time, SSIM was used instead of PSNR. The conducted experiments showed that the proposed SSIV metric is even higher correlated with the results of subjective evaluation (MOS) than other known metrics, including IV-PSNR.**

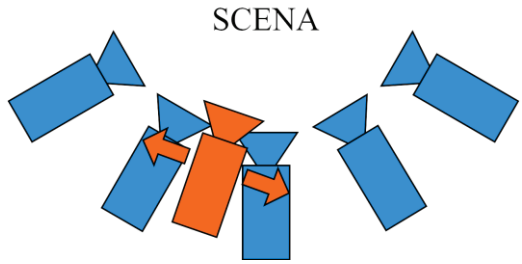
Słowa kluczowe: **wizja wszechogarniająca, ocena jakości wizji, obiektywny pomiar jakości.**

Keywords: **immersive video, video quality evaluation, objective quality assessment.**

1. WSTĘP

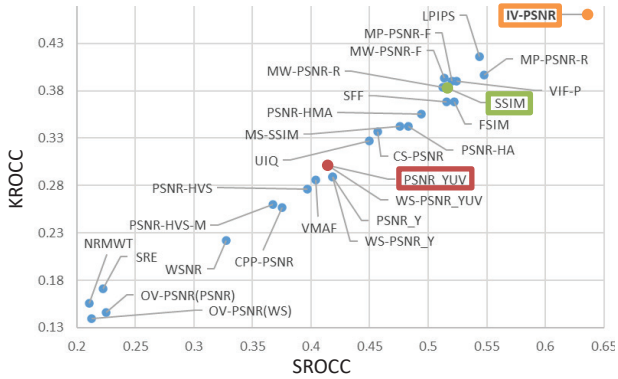
Rozwój technologiczny wpływa na wiele dziedzin naszego życia. W systemach wizyjnych można zauważyć coraz bardziej skomplikowane rozwiązania pozwalające w coraz lepszym stopniu odwzorowywać otaczającą nas rzeczywistość. Najnowsze rozwiązania dotyczą tzw. wizji wszechogarniającej (ang. *Immersive vision*) [1]. Systemy takie pozwalają użytkownikowi/widzowi „zanurzyć się" w prezentowanej treści i dowolnie zmieniać położenie, z którego obserwuje się scenę [2], [4]. Ideę takiego systemu przedstawiono na rysunku 1, gdzie niebieskie kamery oznaczają rzeczywiste kamery, którymi zarejestrowano daną scenę, natomiast kamera pomarańczowa oznacza użytkownika, który może zmieniać punkt obserwacji sceny. Oczywistym jest, że przy takim podejściu, dla większości położeń kamery pomarańczowej nie mamy danych wizyjnych dostępnych bezpośrednio z rzeczywistych kamer, więc dane te należy wytworzyć (wyrenderować) wykorzystując np. algorytm syntezy widoków wirtualnych [2], [4].

Skomplikowanie systemów wizyjnych wraz z nowymi sposobami generowania i prezentowania danych wizyjnych prowadzi do powstawania nowych typów zniekształceń wynikających z przetwarzania wizji wszechogarniakącej. W związku z tym, rzetelna ocena jakości danych wizyjnych staje się coraz trudniejsza.



Rys. 1. Idea systemu wizji wszechogarniającej; niebieskie – kamery rzeczywiste (wejściowe), pomarańczowa – kamera wirtualna.

Aby uwzględnić zniekształcenia typowe dla wizji immersyjnej, wprowadzono metrykę IV-PSNR [3], która uwzględnia typowe dla niej artefakty spowodowane przerzutowaniem informacji między widokami, a więc niewielkie przesunięcie obiektów i globalną zmianę charakterystyki barwnej renderowanego (syntezowanego) widoku wirtualnego.



Rys. 2. Korelacja (mierzona z użyciem metryk SROCC i KROCC [7]) pomiędzy subiektywną oceną jakości a jakością mierzoną obiektywnie, z użyciem metryk opisanych w literaturze. Rysunek z [3].

Jak przedstawiono na rys. 2, dostosowana do charakterystyki wizji wszechogarniającej metryka IV-PSNR, wykazuje wyższą korelację z oceną jakości subiektywnej niż inne metryki (wliczając w to także metryki dedykowane dla obrazów syntezowanych, takie jak MP-PSNR-F [9], MP-PSNR-R [11] oraz MW-PSNR-F [10]). Jednakże,

metryka ta bazuje na PSNR (kolor czerwony, rys. 2), który jest zdecydowanie mniej efektywny niż pomiar podobieństwa strukturalnego (SSIM – kolor zielony, rys. 2).

W niniejszej pracy przedstawiono metrykę łączącą zalety obu technik, dopasowując pomiar podobieństwa strukturalnego do potrzeb wizji wszechogarniającej w sposób analogiczny jak w przypadku metryki IV-PSNR.

2. WSKAŹNIK PODOBIEŃSTWA STRUKTURALNEGO (SSIM)

Wskaźnik podobieństwa strukturalnego – SSIM (*Structural Similarity Index Measure*) [15] jest obiektywną miarą jakości, która od czasu publikacji zyskała bardzo dużą popularność w zastosowaniach związanych z oceną jakości obrazu. Jej implementacje można znaleźć w popularnych narzędziach takich jak MATLAB, bibliotekach programistycznych (np. scikit-image), czy nawet pakietach do uczenia maszynowego (np. TensorFlow). Wiele popularnych implementacji koderów wizyjnych, takich jak x264 [17] (koder AVC [13]), x265 [18] (koder HEVC [12]) czy libaom [16] (koder AV1 [5]) pozwala na użycie SSIM jako miary jakości w procesie optymalizacji przepływność-zniekształcenia RD (*rate – distortion*).

Miara SSIM obliczana jest jako iloczyn ważony, składający się z trzech składników zdefiniowanych jako: luminancja[1] ($\boldsymbol{L}$), kontrast ($\boldsymbol{L}$) i struktura ($\boldsymbol{S}$). Wspomniane składniki wyliczane są na bazie lokalnych statystyk takich jak: wartość średnia, odchylenie standardowe i kowariancja.

Cechą specyficzną dla wskaźnika podobieństwa strukturalnego jest wyliczanie powyższych statystyk w oknie o rozmiarze 11x11, gdzie wartości próbek wejściowych są ważone dwuwymiarową maską Gaussowską $\boldsymbol{\omega}$ (o odchyleniu standardowym równym 1,5) [14].

3. PODOBIEŃSTWO STRUKTURALNE W OCENIE WIZJI WSZECHOGARNIAJĄCEJ (SSIV)

W niniejszym artykule zaproponowano metrykę stworzoną do oceny jakości wizji wszechogarniającej. Zaproponowana metryka jest oparta na oryginalnej koncepcji wskaźnika podobieństwa strukturalnego ale została przeprojektowana w taki sposób aby dopasować ją do opisanych w [3] zniekształceń.

Pierwszym etapem wyznaczania podobieństwa strukturalnego obrazów I i J jest wyznaczenie lokalnych statystyk tychże obrazów: średniej ($\boldsymbol{\mu}^I$, $\boldsymbol{\mu}^J$), odchylenia standardowego ($\boldsymbol{\sigma}^I$, $\boldsymbol{\sigma}^J$) oraz kowariancji ($\boldsymbol{\sigma}^{I,J}$) dla każdej składowej barwnej c obu obrazów.

Dla każdego piksela (x, y) analizowane jest otoczenie $[2k+1] \times [2k+1]$, a wartości pikseli z tegoż otoczenia ważone są z użyciem maski $\boldsymbol{\omega}$. W przypadku obrazu I średnia i odchylenie standardowe wyznaczane są w taki sam sposób, jak w metryce SSIM [15]:

$$\boldsymbol{\mu}_c^I(x,y) = \sum_{i=x-k}^{x+k}\sum_{j=y-k}^{y+k} [\boldsymbol{\omega}(i-x, j-y) \cdot \boldsymbol{I}_c(i,j)] \,, \tag{1}$$

$$\boldsymbol{\sigma}_c^I(x,y) = \sqrt{\sum_{i=x-k}^{x+k}\sum_{j=y-k}^{y+k} \left[\boldsymbol{\omega}(i-x, j-y) \cdot \left(\boldsymbol{I}_c(i,j)\right)^2\right] - \left(\boldsymbol{\mu}_c^I(x,y)\right)^2} \,, \tag{2}$$

Pozostałe lokalne statystyki wyznaczane są w sposób zmodyfikowany, dostosowując się do typowych dla wizji wszechogarniającej artefaktów spowodowanych przerzutowaniem informacji między widokami. Zmiany względem oryginalnej metryki SSIM zaznaczono kolorem zielonym:

$$\boldsymbol{\mu}_c^J(x,y) = \sum_{i=x-k}^{x+k}\sum_{j=y-k}^{y+k} [\boldsymbol{\omega}(i-x, j-y) \cdot \boldsymbol{J}_c(\boldsymbol{i', j'})] \,, \tag{3}$$

$$\boldsymbol{\sigma}_c^J(x,y) = \sqrt{\sum_{i=x-k}^{x+k}\sum_{j=y-k}^{y+k} \left[\boldsymbol{\omega}(i-x, j-y) \cdot \left(\boldsymbol{J}_c(\boldsymbol{i', j'})\right)^2\right] - \left(\boldsymbol{\mu}_c^J(x,y)\right)^2} \,, \tag{4}$$

$$\boldsymbol{\sigma}_c^{\boldsymbol{I\to J}}(x,y) = \sum_{i=x-k}^{x+k}\sum_{j=y-k}^{y+k} [\boldsymbol{\omega}(i-x, j-y) \cdot \boldsymbol{I}_c(i,j) \cdot \boldsymbol{J}_c(\boldsymbol{i', j'})] + - \boldsymbol{\mu}_c^I(x,y) \cdot \boldsymbol{\mu}_c^J(x,y) \,, \tag{5}$$

gdzie:

$$i' = i + s_x(i,j) \,, \quad j' = j + s_y(i,j) \,, \tag{6}$$

a $s_x(i,j)$ i $s_y(i,j)$ określają względną pozycję takiego piksela w obrazie J, który znajduje się w otoczeniu o rozmiarze $B \times B$ wokół piksela (i,j), który jest najbardziej podobny do piksela (i,j) w obrazie I. Domyślnie $B = 2$.

Poszukiwanie najbardziej podobnego piksela w otoczeniu $B \times B$ skutkuje brakiem symetrii wyznaczania kowariancji, przez co konieczne jest zaznaczenie, w którym kierunku kowariancja ta jest wyznaczana. Stąd też oznaczenie $\boldsymbol{\sigma}^{I\to J}$ zamiast typowego $\boldsymbol{\sigma}^{I,J}$ używanego w przypadku SSIM.

W drugim kroku porównywane są trzy podstawowe właściwości obrazów I i J: luminancja $\boldsymbol{L}$, kontrast $\boldsymbol{C}$ i struktura $\boldsymbol{S}$:

$$\boldsymbol{L}_c^{\boldsymbol{I\to J}}(x,y) = \frac{2 \cdot \mu_c^I(x,y) \cdot \left[\mu_c^J(x,y) + s_c^{I\to J}\right] + C_1}{\mu_c^I(x,y)^2 + \left[\mu_c^J(x,y) + s_c^{I\to J}\right]^2 + C_1}, \tag{7}$$

$$\boldsymbol{C}_c^{\boldsymbol{I\to J}}(x,y) = \frac{2 \cdot \sigma_c^I(x,y) \cdot \sigma_c^J(x,y) + C_2}{\sigma_c^I(x,y)^2 + \sigma_c^J(x,y)^2 + C_2}, \tag{8}$$

$$\boldsymbol{S}_c^{\boldsymbol{I\to J}}(x,y) = \frac{\sigma_c^{I\to J}(x,y) + C_3}{\sigma_c^I(x,y) \cdot \sigma_c^J(x,y) + C_3}, \tag{9}$$

gdzie C_1, C_2 i C_3 są stałymi zapewniającymi numeryczną stabilność, zdefiniowanymi jako:

$$C_1 = (K_1 \cdot (2^b - 1))^2, \quad C_2 = (K_2 \cdot (2^b - 1))^2, \quad C_3 = \frac{C_2}{2}, \tag{10}$$

$$K_1 = 0.01, \quad K_2 = 0.03, \tag{11}$$

a $s_c^{I\to J}$ jest globalną różnicą składowej barwnej c obrazów I i J, wyznaczaną jako:

[1] Warto zwrócić uwagę na niefortunny dobór nazwy, który w kontekście wizji może być mylący – luminancja zdefiniowana przez autorów metryki nie ma nic wspólnego z luminancją (lumą) rozumianą jako składowa barwna.

$$s_c^{I\to J} = \frac{1}{W_c \cdot H_c} \sum_{y=0}^{H_c-1} \sum_{x=0}^{W_c-1} (\boldsymbol{I}_c(x,y) - \boldsymbol{J}_c(x,y)) \,, \qquad (12)$$

gdzie W_c i H_c są – odpowiednio – szerokością i wysokością obrazu (a ściślej – jego składowej barwnej c).

Wyznaczone właściwości $\boldsymbol{L}$, $\boldsymbol{C}$ i $\boldsymbol{S}$ są w następnym kroku wymnażane w celu wyznaczenia lokalnego podobieństwa pomiędzy obrazami I i J w punkcie (x, y):

$$\boldsymbol{Q}_c^{I\to J}(x,y) = \boldsymbol{L}_c^{I\to J}(x,y) \cdot \boldsymbol{C}_c^{I\to J}(x,y) \cdot \boldsymbol{S}_c^{I\to J}(x,y) \,. \qquad (13)$$

Następnie, podobieństwo to jest uśredniane po całej powierzchni obrazu:

$$SSIV_c^{I\to J} = \frac{1}{W_c \cdot H_c} \sum_{y=0}^{H_c-1} \sum_{x=0}^{W_c-1} \boldsymbol{Q}_c^{I\to J}(x,y) \qquad (14)$$

i wszystkich składowych barwnych (luminancji i dwóch chrominancji):

$$SSIV_{YUV}^{I\to J} = \frac{SSIV_Y^{I\to J} \cdot w_Y + SSIV_U^{I\to J} \cdot w_U + SSIV_V^{I\to J} \cdot w_V}{w_Y + w_U + w_V} \,, \qquad (15)$$

gdzie wartość wag w_Y, w_U i w_V ustalona jest tak, jak w przypadku miary IV-PSNR [3] i wynosi – odpowiednio – $w_Y = 4, w_U = 1, w_V = 1$.

Ostateczna miara podobieństwa strukturalnego pary obrazów I i J wyznaczana jest jako:

$$SSIV_{YUV}^{I,J} = \min\left(SSIV_{YUV}^{I\to J}, SSIV_{YUV}^{J\to I}\right) \,, \qquad (16)$$

co zapewnia symetryczność zaproponowanej metryki, znacząco zwiększając jej użyteczność i łatwość użycia.

4. EKSPERYMENT

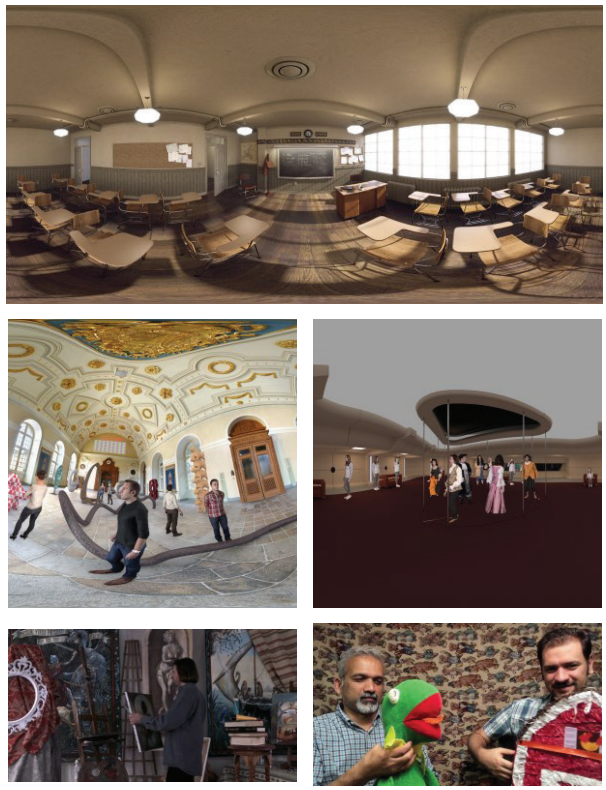
Efektywność zaproponowanej metryki została oceniona na bazie wyników konkursu grupy ISO/IEC MPEG na nowy kodek wizji wszechogarniającej [6], które to wyniki obejmowały subiektywną i obiektywną ocenę pięciu zróżnicowanych sekwencji wielowidokowych (rys. 3) zakodowanych z użyciem siedmiu różnych technik kodowania wizji wszechogarniającej. Szczegółowa metodologia eksperymentu dostępna jest w rozdziale VII artykułu opisującego metrykę IV-PSNR [3].

Zaproponowana metryka została porównana z IV-PSNR, a także SSIM i PSNR. Testowane metryki porównano wykorzystując dwa popularne współczynniki korelacji bazujące na kolejności (*rank-order coefficients*): SROCC i KROCC [7].

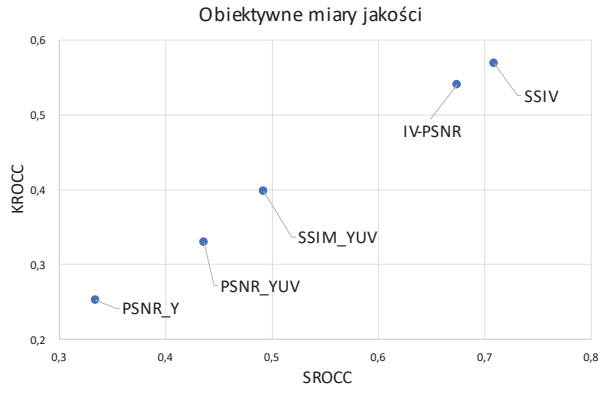
Jak pokazano na rysunku 4, zaproponowana metryka SSIV wykazuje większą korelację z wynikami testów subiektywnych, aniżeli pozostałe testowane metryki, w tym IV-PSNR. Biorąc pod uwagę fakt, iż kompleksowe badania opisane w [3] wykazały, iż inne opisane w literaturze obiektywne metryki jakości radzą sobie z poprawną oceną jakości wizji wszechogarniającej zdecydowanie gorzej niż IV-PSNR, stwierdzić można, iż SSIV przewyższa je wszystkie.

Poza porównaniem z innymi metrykami, eksperymentalnie sprawdzono również efektywność czterech wariantów estymacji metryki SSIV (oraz SSIM).

Miara SSIM opisana w artykule [15] nie jest do końca jednoznacznie zdefiniowana. Dodatkowo, sama implementacja SSIM, pozostawia wiele pytań o czym świadczy obecność publikacji wyjaśniających istotne detale i pominięte szczegóły, takich jak [8] czy [14].



Rys. 3. Zbiór sekwencji testowych.



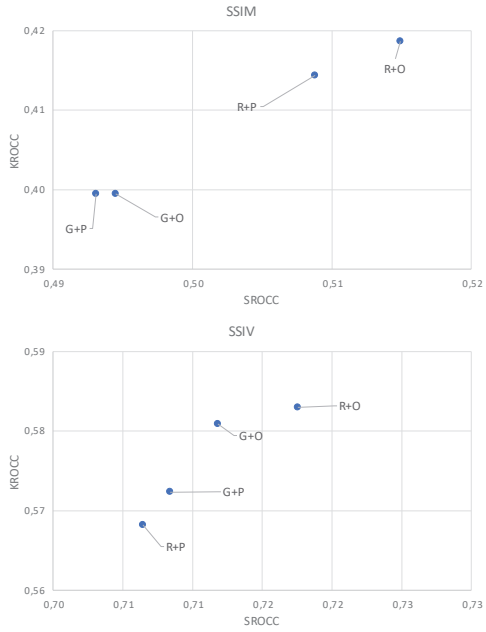
Rys. 4. Korelacja pomiędzy subiektywną oceną jakości a jakością mierzoną obiektywnie dla pięciu testowanych metryk.

Jednym z istotnych kwestii związanych z obliczaniem miary SSIM (i jej pochodnych) jest problem skrajnych próbek obrazu. Zastosowanie maski $\boldsymbol{\omega}$ o rozmiarze 11x11 powoduje konieczność rozszerzenia obrazu o dodatkowe marginesy (ang. *padding*) co potencjalnie może mieć niekorzystny wpływ na skuteczność metryki.

Drugim problemem w wyznaczaniu SSIM jest złożoność obliczeniowa – zastosowanie maski $\boldsymbol{\omega}$ o rozmiarze 11x11 powoduje konieczność wykonania 242 operacji mnożenia na każdy punkt (ok 1,5 miliarda operacji mnożenia na obraz FullHD – 1920x1080).

W związku z powyższym zdecydowaliśmy się przebadać efekty zastosowania dwóch uproszczeń: 1. pomijania punktów dla których maska $\boldsymbol{\omega}$ sięga poza oryginalny obszar obrazu (wariant O) vs użycie rozszerzonego obrazu (wariant P); 2. zastąpienie maski Gaussowskiej (wariant G) maską prostokątną (wariant R).

Jak pokazano na rysunku 5 zastosowanie opisanych uproszczeń (a w szczególności wariantu R+O) nie tylko nie obniżyło sprawności miar SSIM i SSIV ale wręcz minimalnie poprawiło korelację z MOS przy znaczącej kilkunastoprocentowej redukcji czasu obliczeń.



Rys. 5. Korelacja pomiędzy subiektywną oceną jakości a jakością mierzoną obiektywnie dla czterech testowanych wariantów metryk SSIM i SSIV.

5. PODSUMOWANIE

W artykule przedstawiono nową metrykę oceny jakości sekwencji wizji wszechogarniającej. Nowa metryka powstała w sposób analogiczny do tego jak opracowano miarę IV-PSNR na podstawie miary PSNR. Tym razem zamiast PSNR wykorzystano SSIM.

Przeprowadzone eksperymenty pokazały, że zaproponowana miara SSIV jest jeszcze lepiej skorelowana z wynikami oceny subiektywnej (MOS) niż inne znane metryki, w tym IV-PSNR. Co więcej, w pracy przedstawiono wyniki dla czterech wariantów wyznaczania miary SSIM, a w konsekwencji także SSIV, dążące do zredukowania złożoności obliczeniowej przedstawionej propozycji.

PODZIĘKOWANIA

Praca finansowana ze środków przyznanych przez Ministerstwo Nauki i Szkolnictwa Wyższego.

LITERATURA

[1] Boyce Jill, Dore Renaud, Dziembowski Adrian, Fleureau Julien, Jung Joel, Kroon Bart, Salahieh Basel, Kumar Malamal Vadakital Vinod, Yu Lu. 2021. "MPEG Immersive Video coding standard". *Proceedings of the IEEE*, 119 (9): 1521-1536.

[2] Domański Marek, Stankiewicz Olgierd, Wegner Krzysztof, Grajek Tomasz. 2017. "Immersive visual media — MPEG-I: 360 video, virtual navigation and beyond". *24th International Conference on Systems, Signals and Image Processing*, Poznań, Polska.

[3] Dziembowski Adrian, Mieloch Dawid, Stankowski Jakub, Grzelka Adam. 2022. "IV-PSNR – The objective quality metric for immersive video applications". *IEEE Transactions on Circuits and Systems for Video Technology*, 32 (11): 7575-7591.

[4] Dziembowski Adrian, Mieloch Dawid. 2022. "Nowe techniki kompresji wizji dla rzeczywistości wirtualnej – MPEG Immersive Video". *Przegląd Telekomunikacyjny - Wiadomości Telekomunikacyjne.* 4: 116-123.

[5] Han Jingning i in. 2021. "A technical overview of AV1". *Proceedings of the IEEE* 109 (9): 1435-1462.

[6] ISO/IEC. 2019. "Call for Proposals on 3DoF+ Visual". *ISO/IEC JTC1/SC29/WG11 MPEG N18145*, Marakesz, Maroko.

[7] Myers Jerome, Well Arnold. 2003. "Research design and statistical analysis". *Lawrence Erlbaum Associates*, Londyn, UK.

[8] Nilsson Jim, Akenine-Möller Tomas. 2020. "Understanding SSIM". arxiv.org/abs/2006.13846.

[9] Sandić-Stanković Dragana, Kukolj Dragan, Le Callet Patrick. 2015. „DIBR synthesized image quality assessment based on morphological pyramids". *3DTV-CON*, Lizbona, Portugalia.

[10] Sandić-Stanković Dragana, Kukolj Dragan, Le Callet Patrick. 2015. "DIBR synthesized image quality assessment based on morphological wavelets". *International Workshop on Quality of Multimedia Experience QoMEX*, Costa Navarino, Grecja.

[11] Sandić-Stanković Dragana, Kukolj Dragan, Le Callet Patrick. 2016. "Multi–Scale Synthesized View Assessment Based on Morphological Pyramids". *Journal of Electrical Engineering* 67 (1): 1–9.

[12] Sullivan Gary J., Ohm Jens-Rainer, Han Woo-Jin, Wiegand Thomas. 2012. "Overview of the High Efficiency Video Coding (HEVC) Standard". *IEEE Transactions on Circuits and Systems for Video Technology* 22: 1649–1668

[13] Sullivan Gary J., Wiegand Thomas. 2005. "Video Compression – From Concepts to the H.264/AVC Standard". *Proceedings of the IEEE* 93 (1): 18-31.

[14] Venkataramanan Abhinau K., Wu Chengyang, Bovik Alan C., Katsavounidis Ioannis, Shahid Zafar. 2021. "A Hitchhiker's Guide to Structural Similarity". IEEE Access 9: 28872-28896.

[15] Wang Zhou, Bovik Alan C., Sheikh Hamid, Simoncelli Eero. 2004. "Image quality assessment: From error measurement to structural similarity". *IEEE Transactions on Image Processing*, 13 (4): 600-612.

[16] „Koder AOM", dostępny [02.06.2024]: https://aomedia.googlesource.com/aom/

[17] „Koder x264", dostępny [02.06.2024]: https://www.videolan.org/developers/x264.htm

[18] „Koder x265", dostępny [02.06.2024]: https://x265.com/

PORÓWNANIE JAKOŚCI AUDIO WYBRANEJ METODY ZNAKOWANIA WODNEGO DLA SYGNAŁU MOWY, MUZYKI KLASYCZNEJ I POPULARNEJ

COMPARISON OF AUDIO QUALITY OF SELECTED WATERMARKING METHOD FOR SPEECH SIGNAL, CLASSICAL MUSIC, AND POPULAR MUSIC

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: 10.15199/59.2024.4.73

Streszczenie: **W artykule przedstawiono szczegółową analizę oraz porównanie metody znakowania wodnego sygnałów audio za pomocą metody cyfrowego przetwarzania sygnałów bez przetwarzania sygnałów sieciami neuronowymi. Skupiono się na zastosowaniu dyskretnej transformaty falkowej – DWT oraz dyskretnej transformaty kosinusowej - DCT. Głównym celem badań jest ocena i porównanie miar jakości znaku wodnego osadzonego w różnych rodzajach treści audio, takich jak sygnał mowy, muzyka popularna i muzyka klasyczna.**

Abstract: **The article presents a detailed analysis and comparison of watermarking method for audio signals using digital signal processing without signal processing by neural networks. The focus is on the application of discrete wavelet transform - DWT and discrete cosine transform - DCT. The main goal of the research is to evaluate and compare the quality measures of the watermark embedded in various types of audio content, such as speech signal, popular music, and classical music**

Słowa kluczowe: **bezpieczeństwo cyfrowe, przetwarzanie sygnałów, sygnały audio, znakowanie wodne**

Keywords: **audio signals, digital security, signal processing, watermarking**

1. WSTĘP

W dzisiejszym świecie szybki rozwój technologii doprowadził do tworzenia znacznej ilości treści multimedialnych, takich jak obrazy, treści audio i wideo. Stwarza to ogromny problem ochrony tych treści przed nieupoważnionym jak i niepożądanym odbiorem. W związku z tym oczywistym wymogiem jest stworzenie rozwiązań, które przeciwdziałają takim zdarzeniom .

Zastosowanie w dziedzinie ochrony treści multimedialnych znajduje znak wodny (ang.watermark). Termin ten wywodzi się z techniki drukarskiej polegającej na umieszczeniu zazwyczaj niezauważalnego znaku na papierze, aby poświadczyć jego autentyczność, np. banknot, znaczek pocztowy, dokumenty rządowe. Z kolei termin cyfrowy znak wodny (ang. digital watermark) polega na osadzaniu cyfrowego sygnału znakującego w treści multimedialnej, takich jak: sygnał wideo, audio oraz zdjęcia. Taki sposób zapisu informacji może stanowić trudne do usunięcia oraz wykrycia zabezpieczenie plików cyfrowych. Powoduje to jednak pogorszenie jakości zapisanych danych w pliku, co jest jedną z kluczowych cech jakościowych charakteryzujących metody znakowania wodnego.

W świecie cyfrowym, znak wodny znajduje swoje zastosowanie o wiele szerszym zakresie niż w świecie rzeczywistym. Można wyróżnić jego następujące zastosowania:

- Ochrona praw autorskich informacji cyfrowych,
- Monitorowanie mediów,
- Identyfikacja właściciela,
- Kontrola kopiowania,
- Opis treści.

Osadzone w treści dane w formie cyfrowego znaku wodnego, można później wyodrębnić z treści multimedialnych i wykorzystać w danej aplikacji. Należy jednak pamiętać, że w świecie cyfrowym istnieje wiele operacji przetwarzania sygnałów, które w mniejszym lub większym stopniu wpływają na znak wodny, przy czym nawet niewielka zmiana danego parametru sygnału może całkowicie zmienić znaczenie informacji w nim zawartych. W związku z tym, wdrożenie systemu znakowania wodnego uzależnione jest od jego przeznaczenia oraz pełni ważną rolę w ochronie treści multimedialnych w dzisiejszych czasach. Ponadto w systemach znakowania wodnego dużą rolę pełni wydajność obliczeniowa. Definiowana jest ona jako liczba działań arytmetycznych, jakie może wykonać komputer w określonej jednostce czasu. Kluczową rolę pełni ona w systemach wymagających znakowania wodnego w czasie rzeczywistym. Oznacza to, ze proces osadzania oraz ekstrakcji znaku wodnego z treści multimedialnej odbywa się w czasie teraźniejszym. W takich zastosowaniach algorytmy nie mogą być zbyt złożone obliczeniowo.

2. OPIS METODY

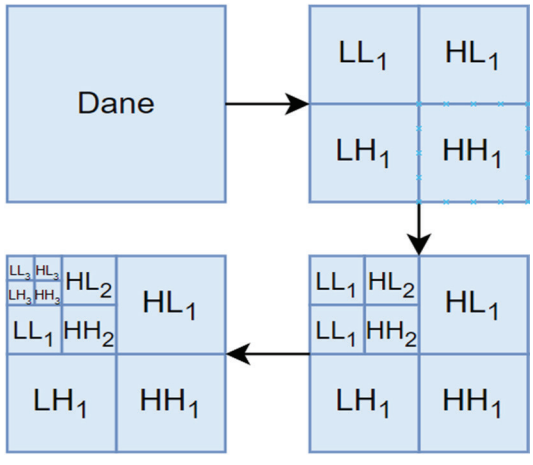
W niniejszym artykule zaimplementowano metodę znakowania opisaną w [6]. Dodatkowo wprowadzono autorskie ulepszenia, które pozwoliły osiągnąć lepsze wyniki w porównaniu do algorytmu przedstawionego w cytowanym artykule. Do znakowania wodnego wybrany został format .wav, ze względu na swoją popularność (kodek PCM), jakość dźwięku (Audio CD) oraz to, że pliki w tym formacie są nieskompresowane. Ze względu na to, przy osadzaniu znaku wodnego, mamy możliwość

wprowadzania modyfikacji na oryginalnych danych, które nie są zmodyfikowane w wyniku przeprowadzenia kompresji stratnej [5].

Działanie algorytmu opiera się na dwóch zasadniczych transformatach: dyskretnej transformacie falkowej (DWT) i dyskretnej transformacie kosinusowej (DCT). Dyskretna transformata falkowa jest wykorzystywana ze względu na jej wyjątkową odporność na większość rodzajów ataków cyfrowych, natomiast dyskretna transformata kosinusowa umożliwia skuteczne zagęszczenie energii w niskoczęstotliwościowych współczynnikach, co jest korzystne zwłaszcza w kontekście kompresji danych.

W ramach algorytmu sygnał audio jest najpierw dzielony na segmenty, które są następnie poddawane analizie w celu wyznaczenia odpowiednich miejsc do osadzenia informacji cyfrowej. Kluczowe jest, że suma współczynników w każdym segmencie nie może wynosić zero, oraz to aby segment ten mógł uczestniczyć w procesie osadzania znaku wodnego. Przy zastosowaniu wielopoziomowej transformacji DWT na segmentach sygnału, dokonuje się rozdzielenia komponentów na częstotliwości niskie i wysokie. Minimalne modyfikacje dokonywane są w komponencie niskiej częstotliwości, aby zminimalizować wpływ na jakość sygnału, podczas gdy zmiany w komponentach wysokiej częstotliwości są bardziej narażone na zewnętrzne zakłócenia.

Aby ulepszyć proces osadzania znaku wodnego, przeprowadzono serię testów polegających na zmianie poziomu dyskretnej transformaty falkowej. W tym celu wykorzystano transformatę folkową trzeciego poziomu, tak aby wybrać najkorzystniejsze pasmo częstotliwości do ukrycia informacji. W ramach eksperymentów stwierdzono, że wykorzystanie pasma z najniższymi częstotliwościami pozwala uzyskać najlepsze rezultaty (pasmo LL3). Dekompozycja DWT została przedstawiona na rysunku poniżej.



Rys1. 1. 3-poziomowa dekompozycja DWT [7]

Ze względu na duże rozbieżności wartości po modyfikacji współczynników według użytego algorytmu, wprowadzono dodatkowe reguły mające na celu wyeliminowanie słyszalności znaku wodnego. Reguły te obejmują:

- Ustalanie dolnej i górnej granicy zmian wartości próbek. Powyższe równania nie będą wykorzystywane dla próbek, których zmiany przekraczają te granice.
- Zapobieganie zbyt dużym zmianom wartości próbek. Dzięki temu unikniemy przesterowania fazy, które skutkuje zakłóceniami w postaci krótko słyszalnego szumu.
- Eliminacja próbek powodujących ciągły szum - Próbki, dla których wprowadzane są małe zmiany, ale w dużej ilości, zostaną wykluczone, aby zapobiec powstawaniu ciągłego szumu.

Wprowadzenie tych reguł pozwala na:

- Zawężenie zakresu współczynników używanych w procesie osadzania znaku wodnego.
- Uniknięcie znacznego spadku uzyskiwanego poziomu bitowej stopy błędów.
- Redukcję ryzyka wprowadzenia zakłóceń dźwiękowych podczas manipulowania wartościami próbek.

Pomimo zastosowanych reguł, manipulowanie wartościami próbek nadal niesie ryzyko wprowadzenia zbyt dużych zmian, które mogą negatywnie wpływać na sygnał akustyczny.

Końcowym etapem algorytmu jest zastosowanie odwrotnej transformaty kosinusowej i odwrotnej transformaty falkowej, które pozwalają na rekonstrukcję zmodyfikowanych segmentów sygnału. Te zmodyfikowane segmenty są następnie łączone w pełny sygnał audio z osadzonym znakiem wodnym.

3. TESTY

W ramach testów przeprowadzone zostały testy jakości sygnału audio oraz pojemności binarnej (ang. payload, capacity) algorytmu znakowania wodnego. Transparentność jest miarą określającą podobieństwo pomiędzy oryginalnym sygnałem akustycznym oraz sygnałem oznaczonym znakiem wodnym. Niezbędnym warunkiem jest zachowanie transparentności (niesłyszalności), aby znak wodny był niesłyszalny w obecności sygnału oryginalnego, w innym przypadku może to prowadzić do niepożądanych operacji takich jak złośliwe usuwanie znaku wodnego, czy próba jego zastąpienia. Z kolei pojemność informacyjna jest miarą określającą ilość bitów osadzonych w danym utworze w jednostce czasu (tzw. stempel binarny), które następnie mogą zostać wyodrębnione w procesie ekstrakcji znaku wodnego [3].

Do przeprowadzenia testów wybranych zostało 10 różnych ścieżek dźwiękowych w formacie .wav (fs=44100 próbek/s, 16 bitów na próbkę). Został w nich osadzony znak wodny o pojemności 25 bit/s oraz 50 bit/s, co sumarycznie daje 20 plików dźwiękowych poddanych testom transparentności. Pierwsze cztery pliki dźwiękowe zawierają muzykę klasyczną, kolejne trzy zawierają muzykę popularną, a pozostałe pliki zawierają nagrania sygnału mowy.

Znakowanie wodne sygnału akustycznego ma na celu osadzenie niesłyszalnego znaku wodnego w orygi-

nalnym sygnale audio. W związku z tym znak wodny powinien być percepcyjnie transparentny. Transparentność znaku wodnego w sygnale akustycznym można ocenić na dwa sposoby:

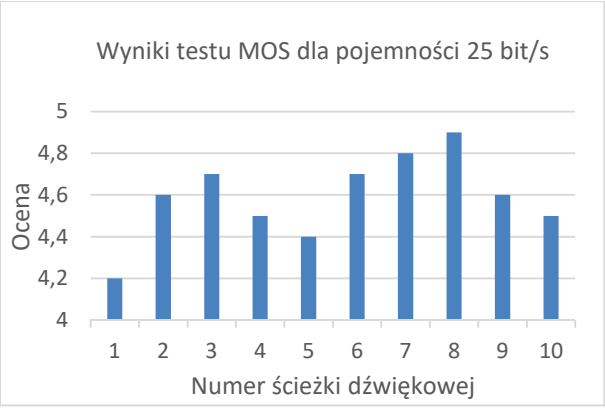
- Testowanie subiektywne,
- Testowanie obiektywne.

Subiektywne testy odsłuchowe są niezbędne dla percepcyjnej oceny jakości sygnału audio. Zostały one przeprowadzone w oparciu o rekomendacje ITU-R BS.1284 [1], która jest zaleceniem mającym służyć jako przewodnik po ogólnej ocenie jakości dźwięku. Wybrana została skala przedstawiona w tabeli poniżej.

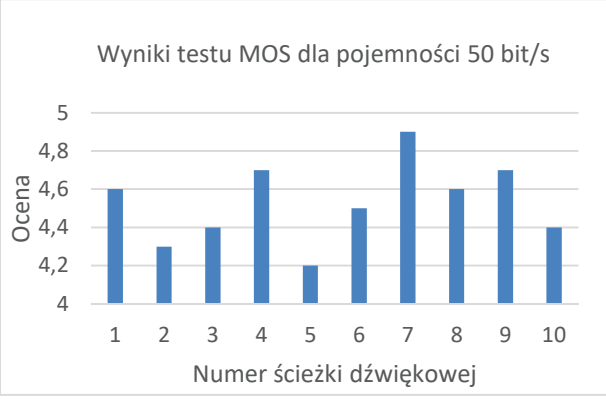
Tab. 1. Skala ocen dla testów subiektywnych

Ocena	*Zniekształcenia*	*Jakość*
5	Niesłyszalne	Doskonała
4	Słyszalne, nie dokuczliwe	Dobra
3	Lekko dokuczliwe	Średnia
2	Dokuczliwe	Słaba
1	Bardzo dokuczliwe	Zła

Subiektywne testy transparentności MOS (Mean Opinion Score) zostały przeprowadzone na 10 słuchaczach. Ich wyniki przedstawione zostały na rysunkach poniżej, kolejno dla pojemności znaku wodnego wynoszącego 25 bit/s oraz 50 bit/s.



Rys. 2. Wyniki testu MOS dla pojemności 25 bit/s



*Rys**Błąd! W dokumencie nie ma tekstu o podanym stylu.** 3. Wyniki testu MOS dla pojemności 50 bit/s*

Ponadto wykonano testy obiektywne, które polegają na wykorzystaniu metryk do oceny transparentności znaku wodnego. Istnieje kilka algorytmów, które w sposób zobiektywizowany obliczają estymowaną wartość oceny subiektywnej dla zadanych sygnałów audio. Do obiektywnej oceny transparentności sygnału audio wybrane zostały:

- SMR (Signal to Mask Ratio) - miara porównująca poziom sygnału użytecznego (informacja) do poziomu sygnału znaku wodnego (niepożądany sygnał). Jest definiowana jako stosunek mocy sygnału użytecznego do mocy znaku wodnego.
- PESQ (Perceptual Evaluation of Speech Quality) - porównuje przetworzony sygnał audio z oryginalnym sygnałem mowy (odtwarzanym po stronie „mówcy"), aby stworzyć w pełni obiektywny wskaźnik rzeczywiści słyszanego dźwięku [4].

Tab. 2. Wyniki obiektywnych testów transparentności dla pojemności 25 bit/s

Numer ścieżki dźwiękowej	*SMR*	*PESQ*
1	59,93	4,3
2	54,41	4,38
3	55,67	4,32
4	42,41	4,43
5	56,64	4,42
6	46,02	4,24
7	45,18	4,36
8	49,09	4,14
9	50,91	4,09
10	46,94	4,43

Wyniki testów obiektywnych uzyskane w wyniku obliczenia SNR oraz przeprowadzenia testu PESQ zostały przedstawione w tabeli 2 oraz 3, kolejno dla pojemności znaku wodnego 25 bit/s oraz 50 bit/s. Według standardu opracowanego przez IFPI znak wodny osadzony w sygnale akustycznym powinien być transparentny, jeśli poziom SNR wynosi powyżej 20 dB [2]. Uzyskane wyniki przekraczają co najmniej dwukrotnie tą granicę, co oznacza, że wspomniane zalecenia zostały spełnione.

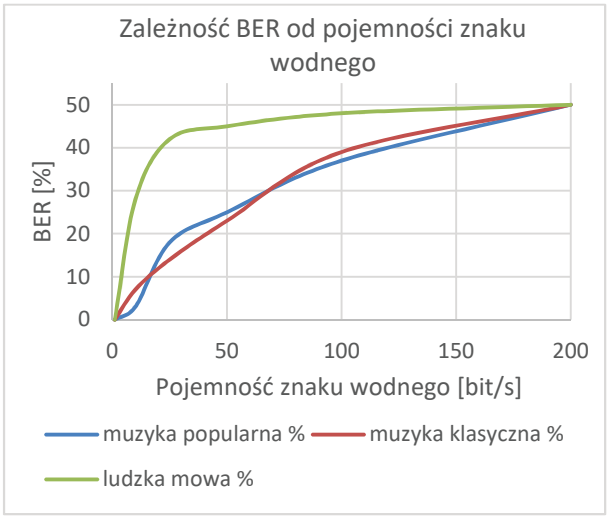
Tab. 3. Wyniki obiektywnych testów transparentności dla pojemności 50 bit/s

Numer ścieżki dźwiękowej	*SMR*	*PESQ*
1	60,2	4,41
2	54,37	4,38
3	56,39	4,37
4	42,91	4,47
5	56,85	4,44
6	46,34	4,35
7	45,86	4,43
8	48,76	4,35
9	49,76	4,37
10	45,96	4,45

Testy PESQ pozwalają na obiektywne odniesienie się do wyników uzyskanych w teście wierności odsłuchowej. W stosunku do testu MOS najlepszą oceną jest 4,5. Według skali ocen PESQ, oceny powyżej 3,8 są równoważne ocenie 5 w skali MOS. Oznacza to, że zniekształcenia są niesłyszalne, a jakość dźwięku doskonała.

W ramach przeprowadzonych testów pojemności binarnej zbadano, jak zmienia się bitowa stopa błędu (BER) w zależności od pojemności binarnej znaku wodnego dla wszystkich trzech typów sygnału dźwiękowe-

go. Badanie miało na celu ocenę, czy zwiększanie ilości informacji zawartych w znaku wodnym wpływa na ilość błędów pojawiających się podczas jego odczytu. Wyniki tych pomiarów zostały zilustrowane na rysunku poniżej.



Rys. 4. Zależność bitowej stopy błędu w funkcji pojemności znaku wodnego

Z rysunku wynika, że osadzenie znaku wodnego w muzyce klasycznej oraz muzyce popularnej powoduje podobny wzrost bitowej stopy błędów dla takich samych pojemności znaku wodnego. Z kolei dla sygnału mowy, bitowa stopa błędów rośnie w bardzo szybkim stopniu już od niskich pojemności. Wynika z tego, że algorytm pozwala na uzyskanie zadowalających wyników dla muzyki, lecz dla nie sprawdziłby się np. w audycjach radiowych. Przy wzroście pojemności znaku wodnego powyżej 150 bit/s, algorytm nie radzi sobie z prawidłowym dekodowaniem znaku wodnego. Bitowa stopa błędów wynosi ok. 50%, co pozwala na stwierdzenie, że wynik jest losowy, a nie uzyskany w wyniku konkretnego algorytmu ekstrakcji znaku wodnego.

4. WNIOSKI

Proponowana metoda znakowania wodnego sygnałów audio bazuje na klasycznym, numerycznym podejściu do osadzania znaków wodnych, wykorzystując transformacje DWT i DCT. Technika ta została poddana testom w zakresie transparentności percepcyjnej i pojemności binarnej. Badania przeprowadzone wśród grupy słuchaczy dziesięciu osób o słuchu normalnym, bez ubytków wykazały, że metoda osiąga dobre wyniki, w zakresie od doskonałej do dobrej jakości, z transparentnością w zakresie od niesłyszalnego zakłócenia (znaku wodnego) do słyszalnego, ale nieuciążliwego. W testach na transparentność, różne rodzaje dźwięków, takie jak muzyka klasyczna, popularna i sygnał mowy, osiągały podobne wyniki, z najniższymi ocenami dla muzyki klasycznej. Natomiast testy pojemności pokazały, że metoda umożliwia osadzenie znaku wodnego o większej pojemności dla muzyki klasycznej i popularnej niż dla mowy, osiągając około 15 bit/s, aby utrzymać satysfakcjonujący poziom błędów bitowych. Dla mowy odpowiedni próg wynosi około 5 bit/s. Wyniki te mogą być poprawione poprzez dalszy rozwój metody, w tym implementację kodów korekcyjnych. Metoda nie wymaga intensywnych obliczeń, co umożliwia znakowanie wodne w czasie rzeczywistym, oraz nie potrzebuje oryginalnego sygnału akustycznego po stronie odbiorczej podczas ekstrakcji znaku, co jest istotne dla praktycznych zastosowań.

Bardziej zaawansowaną perspektywę rozwoju stanowi przejście od transformat do metod wykorzystujących sieci neuronowe. Wymagają one dużej ilości danych i mocy obliczeniowej do procesu nauczania, co stanowi ich wadę. Plusem jest możliwość douczania sieci różnymi zbiorami danych, co pozwala na lepsze wyniki dla konkretnych typów dźwięku, np. sygnału mowy. Zaimplementowana metoda może posłużyć do porównania rezultatów uzyskanych w przyszłości przy użyciu sieci neuronowej.

LITERATURA

[1] BS.1284 : General Methods for the Subjective Assessment of Sound Quality.” n.d. Www.itu.int. Accessed May 7, 2024. https://www.itu.int/rec/R-REC-BS.1284/en.

[2] Li, Wei, Xiangyang Xue, and Peizhong Lu. "Localized Audio Watermarking Technique Robust against Time-Scale Modification." IEEE Transactions on Multimedia 8, no. 1 (February 2006): 60-69. doi: 10.1109/TMM.2005.861291.

[3] Saqib, Manasha, and Sameena Naaz. 2017. "Spatial and Frequency Domain Digital Image Watermarking Techniques for Copyright Protection." 9: 691-699.

[4] Torcoli, Matteo, Thorsten Kastner, and Jurgen Herre. 2021. “Objective Measures of Perceptual Audio Quality Reviewed: An Evaluation of Their Application Domain Dependence.” IEEE/ACM Transactions on Audio, Speech, and Language Processing 29: 1530–41. https://doi.org/10.1109/taslp.2021.3069302.

[5] Wav File Format – How to Open a Wav and Convert Wavs to MP3.” 2021. FreeCodeCamp.org. March 30, 2021. https://www.freecodecamp.org/news/wav-file-format-how-to-open-a-wav-and-convert-wavs-to-mp3/.

[6] Wu, Qiuling, and Meng Wu. 2018. “A Novel Robust Audio Watermarking Algorithm by Modifying the Average Amplitude in Transform Domain” 8 (5): 723–23. https://doi.org/10.3390/app8050723.

[7] 2024. Gstatic.com. 2024. https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcRNc25xgQe0w7_93WsvCRX7SnusA19EdSmx3A&s.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

Adaptacyjne Równoważenie Obciażenia dla Przyszłych Sieci Bezprzewodowych: Podejście Okresowe

Adaptive Load Balancing for Future Wireless Networks: A Periodic Approach

Mustafa Mohammed Hasan Alkalsh[1,2]; Adrian Kliks[1];

[1] Institute of Radiocommunications, Poznan University of Technology, Poznan
[2] Department of Mobile Networks, Nokia Solutions and Networks, Wroclaw

...

DOI: 10.15199/59.2024.4.74

Streszczenie: Nieproporcjonalny rozkład obciażenia spowodowany mobilnościa użytkowników to wyzwanie dla sieci bezprzewo-dowych, które nasila sie przy gestych wdrożeniach małych komórek obniżajacych efektywność sieci. Artykuł proponuje skalowalny algorytm zarzadzania mobilnościa (MM) z uwzglednieniem obciażenia do efektywnego zarzadzania zasobami sieci. Algorytm zbiera informacje o obciażeniu komórek i dynamicznie dostosowuje parametry przekazy-wania (HO) dla zrównoważonego ruchu. Badania te po-prawiaja strategie MM w przyszłych sieciach bezprzewo-dowych.

Abstract: Unbalanced load distribution caused by user mobility is a significant challenge faced by wireless networks. The issue is exacerbated with dense small-cell deployments which regress the network efficiency. This paper proposes a scalable Mobility Management (MM) algorithm with load awareness to manage the network resources efficiently. With a periodic execution, the algorithm gathers cell load information and dynamically adjusts Handover (HO) parameters for balanced traffic distribution. This research contributes to MM strategy improvements in future wireless networks.

Keywords: Small Cells, Future wireless Networks, Load-Aware, Mobility Management, Load Balancing.

1. INTRODUCTION

The growing demand for higher data rates to meet the minimal requirements of innovative technologies results in raising the importance of enhancing network capacity and efficiency. This causes more complex network infrastructure, especially when it comes to densely populated areas such as centers of cities, busy transportation hubs, or stadiums. The deployment of so-called small cells which are a crucial aspect of modern wireless networks, improved the network capacity significantly. Despite the effective benefits of small cells, they are exposed to several limitations due to their smaller coverage areas in comparison to traditional macrocells. However, the utilization of small cells with macro-cells forms a so-called heterogeneous network (HetNet), that leverages the efficiency of the network and user experience [1]. Macro cells provide basic network coverage and can reach several kilometers (km) due to their higher power. Thus, macro-cells are more useful in suburban and rural areas [2]. While in areas with highly dense populations, small cells can offload traffic from macro-cells by deploying them strategically within the macro-cells area forming HetNets (as shown in Figure 1). This deployment led to densifying the network as well as providing several benefits and supported a significantly higher number of users than networks with only macro-cells. However, densifying the network results in major challenges such as increasing overload states in some cells due to unbalanced load distribution. In contrast, other contiguous cells may have unutilized resources that could alleviate the load from those loaded cells. The overloaded state decreases the ability of cells to handle data traffic which leads to a degradation in data rates for connected users (UEs) and increases failure to admit new connections, drop calls, and connection difficulties [3]. Moreover, managing network resources in a network with a high small cell density due to the dynamic nature of user traffic. The challenge is compounded when implementing innovative technologies such as renewable energy-powered base stations (BS) or Unmanned Aerial Vehicles (UAVs) as base stations [4-5].

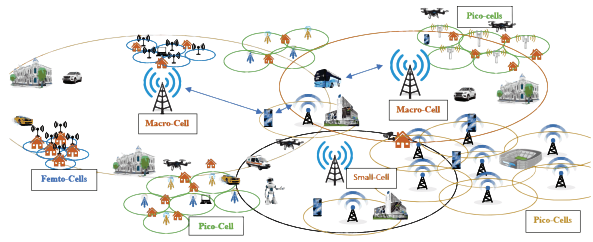


Figure 1: Deployment of Heterogeneous Network

This study proposes a periodic MM algorithm with load awareness to optimize network load in small cells. It achieves this by utilizing a periodically executed algorithm that gathers load information experienced by individual cells. Then, it adjusts network parameters dynamically based on this cell load to achieve a more balanced traffic distribution. This strategy utilizes available resources efficiently which reduces the duration of an overloaded state in a cell and improves overall network performance.

The rest of the paper is organized as follows. Section

2 presents current related work with a brief overview of existing load-balancing techniques, while Sections 3 and 4 provide the proposed algorithm and simulation setup respectively. Section 5 discusses the performance evaluation and results discussion before Section 6 which concludes the paper.

2. RELATED WORK

The increase in data traffic resulted from growth in the number of connected smart devices and applications overwhelming the wireless network. To cope with that HetNet is being utilized by deploying small cells to enhance the system capacity and coverage, especially in populated areas. However, small cell limitations and fluctuating traffic patterns in dynamic environments pose new load management challenges. To address them, researchers have proposed various solutions.

The authors in [6] introduce a Utility-based Mobility Load Balancing (UMLB) algorithm which considers both operator and user utilities for the HO process. With a proposal of a load balancing efficiency factor (LBEF) that prioritizes the overloaded cells for the UMLB algorithm operation, leading to a more balanced network. Next, in [7] the study explores the challenges of load balancing within the small cells of heterogeneous ultra-dense networks (UDNs). The authors address the complexities introduced by the dense deployment of small cells and the need for sophisticated algorithms that can adapt to changing network conditions periodically. The authors in [8] proposed an Enhanced Mobility Load Balancing Algorithm (EMLB) for 5G Small Cell Networks, by adjusting handover parameters according to the overload status of cells. This algorithm aims to improve the load distribution efficiently among neighboring cells. Authors in [9] introduced a novel energy-efficient load-balanced cluster assignment algorithm for small-cell networks. The algorithm showed improvements in requirements of load distribution and energy consumption. Authors in [10] proposed a predictive algorithm that uses machine learning to expect traffic patterns and adjust resources accordingly. This approach shows the potential of a proactive method to manage network load.

Recent works on load-balancing algorithms demonstrate ongoing efforts to improve network efficiency, user experiences, and sustainability in wireless networks. However, further research is needed to address the emerging challenges and develop adaptive techniques that enhance the network efficiency for future demands

3. PROPOSAL SOLUTION

The proposed system model employs a dynamic load-balancing algorithm that improves the network performance by managing the load and applying corrective measures when predefined thresholds are exceeded. The algorithm is sensitive to the load within individual sectors of each cell. The resources of each cell are analyzed to ensure that utilized resources remain below the predefined threshold ThresshT1. Furthermore, the HO process is a critical component of the load-balancing algorithm. based on 3GPP specification HO is triggered by the utilization of A3 events coupled with a Time to Trigger (TTT) parameter, which is used as a delay between the event detection and the HO initiation [11]. This delay is critical for preventing unnecessary HO due to temporary fluctuations in signal strength due to signal propagation combined with factors of path loss. Thus, within the TTT period, an average for signal strength is calculated. If a cell meets the A3 event conditions, it will be reported as a candidate of the target cell. The algorithm is activated when the traffic load in any cell reaches more than $Thressh_{T1}$ of utilized resources and then it's checked. This threshold is an important indicator of the cell's resources that are reserved by active UEs relative to the total available resources. When this limit is exceeded at the end of a specific period, the algorithm calculates the number of resources that need to be released. The resources are calculated according to the load of the neighboring cells according to the following equation

$$Release_{PRB} = (Load_{cell} - \beta) * Total_{PRB} \tag{1}$$

where $Load_{cell}$ is the current load of the cell that exceeded the threshold, $Total_{PRB}$ represents the total available resources in that cell, and β is the average load of the contiguous cell and calculated according to:

$$\beta = \frac{\sum_{i=1}^{n} load(i)}{n}, \tag{2}$$

where n is the number of neighboring cells in addition to the cell exceeding the predefined threshold, based on the released resources, the algorithm identifies the most loaded sector of the loaded cell. It performs the necessary calculations to determine the necessary offsets that will be used to redistribute the load. Each cell's coverage area is divided into three sectors representing 120 degrees. The algorithm iterates through the UEs in the identified sector to determine the UEs which handing them over to the neighboring cells would balance the load with the neighboring cells. According to conditions of A3 event, to trigger the HO process, the reference signal received power (RSRP) received by the UE from the source cell is compared with the RSRP of neighboring cells during TTT time to ensure the successful identification of target cell candidates. The calculated offsets are applied to that comparison at UEs located in the loaded sector to adjust their coverage areas and balance the load.

4. SIMULATION SETUP

To assess the effectiveness of the proposed algorithm in real-world situations, detailed simulations replicating real-life scenarios were conducted. It simulates a 10-minute duration, where each step represents one millisecond, providing a granular view of the network's behavior under various load conditions. The simulation is based on the city center of Poznan (as shown in Figure 2), chosen for its urban challenges and network demand during peak times with a total of 15 NR small cells deployed to simulate a 5G network with heterogeneous network deployment.

The movement of users within the simulation environment follows a non-uniform random walk model, potentially reflecting real-world movement patterns observed in urban environments. The key parameters used in the simulation are outlined in the Table. 1.

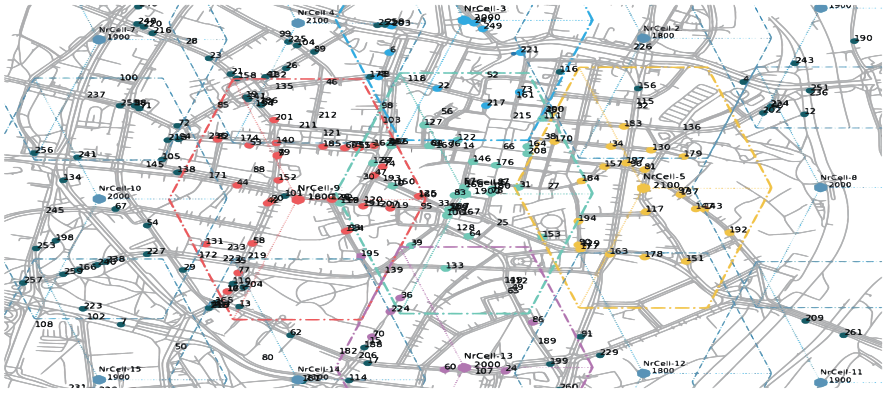


Figure 2: Simulated Poznan City Center Environment for Algorithm Evaluation

Table 1: Simulation Parameters

Parameter	*Value*	*Unit*
Carrier Freq.	[1.8 to 2.1]	[GHz]
Antenna Gain	12	[dBi]
Bandwidth	25	[MHz]
Antenna Height	10	[m]
UE Height	1.5	[m]
UE Speed	[3 to 20]	[km/h]
Shadowing STD	[4, 7.82]	[dB]
TTT values	[360, 450]	[ms]

5. PERFORMANCE EVALUATION AND RESULTS DISCUSSION

In this section, the performance evaluation for the proposed algorithm is conducted through comparative analysis across three identically configured environments: The first environment serves as a control without the proposed algorithm. The remaining environments implement the algorithm at 1-minute and 3-minute intervals, respectively, to explore the trade-off between responsiveness to rapid traffic changes (1 minute) and efficient handling of sustained loads (3 minutes). These environments include the same network configurations and user distributions, ensuring a controlled and repeatable evaluation. The load average of individual cells shown in Figure 3, demonstrates the challenge of load distribution in wireless networks.

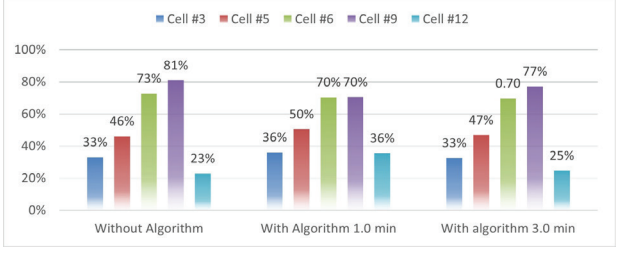


Figure 3: The Average load percentage of each tested cell across the environments

While some cells experience lower load percentages in the environment without the algorithm, other cells face significant congestion. The proposed load-balancing algorithm addresses this challenge. As observed in Figure 3, the cells with low load in the baseline environment may take on a higher value of average load in the environment with the proposed algorithm (1.0-minute and 3.0-minute intervals). Conversely, previously overloaded cells experience a decrease in their average load. This dynamic redistribution of traffic is crucial for achieving more efficient resource utilization and balanced load distribution across the network. This concept is further supported by the average overload state metric percentage. Cells in the overload state, where 100% of their resources are utilized by connected UEs, represent a temporary inefficiency in network resource utilization. As shown in Figure 4, in the environment without the proposed algorithm, the average overload percentage is 13%. This indicates inefficient utilization of the network resources. However, with the load-balancing algorithm active (both 1.0-minute and 3.0-minute intervals), the overall overload percentage drops to 7% and 8%, respectively.

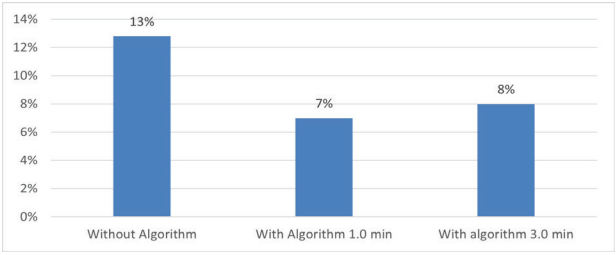


Figure 4: The Average Overload state metric

This represents a reduction of 46.2% and 38.5% in the overload state percentage compared to the baseline scenario. Furthermore, Figure 5 shows the overload state that occurred per individual tested cells among the simulated environments.

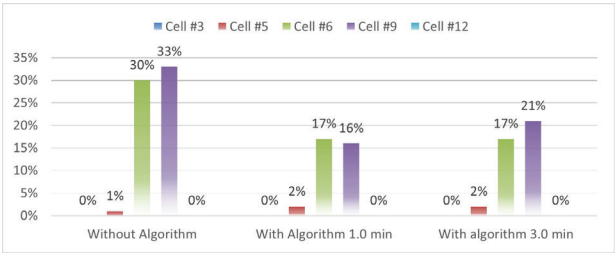


Figure 5: The Overload state percentage of each tested cell across the environments

It highlights the effectiveness of the algorithm in mitigating congestion across the network. Moreover, Figures 6-8 provide a comprehensive view of load distribution during the simulation across tested cells.

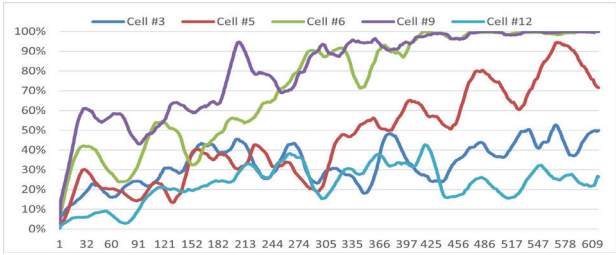


Figure 6: The Load distribution among tested cells for the base-line environment.

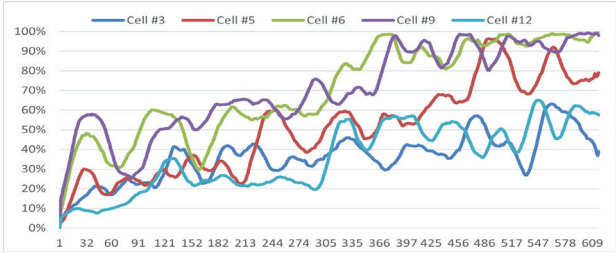


Figure 7: The Load distribution among tested cells for the environment with the proposed algorithm with an interval of 1.0 minute.

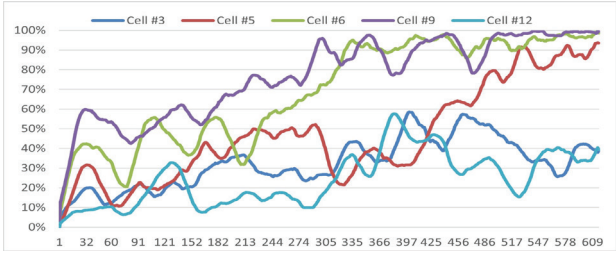


Figure 8: The Load distribution among tested cells for the environment with the proposed algorithm with an interval of 3.0 minutes.

6. CONCLUSION

This study investigated a load-balancing algorithm for wireless networks in a dense urban environment. The results demonstrate that the environment without the proposed algorithm suffered from insufficient resource utilization, and a high average of overload state among available cells. However, the proposed algorithm showed a significant reduction in overload state (up to 46.2%) through dynamic load distribution, demonstrating improved resource utilization and network efficiency. The periodic intervals suggest a potential trade-off for further exploration. Overall, the dynamic load balancing algorithm offers a promising solution to the challenges of modern network traffic management.

ACKNOWLEDGMENTS

The author would like to acknowledge the financial support provided by the MEiN Program "Doktorat Wdrożeniowy" (Agreement no. DWD/6/0057/2022, date: 23.12.2022) administered by the Ministry of Science and Higher Education, Republic of Poland. The author would also like to acknowledge the affiliation with Nokia Solutions and Networks about the implementation of the cooperation agreement for the "Doktorat Wdrożeniowy" program

REFERENCES

[1] Sathya, Vanlin, Srikant Manas Kala, and Kalpana Naidu. "Heterogenous networks: From small cells to 5G NR-U." Wireless Personal Communications 128, no. 4 (2023): 2779-2810.

[2] Rehman, Aziz Ur, Mardeni Bin Roslee, and Tiang Jun Jiat. "A survey of handover management in mobile HetNets: current challenges and future directions." Applied Sciences 13, no. 5 (2023): 3367.

[3] Alkalsh, Mustafa Mohammed Hasan. "An Overview of Mobility Management Mechanisms and the Related Challenges in 5G Networks and Beyond." Journal of Telecommunications and Information Technology 2 (2023): 53-62.

[4] Janji, Salim, and Adrian Kliks. "Drone Base Stations Transmission Power Control and Localization." In International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services, pp. 356-377. Cham: Springer Nature Switzerland, 2022.

[5] Samorzewski, Adam, and Adrian Kliks. "5G Networks Supported by UAVs, RESs, and RISs." In 2023 International Conference on Software, Telecommunications and Computer Networks (SoftCOM), pp. 1-6. IEEE, 2023.

[6] Addali, Khaled M., Suhib Younis Bani Melhem, Yaser Khamayseh, Zhenjiang Zhang, and Michel Kadoch. "Dynamic mobility load balancing for 5G small-cell networks based on utility functions." IEEE Access 7 (2019): 126998-127011.

[7] Xu, Yue, Wenjun Xu, Zhi Wang, Jiaru Lin, and Shuguang Cui. "Load balancing for ultradense networks: A deep reinforcement learning-based approach." IEEE Internet of Things Journal 6, no. 6 (2019): 9399-9412.

[8] Addali, Khaled, and Michel Kadoch. "Enhanced mobility load balancing algorithm for 5G small cell networks." In 2019 IEEE Canadian conference of electrical and computer engineering (CCECE), pp. 1-5. IEEE, 2019.

[9] Dey, B., Sivaji Bandyopadhyay, and Sukumar Nandi. "A novel energy efficient load balanced cluster assignment algorithm for IoT based sensor network." In AIP Conference Proceedings, vol. 2930, no. 1. AIP Publishing, 2023.

[10] Downey, Alexander, Evren Tuna, and Alkan Soysal. "An Adaptive Framework for Generalizing Network Traffic Prediction towards Uncertain Environments." arXiv preprint arXiv:2311.18824 (2023).

[11] 3GPP. "TS 36.331: Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification." 3rd Generation Partnership Project (3GPP) Technical Specification Group Radio Access Network (2021).

IDENTYFIKACJA OPERATORA NA PODSTAWIE WSKAŹNIKÓW JAKOŚCI SYGNAŁU RADIOWEGO – ZARYS PODEJŚCIA

MOBILE NETWORK OPERATOR IDENTIFICATION BASED ON RF SIGNAL QUALITY METRICS – APPROACH OUTLINE

Dariusz Zmysłowski[1]; Jan M. Kelner[2]

[1,2] Wojskowa Akademia Techniczna im. Jarosława Dąbrowskiego, Wydział Elektroniki, Instytut Systemów Łączności
ul. Gen. Sylwestra Kaliskiego 2, 00-908 Warszawa, {[1]dariusz.zmyslowski, [2]jan.kelner}@wat.edu.pl

DOI: 10.15199/59.2024.4.75

Streszczenie: **Ocena jakości usług QoS (*Quality of Service*) w sieciach 5G jest zwykle realizowana na podstawie wartości kluczowych wskaźników wydajności KPI (*Key Performance Indicator*) takich jak przepustowość lub opóźnienie. Parametry te można wyznaczać przy realizacji konkretnej usługi lub testów wydajnościowych, np. z wykorzystaniem aplikacji iPerf. Bardziej powszechnym rozwiązaniem jest wykorzystanie parametrów referencyjnego sygnału radiowego, np. RSRP, SINR, RSRQ, które mogą być wyznaczane w sposób ciągły, niezależnie od realizowanych usług. Autorzy we wcześniejszej pracy przedstawili analizę korelacyjną pomiędzy parametrami sygnałów radiowych i KPI usług. Na podstawie tych badań dla scenariusza iPerf, wyselekcjonowali dwa parametry RSRQ i SINR wyznaczane dla standardu 5G, które pozwalają jednoznacznie identyfikować operatora. W tym przypadku zastosowali regresję liniową dla danych pomiarowych. W tym artykule przedstawiono analogiczną metodę identyfikacji operatora dla scenariusza transmisji strumieni wideo realizowanych w serwisie YouTube wraz z propozycją wykorzystania regresji kwadratowej.**

Abstract: **Assessment of quality of service (QoS) in 5G networks is usually based on the key performance indicators (KPIs) such as throughput or latency. However, these parameters can be determined when implementing a specific service or performance test, e.g., using the iPerf application. A more common solution is to use the parameters of the reference radio signal, e.g., RSRP, SINR, or RSRQ, which can be determined continuously, regardless of the services provided. In our previous work, we presented a correlation analysis between radio signal parameters and service KPIs. Based on this research, for the iPerf scenario, we selected two parameters RSRQ and SINR determined for the 5G standard, which allow for the unambiguous identification of the mobile network operator (MNO). In this case, linear regression for the measurement data is used. In this paper, we present this method for video streaming scenario from the YouTube service, along with a proposal to use square regression.**

Słowa kluczowe: **jakość usług, 5G New Radio, sieć mobilna, reference signal received quality (RSRQ), signal-to-interference-plus-noise ratio (SINR), regresja.**

Keywords: **5G New Radio, mobile network, quality of service (QoS), reference signal received quality (RSRQ), signal-to-interference-plus-noise ratio (SINR), regression.**

1. WPROWADZENIE

W dobie dynamicznego rozwoju technologii 5G ocena jakości usług QoS (ang. *Quality of Service*) świadczonych przez poszczególnych operatorów sieci mobilnych MNO (ang. *Mobile Network Operator*) ma kluczowe znaczenie, szczególnie dla użytkowników końcowych oraz instytucji regulujących rynki telekomunikacyjne [2]. Istotne jest również ustalenie minimalnych, akceptowalnych poziomów odczuwalnej jakości QoE (ang. *Quality of Experience*) dla różnych aplikacji i serwisów obecnie oferowanych. MNO uczestnicząc w grze rynkowej o klienta końcowego coraz częściej bieżę pod uwagę parametry bieżące wartości kluczowych wskaźników wydajności KPI (ang. *Key Performance Indicator*) oraz statystyki je opisujące.

Regulator rynku może być zainteresowany pozyskaniem wiedzy o tym, który MNO świadczy usługi na danym poziomie QoS i QoE [2, 3], w związku z tym szybka identyfikacja takiego operatora spośród dużej ilości danych, w tym tych pozyskiwanych z testów sieci 5G jest dzisiaj potrzebą chwili [6]. Proces identyfikacji i selekcji MNO może się odbywać w oparciu o analizę korelacyjną pomiędzy metrykami QoS a parametrami referencyjnego sygnału radiowego takimi jak RSRP (ang. *Reference Signal Received Power*), SINR (ang. *Signal-to-Interference-plus-Noise Ratio*) czy RSRQ (ang. *Reference Signal Received Quality*). Analizę taką wykorzystującą regresję liniową przedstawiono [4,5]. Podobnie, korelacja QoS pomiędzy usługami i wymaganiami użytkowników może stanowić mechanizm oceny ich wzajemnej przystawalności [1].

Artykuł przedstawia zastosowanie korelacji opartej na regresji kwadratowej między parametrami sygnału referencyjnego takimi jak RSRQ i SINR do identyfikacji MNO sieci 5G oraz porównanie jej użycia z wykorzystaniem korelacji liniowej dla scenariuszy testowych iPerf i YouTube.

2. IDENTYFIKACJA OPERATORA NA PODSTAWIE REGRESJI LINIOWEJ

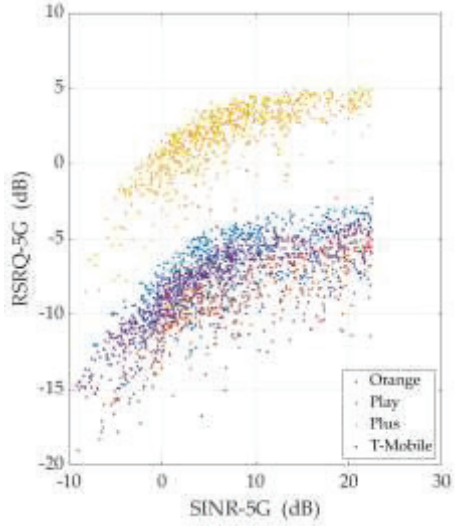
W [4] przeprowadzono analizę i ocenę związku między parametrami jakości odbieranego radiowego sygnału

referencyjnego (tj. RSRP, SINR, RSRQ) a metrykami QoS (tj. przepustowością i jitterem) transferowanych danych w sieci 5G dla scenariusza iPerf. Autorzy uzyskali wówczas obiecujące rezultaty wskazujące, że jest możliwe dokonanie identyfikacji MNO na podstawie przebiegu prostych regresji liniowej parametrów RSRQ i SINR dla sieci 5G.

Zastosowanie podobnego podejścia dla scenariusza YouTube nie daje już tak dobrej separacji prostych, które można byłoby wykorzystać dla identyfikacji MNO. Stąd też, uwzględniając charakter zmian danych pomiarowych RSRQ w funkcji wzrastającej wartości SINR, zaproponowano modyfikację metody bazującą na zastosowaniu regresji kwadratowej. Poniżej przedstawiono metodę regresji liniowej i kwadratowej parametrów RSRQ i SINR dla dwóch scenariuszy usługowych: iPerf (transfer danych) i YouTube (transmisja strumienia wideo). Na tej podstawie dokonano porównania błędów regresji i możliwości identyfikacji MNO za pomocą dwóch wspomnianych metod.

2.1. Scenariusz iPerf

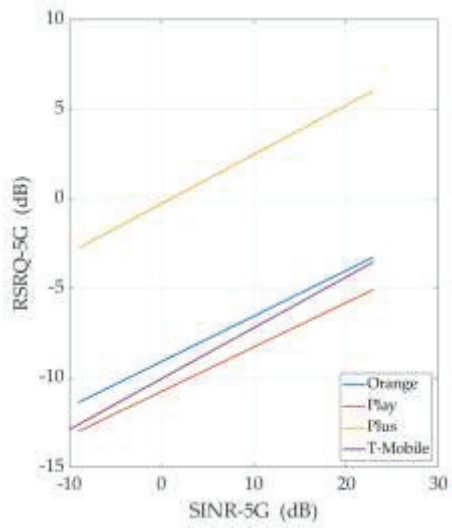
Pomiary z wykorzystaniem aplikacji iPerf zostały wykonane przez firmę Systemics-PAB w Warszawie, w dniach 23-27 września 2021 r. dla czterech MNO (Orange, T-Mobile, Play i Plus) podczas testów mobilnych (ang. *drive tests*) z użyciem sprzętu firm Rohde&Schwarz [4,5]. Kampania testowa trwała 33,5 h i była realizowana dla pasm częstotliwości: 2,1 GHz i 2,6 GHz. Testy miały na celu sprawdzenie stabilności i wydajności radiowego łącza transmisyjnego. Przeprowadzono je dla dwóch typów protokołów warstwy transportowej: TCP (ang. *Transmission Control Protocol*) i UDP (ang. *User Datagram Protocol*). Zebrane dane podzielono na dwa strumienie transmisji: w górę UL (ang. *uplink*) i w dół DL (ang. *downlink*). Zmierzono wartość przepustowości i ilość utraconych pakietów odpowiednio dla TCP i UDP oraz wartości jitter dla UDP, jak również wartości RSRP, SINR i RSRQ [4]. Przykładowe dane pomiarowe RSRQ w funkcji SINR dla protokołu UDP w łączu w dół zilustrowano na Rys 1.



Rys. 1. Dane pomiarowe RSRQ w funkcji SINR-5G dla protokołu UDP w łączu w dół, dla czterech operatorów w scenariuszu iPerf [4]

2.2. Metoda identyfikacji operatora na podstawie regresji liniowej

Badania jakości usług realizowane przez operatorów sieci mobilnych MNO i regulatorów rynków telekomunikacyjnych dostarczają dużych ilości danych, w tym wyników pomiarów realizowanych w ruchu tzw. *drive tests*. Analiza tych danych pozwala nie tylko na ocenę, która sieć mobilna oferuje lepsze usługi, ale daje również możliwości przeprowadzenia oceny statystycznej tych sieci pod kątem ich identyfikacji, co autorzy przedstawili w [4]. Dokonano tam analizy regresji liniowej pomiędzy KPI usług transferu danych a parametrami sygnałowymi dla scenariusza pomiarowego iPerf. Analiza prostych regresji dla RSRQ-SINR pokazała, że na ich podstawie można jednoznacznie identyfikować MNO, których dotyczyło badanie. Na rysunku 2 przedstawiono przykładowy wynik regresji dla danych pomiarowych zilustrowanych na rys. 1. Szersza analiza została omówiona w [4].

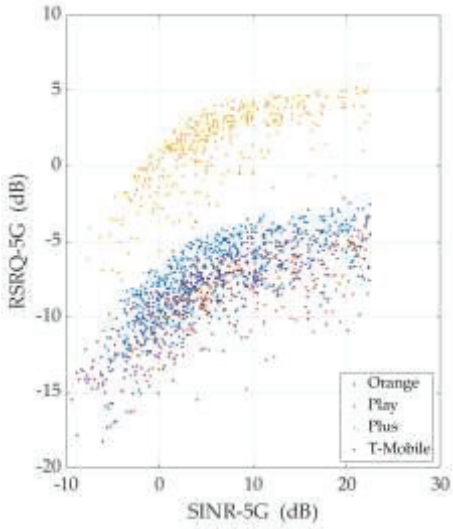


Rys. 2. Regresja liniowa dla danych iPerf przedstawionych na rys. 1 [4]

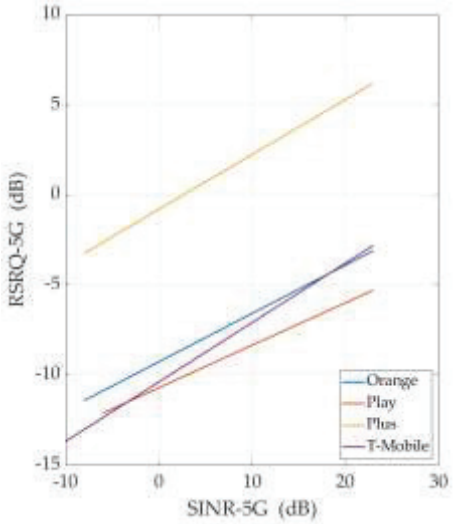
Bazując na rezultatach analizy można zauważyć, że oprócz identyfikacji operatorów można również wnioskować co do zasobów widmowych, którymi oni dysponują. Zaobserwowana duża separacja na poziomie sygnałowym pomiędzy Plus a pozostałymi operatorami odzwierciedla bowiem fakt dysponowania przez tego operatora większymi zasobami częstotliwości.

2.3. Zastosowanie metody dla scenariusza YouTube

Autorzy postanowili sprawdzić, czy zaproponowana metoda identyfikacji MNO jest uniwersalna, tj. czy jest niezależna od badanego scenariusza lub usługi. W tym artykule, weryfikowana jest uniwersalność tej metody w odniesieniu do usługi streamingu wideo z serwisu YouTube. Na rys. 3 przedstawiono dane pomiarowe RSRQ i SINR-5G dla takiego scenariusza. Natomiast rys. 4 ilustruje otrzymane proste regresji. W tym przypadku proste dla trzech MNO nachodzą na siebie, co uniemożliwia ich jednoznaczną identyfikację.



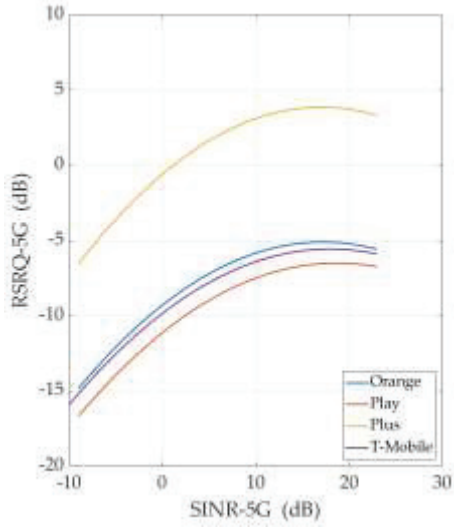
Rys. 3. Dane pomiarowe RSRQ w funkcji SINR-5G dla czterech operatorów w scenariuszu YouTube



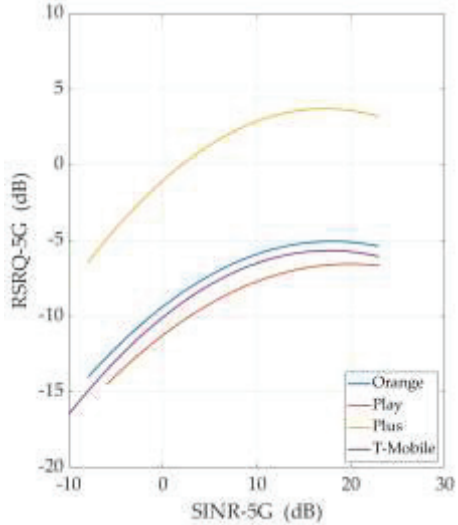
Rys. 4. Regresja liniowa dla danych YouTube przedstawionych na rys. 3

3. IDENTYFIKACJA OPERATORA NA PODSTAWIE REGRESJI KWADRATOWEJ

Rozmieszczenie danych pomiarowych na płaszczyźnie RSRQ-5G – SINR-5G (patrz rys. 1 i 3) pokazuje, że punkty pomiarowe dla poszczególnych MNO nie układają się wzdłuż prostej. Stąd też można wnioskować, że zastosowanie innej krzywej regresji może pozwolić na uzyskanie lepszego dopasowania. Dlatego też, autorzy proponują wykorzystanie regresji kwadratowej. Takie podejście powinno zapewnić lepsze odzwierciedlenie charakteru zmian danych pomiarowych oraz zapewnić mniejszy błąd aproksymacji. Bazując na tym podejściu, wyznaczono krzywe regresji (parabole) dla danych pomiarowych w scenariuszu iPerf (rys. 5) i YouTube (rys. 6).



Rys. 5. Regresja kwadratowa dla danych iPerf przedstawionych na rys. 1



Rys. 6. Regresja kwadratowa dla danych YouTube przedstawionych na rys. 3

Uzyskane rezultaty wskazują, że zastosowanie regresji kwadratowej w analizowanej płaszczyźnie RSRQ–SINR dla standardu 5G poprawia identyfikację operatorów oraz zwiększa zakres wartości RSRQ, dla których ona zachodzi.

Szczegółowe zestawienie wartości błędów aproksymacji wyrażonych wartościami odchylenia standardowego zawarto w Tab.1.

Tab. 1. Porównanie błędów aproksymacji danych pomiarowych dla czterech operatorów i scenariuszy iPerf i YouTube

Scenariusz	*iPerf*		*YouTube*	
Regresja	*liniowa*	*kwadratowa*	*liniowa*	*kwadratowa*
Orange	1.93 dB	1.69 dB	1.88 dB	1.73 dB
Play	2.13 dB	2.02 dB	1.84 dB	1.74 dB
Plus	1.96 dB	1.79 dB	2.20 dB	2.03 dB
T-Mobile	1.86 dB	1.64 dB	1.79 dB	1.55 dB

Dzięki zastosowaniu regresji kwadratowej uzyskuje się zmniejszenie błędów aproksymacji w stosunku do regresji liniowej, odpowiednio dla scenariusza:

- iPerf o: 12,44% (dla Orange), 5,16% (Play), 8,67% (Plus), 11,83% (T-Mobile);
- YouTube o: 7,98% (dla Orange), 5,44% (Play), 7,73% (Plus), 13,41% (T-Mobile).

4. PODSUMOWANIE

W artykule przedstawiono metodę identyfikacji operatora sieci 5G. bazującą na analizie korelacji liniowej i kwadratowej pomiędzy parametrami RSRQ i SINR, wyznaczanymi dla radiowych sygnałów referencyjnych odbieranych w trakcie trwania dwóch scenariuszu pomiarowych iPerf i YouTube. Analiza uzyskanych wyników wskazuje, że zastosowanie regresji kwadratowej zapewnia dokładniejszą aproksymację danych pomiarowych niż w regresja liniowej. Jednocześnie, regresja kwadratowa pozwala na jednoznaczną identyfikację operatora dla dwóch analizowanych scenariuszy, podczas gdy regresja liniowa umożliwia taką identyfikację tylko w scenariuszu iPerf.

Autorzy planują w przyszłości sprawdzenie opisanej metody dla innych scenariuszy testowych QoS w sieci 5G, np. przeglądanie stron www, usługi interakcyjne. Opisane metoda może znaleźć zastosowanie w szybkiej identyfikacji operatora w oparciu o duże zbiory danych pomiarowych, co może być przydatne dla regulatorów rynków telekomunikacyjnych, MNO czy analityków QoS w sieciach 5G.

PODZIĘKOWANIE

Autorzy pragną podziękować firmie Systemics-PAB Sp. z o.o. za udostępnienie danych z kampanii pomiarowej, które zostały wykorzystane w niniejszej publikacji.

Artykuł powstał w ramach uczelnianego grantu badawczego nr UGB/22-748/2024/WAT finansowanego przez Wojskową Akademię Techniczną.

LITERATURA

[1] Li D., D. Ye, N. Gao, S. Wang. 2019. "Service selection with QoS correlations in distributed service-based systems". *IEEE Access*, 7: 88718–88732.

[2] Mongay Batalla, J., *et al.* On studying active radio measurements estimating the mobile network quality of service for the Regulatory Authority's purposes. *Computer Networks*, 2023, 235: 109980.

[3] Mongay Batalla J., S. Sujecki, J. Oko, J.M. Kelner. 2022. "Cost-effective measurements of 5G radio resources allocation for telecom market regulator's monitoring". *2022 19th ACM PE-WASUN*, 83–90.

[4] Zmysłowski, D.; Kelner, J. M. Mobile Network Operators' Assessment Based on Drive-Test Campaign in Urban Area for iPerf Scenario. *Applied Sciences*, 2024, 14(3): 1268.

[5] Zmysłowski D., J.M. Kelner. 2022. Drive test-based correlation assessment of QoS parameters for exemplary measurements scenario in suburban environment. *2022 18th WEBIST*, 485–494.

[6] Zmysłowski D., J.M. Kelner. 2022. The concept of application of the Wroclaw taxonomy for QoS assessment in mobile networks. *2022 18th WEBIST*, 485–494.

KRiT 2024
01001011 01010010 01101001 01010100

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

WYKORZYSTANIE INFORMACJI O ROZMIESZCZENIU BUDYNKÓW DO ZARZĄDZANIA ZASOBAMI KOMUNIKACJI BEZPOŚREDNIEJ

MAP-BASED RESOURCE MANAGEMENT METHOD FOR DEVICE-TO-DEVICE COMMUNICATIONS

Marcin Rodziewicz[1]

[1] Politechnika Poznańska, Instytut Radiokomunikacji, Poznań,
email marcin.rodziewicz@put.poznan.pl

DOI: 10.15199/59.2024.4.76

Streszczenie: **W artykule rozpatrywany jest system, w którym umożliwiono komunikację bezpośrednią przy współdzieleniu zasobów łącza w górę z użytkownikami komórkowymi. W rezultacie zarówno użytkownicy komunikacji bezpośredniej, jak i użytkownicy komórkowi doświadczają dodatkowych zakłóceń. W celu zmniejszenia ich wpływu na działanie systemu przedstawiono metodę alokacji zasobów radiowych dla urządzeń komunikacji bezpośredniej wykorzystującą informację kontekstową.**

Abstract: **The paper considers a system that allows direct communication while sharing uplink resources with cellular users. As a result, both types of users experience additional interference in the system. In order to reduce their impact on system performance, a method for allocating radio resources for direct communication devices using contextual information is presented.**

Słowa kluczowe: Komunikacja bezpośrednia, Świadomość otoczenia, Rozdział zasobów, System komórkowy.

Keywords: Device-to-Device, Context-awareness, Resource allocation, Cellular system.

1. WSTĘP

W ostatnich latach komunikacja bezpośrednia (D2D, ang. *Device-to-Device*) działająca jako podsystem w sieciach komórkowych przyciągnęła wiele uwagi społeczności badawczej. Uważa się, że wprowadzenie tego typu komunikacji może poprawić wydajność widmową systemu, a także zmniejszyć zużycie energii, a tym samym poprawić efektywność energetyczną. Ponadto nowy rodzaj komunikacji może umożliwić nowe rodzaje usług multimedialnych [4, 8, 16].

Działając jako podsystem sieci komórkowej, komunikacja D2D może wykorzystywać te same zasoby radiowe co użytkownicy systemu komórkowego. Pozwala to na zwiększenie współczynnika ponownego wykorzystania częstotliwości (FRF, ang. *Frequency Reuse Factor*) nawet powyżej 1 [4]. Jednak umożliwienie komunikacji D2D, która współdzieli widmo z systemem komórkowym, stawia nowe wyzwania. Jednym z nich jest wprowadzenie nowego rodzaju zakłóceń wewnątrz komórkowych, których wpływ należy minimalizować, zazwyczaj za pomocą mechanizmów kontroli mocy lub alokacji zasobów. Dzięki czemu można poprawić wydajność sieci pod względem widmowym i energetycznym.

W niniejszym artykule uwaga skupiona jest na rozwiązania łagodzące zakłócenia oparte na alokacji zasobów. Przegląd literatury skupiającej się na tym aspekcie można znaleźć w [5]. W wielu rozwiązaniach wykorzystywane są wolno zmienne parametry kanału, takie jak straty na ścieżce lub zacienienie (ang. *shadowing*) [7]. Niektóre prace badają możliwość wykorzystania informacji o lokalizacji do alokacji zasobów i wyboru użytkowników współdzielących [2, 3, 11, 14, 15, 17]. Dla przykładu w pracach [2] i [11], zaproponowano mechanizmu kontroli mocy oraz strategię kontroli zakłóceń wykorzystującą ograniczenie obszaru do współdzielenia (ILA, ang. *Interference Limited Area*). Ograniczenie obszaru współdzielenia ma za zadanie zapewnić, że prawdopodobieństwo blokady komunikacji D2D spowodowane zakłóceniami ze strony użytkowników komórkowych jest mniejsze niż z góry określony próg. W pracy [17] wprowadzono kryterium współdzielenia zasobów z ograniczeniem odległości, tak aby ograniczyć zestaw użytkowników komórkowych, którzy mogą współdzielić zasoby z użytkownikami D2D. Skutkiem jest zmniejszenie prawdopodobieństwa blokady łącza D2D. Dodatkową zaletą proponowanego w [17] rozwiązania jest to, że nie wymaga od użytkowników komórkowych zmniejszenia mocy transmisji, dzięki czemu unika się pogorszenia jakości łącza komórkowego.

Komunikacja D2D jest często rozpatrywana w kontekście komunikacji między pojazdami lub między pojazdami a infrastrukturą. W pracy [10] przestawiono schemat alokacji zasobów dla komunikacji D2D, które zapewnia proporcjonalną sprawiedliwość między użytkownikami przy jednoczesnej maksymalizacji ogólnej przepustowości systemu. Proponowana metoda wykorzystuje alokacje w długich szczelinach czasowych w celu zwiększenia wydajności systemu i sprawiedliwości. W [9] przedstawiono metodę alokacji zasobów w scenariuszach komunikacji V2X (ang. *Vehicle-to-Everything*) opartych na D2D, biorąc pod uwagę realistyczne założenie niedoskonałej informacji o stanie kanału (CSI, ang. *Channel State Information*). Proponowany algorytm ma na celu maksymalizację sumarycznej pojemności ergodycznej urządzeń użytkownika znajdujących się w pojeździe przy jednoczesnym spełnieniu wymagań dotyczących jakości usług.

2. MODEL SYSTEMU

W tym artykule rozważany jest system komórkowy o współczynniku FRF wynoszącym 1, w którym umożliwiono działanie komunikacji D2D przy współdzieleniu zasobów łącza w górę z użytkownikami komórkowymi (CUE, ang. *Cellular User Equipment*). Współdzielenie wprowadza dodatkowe zakłócenia w systemie zarówno dla komunikacji komórkowej jak i bezpośredniej. Zakłócenia w kierunku odbierającego urządzenia D2D (DUE, ang. *Device-to-Device User Equipment*) są powodowane, gdy użytkownicy komórkowi nadają na tych samych zasobach radiowych. Z drugiej strony, z perspektywy CUE, zakłócenia są powodowane przez nadające DUE i są odbierane przez obsługującą stację bazową (BS, ang.*Base Station*). Stosunek sygnału do zakłóceń (SIR, ang. *Signal to Interference Ratio*) dla odbiornika DUE (γ_D) i stacji bazowej (BS) dla k-tego użytkownika komórkowego (γ_{C_k}) są dane przez:

$$\gamma_D = \frac{h_D(d)P_D}{\sum_{i=1}^{N} h_{DC_i}(d)P_{C_i}} \tag{1}$$

i

$$\gamma_{C_k} = \frac{h_{C_k}(d)P_{C_k}}{\sum_{i=1, i\neq k}^{N} h_{C_i}(d)P_{C_i} + h_D(d)P_D} \tag{2}$$

gdzie N to liczba sąsiednich komórek korzystających z tej samej częstotliwości, w tym komórki, w której znajduje się para D2D. Parametry $h_D(d)$ i $h_{DC_i}(d)$ to zależne od odległości straty na ścieżce między użytkownikami D2D oraz między odbiornikiem DUE, a i-tym nadajnikiem CUE, $h_{C_k}(d)$, $h_{C_i}(d)$ i $h_D(d)$ to odpowiednio straty na ścieżce między nadajnikiem komórkowym k i i a stacją bazową oraz na ścieżce między nadajnikiem DUE a stacją bazową. P_D to moc sygnału nadawanego przez DUE, a P_{C_i} to moc sygnału nadawanego przez i-ty nadajnik CUE. W rozważanym systemie do ustawień mocy transmisji wykorzystywany jest mechanizm sterowania w pętli otwartej (OLPC, ang. *Open Loop Power Control*):

$$P_C = \min\left(P_0 + Ah(d), P_{max}\right) \tag{3}$$

gdzie P_0 to początkowy poziom mocy urządzenia, A to współczynnik kompensacji strat na ścieżce, a $h(d)$ to strata na ścieżce między nadajnikiem a odbiornikiem. Maksymalna moc nadawania jest ograniczona przez P_{max}

3. ZARZĄDZANIE ZASOBAMI KOMUNIKACJI D2D

3.1 Założenia

Przy tworzeniu metody zarządzania zasobami dla komunikacji D2D, przyjęto szereg założeń. Po pierwsze celem proponowanego rozwiązania jest zminimalizowanie wpływu komunikacji komórkowej na komunikację bezpośrednio, założenie to wynika z tego, że stacja bazowa ma większe możliwości przetwarzania pozwalające na bardziej zaawansowane mechanizmy minimalizacji wpływu interferencji pochodzących od komunikacji bezpośredniej. Drugie założenie, to centralizacja rozwiązania, oznacza to, że w systemie wprowadzany jest węzeł sterujący, który jest powiązany z pewnym zbiorem stacji bazowych obsługujących dany obszar i posiada wiedzę o lokalizacji urządzeń, które obsługuje oraz o rozmieszczeniu budynków na obsługiwanym obszarze. Rozdział zasobów dla urządzeń D2D jest realizowany nad rozdziałem zasobów dla urządzeń CUE i jest zależny od niego. Mechanizmy bazują na wyborze kandydatów do współdzielenia zasobów spośród urządzeń, które są zaplanowane do transmisji w danych momentach transmisyjnych.

W badaniach wzięto pod uwagę dwa sposoby podejścia do problemu rozdziału zasobów, w obu celem jest minimalizacja interferencji w odbiorniku D2D. Pierwsze podejście działa z wykorzystaniem pomiarów i raportowania. Takie rozwiązanie wymaga wiedzy o stanie kanału pomiędzy wszystkimi węzłami w systemie, nie tylko pomiędzy użytkownikami a stacją bazową, co sprawia, że nakład sygnalizacyjny będzie znaczący. Z tego względu jest to rozwiązanie niepraktyczne ale służące jako punkt odniesienia dla drugiego rozważanego sposobu, w którym do zarządzania zasobami wykorzystywana jest informacja kontekstowa. Rozpatrywane są dwie metody, pierwsza bazująca tylko na informacji o pozycjach użytkowników w sieci komórkowej, oraz druga, która oprócz pozycji wykorzystuje wiedzę o rozmieszczeniu budynków w obsługiwanym obszarze.

3.2 Rozpatrywane rozwiązania

We wszystkich przedstawionych rozwiązaniach procedura alokacji zasobów jest taka sama. Różnica jest jedynie w sposobie odnajdywania kandydatów do współdzielenia zasobów.

Dla każdego odstępu transmisyjnego (TTI, ang. *Transmission Time Interval*) rozdzielane są najpierw zasoby dla użytkowników komórkowych w każdej stacji bazowej podłączonej do jednostki sterującej D2D. Następnie wybieramy jest zbiór par D2D zgodnie z algorytmem karuzelowym (ang. *Round Robin*). Liczebność tego zbioru zależy od długości rozdziału dla użytkowników CUE w stacjach bazowych. W dalszej kolejności dla każdej pary D2D tworzona jest sortowana lista kandydatów do współdzielenia z każdej stacji bazowej. W zależności od przyjętego podejścia lista ta jest tworzona na różne sposoby:

A. Metoda wykorzystująca raportowanie stanu kanału

   W tej metodzie zakłada się znajomość strat w kanale oraz spodziewanych mocy sygnałów nadawanych. Na tej podstawie wyznaczany jest poziom interferencji pomiędzy urządzeniami, konkretnie interferencja od kandydata do współdzielenia CUE do odbiornika D2D. Wyznaczony poziom interferencji wykorzystywany jest do sortowania, w sposób rosnący, listy kandydatów do współdzielenia zasobów.

B. Metoda wykorzystująca położenie

   W tej metodzie zakładana jest znajomość pozycji urządzeń końcowych, na podstawie których wyznaczane są odległości między nimi. W metodzie zakładana jest maksymalizacji odległości pomiędzy urządzeniami ponieważ jej maksymalizowanie potencjalnie minimalizuje interferencje. Rozpatrywane są cztery odległości: odległość między urządzeniami D2D, odległość nadajnika D2D do stacji bazowej, odległość kandydata do współdzielenia zasobów CUE do odbiornika D2D oraz odległość tego kandydata do obsługującej go stacji bazowej. Te odległości są sprawdzane dla każdego zaplanowanego CUE z każdej stacji bazowej, jeżeli odległość nadajnika D2D do stacji bazowej jest mniejsza od odległości D2D oraz odległość kandydata do współdzielenia do odbiornika jest większa niż

połowa jego odległości do jego stacji bazowej, to taki kandydat dodawany jest do listy. Po odnalezieniu wszystkich kandydatów, którzy spełniają te warunki lista jest sortowana względem odległości kandydata do odbiornika D2D w sposób malejący.

C. Metoda wykorzystująca położenie i rozmieszczenie budynków

W tej metodzie, będącej rozszerzeniem metody lokalizacyjnej, oprócz informacji o położeniu użytkowników wykorzystywana jest wiedza o rozmieszczeniu przeszkód w rozpatrywanym obszarze. W tym przypadku warunek dodania danego CUE do listy kandydatów do współdzielenia zasobów jest dodatkowo ograniczony warunkami widoczności. W metodzie zakłada się, że tylko kandydaci bez bezpośredniej ścieżki widoczności są dodawaniu do listy. Podobnie jak w przypadku metody lokalizacyjnej po rozpatrzeniu wszystkich kandydatów powstała lista jest sortowana względem odległości kandydata CUE do odbiornika D2D.

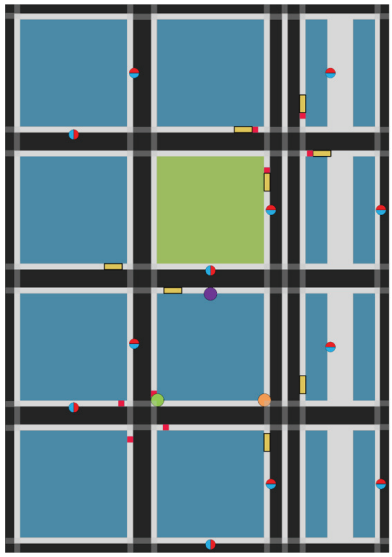
Listy, które powstają w ramach wspomnianych metod są w kolejnym kroku procedury alokacji wykorzystywane do wyboru konkretnych urządzeń CUE dla rozpatrywanych w rundzie alokacji par D2D. Kolejność rozpatrywania par D2D jest ustalana na podstawie odległości między urządzeniami w parze. W pierwszej kolejności przydzielane są zasoby do współdzielenia parom o największej odległości między urządzeniami. Jest to podyktowane tym, że prawdopodobieństwo niskiego SIR odbiornika w takiej parze, ze względu na większe tłumienie na ścieżce między parą, jest większe niż w przypadku par o małej odległości między urządzeniami. Może się zdarzyć, że ten sam kandydat do współdzielenia przypada dla wielu par D2D, w takim przypadku zastosowano prosty mechanizm rozwiązywania konfliktów polegający na wybieraniu kolejnej pozycji z listy kandydatów. Jeżeli lista kandydatów zostanie wyczerpana to taka para D2D jest wykluczana z rozpatrywanej rundy przydziału zasobów. Po przydzieleniu wszystkich kandydatów do zaplanowanych par D2D, ostatnim etapem procedury jest zmiana pierwotnego planu transmisji CUE tak, aby kandydaci w każdej parze korzystali z tych samych zasobów.

4. SYMULACJE I WYNIKI

4.1 Środowisko symulacyjne

Proponowane w sekcji 3 rozwiązania alokacji zasobów zbadane zostały za pomocą symulacji systemowych sieci komórkowej o współczynniku FRF wynoszącym jeden i wykorzystującej OFDMA (ang. *Orthogonal Frequency Division Multiple Access*) jako technikę wielodostępu. Narzędzie wykorzystane w symulacjach było współtworzone przez autora i zostało zaimplementowane zgodnie z wytycznymi ustanowionymi w projekcie METIS [12]. Więcej szczegółów na temat narzędzia można znaleźć w [1].

W narzędziu symulacyjnym zaimplementowano modele kanałów zdefiniowane przez METIS [13], które są używane dla użytkowników komórkowych. Modele te, w przeciwieństwie do najczęściej stosowanych modeli, wykorzystują oparte na mapach 3D metody czasu rzeczywistego do oceny warunków widoczności między węzłami. Dla użytkowników D2D stosowana jest zmodyfikowana wersja modelu D2D zdefiniowanego przez ITU-R [6].



Rysunek 1: Model siatki madryckiej

Modyfikacja polega na wykorzystaniu mapy, zamiast podejścia statystycznego, do wyznaczania warunków widoczności pomiędzy urządzeniami.

W badaniach rozpatrywana jest sieć komórkowa (rys. 1) rozmieszczona w terenie miejskim zgodnie z modelem siatki madryckiej (ang. *Madrid Grid Model*) [12]. Składa się ona ze stacji bazowej typu makro działającej w trybie dupleksu z podziałem częstotliwości w 3 sektorach, każdy sektor działa na częstotliwości nośnej 2,6 MHz z wykorzystaniem 80 MHz pasma. Do rozdziału zasobów użytkownikom CUE wykorzystano metodę Round Robin gdzie w danej rundzie kolejne bloki zasobów przydzielane są kolejnym urządzeniom CUE.

W scenariuszu symulacyjnym, 800 użytkowników zostało równomiernie rozmieszczonych na zewnątrz budynków, na chodnikach lub w samochodach. Wśród tych 800 użytkowników znajdowało się 80 par D2D, przy czym odległość między każdym urządzeniem z pary D2D było losowane z zakresu od 0 do 50 m zgodnie z rozkładem równomiernym. Spośród wszystkich par, 30 to pary składające się z użytkowników pieszych, a pozostałe 50 par D2D rozmieszczono w samochodach (odległość między użytkownikami w samochodach jest ograniczona zakładanymi wymiarami pojazdu). Mobilność użytkowników (w tym samochodów) jest również modelowana zgodnie z wytycznymi METIS. Symulacje były powtarzane pięćdziesięciokrotnie i każda trwała 10 s.

4.2 Wyniki

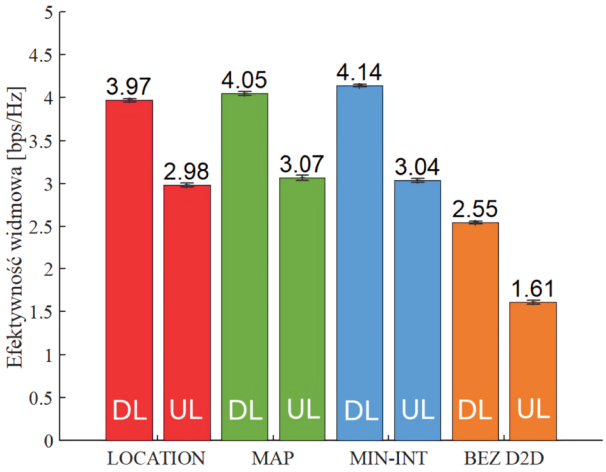
Pierwszy zbiór wyników (rys. 2) przedstawia ogólną efektywność widmową w systemie dla łącza w dół (lewy słupek) i łącza w górę (prawy słupek) dla każdej z przedstawianych metod alokacji zasobów. Dodatkowo dla odniesienia dodano wyniki symulacji, gdzie komunikacja bezpośrednia nie była dozwolona (oznaczenie BEZ D2D). Analizując wykres, skupiając się na wynikach dla łącza w dół, można zauważyć, że metoda referencyjna (MIN-INT), oparta na pomiarach, osiąga najlepsze rezultaty. Druga jest metoda wykorzystująca mapy (MAP), a trzecia metoda tylko lokalizacyjna (LOCATION). Jednak, warto zaznaczyć, że różnice między nimi są rzędu kilku procent. Analizując wyniki względem łącza w górę widzimy, że różnice są jeszcze mniejsze a najlepiej wypada me-

toda wykorzystująca mapę. Porównując wszystkie wyniki z przypadkiem bez komunikacji bezpośredniej wyraźnie widać korzyści wynikające z wprowadzenia komunikacji D2D. Wzrost efektywności widmowej systemu wynika z odciążenia sieci i zwiększenia współczynnika wykorzystania pasma ponad jeden.

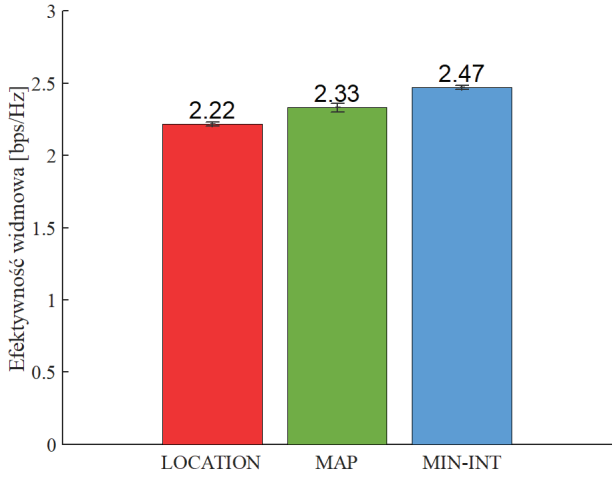
W ramach badań zebrano również wyniki dotyczące efektywności widmowej urządzeń mogących tworzyć pary D2D (rys. 3). Na tym wykresie możemy zauważyć podobny trend jak na wykresie z rys. 2. Jednak w tym przypadku różnice różnice są tutaj większe i wynoszą maksymalnie około 12%. Jak poprzednio lokalizacyjna metoda wypada najgorzej, natomiast możemy jej działanie poprawić dzięki uwzględnieniu map. Metoda z pomiarem i raportowaniem stanu kanału pozwala osiągnąć najlepsze rezultaty ponieważ ma najdokładniejszą wiedzę o warunkach transmisyjnych, ale tak jak zostało to wcześniej wspomniane, prowadzenie pomiarów i raportowanie stanu kanału pomiędzy wszystkimi urządzeniami w sieci wiązałoby się ze zbyt dużym narzutem sygnalizacyjnym.

Ocenie poddano także jaki wpływ na stosunek sygnału do interferencji w stacji bazowej ma komunikacja D2D (rys. 4). Wykres przedstawia zestaw dystrybuant SIR w stacji bazowej. Różnice pomiędzy rozpatrywanymi przypadkami nie są bardzo wyraźne, mediana SIR wynosi kolejno, zgodnie z legendą, 22,1, 22,5, 19 i 25,6 dB. Można zauważyć, że metoda z raportowaniem kanału ma największy wpływ na SIR w stacji bazowej. Wynika to głównie z tego, że nie było dodatkowych ograniczeń w postaci odległości czy widoczności przy dodawaniu urządzeń do listy kandydatów. Widzimy też, że metoda lokalizacyjna i ta wykorzystująca mapy mają bardzo podobny wpływ na działanie stacji bazowej.

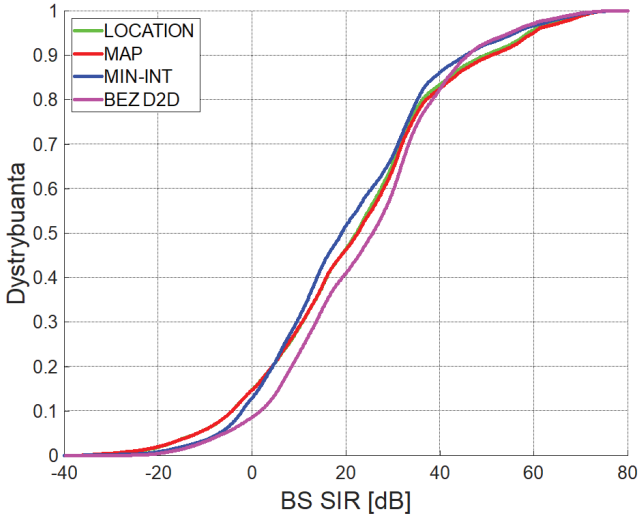
Podobnie można poddać analizie SIR po stronie urządzeń będących w parach D2D (rys. 5). Można zauważyć, że wszystkie 3 metody w podobnym stopniu chronią komunikację D2D, przy czym metoda lokalizacyjna nieznacznie ustępuje pozostałym w przedziale od mediany do 90 percentyla. Dla przykładu, różnica w medianie względem metody z mapą, wynosi około 2dB (30,25 dB dla metody z mapą i 28,36 dB dla metody lokalizacyjnej). Ta różnica może wynikać z faktu nałożenia dodatkowego zabezpieczenia w postaci warunków widoczności w metodzie wykorzystującej mapę. Takie podejście pozwala zwiększyć SIR ale nie musi konieczne oznaczać lepszego działania systemu ponieważ takie ograniczenie może się wiązać z mniejszą liczbą okazji transmisyjnych.



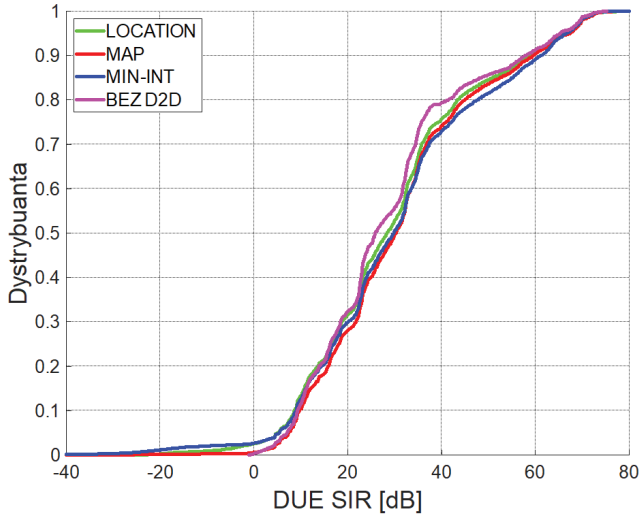
Rysunek 2: Efektywność widmowa całego systemu



Rysunek 3: Efektywność widmowa urządzeń D2D



Rysunek 4: SIR stacji bazowej



Rysunek 5: SIR odbiornika D2D

5. PODSUMOWANIE

W artykule przedstawiono metodę alokacji zasobów wykorzystującą informacje kontekstowe w postaci położenia użytkowników oraz przeszkód. Porównano ją z metodą wyłącznie lokalizacyjną zaproponowaną między innymi w [14] oraz metodą referencyjną działającą z wykorzystaniem pomiarów i raportowania stanu kanału. Pokazano, że metody wykorzystujące informacje kontekstową z powodzeniem mogą być wykorzystywane do alokacji zasobów dla komunikacji bezpośredniej.

Dodatkowo pokazano, że wprowadzenie komunikacji bezpośredniej w systemie komórkowym niesie za sobą szereg korzyści, np. zwiększenie w efektywności widmowej, a jej wpływ na działanie systemu komórkowego okazał się w badanych scenariuszach nieznaczny.

LITERATURA

[1] K. Bakowski, K. Wesolowski, M. Rodziewicz. 2014. „Simulation Tools for the Evaluation of Radio Interface Technologies for IMT-Advanced and Beyond". Simulation Technologies in Networking and Communications: Selecting the Best Tool for the Test. s. 365-391, CRC Press.

[2] P. Bao, G. Yu. 2012. „An interference management strategy for device-to-device underlaying cellular networks with partial location information". Proc. Personal Indoor and Mobile Radio Communications (PIMRC12). pp. 465-470.

[3] X. Chen, L. Chen, M. Zeng, X. Zhang, D. Yang. 2012. „Downlink resource allocation for Device-to-Device communication underlaying cellular networks". Proc. IEEE 23rd International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), pp.232-237.

[4] K. Doppler, M. Rinne, C. Wijting, C. Riberio, K. Hugl. 2009. „Device-to device communication as an underlay to LTE-advanced networks". IEEE Commun. Mag. , vol. 47, pp. 42-49.

[5] Islam, T., Kwon, C. 2022. „Survey on the state-of-the-art in device-to-device communication: A resource allocation perspective". Ad Hoc Networks, 136, 102978.

[6] ITU-R Recommendation P.1411. 2013 „Propagation data and prediction methods for the planning of short-range outdoor radiocommunication systems and radio local area networks in the frequency range 300 MHz to 100 GHz".

[7] P. Janis, V. Koivunen, C. Ribeiro, J. Korhonen, K. Doppler, K. Hugl. 2009. „Interference-aware resource allocation for device-to-device radio underlaying cellular networks". IEEE 69th Vehicular Technology Conference Spring (VTC Spring'09). pp. 1-5.

[8] P. Janis, C.-H. Yu, K. Doppler, C. Ribeiro, C. Wijting, K. Hugl, O. Tirkkonen, and V. Koivunen. 2009. „Device-to-Device communication underlaying cellular communications systems". International J. Commun. Network Syst. Sciences, vol. 2, no. 15.

[9] X. Li, L. Ma, Y. Xu, R. Shankaran. 2020. „Resource allocation for D2D-based V2X communication with imperfect CSI". IEEE Internet of Things Journal, 7(4), 3545-3558.

[10] X. Li, R. Shankaran, M.A. Orgun, G. Fang, Y. Xu 2018. „Resource allocation for underlay D2D communication with proportional fairnesss". IEEE Transactions on Vehicular Technology. 67(7). 6244-6258.

[11] H. Min, J. Lee, S. Park, D. Hong. 2011. „Capacity enhancement using an interference limited area for device-to-device uplink underlaying cellular networks". IEEE Transactions on Wireless Communications 10(12). pp. 3995-4000.

[12] METIS deliverable D6.1. 2013. „Simulation guidelines", ICT-317669-METIS/D6.1.

[13] METIS deliverable D1.2. 2014. „Initial channel models based on measurements", ICT-317669-METIS/D1.2.

[14] M.Rodziewicz. 2015 „Location-based mode selection and resource allocation in cellular networks with D2D underlay". Proceedings of European Wireless 2015.

[15] M.Rodziewicz. 2023. „Location-based Power Control Mechanism for D2D Communication Underlaying a Cellular System". Journal of Telecommunications and Information Technology. vol. 3, pp. 49–53.

[16] J. Seppala, T. Koskela, T. Chen, S. Hakola. 2011. „Network controlled Device-to-Device (D2D) and cluster multicast concept for LTE and LTE-A networks". IEEE Wireless Communications and Networking Conference (WCNC), pp.986-991.

[17] H. Wang, X. Chu. 2012. „Distance-constrained resource-sharing criteria for device-to-device communications underlaying cellular networks". IET Electron. Lett., vol. 48, no. 9, pp. 528-530.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

OPTIMAL RESOURCE ALLOCATION FOR 6G V2X COMMUNICATION SYSTEMS

OPTYMALNA ALOKACJA ZASOBÓW DLA SYSTEMÓW ŁĄCZNOŚCI V2X W SIECIACH 6G

Faysal Marzuk[1], Andres Vejar[2], and Piotr Chołda[3]

AGH University of Krakow, Institute of Telecommunications, Kraków, Poland.
[1]marzuk@agh.edu.pl, [2]avejar@agh.edu.pl, [3]piotr.cholda@agh.edu.pl

DOI: 10.15199/59.2024.4.77

Streszczenie: **Artykuł dotyczy problematyki zarządzania zasobami w łączności V2X (vehicle-to-everything), która korzysta z otwartych sieci dostępu radiowego Open-RAN w systemach 6G. Skupiono się na przetwarzaniu zadań obliczeniowych przekazywanych przez pojazdy do infrastruktury (j. ang. *offloading tasks*), która jest połączona z jednostkami RSU zlokalizowanymi w ciągach komunikacyjnych użytkowników końcowych (pojazdów). Sformułowano problem alokacji zasobów jako zadanie optymalizacji skupione na minimalizowaniu zużycia energii przy jednoczesnym spełnieniu ograniczeń dotyczących szybkości przesyłania danych w górę oraz opóźnień. W celu uzyskania rozwiązania optymalnego posłużono się profesjonalnym solwerem CPLEX, dzięki czemu zminimalizowano liczbę aktywnych RSU oraz ustalono optymalne powiązania między pojazdami a infrastrukturą. Wydajność modelu została przebadana dla kilku scenariuszy z różną gęstością pojazdów i RSU.**

Abstract: **In this paper, we consider the resource management problem in 6G vehicle-to-everything (V2X) Open-Radio access networks (O-RANs), where roadside units (RSUs) are deployed near end users to process the vehicles offloaded tasks. We formulate the resource allocation problem as an optimization problem to minimize energy consumption while satisfying uplink data rates and delay constraints. We use the CPLEX solver to find the optimal solutions by jointly minimizing the number of active RSUs and finding optimal associations between vehicles and RSUs. To examine the performance of our model, we analyze 6G V2X O-RANs for several scenarios with various densities of vehicles and RSUs.**

Słowa kluczowe: **6G, alokacja zasobów, inteligentne systemy transportowe ITS, optymalizacja, systemy łączności V2X (pojazd-cokolwiek).**

Keywords: **6G, Intelligent Transportation Systems (ITS), Optimization, Resource Allocation, Vehicle to Everything (V2X) Communications.**

1. INTRODUCTION

In the forthcoming years, fifth-generation (5G) networks are expected to face obstacles in fulfilling the required quality of service (QoS) in vehicle-to-everything (V2X) communication systems due to constraints such as available radio resources, huge amounts of data, developed applications, and connected devices. Therefore, pioneer technologies in intelligent transportation systems (ITS) are needed to satisfy V2X communications requirements of low power consumption, ultralow latency, and ultrahigh data rates, among others [1].

Sixth-generation (6G) networks are expected to enhance the performance of ITS and meet the wide variety of V2X communications requirements [2]. V2X communications comprise vehicle-to-network (V2N), vehicle-to-infrastructure (V2I), such as roadside units (RSUs), vehicle-to-roadside unit (V2R), vehicle-to-vehicle (V2V) and vehicle-to-pedestrian (V2P) communications [3]. The demands of V2X 6G communications include data rates (in Gbps), availability (99.999%), low delays (1 ms), high mobility and omnipresent connectivity. To meet these requirements, technological enhancements are needed, including artificial intelligence (AI) and machine learning (ML) techniques, intelligent edge, terahertz communications, low wavelengths, and high frequencies [4].

Open radio access networks (O-RANs) are promising architectures for developing and improving the implementation of future 6G V2X systems. The O-RAN deployment is based on disaggregated, virtualized, and software-based units, connected by open interfaces among various vendors [5]. The next generation Node-B (gNB) functionalities in O-RAN are disaggregated into a central unit (CU), a distributed unit (DU), and a radio unit (RU), and connected by open interfaces to management intelligent controllers at near real-time and non real-time [6]. The deployment of multi-access edge computing (MEC) on the O-RAN edge and closer to end users provides vehicles with network capabilities to perform computing and storage services with reduced delay and energy consumption [7]. Due to the limited capabilities of MEC servers such as RSUs, optimal resource allocation is considered one of the key challenges that play a role in handling the energy and latency requirements for applications of 6G V2X O-RANs [8].

Several research publications investigated optimal resource allocation for MEC-based V2X networks, including various ML tactics such as reinforcement learning (RL). The work presented in [9] introduced a collaborative Internet of Vehicles (IoV) architecture with integrated software-defined networking (SDN) based on cloud-edge clusters. The resource management problem was formulated to minimize latency and maximize profits of edge service providers. A clustering algorithm was employed to make the offloading decisions while the double deep Q network (DDQN) algorithm was used to find the optimal strategy for computational resource allocation.

The authors of [10] proposed multitask offloading and broadcasting in a stochastic MEC system considering the coexistence of uplink-downlink streams. An optimization problem was formulated to maximize the system utility functionas a combination of latency, energy,

and achievable data rate. To solve the resource management problem, the whale optimization algorithm (WOA), improved WOA, and particle swarm optimization (PSO) algorithms were integrated with a unified natural-inspired optimization framework. The work reported in [11] investigated the use of a rate splitting multiple access technique in the MEC network to help users offload tasks to multiple MEC servers. The resource allocation problem was formulated to minimize the sum of the maximum delays experienced by user equipments (UEs).

The authors of [12] considered a sustainable edge computing and formulated a multi objective optimization problem to minimize delay and energy consumption. A two-stage hybrid evolutionary algorithm based on the swarm optimizer and the immune algorithm was proposed to solve the computing offloading problem. The work reported in [13] proposed a multiaccess resource allocation scheme for V2X communication network. This scheme was formulated to maximize the utilization of the available spectrum resources.

In this work, we aim to meet the 6G V2X O-RAN challenges. Namely, we examine the problem of how to perform an effective resource allocation in MEC to optimize resource allocation and save energy while meeting QoS requirements for 6G V2X O-RANs. We formulate the resource management problem as an optimization problem to minimize the number of active RSUs and energy consumption. In particular, given several RSUs with available radio resource blocks (RBs) and several vehicles with required RBs for uploading tasks, we need to find the best vehicle to RSU association and whether to turn the RSU on or off to reduce the number of RSUs to be active and minimize the energy consumption while satisfying the constraints on the number of RBs, data rates, signal-to-interference and noise ratio (SINR), and latency. To evaluate our proposed model, we utilize a CPLEX implementation to solve our optimization problem. To examine the performance of our proposed model, we examined 6G V2X O-RANs for different scenarios with various densities of vehicles and RSUs.

The remainder of this paper is organized as follows. Section 2 describes the system model and the formulation of the optimization problem. Section 3 presents the model performance and the numerical results. Section 4 concludes the article and points out future work.

2. SYSTEM MODEL

We consider a 6G V2X O-RAN that includes sets of vehicles and RSUs. RSUs are densely deployed closer to vehicles and provide computation and communication capabilities toward end users. In our computing infrastructure, as shown in Figure 1, vehicles need to upload their tasks to RSUs for processing. We investigate a scenario consisting of a set of RSUs ($n = 1, 2, \ldots, N$) and a set of vehicles ($v = 1, 2, \ldots, V$). Each RSU n has several available RBs per time interval, denoted by M_n. Each vehicle v, if it is associated with RSU n, will require several RBs to upload its data, indicated by $R_{n,v}$. The required number of RBs depends on SINR values and uplink data rates.

We need to determine the optimal assignments between RSUs and vehicles in order to decide whether to turn on or off the RSU.

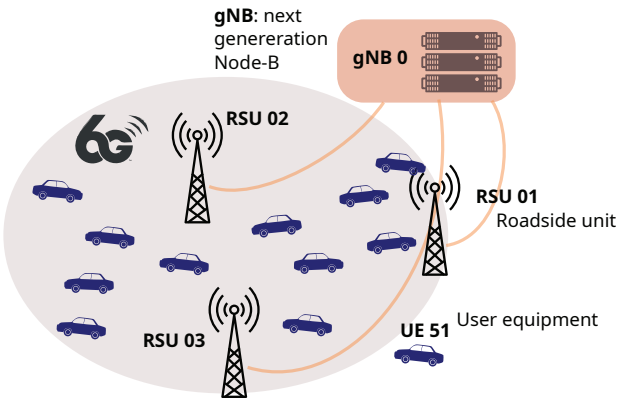


Figure 1: An example of a V2X network system.

Our objective is to minimize energy consumption and reduce the number of active RSUs depending on the required number of RBs and subject to task uplink time and uplink bandwidth constraints exemplified by SINR and inter-cell interference (ICI). We approximate the uplink ICI from the RSU n with a log-normal distribution by analytically determining the statistical parameters [14]. Then, we calculate SINR values for uplink transmission depending on aggregate uplink ICI, RSU coverage, and distance between vehicle and interfering RSU [15].

After calculating the SINR values for each vehicle, we calculate the RB data rates based on different modulation orders, SINR ranges, and efficiencies from the mapping table given in [16]. We use this mapping to find the number of required RBs where an RB per time interval of 0.25 ms consists of 12 subcarriers of 60 kHz spacing and each subcarrier consists of 14 OFDM symbols. The data rate of an RB (bits/ms) is calculated as $dr_{RB} = 12 \times 14 \times$ efficiency. The number of RBs requested ($R_{n,v}$) by each vehicle to process its task is calculated as $R_{n,v} = \frac{U_{n,v}}{dr_{RB}}$.

The task of vehicle v is represented by the uplink demand D_v and the processing demand Φ_v. We define $U_{n,v}$ as the maximum achievable data transmission rate from vehicle v to RSU n. It is calculated from the SINR mapping table into the throughput of the 5G new radio (NR) with 100 MHz bandwidth and 60 kHz sub-carrier spacing [17]. Also, $f_{n,v}$ is the computing capacity assigned from RSU n to vehicle v to process its uploaded task. We denote the uplink time of the task by $L_{n,v}$, which includes the upload time $\ell^u_{n,v} = D_v \div U_{n,v}$ and the processing time $\ell^p_{n,v} = (D_v \times \Phi_v) \div f_{n,v}$. We calculate the uplink time as $L_{n,v} = \ell^u_{n,v} + \ell^p_{n,v}$.

The energy consumption of the computation task $E_{n,v}$ includes the energy consumed in transmission ($e^u_{n,v} = P_v \times \ell^u_{n,v}$) and the energy consumed in processing ($e^p_{n,v} = \psi^V_n \times \ell^p_{n,v} \times f^2_{n,v}$). P_v stands for transmitting power of vehicle v and ψ^V_n is the energy coefficient associated with the chip architecture of the RSU server. We can calculate the energy consumption as $E_{n,v} = e^u_{n,v} + e^p_{n,v}$.

We formulate the resource allocation problem as an

optimization problem to minimize the energy consumption of the computation task and reduce the number of active RSUs by determining the following binary variables: y_n^v that indicates whether the vehicle v is associated with RSU n or not; and x_n that denotes whether to turn on or off RSU n. We formulate the optimization task as given below.

$$\min \quad \omega_1 \sum_{n\in N} x_n + \omega_2 \sum_{n\in N}\sum_{v\in V} E_{n,v}\, y_{n,v} \tag{1}$$

$$\text{s.t.} \quad \sum_{v\in V} R_{n,v}\, y_{n,v} \le M_n, \quad n \in N, \tag{2}$$

$$\sum_{v\in V} \gamma_{n,v}\, y_{n,v} \ge H_{n,v}, \quad n \in N, \tag{3}$$

$$\sum_{v\in V} f_{n,v}\, y_{n,v} \le F_n, \quad n \in N, \tag{4}$$

$$\sum_{n\in N} L_{n,v}\, y_{n,v} \le \mathcal{L}_v, \quad v \in V, \tag{5}$$

$$\sum_{n\in N} y_{n,v} = 1, \quad v \in V, \tag{6}$$

$$x_n \ge y_{n,v}, \quad n \in N, v \in V, \tag{7}$$

$$x_n,\ y_{n,v} \in \{0,1\}, \quad n \in N, v \in V. \tag{8}$$

Our objective (1) is to jointly minimize the weighted sum of active RSUs and overall energy consumption. It is subject to the following constraints: constraint (2) indicates that the number of RBs required to upload the task from vehicle v to RSU n should not exceed the maximum number of available RBs in RSU n, constraint (3) ensures that vehicle v can be assigned to RSU n if the predicted SINR is greater than the minimum threshold H, constraint (4) guarantees that the computational capacity assigned from RSU n to vehicle v should not exceed the maximum computation capacity of RSU n, constraint 5 insures that processing time for vehicle's v task should not exceed the maximum allowed latency, constraint (6) states that each vehicle v should be associated with one and only one RSU n, constraint (7) ensures that UE (vehicle) v can be connected to RSU n only if RSU n is active, and constraint (8) define the binary decision variables.

3. NUMERICAL RESULTS

In this section, we analyze the performance of our proposed model. In particular, we define seven scenarios with maximum number of RSUs and deployment area per scenario, as shown in Table 1, where the number of RSUs varies between 4 and 207 with various vehicle densities ranging from 153.4 to 1102.5 vehicles per km^2. Table 2 lists the parameter values that are used in the calculations and evaluations. Where the values are assumed according to the service requirements for 6G V2X services and to guarantee the QoS requirements of the communications system [18].

Figure 2 shows the comparison of energy consumption, number of active RSUs and uplink data rates with different density of vehicles for each scenario. As we can see, for 153 vehicles per km^2, the energy consumed ranges from 0.36 [mJ] to 12.35 [mJ] and the number of active RSUs ranges from 2 to 21. For 1100 vehicles per km^2, the energy consumed ranges from 2.65 mJ to 69.66 mJ and the number of active RSUs ranges from 2 to 92. Thus, when the vehicle density increases by a factor of six, the energy consumption increases by 4.5 times and the number of active RSUs increases by 3.5 times. Hence, it is advisable to take into account congestion-aware allocation strategies in locations where vehicle density fluctuates significantly on a daily, weekly, or seasonal basis.

Table 1: Configuration of scenarios.

Scenario	0	1	2	3	4	5	6
# RSU max	4	8	15	29	56	107	207
Area (km^2)	0.03	0.06	0.12	0.23	0.44	0.85	1.64

Table 2: Parameter values used in the evaluation.

Parameter	Value	Parameter	Value
RSU coverage	200 m	$\mathcal{P}_v$	23 – 33 dBm
M_n	135 RBs	$H_{n,v}$	-7 dB
F_n	20 GHz	$U_{n,v}$	50 – 100 Mbps
$\mathcal{L}_v$	30 ms	D_v	10 – 60 kbit

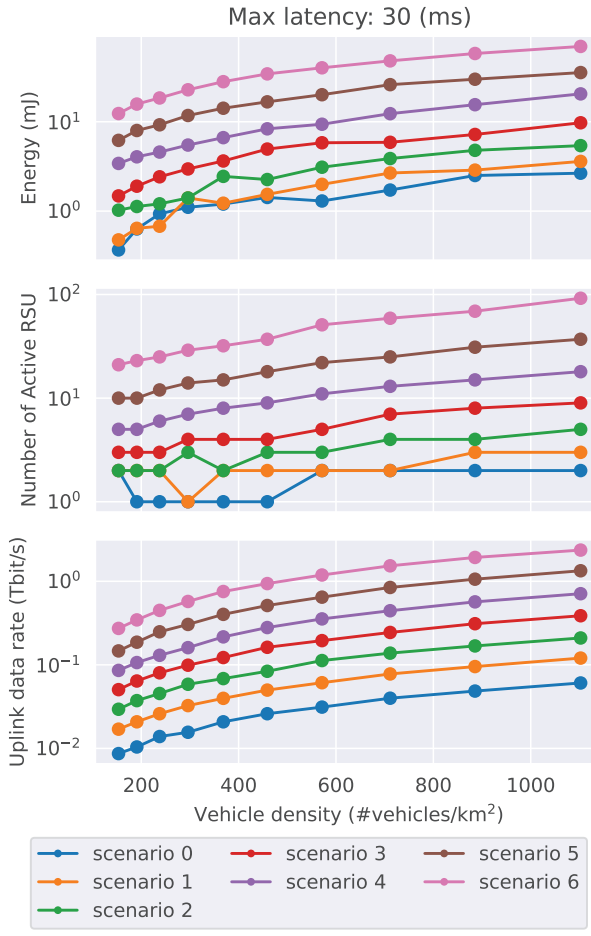


Figure 2: Comparison of energy consumption (top), number of active RSU (center) and uplink data rate (bottom) for the seven V2X communication scenarios, considering an area of deployment ranging from 0.03 km^2 to 1.64 km^2.

4. CONCLUSIONS

In this work, we investigate the problem of resource allocation in 6G V2X O-RAN. The resource allocation problem is mathematically formulated as an optimization problem based on integer linear programming and then solved using a professional solver. Our numerical results show that the MEC-based model is suitable for V2X applications and provides dynamic resource allocation while reducing energy consumption. For future work, we plan to investigate the impacts of more factors (e.g., various maximum allowed delays) on the model performance. We also plan to apply ML tactics (DRL) to overcome the long convergence time of traditional optimization.

ACKNOWLEDGMENTS

This research was supported by the National Research Institute, grant number POIR.04.02.00-00-D008/20-01, on "National Laboratory for Advanced 5G Research" (acronym PL-5G) as part of the Measure 4.2 Development of modern research infrastructure of the science sector 2014–2020 financed by the European Regional Development Fund.

REFERENCES

[1] Van-Linh Nguyen, Ren-Hung Hwang, Po-Ching Lin, Abhishek Vyas, and Van-Tao Nguyen. Toward the age of intelligent vehicular networks for connected and autonomous vehicles in 6G. *IEEE Network*, 37(3):44–51, 2023.

[2] Md. Noor-A-Rahim, Zilong Liu, Haeyoung Lee, Mohammad Omar Khyam, Jianhua He, Dirk Pesch, Klaus Moessner, Walid Saad, and H. Vincent Poor. 6G for vehicle-to-everything (V2X) communications: Enabling technologies, challenges, and opportunities. *Proceedings of the IEEE*, 110(6):712–734, 2022.

[3] Hamza Ouamna, Zhour Madini, and Younes Zouine. 6G and V2X communications: Applications, features, and challenges. In *2022 8th International Conference on Optimization and Applications (ICOA)*, pages 1–6, 2022.

[4] Atif Shakeel, Adeel Iqbal, Ali Nauman, Riaz Hussain, Xingwang Li, and Khaled Rabie. 6G driven vehicular tracking in smart cities using intelligent reflecting surfaces. In *2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring)*, pages 1–7, 2023.

[5] Michele Polese, Leonardo Bonati, Salvatore D'Oro, Stefano Basagni, and Tommaso Melodia. Understanding O-RAN: Architecture, interfaces, algorithms, security, and research challenges. *IEEE Communications Surveys & Tutorials*, 25(2):1376–1411, 2023.

[6] Mohammadreza Kouchaki and Vuk Marojevic. Actor-critic network for O-RAN resource allocation: xapp design, deployment, and analysis. In *2022 IEEE Globecom Workshops (GC Wkshps)*, pages 968–973, 2022.

[7] Azadeh Arnaz, Justin Lipman, Mehran Abolhasan, and Matti Hiltunen. Toward integrating intelligence and programmability in open radio access networks: A comprehensive survey. *IEEE Access*, 10:67747–67770, 2022.

[8] Anselme Ndikumana, Kim Khoa Nguyen, and Mohamed Cheriet. Age of processing-based data offloading for autonomous vehicles in MultiRATs Open RAN. *IEEE Transactions on Intelligent Transportation Systems*, 23(11):21450–21464, 2022.

[9] Xianhao Shen, Li Wang, Panfeng Zhang, Xiaolan Xie, Yi Chen, and Shaofang Lu. Computing resource allocation strategy based on cloud-edge cluster collaboration in internet of vehicles. *IEEE Access*, 12:10790–10803, 2024.

[10] Thien Hieu Hoang, Chuyen T. Nguyen, Tri Nhu Do, and Georges Kaddoum. Joint task offloading and radio resource management in stochastic mec systems. *IEEE Transactions on Communications*, 72(5):2670–2686, 2024.

[11] Maria Diamanti, Christos Pelekis, Eirini Eleni Tsiropoulou, and Symeon Papavassiliou. Delay minimization for rate-splitting multiple access-based multi-server mec offloading. *IEEE/ACM Transactions on Networking*, 32(2):1035–1047, 2024.

[12] Lingjie Li, Qijie Qiu, Zhijiao Xiao, Qiuzhen Lin, Jiongjiong Gu, and Zhong Ming. A two-stage hybrid multi-objective optimization evolutionary algorithm for computing offloading in sustainable edge computing. *IEEE Transactions on Consumer Electronics*, 70(1):735–746, 2024.

[13] Ali. M. A. Ibrahim, Zhigang Chen, Yijie Wang, Hala A. Eljailany, and Aridegbe A. Ipaye. Optimizing v2x communication: Spectrum resource allocation and power control strategies for next-generation wireless technologies. *Applied Sciences*, 14(2), 2024.

[14] Jianhua He, Zuoyin Tang, Hsiao-Hwa Chen, and Wenqing Cheng. Statistical model of OFDMA cellular networks uplink interference using lognormal distribution. *IEEE Wireless Communications Letters*, 2(5):575–578, 2013.

[15] Ren-Hung Hwang, Faysal Marzuk, Marek Sikora, Piotr Chołda, and Ying-Dar Lin. Resource management in ladns supporting 5G V2X communications. In *2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall)*, pages 1–6, 2020.

[16] 3GPP Technical Specification Group Radio Access Network. Evolved universal terrestrial radio access (E-UTRA)–physical layer procedures. 3GPP TS 36.213 V17.6.0(Release 17):1–551, 2024.

[17] Anderson R. Ramos, Bruno C. Silva, Marisa S. Lourenço, Emanuel B. Teixeira, and Fernando J. Velez. Mapping between average sinr and supported throughput in 5G new radio small cell networks. In *2019 22nd International Symposium on Wireless Personal Multimedia Communications (WPMC)*, pages 1–6, 2019.

[18] Khabaz Sehla, Thi Mai Trang Nguyen, Guy Pujolle, and Pedro Braconnot Velloso. Resource allocation modes in c-v2x: From lte-v2x to 5g-v2x. *IEEE Internet of Things Journal*, 9(11):8291–8314, 2022.

WPŁYW NORMY IEC 62232 I OCENY ZGODNOŚCI NA EFEKTYWNOŚĆ REALIZACJI INWESTYCJI RADIOKOMUNIKACYJNYCH

THE IMPACT OF THE IEC 62232 STANDARD AND CONFORMITY ASSESSMENT ON THE EFFICIENCY OF IMPLEMENTING RADIOCOMMUNICATIONS INVESTMENTS

Waldemar Karcz

Emitel S.A. Szczecin, waldemar.karcz@emitel.pl

DOI: 10.15199/59.2024.4.78

Streszczenie**: Zmiany dopuszczalnych poziomów PEM w środowisku miały ułatwić realizację inwestycji w Polsce. Efekt ten nie został osiągnięty, ponieważ w przepisach wykonawczych nadal jako kryterium instalacji mogących negatywnie oddziaływać na środowisko obowiązuje aksjomat „moc EIRP niemniejsza niż 15 W". Norma PN-EN IEC 62232:2023-07 otwiera możliwości wyznaczania gęstości mocy i natężenia pola wokół instalacji stacji bazowych pracujących w zakresie od 100 MHz do 300GHz oraz wokół anten parabolicznych .**

Abstract**: Changes to the permissible levels of EMF in the environment were intended to facilitate the implementation of investments in Poland. This effect has not been achieved because the implementing regulations still use the axiom "EIRP power not less than 15 W" as a criterion for installations that may have a negative impact on the environment. The PN-EN IEC 62232:2023-07 standard opens the possibility of determining power density and field intensity around the installation of base stations operating in the range from 100 MHz to 300 GHz and around parabolic antennas.**

Słowa kluczowe**: gęstość mocy, stacja bazowa, uproszczone procedury oceny, zgodność produktu.**

Keywords**: base station, power density, product compliance, simplified assessment procedures**

1. WSTĘP

Norma EN IEC 62232:2022 została uznana przez Polski Komitet Normalizacyjny za polską normę [6] PN-EN IEC 62232:2023-07 *Wyznaczanie natężania pola RF, gęstości mocy i SAR w otoczeniu stacji bazowych dla oceny poziomu ekspozycji człowieka.* Została ona zatwierdzona przez Prezesa PKN 30 maja 2023 r. Dokument ten opiera się na obecnym międzynarodowym konsensusie dotyczącym najlepszej praktyki inżynierskiej w zakresie oceny zgodności ekspozycji na pola o częstotliwościach radiowych (RF) z obowiązującymi poziomami dopuszczalnymi. 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 [1], 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. (ICNIRP 1998) oraz uzupełniają wytyczne z 2010 r. (ICNIRP 2010) dotyczące pól niskich częstotliwości w zakresie częstotliwości od 100 kHz do 10 MHz.

2. NORMY, PRZEPISY I ZALECENIA

2.1 Norma PN-EN IEC 62232:2023-07

Polska Norma PN-EN IEC62232 [6] *Wyznaczanie natężenia pola RF, gęstości mocy i SAR w otoczeniu stacji bazowych dla oceny poziomu ekspozycji człowieka* zawiera wytyczne zgodne z ICNIRP pozwalające wyznaczać natężenia pola, gęstości mocy oraz SAR w otoczeniu radiokomunikacyjnych stacji bazowych w celu oceny poziomu ekspozycji człowieka.

Dokument ten:

a) dotyczy promieniowania użytecznego stacji bazowej (BS), nadawanego z jednej lub wielu anten, na jednej lub wielu częstotliwościach w zakresie częstotliwości od 110 MHz do 300 GHz;

b) rozpatruje wpływ źródeł zewnętrznych w otoczeniu BS na ekspozycję na pola RF co najmniej w zakresie częstotliwości od 100 kHz do 300 GHz;

c) określa metody stosowane do oceny ekspozycji na pola RF dla oceny zgodności zastosowań, a mianowicie:

1. zgodność wyrobu – określenie granic obszaru zgodności z wartościami dopuszczalnymi dla produktu BS przed wprowadzeniem go na rynek;
2. zgodność wyrobu do instalacji – określenie całkowitych poziomów ekspozycji na pola RF od wyrobu BS i innych źródeł promieniowania w obszarach dostępnych dla ludzi, przed wprowadzeniem wyrobu do eksploatacji;
3. ocena ekspozycji na pola RF w miejscu zainstalowania – metody przeprowadzania pomiaru poziomów ekspozycji na pola RF w otoczeniu miejsca instalacji BS po oddaniu wyrobu do eksploatacji;

d) określa sposób przeprowadzania oceny ekspozycji na pola RF w oparciu o faktyczne maksymalne poziomy ekspozycji;

e) opisuje szereg metod pomiaru oraz obliczeń natężenia pola RF, gęstości mocy i SAR wraz z wytycznymi dotyczącymi zakresu ich stosowania do celów oceny stacji (BS) zarówno po ich zainstalowaniu jak i w trakcie badań laboratoryjnych;

f) opisuje, jak kontrolujący ustalają szczegóły własnych procedur oceny odpowiednich dla swoich celów;

g) przedstawia wytyczne dotyczące sposobu opisywania, interpretacji i porównywania wyników z różnych metodologii oceny oraz, tam gdzie wymaga tego cel oceny, określa i uzasadnia sposób podejmowania decyzji dotyczącej zgodności z poziomami dopuszczalnymi;
h) przedstawia metody oceny ekspozycji na pola RF od BS wykorzystujących technologie wąskich wiązek promieniowania o sterowanym i zmiennym w czasie kierunku promieniowania, takich jak new radio (NR) BS wykorzystujących technologie bardzo wielu wejść i wielu wyjść (MIMO). Technologia 5G wymagała opracowania nowego interfejsu radiowego, czyli sposobu łączenia się urządzeń za pośrednictwem fal elektromagnetycznych.

Interfejs 5G jest zupełnie nowy, ale będzie w stanie współpracować z dotychczasową technologią 4G/LTE /LTE-Advanced. Zasadniczą różnicą jest to, że interfejs 5G znany, jako NR czyli Nowe Radio, obsługuje o wiele szerszy zakres częstotliwości, niż dotychczasowe sieci mobilne. Do tej pory, sieci komórkowe korzystały zazwyczaj z fal radiowych o zakresie częstotliwości od 300 MHz do 3 GHz. 5G NR wykorzystuje tak zwane spektrum fal milimetrowych, czyli fale o bardzo wysokiej częstotliwości, jak na przykład 26 GHz. Częstotliwości w paśmie 26 GHz pozwolą przesyłać niespotykaną dotychczas ilość danych. Ich ograniczeniem jest niewielki zasięg i wyższa podatność na zakłócenia środowiskowe, takie jak np. deszcz. Jednak technologie wykorzystywane w sieci 5G, takie jak Massive MIMO pozwalają częściowo rozwiązać te problemy. Dla zapewnienia jak najlepszych parametrów, każde urządzenie będzie próbowało w pierwszej kolejności połączyć się z siecią właśnie za pomocą wysokich częstotliwości NR. Jeśli okaże się, że to niemożliwe, połączenie nastąpi za pomocą istniejącej dziś sieci 4G/LTE.

2.2 Zmiany w rozporządzeniu RM z 10.09.2019

Rozporządzenie Rady Ministrów z 5 maja 2022 r.[8] zmieniające rozporządzenie w sprawie przedsięwzięć mogących znacząco oddziaływać na środowisko (Dz.U. z 2022 r. poz. 1071) weszło w życie 4.06.2022 r. W Rozporządzeniu Rady Ministrów z dnia 10 września 2019 r. w sprawie przedsięwzięć mogących znacząco oddziaływać na środowisko(Dz.U. z 2019 r. poz. 1839) wprowadzono następujące zmiany: z wykazu przedsięwzięć mogących zawsze znacząco oddziaływać na środowisko oraz mogących potencjalnie znacząco oddziaływać na środowisko wykreślono instalacje radiokomunikacyjne, radionawigacyjne i radiolokacyjne. Podstawę stanowił art. 60 ustawy z dnia 3 października 2008 r. o udostępnianiu informacji o środowisku i jego ochronie, udziale społeczeństwa w ochronie środowiska oraz o ocenach oddziaływania na środowisko (Dz. U. z 2022 r. poz. 1029). Wprowadzone nowelizacją zmiany oznaczają, że budowa stacji telefonii komórkowej nie będzie wymagać uzyskania decyzji o uwarunkowaniach środowiskowych. Instalacje radiokomunikacyjne nie zaliczają się do instalacji mogących znacząco negatywnie oddziaływać na środowisko.

2.3 Rozporządzenie M.Ś. z 2.07.2010 r. poz. 879

Rozporządzenie Ministra Środowiska z dnia 2 lipca 2010r.[9] w sprawie zgłoszenia instalacji wytwarzających pola elektromagnetyczne Dz.U. 2010 nr 130 poz. 879 zostało uchylone 02.01.2021r. Operatorzy stacji bazowych składają zgłoszenie instalacji na formularzu zgłoszenia instalacji wytwarzających pola elektromagnetyczne stanowiącym załącznik nr 1 do Rozporządzenia Ministra Środowiska z dnia 2 lipca 2010 roku, które zostało uchylone 2 stycznia 2021 r.

2.4 Rozporządzenie Ministra Zdrowia z dnia 17 grudnia 2019 r. [12]

Zmiany dopuszczalnych poziomów pól elektromagnetycznych w środowisku weszły w życie z dniem 1 stycznia 2020 r. Pozytywna zmiana dopuszczalnych poziomów pól elektromagnetycznych w środowisku wprowadziła nowe limity PEM dla poszczególnych częstotliwości zgodne z wytycznymi ICNIRP[ ].

2.5 Rozporządzenie Ministra Klimatu z dnia 17 lutego 2020 r.

Rozporządzenie Ministra Klimatu z dnia 17 lutego 2020 r.[10] w sprawie sposobów sprawdzania dotrzymania dopuszczalnych poziomów pól elektromagnetycznych w środowisku (Dz. U. poz. 258) [10]. Sposoby sprawdzania dotrzymania dopuszczalnych poziomów pól elektromagnetycznych w środowisku określa załącznik do rozporządzenia. Załącznik został zmieniony Rozporządzeniem Ministra Klimatu i Środowiska z dnia 6 maja 2022 r.[11]

2.6 Ustawa z 28 07 2005 r. o lecznictwie uzdrowiskowym

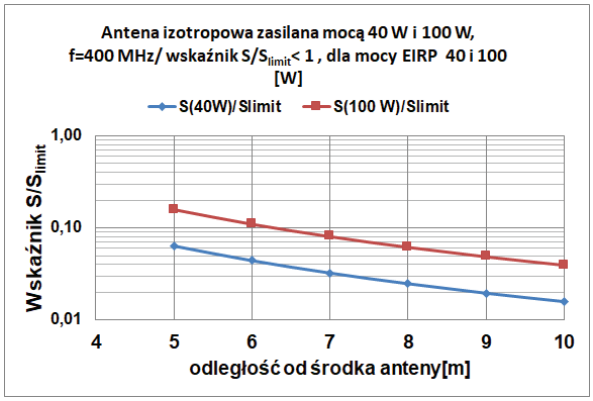
Ustawa o uzdrowiskach [13] zabraniała lokalizacji instalacji radiokomunikacyjnych w obrębie strefy ochrony uzdrowiskowej B w odległości mniejszej niż 500 m od strefy ochrony uzdrowiskowej A oraz całkowicie zabraniała lokalizacji instalacji radiokomunikacyjnych w strefie A. Zmiana nastąpiła dopiero od 8 grudnia 2022 r i została opublikowana w Dz. U. 19 stycznia 2023 r. Poz. 151. W rozdziale 5 ustawy[13] w Art. 38a.1 został usunięty podpunkt h), w którym był zapis zabraniający lokalizacji instalacji radiokomunikacyjnych w obrębie strefy ochrony uzdrowiskowej A. Po długim oczekiwaniu na taką zmianę można wreszcie instalować stacje bazowe telefonii ruchomej, stacje nadawcze radiowe i telewizyjne, stacje radiolokacyjne i inne instalacje emitujące fale elektromagnetyczne.

3. KRYTERIUM MOCY EIRP ≥15W

3.1. 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 instala-

cję 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 równej 15 W lub wyższej bez merytorycznego uzasadnienia, to wprowadzenie art.122a powinno być uzasadnione w oparciu o dopuszczalne poziomy PEM w miejscach dostępnych dla ogółu ludności. Pomimo zmian dopuszczalnych poziomów PEM w rozporządzeniach, dogmat 15 W EIRP nie uległ zmianie. Wskaźnik S/S_{limit} przy mocy EIRP 40 i 100 [W] w odległości od 5 do 10 metrów od anteny przedstawiony jest na rys. 1. Wskaźnik S/S_{limit} jest znacznie niższy od 1.



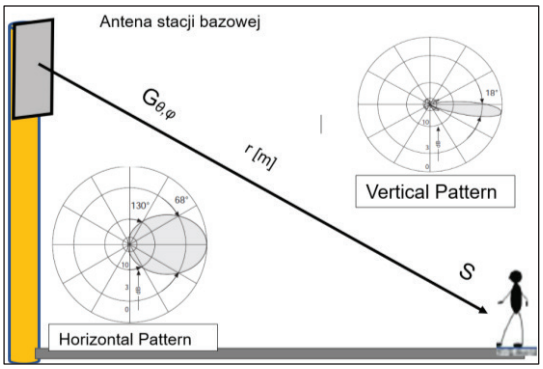
Rys. 1. Wskaźnik S/S_{limit}, moc EIERP 40W i 100W

Moc EIRP występuje tylko w kierunku wiązki głównej, a w kierunku 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 ostał 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) lub równaniu (3).



Rys. 2. Typowy przypadek oceny narażenia na RF

$$S = \frac{P \cdot G_{\theta,\varphi}}{4\pi r^2} \qquad (2)$$

$$S = \frac{EIRP_{\theta,\varphi}}{4\pi r^2} \qquad (3)$$

W przeciwieństwie do art. 122a i art.338a i338b[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 zapewnia łącze o bardzo dobrych parametrach zestawionych w Tabeli 1.

Tab. 1. Parametry łącza LR 18 GHz z antenami 0,3 m

Parametr	*Wartość*	*Jednostka*
Moc P_t	2 1,4	[dBm]
FSl	137,5	[dB]
Moc P_r	-46,9	[dBm]
Gi	34,6	[dBi]
Niedostępność prop.	65	[sek/rok]

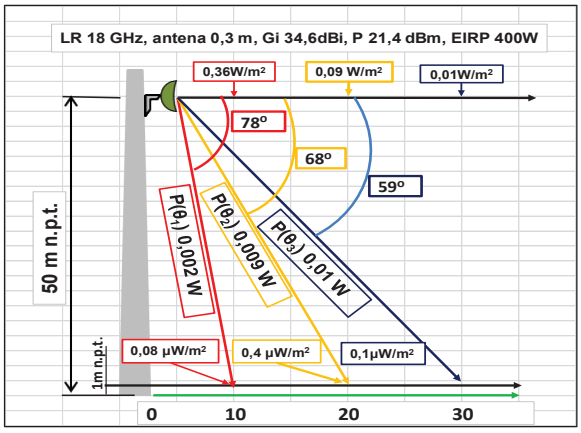
Poziomy gęstości mocy w osi głównej i na wysokość 1 m n.p.t. oraz poziomy mocy są zestawione w Tabeli 2. i na Rys. 3. Moc P**f(θ)* << 15 W. Art.122a nie ma tu zastosowania.

Tabela 2. Gęstości mocy w pionach wys.50 m ina wys. 1m

Parametr	*Pion 10 m*	*Pion 20 m*	*Pion 30 m*	
EIRP oś gł.	400	400	400	
S [W/m^2]	0,36	0,09	0,01	*h 50m*
P*f(θ) [W]	**0,002**	**0,009 W**	**0,01**	
S[μW/m^2]	8,9	20	40,0	*h 1m*

Wartości gęstości mocy na poziomie 1 m n.p.t. są niemierzalne miernikami z sondami pomiarowymi, którymi pomiary wykonują akredytowane laboratoria. W normie [6] na str. 162 znajduje się schemat blokowy uproszczonej oceny granicy zgodności. Wyznaczona ocena zgodności

dla anteny z Rys. 3. ograniczona jest powierzchnią cylindra o średnicy anteny D=0,3m i długości walca CD_{los}=0,68 m.



Rys. 3. Rozkład gęstości mocy na wys. 1m n.p.t.

3.2. klasy instalacji stacji bazowych

Uproszczone procedury oceny [4] i [6] służą do identyfikacji testowanego urządzenia EUT, o którym wiadomo, że spełnia odpowiednie wartości graniczne narażenia, bez konieczności przestrzegania ogólnych lub kompleksowych procesów oceny narażenia. Jest to istotne na przykład ze względu na przesyłaną małą moc lub położenie nadajników lub anten i odpowiednich źródeł w stosunku do miejsc dostępnych dla ogółu ludności. Uproszczone procedury oceny opierają się na znajomości wartości poziomu mocy EIRP, charakterystyki anteny, instrukcji producenta oraz wysokości zainstalowania anteny w stosunku do miejsc dostępnych dla ludności. Uproszczone kryteria procedury oceny w celu spełnienia wymagań zgodności produktów instalowanych w BS przedstawia Rys. 4, a poniżej opis dla E100 i E+[6] i [3]

PN-EN 62232 Montaż należy wykonać zgodnie z instrukcją producenta lub podmiotu wprowadzającego do eksploatacji			2,2 m	CDm 2,5	Hm
Klasa instalacji	E0	E2	E10	E100	E+
Total EIRP	Nie dotyczy	≤ 2	≤ 10 W	≤ 100 W	Bez limitu
Minimalna wysokość	-	-	2,2 m	2,5 m	Hm (obliczone)
Strefa wykluczenia	Brak, zgodna z dotykiem	Podana w instrukcji producenta mała odległość CDm		W instrukcjach producenta podano CDm w kierunkach listków głównych	
Sprawdź istniejące źródła	Nie dotyczy	Nie dotyczy	Nie dotyczy	5 CDm w długości płata gł. CDm w innych kierunkach	

Rys. 4 . Uproszczone zasady instalacji

E100 - EIRP ≤ 100 W- Produkt jest instalowany tak, że:

a) najniższa część promieniująca anteny znajduje się na wysokości co najmniej 2,5 m nad ogólnodostępnymi miejscami dla ogółu ludności.
b) minimalna odległość od obszarów dostępnych dla ludność w kierunku głównej wiązki wynosi CD_m;
c) żadne inne źródła o częstotliwości radiowej o EIRP powyżej 10 W nie znajdują się w odległości 5 CD_m w kierunku głównego listka (określonej z uwzględnieniem szerokości wiązki połowy mocy) oraz w odległości CD_m [m] w innych kierunkach. Jeżeli CD_m nie jest dostępna, można zastosować wartość 2 m lub 1 m, jeżeli wszystkie częstotliwości nadawania produktu są równe lub wyższe niż 1500 MHz.

E+ > 100 W – Produkt jest instalowany tak, że:

a) najniżej promieniująca część anteny (anten) znajduje się na minimalnej wysokości H_m [m] nad ogólnodostępnym miejscem dla ludności;
b) minimalna odległość od obszarów dostępnych dla ogółu ludności w kierunku wiązki głównej wynosi CD_m [m];
c) żadne inne źródła o częstotliwościach radiowych, o EIRP powyżej 100 W nie znajdują się w odległości 5CD_m [m] w wiązce głównej i w CD_m [m] w innych kierunkach.

Kryteria dotyczące wysokości montażu dla wartości EIRP ≤100 W zostały opracowane przy użyciu wzorów częstotliwości między 100 MHz a 400 MHz:

$$H_m = max\begin{cases} 2+\sqrt{\frac{EIRP\cdot A_{sl}}{2\cdot\pi}} \\ 2+\sqrt{\frac{EIRP}{2\cdot\pi}}\cdot\sin(\alpha+1{,}129\cdot\Delta\theta_{3dB}) \end{cases} \quad (4)$$

$$CD_m = \sqrt{\frac{EIRP}{2\cdot\pi}} \quad (5)$$

Dla częstotliwości między 400 MHz a 2000 MHz:

$$H_m = max\begin{cases} 2+\sqrt{\frac{EIRP\cdot 200\cdot A_{sl}}{f\cdot\pi}} \\ 2+\sqrt{\frac{200\cdot EIRP}{f\cdot\pi}}\cdot\sin(\alpha 1{,}129\cdot\Delta\theta_{3dB}) \end{cases} \quad (6)$$

$$CD_m = \sqrt{\frac{EIRP\cdot 200}{f\cdot\pi}} \quad (7)$$

Dla częstotliwości między 2000 MHz a 40000 MHz

$$H_m = max\begin{cases} 2+\sqrt{\frac{EIRP\cdot A_{sl}}{10\cdot\pi}} \\ 2+\sqrt{\frac{EIRP}{10\cdot\pi}}\cdot\sin(\alpha+1{,}129\cdot\Delta\theta_{3dB}) \end{cases} \quad (8)$$

$$CD_m = \sqrt{\frac{EIRP}{10\cdot\pi}} \quad (9)$$

W normie PN-EN IEC 62232:2023-07[6] do uproszczonych procedur w stosunku do normy EN 62232:2017 zostały dodane stacje bazowe z antenami masive MIMO lub z antenami ze sterowanymi wiązkami pracujące w zakresie częstotliwości od 10 000 MHz do 300 000 MHz

Dla częstotliwości między 2000 MHz a 40000 MHz

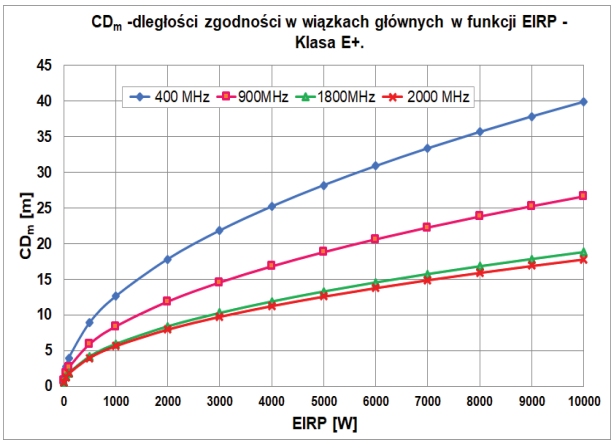
$$H_m = max\begin{cases} 2+\sqrt{\frac{EIRP\cdot A_{sl}}{65\cdot\pi}} \\ 2+\sqrt{\frac{EIRP}{65\cdot\pi}}\cdot\sin(\alpha+1{,}129\cdot\Delta\theta_{3dB}) \end{cases} \quad (10)$$

$$CD_m = \sqrt{\frac{EIRP}{65\cdot\pi}} \quad (11)$$

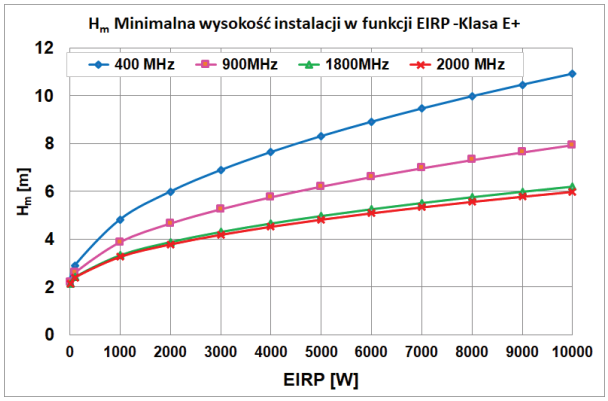
gdzie:

f – jest częstotliwością pracy stacji bazowej w MHz;
Asl – jest wartością tłumienia listków bocznych w skali liniowej;
α – jest kątem pochylenia wiązki głównej w radianach (zarówno elektrycznym, jak i mechanicznym);
$3dB$ – to pionowa szerokość wiązki połowy mocy [radian]

Na Rys. 5. został załączony wykres zależności odległości CD_m od mocy EIRP odpowiednio dla częstotliwości 400 MHz, 900MHz, 1800 MHz i 2000 MHz oraz na Rys. 6. został załączony wykres zależności minimalnej dopuszczalnej wysokości H_m instalowanych anten w zależności mocy EIRP większej niż 100W.



Rys.5. Zależność CD_m od EIRP dla 4 częstotliwości

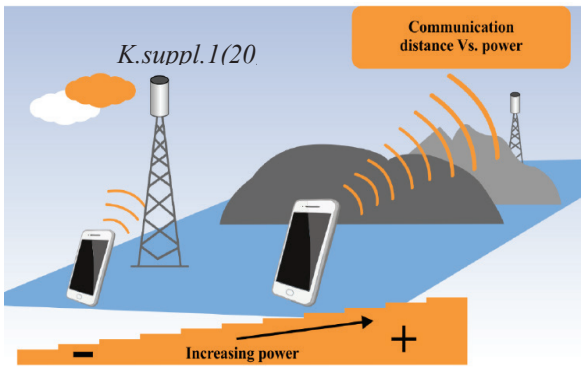


Rys.6. Zależność H_m od EIRP dla 4 częstotliwości

3.3 Ustawa o uzdrowiskach w Polsce

Ustawa o uzdrowiskach [13] zabraniała lokalizacji instalacji radiokomunikacyjnych w strefach ochrony uzdrowiskowej B w odległości mniejszej niż 500 m od strefy ochrony uzdrowiskowej A oraz całkowicie zabraniała lokalizacji instalacji radiokomunikacyjnych w strefie A.

Jeśli między telefonem komórkowym a stacją bazową znajduje się budynek, ściana, wzgórze, drzewo lub inna przeszkoda, sygnał odbierany przez stację bazową może być również słabszy, co oznacza, że natężenie pola RF z telefonu komórkowego musi wzrosnąć, aby telefon komórkowy mógł nadal komunikować się ze stacją bazową. Wykonywanie połączenia głosowego z telefonu komórkowego może prowadzić do większego narażenia niż wysyłanie i odbieranie danych lub uzyskiwanie dostępu do Internetu. Dzieje się tak, ponieważ połączenia głosowe są zazwyczaj wykonywane, gdy telefon komórkowy znajduje się obok głowy, a podczas wysyłania i odbierania danych zwykle trzyma się go z dala od ciała.



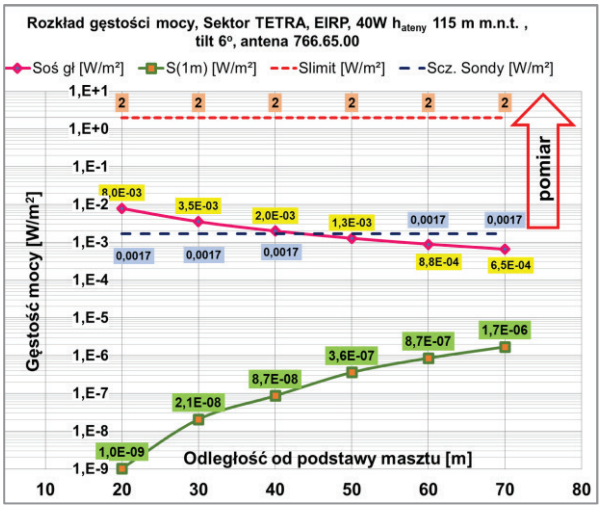
Rys. 7. Wzrost mocy telefonu ze wzrostem odległości od stacji bazowej

Połączenia mogą również trwać dłużej niż wysyłanie danych, co ponownie zwiększa ryzyko. Telefony komórkowe są również zaprojektowane tak, aby wykorzystywać możliwie najniższą moc do łączenia się z najbliższą stacją bazową i automatycznie dostosowywać moc w zależności od otoczenia. Ustawodawca podjął dobrą decyzję dla uzdrowisk i dla kuracjuszy, szkoda tylko, że tak późno.

4. STACJE BAZOWE TETRA

4.1. Procedura pełna ,czy uproszczona?

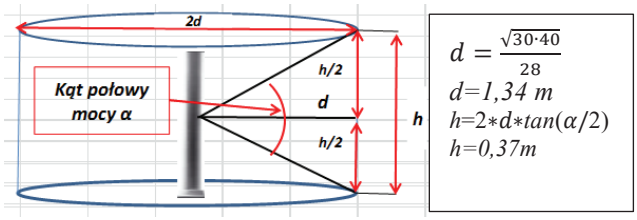
1.Stacja bazowa TETRA z anteną zainstalowaną na wysokości 115 m n.p.t. z mocą EIRP 40 W.



Rys. 8. Rozkład gęstości mocy, antena sektorowa tilt 6°

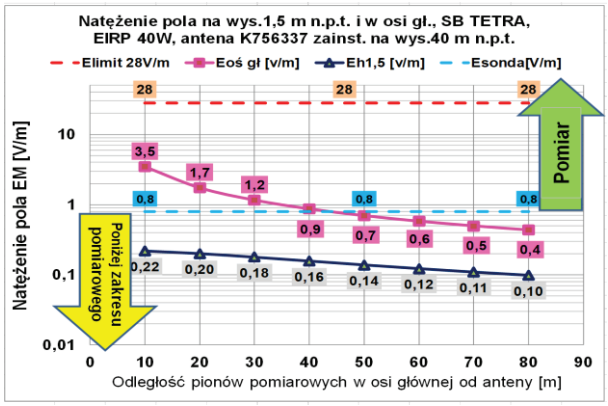
Wszystkie istotne parametry są na Rys. 8., H_m << 115m, CD_m =2,5 m w miejscach niedostępnych, bo na wys. 115 m. Taka instalacja na podstawie normy[6]kwalifikuje się do procedury uproszczonej E100.

2. Stacja bazowa TETRA z anteną K756337 zainstalowaną na wysokości 40 m n.p.t. z mocą EIRP=40W.



Rys. 9. Wyznaczenie granicy zgodności w polu bliskim anteny K756337.

Wszystkie istotne parametry SB TETRA z anteną dookólną zainstalowaną na szczycie wieży o wys. 40 m n.p.t. są na Rys. 10., Hm<< 40 m, CDm =2,5 w miejscu niedostępnym, bo na wys. 40 m n.p.t.



Rys. 10. Rozkład gęstości mocy w polu dalekim

Taka instalacja na podstawie normy[6] kwalifikuje się do procedury uproszczonej E100. Natężenie pola EM na wysokości h =1,5 m n.p.t. są mniejsze niż 0,23 V/m. i są mniejsze od dolnej granicy czułości sondy pomiarowej.

5. PODSUMOWANIE

1. 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-EN 50383:2011. Norma ta nie tylko otwiera możliwości wyznaczania gęstości mocy i natężenia pola wokół instalacji klasyczny stacji bazowych, ale rozwiązuje też problemy zgodności nowych instalacji z antenami massive MIMO i SMART Antena.

2. W okresie od 2018 do 2024 mamy polską norm PN-EN 62232, ale w ustawach i rozporządzeniach pozytywne zmiany postępują z bardzo dużym opóźnieniem. Przykładem są artykuły w ustawie [14] 122a i 338a. Minimalny krok jaki należy zrobić, to 15 EIRP należy zastąpić zamiast EIRP (czyli w osi głównej) zapisem EIRP(θ,φ) w kierunku od anteny do punktu pomiarowego, tak jak to jest w normie [6] i jak jest na Rys. 2 w tym referacie.

3. W normie [6] na str. 162 znajduje się schemat blokowy uproszczonej oceny granicy zgodności. Taki sam schemat jest w ETSI TR 102 457 v2.1.0 (2018-09). W taki sposób należałoby postępować w stosunku do anten paraboliczych, zamiast stosować się do artykułu 122a.

4. WHO i Komitet Naukowy ds. Zagrożeń dla Zdrowia stwierdziły, że narażenie związane z sieciami bezprzewodowymi i z 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, środowiska akademickie, 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 RF-PEM 5G są podobne do innych technologii bezprzewodowych i nie przekraczają 1% gęstości mocy dopuszczalnych limitów przez ICMIRP [1]. To była również podstawa do zmian w ustawie [13]

LITERATURA

[1] ICNIRP *Guidelines for limiting exposure to electromagnetic fields (100 kHz-300 GHz)*, Health Phys, March 2020

[2] 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

[3] 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

[4] ITU-T K Supplement 9 2017, 5Gtechnology and human exposure to RF EMF

[5] ITU-T K Supplement 16 2022 Electromagnetic field compliance assessments for 5G wireless network

[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] Rozporządzeniu Rady Ministrów z dnia 10 września 2019 r. w sprawie przedsięwzięć mogących znacząco oddziaływać na środowisko(Dz.U. z 2019 r. poz. 1839)

[8] 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)

[9] Rozporządzenie Ministra Środowiska z dnia 2 lipca 2010 r.[ ] w sprawie zgłoszenia instalacji wytwarzających pola elektromagnetyczne Dz.U. 2010 nr 130 poz. 879

[10] Rozporządzenie Ministra Klimatu z dnia 17 lutego 2020 *r w sprawie sposobów sprawdzania dotrzymania dopuszczalnych poziomów pól elektromagnetycznych w środowisku* (Dz. U. poz. 258)

[11] 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

[12] Rozporządzenie Ministra Zdrowia z dnia 17 grudnia 2019 r., *w sprawie dopuszczalnych poziomów pól elektro-magnetycznych w środowisku,* Dz. U. 2019 poz. 2448

[13] Ustawa z dnia 28 lipca 2005 r. *o lecznictwie uzdrowiskowym, uzdrowiskach i obszarach ochrony uzdrowiskowej oraz o gminach uzdrowiskowych* Dz. U. z 2023 r. poz. 151, 1688. 1692.

[14] Ustawa z dnia 27 kwietnia 2001 r., P*rawo ochrony środowiska*. Dz. U. 2022 poz. 2556

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

Wykorzystanie powietrznych przełączalnych inteligentnych powierzchni do komunikacji międzypojazdowej

Deploying an Aerial Reconfigurable Intelligent Surface for Vehicle-to-Vehicle Communications

Salim Janji

Institute of Radiocommunications, Poznan
salim.janji@put.poznan.pl

DOI: 10.15199/59.2024.4.79

Streszczenie: Niniejszy artykuł omawia wdrożenie drona wyposażonego w rekonfigurowalną inteligentną powierzchnię (Reconfigurable Intelligent Surface, RIS) celem stworzenia dronowej stacji przekaźnikowej (Drone Relay Station, DRS), aby zwiększyć łączność między pojazdami (Vehicle-To-Vehicle, V2V) na ziemi. Trajektoria DRS jest optymalizowana w taki sposób, aby możliwie najszybciej osiągnąć położenie maksymalizujące przepustowość. Dodatkowo uwzględniana jest obecność węzła zakłócającego, a rozwiązanie analityczne jest wyprowadzone w celu określenia optymalnej orientacji DRS w każdym kroku czasowym, minimalizując zakłócenia dla odbiornika. Wyniki symulacji potwierdzają skuteczność proponowanego modelu.

Abstract: This paper addresses the deployment of a drone equipped with a reconfigurable intelligent surface (RIS), creating a drone relay station (DRS) to enhance the connectivity of vehicle-to-vehicle (V2V) pairs on the ground. The trajectory of the DRS is optimized to quickly reach the best location for maximizing throughput. Additionally, the presence of an interfering node is considered, and an analytical solution is derived to determine the optimal orientation of the DRS at each time step, minimizing interference to the receiver. Simulation results confirm the effectiveness of the proposed framework.

Słowa kluczowe: bezzałogowy statek powietrzny, rekonfigurowalna inteligentna powierzchnia, niezawodność komunikacji, zakłócenia

Keywords: UAV, RIS, communication reliability, interference

1. INTRODUCTION

The emergence of intelligent transportation systems, including autonomous vehicles, has propelled Vehicle-to-Everything (V2X) technology as a key enhancer of road safety, traffic efficiency, and passenger connectivity. V2X incorporates various communication modes such as Vehicle-to-Vehicle (V2V), Vehicle-to-Infrastructure (V2I), Vehicle-to-Pedestrian (V2P), and Vehicle-to-Network (V2N), all demanding robust, low-latency, and high-throughput communication capabilities for applications like collision avoidance, traffic management, and autonomous driving [1, 2].

In areas with connected vehicles, efficient V2V networks are essential, particularly when spectrum availability is limited or vehicle traffic is high. Obstacles blocking the line-of-sight (LOS) path between vehicles pose significant challenges by reducing signal strength. The deployment of drones, or unmanned aerial vehicles (UAVs), has been explored as a solution to support V2V communication by acting as relay nodes [3, 4].

Additionally, reconfigurable intelligent surfaces (RISs) are becoming a popular tool for network enhancement by modulating signal strength, coverage, and energy efficiency [5]. RISs, which are embedded with elements that control the amplitude and phase of electromagnetic waves, can direct signals to specific areas, improving desired signal reception and reducing interference. The integration of RISs on drones to support ground connectivity has recently been investigated, offering a potential solution to bandwidth limitations and signal blockages [6, 7, 8, 9].

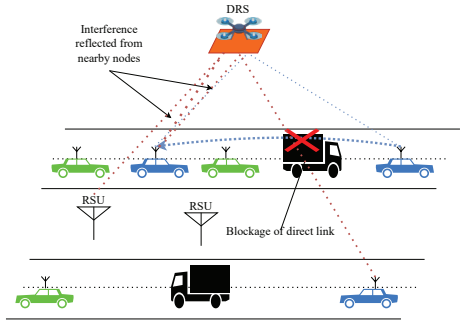


Figure 1: Depiction of the examined scenario where a V2V pair is engaged in communication with potential obstacles impeding direct connections between vehicles. A DRS is deployed to enhance connectivity, yet interference reflected by the DRS from other nodes must also be accounted for.

This paper examines the deployment of a drone relay station (DRS) equipped with an on-board RIS to aid V2V communications on the ground. The previous study in [9] addressed this problem for a single pair of vehicles without considering interference from other nodes. In contrast, this work considers the interference from a nearby

transmitting vehicle or road-side unit (RSU) and includes the optimization of the RIS orientation to minimize interference to the receiver of the serviced pair. The scenario is illustrated in Fig. 1.

The structure of the paper is as follows: The next section introduces the system model and discusses the channel model. Section 3 presents the problem formulation, followed by Section 4, which details the proposed solutions for trajectory planning and orientation control. An analysis of the simulation results follows. The paper concludes with a summary of the findings.

2. SYSTEM MODEL

This study explores a V2V communication scenario with vehicles traveling in two lanes, as depicted in Fig. 1. It is assumed that $K = 1$ pair of vehicles communicate using the same frequency band. The 3D locations of the vehicles at time t are given by $(\mathbf{L}_1(t), \mathbf{L}_2(t))$. The position of each vehicle or RSU is represented as $\mathbf{L}_j(t) = (x_j(t), y_j(t), h_j(t))$, where $x_j(t)$, $y_j(t)$, and $h_j(t)$ denote the lateral, longitudinal, and vertical positions, respectively. Similarly, an interfering node, which can be either an RSU or another vehicle, is identified with its location, $\boldsymbol{L}_{\mathrm{I}}(t)$. All vehicles maintain a constant velocity v_j, and their antennas are positioned between 1.5 and 2 meters high.

Direct V2V links may not provide sufficient reliability or throughput. Therefore, the use of a UAV equipped with a nearly-passive RIS, referred to as a DRS, is considered to enhance signal reflection. The UAV's position is defined by $\mathbf{L}_D(t) = (x_D(t), y_D(t), h_D(t))$, and its velocity v_D is greater than or equal to any vehicle's velocity v_i. The RIS's reflection capabilities depend on its orientation matrix $\mathbf{O}(t)$, a 3×3 matrix determining the alignment of the RIS.

To model RIS-aided transmission, the methodologies from [10] for far-field or near-field beamforming, based on the relationship between transmitter-to-RIS distance d_1, RIS-to-receiver distance d_2, and the Faunhofer distance $d_{Fr} = \frac{2D^2}{\lambda}$, are employed. Given the setup, all transmissions use far-field beamforming. The effective path-loss in this scenario is influenced by the alignment of the transmission path with the RIS orientation, accounted for by the azimuth and elevation angles of both transmitter and receiver. The applicable path-loss formula is presented as:

$$\begin{aligned} PL_{f-f} &= \frac{64\pi^3 d_1^2 d_2^2}{G_t G_r G M^2 N^2 d_x d_y \lambda^2 F(\theta_t), F(\theta_r) A^2 \left|\Psi\right|^2} \\ \Psi &= \frac{sinc(\frac{M\pi}{\lambda}(\sin\theta_t \cos\varphi_t + \sin\theta_r \cos\varphi_r) d_x)}{sinc(\frac{\pi}{\lambda}(\sin\theta_t \cos\varphi_t + \sin\theta_r \cos\varphi_r) d_x)} \\ &\times \frac{sinc(\frac{N\pi}{\lambda}(\sin\theta_t \sin\varphi_t + \sin\theta_r \sin\varphi_r) d_y)}{sinc(\frac{\pi}{\lambda}(\sin\theta_t \sin\varphi_t + \sin\theta_r \sin\varphi_r) d_y)}, \end{aligned} \tag{1}$$

where G_t, G_r, and G denote the antenna gains of the transmitter, receiver, and RIS, respectively; M and N represent the rows and columns of RNA elements; d_x and d_y are the element spacings; $F(\theta_t)$ and $F(\theta_r)$ are the normalized radiation patterns directed towards the transmitter and receiver. All RIS elements share a uniform radiation pattern, defined as:

$$F(\theta) = \begin{cases} \cos^3\theta & \theta \in \left[0, \frac{\pi}{2}\right] \\ 0 & \theta \in \left(\frac{\pi}{2}, \pi\right] \end{cases} \tag{2}$$

The throughput for each pair is calculated using the modified Shannon formula, incorporating the signal-to-interference-plus-noise ratio (SINR), bandwidth, and path-loss:

$$R = \eta B_{\text{eff}} \log_2(1 + SINR) \tag{3}$$

$$SINR = \frac{P_t}{PL\sigma_N^2 + \frac{P_t}{PL_I}} \tag{4}$$

where η and B_{eff} denote the link effectiveness and effective bandwidth, respectively, PL represents the path-loss between the communicating pair of vehicles, PL_I is the path loss from the interfering node to the receiver (the receiver is identified by L_2), and P_t is the transmission power.

3. PROBLEM FORMULATION

The instantaneous positions of a V2V communication pair, represented as $(\boldsymbol{L}_1(t), \boldsymbol{L}_2(t))$, along with the position of the DRS, $\boldsymbol{L}_{\mathrm{D}}(t)$, the position of the interfering node, $\boldsymbol{L}_{\mathrm{I}}(t)$, and the 3x3 orientation matrix $\boldsymbol{O}(t)$, are considered. The task is to design the future trajectory of the DRS, specifically determining $\boldsymbol{L}_{\mathrm{D}}(t+nT_s)$ for $n \in \{1, 2, \ldots, N\}$, where T_s is the planning time step, and N is the number of steps, determined by the duration of the V2V pair's transmission or their presence within the designated highway segment, which is confined by $(x_{\min}, x_{\max})$ on the x-axis and $(y_{\min}, y_{\max})$ on the y-axis. Additionally, the DRS is capable of modifying the orientation of the RIS along the xy-plane, denoted by $\boldsymbol{O}(t)$, at a rotation rate of $\Gamma_D\ [\frac{rad}{s}]$. It is essential to establish not only the trajectory but also the orientation $\boldsymbol{O}^{(n)}$ at each time step. For clarity and consistency throughout this paper, the time index is represented as a superscript in all subsequent references. Uniform direction and speed $[\frac{m}{s}]$ within each time step are assumed, constrained by the maximum speed v_D.

The objective is to maximize the achieved rate through minimizing the path loss of the virtual LOS link, established through the RIS, while also minimizing the reflected interference from a nearby node. The transmission rate $R^{(n)}$ is determined using (3) based on the vehicles' locations and the DRS's position and orientation. The optimization problem is thus formulated as:

$$\max_{\boldsymbol{O}^{(n)}, \boldsymbol{L}_{\mathrm{D}}^{(n)} \forall n \in \{1,\ldots,N\}} \sum_{n=1}^{(n)} R^{(n)} \tag{5a}$$

$$\text{subject to } \forall n \in \{1, \ldots, N\}: \tag{5b}$$

$$\cos^{-1}\left(\frac{\mathrm{tr}\left(\boldsymbol{O}^{(n)}\left(\boldsymbol{O}^{n+1}\right)^{-1}\right)}{2}\right) \leq \Gamma_D \times T_s \tag{5c}$$

$$\left\|\left(x_D^{n+1} - x_D^{(n)}, y_D^{n+1} - y_D^{(n)}, z_D^{n+1} - z_D^{(n)}\right)\right\| \leq v_D \times T_s \tag{5d}$$

$$x_{\min} \leq x^{(n)} \leq x_{\max} \tag{5e}$$

$$y_{\min} \leq y^{(n)} \leq y_{\max} \tag{5f}$$

$$z_{\min} \leq z^{(n)} \leq z_{\max} \tag{5g}$$

The goal, as defined in (5), is to optimize the cumulative throughput over the time span of NT_s. The constraints are:

- (5c) uses the matrix trace operator tr(·) to compute the rotation angle from the rotation matrices and restrict it to a maximum value, thereby adhering to the DRS's maximum rotational speed limit.
- (5d) confirms the DRS's displacement per time step stays within its maximum translational speed.
- Constraints (5e), (5f), and (5g) ensure the DRS's movement is confined within the designated 3D boundary.

The problem is divided into two subproblems: the first problem deals with deciding the trajectory of the DRS, $\boldsymbol{L}^{\boldsymbol{n}}{}_{\mathrm{D}}(t)$, by guiding it towards the optimal locations for maximizing the received power from the transmitter at $\boldsymbol{L}_1(t)$ to the receiver at $\boldsymbol{L}_2(t)$; and the second problem is concerned with determining the DRS orientation at each time step to minimize the interference from a nearby node, identified by its location $\boldsymbol{L}_I(t)$. The solutions are introduced below.

4. PROPOSED SOLUTION

To understand the solutions, note that the path-loss formula in (1) can be divided into two main components: the first depends on the locations, involving the elevation angles and distances from the RIS to each vehicle, while the second part (Ψ in (1)) is influenced by the orientation, primarily depending on the azimuth angles.

The first part of the path-loss in (1) is influenced only by changes in the DRS's position relative to the V2V pair. Therefore, it can be optimized independently as follows.

4.1 Setting the trajectory of the DRS

The optimal location is derived similarly to the approach in [9]. Initially, the elevation angle θ for any vehicle relative to the RIS is given by:

$$\theta = \tan^{-1}\left(\frac{d_{2\mathrm{d}}}{h_D}\right), \tag{6}$$

where $d_{2\mathrm{d}}$ denotes the horizontal distance on the xy-plane between the vehicle and the DRS. The optimal (x, y) location of the DRS, which minimizes (1), is identified as the midpoint between the V2X pairs. The optimal height is determined by minimizing the terms in (1) excluding Ψ, resulting in:

$$f(h_D) = \frac{d_{2\mathrm{D}}^2 + h_D^2}{\cos^6\left(\tan^{-1}\left(\frac{d_{2\mathrm{D}}}{h_D}\right)\right)}, \tag{7}$$

where equations (4.1) and (6) simplify the function to its current form, and $d_{2\mathrm{D}}$ is the 2D distance between the vehicle and the DRS at the midpoint. Gradient-based optimization methods such as truncated Newton conjugate-gradient (TNC) are employed to find the optimal height $h_{D,\mathrm{opt}}$, using Scipy's implementation of TNC to optimize within the bounded space defined by (5g).

The optimal location is thus:

$$\boldsymbol{L}_{\mathrm{opt}}^{(n)} = \left(\frac{x_1^{(n)} + x_2^{(n)}}{2}, \frac{y_1^{(n)} + y_2^{(n)}}{2}, h_{D,\mathrm{opt}}\right), \tag{8}$$

which informs the DRS's trajectory at each time step, calculated as:

$$\boldsymbol{L}^{n+1} = \boldsymbol{L}^{(n)} + v_D T_s \frac{\boldsymbol{L}_{\mathrm{opt}}^{(n)} - \boldsymbol{L}^{(n)}}{\left|\left|\boldsymbol{L}_{\mathrm{opt}}^{(n)} - \boldsymbol{L}^{(n)}\right|\right|^2}. \tag{9}$$

This strategy ensures the DRS swiftly reaches the optimal position.

4.2 Reducing interference by controlling orientation

To reduce interference, the orientation of the DRS, $\boldsymbol{O}^{(n)}$, is varied within the limits of its speed, Γ_D, as follows. Given the current location of the DRS, the location of the interfering node, and the receiver, then rotating the DRS with a value of α [rad] results in modifying the Ψ term according to the following relation.

$$\Psi_I^{(n)}(\alpha) = \frac{sinc(\frac{M\pi}{\lambda}(\sin\theta_I\cos(\varphi_I+\alpha)+\sin\theta_r\cos(\varphi_r+\alpha))d_x)}{sinc(\frac{\pi}{\lambda}(\sin\theta_I\cos(\varphi_I+\alpha)+\sin\theta_r\cos(\varphi_r+\alpha))d_x)} \times \frac{sinc(\frac{N\pi}{\lambda}(\sin\theta_I\sin(\varphi_I+\alpha)+\sin\theta_r\sin(\varphi_r+\alpha))d_y)}{sinc(\frac{\pi}{\lambda}(\sin\theta_I\sin(\varphi_I+\alpha)+\sin\theta_r\sin(\varphi_r+\alpha))d_y)}. \tag{10}$$

Note that guiding this term to zero results in a large path-loss value ((1)), thus cancelling the interference from the interfering location. Then we need to solve for an $\alpha < \Gamma_D \times T_s$ that will make $\Psi_I^{(n)}(\alpha) = 0$. Since forcing any of the two numerators in (10) to zero will effetively guide the path-loss to ∞, then our goal is to achieve one of the following conditions:

$$C_1 = \frac{M\pi}{\lambda}(\sin\theta_I\cos(\varphi_I+\alpha)+\sin\theta_r\cos(\varphi_r+\alpha))d_x = N\pi \tag{11}$$

or

$$C_2 = \frac{N\pi}{\lambda}(\sin\theta_I\sin(\varphi_I+\alpha)+\sin\theta_r\sin(\varphi_r+\alpha))d_y = N\pi. \tag{12}$$

for any $N \in \{1, 2, ...\}$ Through leveraging the trigonometric law of product, and the formula

$$A\sin(x) + B\cos(x) = R\cos(x-\phi) \tag{13}$$

where $R = \sqrt{A^2+B^2}$ and $\phi = \tan^{-1}\left(\frac{A}{B}\right)$, C_1 and C_2 can be rewritten as

$$C_{1,2} = D_{1,2}\sqrt{A_{1,2}^2 + B_{1,2}^2}\cos\left(\alpha - \tan^{-1}\left(\frac{A_{1,2}}{B_{1,2}}\right)\right) \tag{14}$$

where

$$D_1 = \frac{M\pi d_x}{\lambda} \tag{15a}$$
$$A_1 = \sin(\theta_I)\cos(\varphi_I) + \sin(\theta_r)\cos(\varphi_I) \tag{15b}$$
$$B_1 = -\sin(\theta_I)\sin(\varphi_I) - \sin(\theta_r)\sin(\varphi_I), \tag{15c}$$

and

$$D_2 = \frac{N\pi d_y}{\lambda} \tag{16a}$$
$$A_2 = \sin(\theta_I)\sin(\varphi_I) + \sin(\theta_r)\sin(\varphi_I) \tag{16b}$$
$$B_2 = \sin(\theta_I)\cos(\varphi_I) + \sin(\theta_r)\cos(\varphi_I). \tag{16c}$$

Then, to find a solution to the problem, α can be selected within $[-\Gamma_D, \Gamma_D]$ to satisfy any of the two constraints defined by

$$\alpha_{1,2} = \cos^{-1}\left(\frac{N\pi}{D_{1,2}\sqrt{A_{1,2}^2+B_{1,2}^2}}\right) + \tan^{-1}\left(\frac{A_{1,2}}{B_{1,2}}\right) \tag{17}$$

5. SIMULATION RESULTS

5.1 Simulation setup

The scenario assumes two parallel lanes with opposite directions and a single RSU in the middle of the area. The arrival times of vehicles entering the lanes are generated according to an exponential distribution with a rate parameter λ_{arrival}, and the number of V2V communication events starting at time step n is Poisson distributed with densities λ_{v2v}, and vehicles are randomly selected. The lanes are situated parallel to the y-axis at $x = x_{\min} = 0$ m and $x = x_{\max} = 500$ m, and they stretch across the y axis between $y = y_{\min} = 0$ m and $y = x_{\max} = 5000$ m. It is assumed that the DRS can fly as low as $z_{\min} = 100$ m and $z_{\max} = 600$ m. Also, time step (T_s), DRS speed (v_D, Γ_D), and vehicle speed (v_t) are 0.5 s, 18 m/s, 0.1745 rad/s, and 15 m/s, respectively.

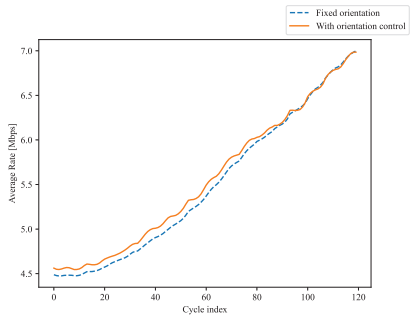


Figure 2: Received throughput through the DRS at each cycle index averaged across all simulated V2V pairs. Optimizing orientation seems to slightly improve the performance since it cancels the interference.

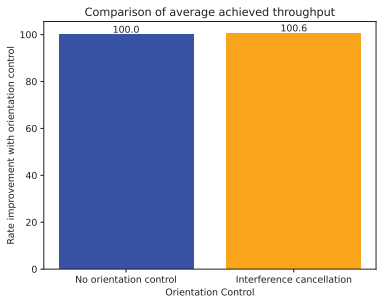


Figure 3: Cancelling the interference seems to improve the rate by 0.5%.

Fig. 2 plots the average rate received through the DRS across the simulation cycles with and without using orientation control. Optimizing the orientation with the described method seems to slightly improve the received rate by around 0.6% across all simulation runs as highlighted in Fig. 3. Although this value is small for a single simulated interference source, it nevertheless sheds light on the presented problem, and proves the formulation described in the previous section.

6. CONCLUSIONS

This paper investigated the application of DRS equipped with RIS to support V2V communications while also considering interference from a single source. A heuristic trajectory that guides the DRS to the optimal location was proposed along with an analytical solution to control the orientation of the DRS in order to cancel interference. The presented technology is simple to implement as no complex hardware is needed to be mounted on the drone. It is only needed to control the location and orientation of the drone as described. This is possible with several drones on the market. In future work, larger problem size will be considered with multiple V2V pairs and interference sources.

The presented work has been funded by the National Science Centre in Poland within the project (no. 2021/43/B/ST7/01365) of the OPUS programme.

LITERATURA

[1] Garcia, M. H. C., A. Molina-Galan, M. Boban, J. Gozalvez, B. Coll-Perales, T. Şahin, and A. Kousaridas (2021). A tutorial on 5g nr v2x communications. *IEEE Communications Surveys & Tutorials 23*(3), 1972–2026.

[2] Gyawali, S., S. Xu, Y. Qian, and R. Q. Hu (2021). Challenges and solutions for cellular based v2x communications. *IEEE Communications Surveys & Tutorials 23*(1), 222–255.

[3] Pourbaba, P., K. B. S. Manosha, S. Ali, and N. Rajatheva (2019). Full-duplex uav relay positioning for vehicular communications with underlay v2v links. In *2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring)*, pp. 1–6.

[4] Deng, L., G. Wu, J. Fu, Y. Zhang, and Y. Yang (2019). Joint resource allocation and trajectory control for uav-enabled vehicular communications. *IEEE Access 7*, 132806–132815.

[5] Liu, Y., X. Liu, X. Mu, T. Hou, J. Xu, M. Di Renzo, and N. Al-Dhahir (2021). Reconfigurable intelligent surfaces: Principles and opportunities. *IEEE Communications Surveys & Tutorials 23*(3), 1546–1577.

[6] Li, Y., C. Yin, T. Do-Duy, A. Masaracchia, and T. Q. Duong (2021). Aerial reconfigurable intelligent surface-enabled urllc uav systems. *IEEE Access 9*, 140248–140257.

[7] Shang, B., E. S. Bentley, and L. Liu (2023). Uav swarm-enabled aerial reconfigurable intelligent surface: Modeling, analysis, and optimization. *IEEE Transactions on Communications 71*(6), 3621–3636.

[8] Lin, C.-H. and K.-H. S. Liu (2023). Robust reflective beamforming for aerial reconfigurable intelligent surface. In *ICC 2023 - IEEE International Conference on Communications*, pp. 1300–1306.

[9] Janji, S., P. Sroka, and A. Kliks (2024). Enhancing v2x communications with uav-mounted reconfigurable intelligent surfaces. In *2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pp. 708–713.

[10] Tang, W., M. Z. Chen, X. Chen, J. Y. Dai, Y. Han, M. Di Renzo, Y. Zeng, S. Jin, Q. Cheng, and T. J. Cui (2021). Wireless communications with reconfigurable intelligent surface: Path loss modeling and experimental measurement. *IEEE Transactions on Wireless Communications 20*(1), 421–439.

OCENA ZASTOSOWANIA BSP W PROCESIE MONITOROWNIA WIDMA W WALCE RADIOELEKTRONICZNEJ

Michał Konieczka[1]; Paweł Skokowski[2]

[1] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, michal.konieczka@student.wat.edu.pl
[2] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, pawel.skokowski@wat.edu.pl

DOI: 10.15199/59.2024.4.80

Streszczenie: **Zdobycie przewagi informacyjnej jest jednym z kluczowych elementów prowadzenia działań operacyjnych. Wykorzystanie nowoczesnej technologii w konfliktach powoduje, że funkcjonujące systemy rozpoznawcze mogą być niewystraczająco skuteczne by zaspokoić bieżące zapotrzebowanie informacyjne. Aby sprostać stawianym wymaganiom, poszukiwane są nowe rozwiązania, które będą w stanie szybko i skutecznie zbierać niezbędne informacje. W artykule przedstawiono ocenę zastosowania BSP w procesie monitorowania widma w walce radioelektronicznej. Przeprowadzone badania, wykazały niepodważalne zalety wykorzystania dronów w procesie monitorowania widma w walce radioelektronicznej.**

Abstract: **Achieving information supremacy is one of the key elements of executing operations. The use of modern technology in conflicts means that functioning reconnaissance systems may be ineffective in meeting current information needs. New solutions are being developed to meet the demands that can quickly and effectively collect the necessary information. The article presents an assessment of the application of BSP in spectrum monitoring in radio combat. Studies conducted showed the indisputable advantages of using drones in spectrum monitoring in radio-electronic warfare.**

Słowa kluczowe: **bezzałogowy statek powietrzny, detekcja sygnałów, monitorowanie widma, radio programowalne, walka radioelektroniczna**

Keywords: **unmanned aerial vehicle, detection, spectrum sensing, software-defined radio, electronic warfare**

1. WSTĘP

Z praktycznego punktu widzenia pojęcie „monitorowania widma" obejmuje dwa zasadnicze pojęcia: detekcja i klasyfikacja. Detekcja to działanie polegające na wykryciu sygnału użytecznego. Zabieg ten zakończony jest podjęciem decyzji o obecności sygnału lub jego braku. Natomiast klasyfikacja sygnału jest działaniem bardziej złożonym, mającym na celu identyfikacje i estymację parametrów sygnału, np. typ modulacji czy pasmo sygnału. Jak już wcześniej wspomniano, rola detektora polega na wyborze jednej z dwóch hipotez: H_1 oraz H_0. Hipotezy te odpowiadają przypadkom występowania i braku emisji radiowej [8]. Dla każdej z hipotez możliwe są dwa wyniki: decyzja została podjęta prawidłowo lub nieprawidłowo.

2. TECHNIKI DETEKCJI

Raport ITU [6] przedstawia podział technik detekcji na techniki wąsko- i szerokopasmowe. Do najpopularniejszych metod wąskopasmowych należy zaliczyć: detektor energii, detektor cyklostacjonarności, filtr dopasowany, detektor falkowy oraz detektor hybrydowy [2][3][7][8]. W przeciwieństwie do metod wąskopasmowych, techniki wykrywania widma szerokopasmowego mają na celu wykrywanie pasma częstotliwości, które przekracza pasmo koherencji kanału. Warto zauważyć, że techniki wykrywania wąskopasmowego nie mogą być bezpośrednio wykorzystywane do wykrywania szerokopasmowego widma, ponieważ podejmują one pojedynczą decyzję binarną dla całego widma, a zatem nie są w stanie zidentyfikować pojedynczych sygnałów leżących w widmie szerokopasmowym. Metody szerokopasmowe można zasadniczo podzielić na dwie grupy: metody pracujące z częstotliwością próbkowania równą lub większą od częstotliwości Nyquista i metody, które przetwarzają sygnały z częstotliwością mniejszą od częstotliwości Nyquista. Powyższe rozwiązania, zaliczane są do metod detekcji autonomicznej (niezależnej). Skuteczniejszą metodą monitorowania widma jest sensing kooperacyjny, który polega na podjęciu decyzji, o obecności lub absencji sygnału na podstawie danych otrzymanych z wielu sensorów [10][11][12].

3. PROPOZYCJA ROZWIĄZANIA

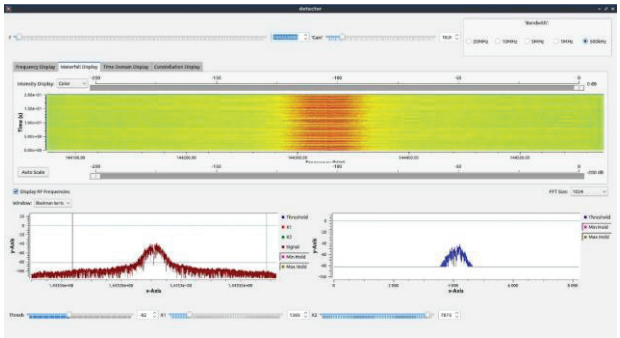
Powszechnie wiadomo, że rozpoznanie powietrzne, ze względu na szybkość i skuteczność działania, stanowi jeden z najlepszych sposobów pozyskiwania informacji. Wykorzystanie dronów w procesie monitorowania widma niesie za sobą wiele korzyści. System oparty na BSP charakteryzuje się wysoką mobilnością oraz krótkim czasem osiągania gotowości do pracy. Wykorzystanie BSP sprawia, że wysokość zawieszenia sensora ograniczona jest jedynie poprzez maksymalny pułapu na jakim może latać wybrany BSP. Dzięki temu operator ma możliwość umieszczenia sensora na wybranej wysokości, w zależności od potrzeb i panujących warunków. Ważną cechą systemów opartych na BSP jest brak specjalnych wymagań dotyczących terenu potrzebnego do rozstawienia systemu. Powoduje to, że sensor taki charakteryzuje się większą skrytością i bezpieczeństwem dla operatora, który może pozostawać w ukryciu w trakcie wykonywania zadań. Kolejną zaletą są niskie koszty eksploatacji w porównaniu ze

statkami załogowymi. Drony są również nieporównywalnie tańsze w zakupie niż klasyczne statki powietrzne, jak również zużywają mniej paliwa, a obsługa techniczna jest zdecydowanie prostsza i mniej kosztowna [1][14]. Małe wymiary i niewielka waga sprawiają, że cały system nie wymaga angażowania dodatkowych środków logistycznych potrzebnych do transportu. Cechy te, w połączeniu z prostotą obsługi, stwarzają możliwość użycia sensora przez żołnierzy znajdujących się bezpośrednio na polu walki, dając tym samym możliwość zdobycia informacji i uzyskania przewagi nad przeciwnikiem bez konieczności angażowania innych systemów. Zależnie od konstrukcji, drony mogą różnić się między sobą zasięgiem, udźwigiem lub sposobem startu. Umożliwia to konfigurację wyposażenia platformy w zależności od warunków czy potrzeb wykonywanego zadania. Ważną zaletą zaproponowanego rozwiązania jest możliwość stworzenia wielosensorowej sieci monitorującej opartej na metodzie sensingu kooperacyjnego. Rozbudowa o kolejne sensory zapewni wzrost wydajności całego systemu oraz zwiększy obszar, z którego mogą być zbierane dane.

3.1. Implementacja rozwiązania

Do implementacji wybrany został detektor energii. Jest to najpopularniejsza metoda charakteryzująca się łatwością implementacji, niewielkimi wymaganiami obliczeniowymi oraz uniwersalnością. Dodatkowo technika ta nie wymaga informacji a'priori dotyczących sygnałów. Detektor energii wykrywa sygnał poprzez pomiar energii sygnału nadawanego w zadanym paśmie, a następnie podejmuje decyzję o obecności transmisji. Decyzja zapada w układzie progowym, gdzie oszacowana wartość energii sygnału porównywana jest z progiem detekcji zależnym od wartości SNR. Jeżeli odebrany sygnał jest większy od wartości progowej, wówczas podejmuje się decyzję o obecności sygnału.

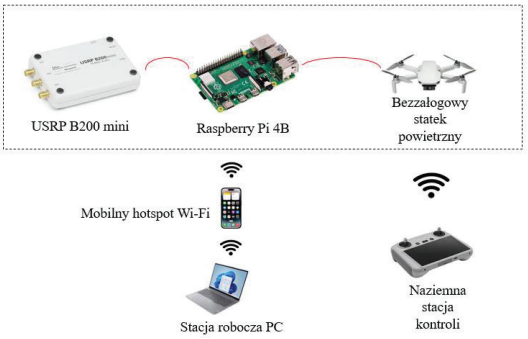
Do opracowania detektora wykorzystano program GNU Radio Companion. Jest to darmowy zestaw narzędzi programistycznych o otwartym kodzie źródłowym, umożliwiających implementację na module radia programowalnego. Interfejs graficzny użytkownika opracowanej aplikacji detektora przedstawia rysunek 1. Zważywszy na wybraną platformę, opracowany sensor wykorzystuje radio programowalne USRP B200 mini. Wybrano to urządzenie ze względu na szerokie pasmo pracy 70MHz-6GHz a także małą wagę i wymiary [4][5][8][13].



Rys. 1 GUI opracowanej aplikacji detektora

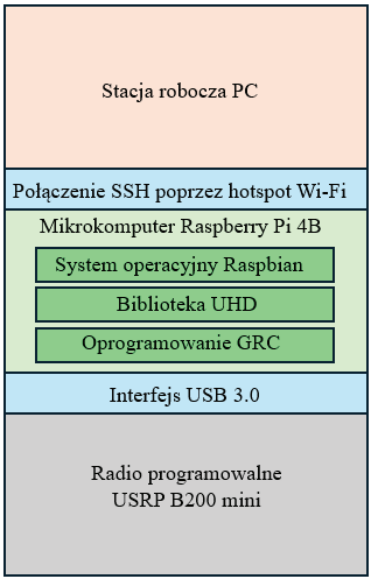
3.2. Integracja detektora z BSP

Integracja opracowanego sensora z dronem, w głównej mierze sprowadza się do bezprzewodowego połącznia komputera z sensorem, w taki sposób, aby zapewnić użytkownikowi możliwość podglądu środowiska radiowego oraz obsługi detektora w czasie rzeczywistym (rys. 2). By spełnić powyższe założenia zdecydowano się na zastosowanie technologii zdalnego pulpitu, która umożliwia zarówno podgląd, jak i sterowanie aplikacją sensora. Do komunikacji między sensorem umieszczonym na BSP i stacją roboczą zastosowano mikrokomputer Raspberry Pi 4B, który dzięki wybudowanemu modułowi WiFi, zapewnił bezprzewodową łączność między sensorem i stacją roboczą. Schemat połączeń pomiędzy poszczególnymi elementami został przedstawione na rysunku 3.



Rys. 2 Schemat integracji z BSP [4]

Zastosowanie technologii zdalnego pulpitu stwarza możliwość uruchominia opracowanej aplikacji na dowolnym urządzeniu mobilnym, np. telefonie komórkowym, zmniejszając tym samym gabaryty całego systemu (rys. 4).



Rys. 3 Schemat blokowy połączeń

Rys. 4 Uruchomiona aplikacja detektora

Cały system jest niewielki i wygodny w transporcie. Mieści się w plecaku, a małe gabaryty powodują, że całość może być przenoszona przez jednego żołnierza bez nadmiernego obciążenia. Sensor zintegrowany z BSP przedstawiono na rysunku 5.

Rys. 5 Zintegrowany z BSP sensor

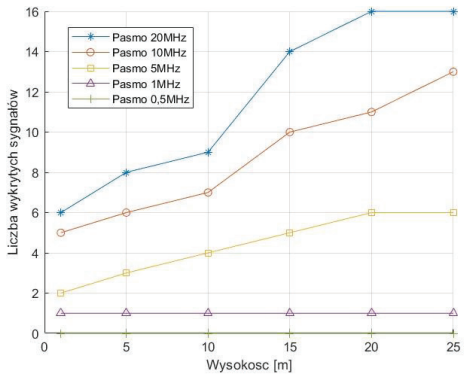
4. BADANIA ROZWIĄZANIA

W celu sprawdzenia efektywności opracowanego rozwiązania przygotowano scenariusze badawcze, w których starano się odwzorować rzeczywiste warunki pracy sensora. Badaniom podlegał wpływ zmiany wysokości zawieszenia sensora na skuteczność rozpoznania oraz wpływ zjawiska przesłaniania na monitorowanie widma. Przedstawiono także koncepcję przeprowadzenia patrolu z wykorzystaniem sensora opartego na BSP. Badania prowadzono empirycznie oraz symulacyjnie.

4.1. Scenariusz nr 1.

W tym scenariuszu badano wpływ zwiększenia wysokości zawieszenia sensora na skuteczność prowadzonego rozpoznania. Jako, że celem badania było wykrycie jak największej liczby sygnałów, zdecydowano się na prowadzenie sensingu w paśmie rozgłoszeniowym systemu radiofonii, charakteryzującym się wysokim poziomem zajętości widma. Proces badawczy polega na prowadzeniu sensingu na ustalonych wysokościach i zdefiniowanych w aplikacji pasmach obserwacji.

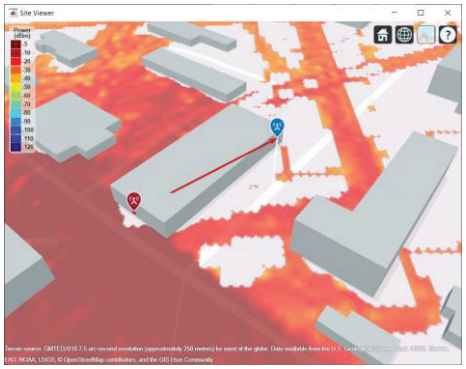
Na podstawie otrzymanych wyników zauważyć można wzrost liczby wykrytych sygnałów wraz ze wzrostem wysokości zawieszenia sensora (rys. 6). Pokazuje to, że umieszczając sensor na BSP można w prosty sposób zwiększyć zasięg rozpoznawania.



Rys. 6 Zależność liczby wykrytych sygnałów w funkcji wysokości zawieszenia sensora

4.2. Scenariusz nr 2.

Badanie wpływu zjawiska przesłaniania na monitorowanie widma zostało przeprowadzone w sposób symulacyjny i empiryczny. Do badań wybrano jeden z budynków znajdujących się na terenie Wojskowej Akademii Technicznej. Pomiary odbywały się w trzech miejscach: 50 m od rogu budynku, 25 m od rogu budynku oraz na wysokości rogu budynku. Dla urządzenia nadawczego przewidziano dwa położenia: na rogu budynku oraz 15 m od rogu na wysokości 1 m. Badania empiryczne poprzedziły badania symulacyjne wykonane w środowisku Matlab (rys. 7). W przypadku gdy urządzenie nadawcze znajdowało się na rogu budynku, wykrycie transmitowanego sygnału było możliwe w dowolnym miejscu znajdującym się na trasie przelotu BSP. W momencie, kiedy nadajnik został oddalony od narożnika budynku konieczne było przemieszczenie sensora w celu wykrycia sygnału. Dodatkowo, na podstawie wyników zawierających poziom mocy odbieranego sygnału, zauważyć można, że zmniejszenie odległości między sensorem i urządzeniem nadawczym powoduje wzrost poziomu mocy odbieranego sygnału.



Rys. 7 Przykładowy wynik przeprowadzonej symulacji

4.3. Scenariusz badawczy nr 3.

Scenariusz badawczy przedstawia koncepcję przeprowadzenia patrolu z wykorzystaniem BSP. Badanie polegało na wykonaniu przelotu, po wcześniej ustalonej trasie z wykorzystaniem sensora zintegrowanego z dronem. Zadaniem operatora było wykrycie trzech emisji prowadzonych na znanych częstotliwościach. Operator wraz ze stanowiskiem dowodzenia znajdował się stałym miejscu, natomiast emisje przeprowadzane były w losowych lokalizacjach znajdujących się na zadanym obszarze placu ćwiczeń taktycznych WAT (rys. 8).

Rys. 8. Trasa przelotu BSP w czasie patrolu

W trakcie trwania patrolu sensor został zawieszony na wysokości 20 m. W trakcie badania operator wykrył wszystkie trzy przeprowadzone transmisje. Potwierdza to skuteczność opracowanego rozwiązania. Operator może pozostać w ukryciu, w trakcie wykonywania patrolu, co znacząco zwiększa bezpieczeństwo oraz zmniejsza ryzyko jego wykrycia.

5. PODSUMOWANIE

Przeprowadzone badania symulacyjne i empiryczne pozwoliły stwierdzić słuszność zastosowania bezzałogowych statków powietrznych w procesie monitorowania. System oparty na dronach spełnia wymagania współczesnego pola walki stawiane narzędziom rozpoznania radioelektronicznego. Zapewnia mobilność, skrytość oraz bezpieczeństwo operatora. Sensory zintegrowane z BSP mogą stanowić rozwiązanie problemu przesłaniania, utrudniającego prowadzenie monitoringu środowiska radiowego szczególnie w terenie zurbanizowanym. Do wad zaproponowanego rozwiązania, wynikających z ograniczeń platformy, zaliczyć można podatność na warunki atmosferyczne oraz ograniczenie wynikające z pojemności baterii. W przyszłości sensory oparte na bezzałogowych platformach mogą stanowić podstawowe narzędzie zapewniające świadomość sytuacyjną w środowisku radiowym. W ramach dalszego rozwoju, opracowany system będzie rozszerzony o większą liczbę sensorów i zastosowanie sensingu kooperacyjnego. Pozwoli to zwielokrotnić wydajność całego systemu oraz zwiększy obszar, z którego mogą być zbierane informacje. Dowodzenie takim systemem odbywać się będzie z jednego miejsca, pełniącego jednocześnie funkcję centrum fuzji danych, do którego dostarczane będą wyniki sensingu. Wykorzystanie roju dronów umożliwi również lokalizację wykrytych źródeł emisji zapewniając tym samym więcej informacji niezbędnych do dalszego planowania i prowadzenia działań. Ponadto system można przystosować do prowadzenia zakłóceń, czyniąc tym samym SBSP pełnoprawnym elementem systemu walki radioelektronicznej.

LITERATURA

[1] Gołąbek Adam, Andrzej Wrona. 2022. „Bezzałogowe systemy powietrzne w Siłach Zbrojnych RP" Lotnictwo Aviation International 6/2022. https://zbiam.pl/artykuly/bezzalogowe-systemy-powietrzne-w-silach-zbrojnych-rp/ (dostęp: 04.2024).

[2] Malon Krzysztof, Skokowski Paweł, Łopatka Jerzy. 2017. „The use of a wireless sensor network to monitor the spectrum in urban areas". Proc. SPIE 10418, XI Conference on Reconnaissance and Electronic Warfare Systems, 1041809. https://doi.org/10.1117/12.2269709

[3] Malon Krzysztof, Skokowski Paweł, Łopatka Jerzy. 2018. „Optimization of wireless sensor network deployment for electromagnetic situation monitoring". International Journal of Microwave and Wireless Technologies. 10(7):746-753. doi:10.1017/S1759078718000211.

[4] Mazuro Maciej. 2022. „Koncepcja sensora monitoringu widma na bezzałogowym statku powiertrznym". Elektronika (8/2022): 20-23.

[5] Mazuro Maciej, Skokowski Paweł. 2022. „Sensor monitoringu widma z wykorzystaniem radia programowalnego". Przegląd Telekomunikacyjny - Wiadomości Telekomunikacyjne.

[6] Report ITU-R SM.2405-1 (12/2020): „Spectrum management principles, challenges and issues related to dynamic access to frequency bands by means of radio systems employing cognitive capabilities".

[7] Skokowski Paweł. 2011. „Efficient Method of 3G Signal Detection." PIERS Online 7.1: 51-54.

[8] Skokowski Paweł. 2021. „Electromagnetic situation awareness building in ad-hoc networks with cognitive nodes". Military University of Technology, ISBN: 978-83-7938-323-8.

[9] Skokowski Paweł, Dułowicz Jan. 2023. „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. pp. 149-151.

[10] Skokowski Paweł, Łopatka Jerzy, Malon Krzysztof. 2020. „Evidence Theory Based Data Fusion for Centralized Cooperative Spectrum Sensing in Mobile Ad-hoc Networks", Baltic URSI Symposium (URSI), Warsaw, Poland, pp. 24-27, doi: 10.23919/URSI48707.2020.9254038.

[11] Skokowski Paweł, Malon Krzysztof, Łopatka Jerzy. 2017. „Properties of centralized cooperative sensing in cognitive radio networks". Proc. SPIE 10418, XI Conference on Reconnaissance and Electronic Warfare Systems, 1041807. https://doi.org/10.1117/12.2269996.

[12] Skokowski Paweł, Malon Krzysztof, Łopatka Jerzy. 2017. „Sensing kooperacyjny scentralizowany vs. Lokalny w kognitywnych sieciach radiowych". Przegląd Telekomunikacyjny-Wiadomości Telekomunikacyjne, wydanie 8-9.

[13] Skokowski Paweł, Marszałek Piotr. 2012. „Wykorzystanie technologii radia programowalnego SDR do monitorowania widma". IX Konferencja Naukowo-Techniczna Systemy Rozpoznania i Walki Elektronicznej KNTWE. Kazimierz Dolny.

[14] Zmysłowski Dariusz, Skokowski Paweł, Kelner Jan. 2023. „Anti-drone Sensors, Effectors, and Systems – A Concise Overview", TransNav, the International Journal on Marine Navigation and Safety of Sea Transportation 2023 | Journal article DOI: 10.12716/1001.17.02.23.

REKURENCYJNY ALGORYTM DETEKCJI SYGNAŁU

RECURSIVE SIGNAL DETECTION ALGORITHM

Bartosz Bossy[1], Grzegorz Fotyga[1,2], Aleksander Nowak[1], Paweł Żukowski[1]

[1] Hertz Systems Ltd sp. z o. o., Zjednoczenia 118A, 65-120 Zielona Góra
b.bossy@hertzsystems.com, a.nowak@hertzsystems.com, p.zukowski@hertzsystems.com
[2] Faculty of Electronics, Telecommunications, and Informatics, Gdansk University of Technology, Gdansk, Poland
g.fotyga@hertzsystems.com

DOI: 10.15199/59.2024.4.81

Streszczenie: W niniejszym artykule przedstawiono nowy rekurencyjny algorytm detekcji sygnału, bazując na analizie częstotliwościowo czasowej. Algorytm przetestowano dla szerokiej gamy rzeczywistych sygnałów zarejestrowanych z wykorzystaniem stanowiska pomiarowego. Wyniki pokazują, że zaproponowany algorytm bardzo dobrze wykrywa sygnał użyteczny, a w dodatku jest skalowany. Ponadto w obecnej postaci detekcja zajmuje dwukrotność czasu trwania sygnału, a możliwe jest zmniejszenie jej czasu poprzez m.in. zrównoleglenie obliczeń.

Abstract: This article presents the proposed recursive signal detection algorithm based on time-frequency analysis. The algorithm has been tested for a wide range of real signals recorded using a measurement station. The results show that the proposed algorithm detects the useful signal very well and is also scalable. Moreover, in its current form, detection takes twice the signal duration and it is possible to reduce its time by, among others, parallelization of calculations.

Słowa kluczowe: detekcja, dron, pomiary, zagłuszanie

Keywords: detection, drone, jamming, measurements

1. WSTĘP

W ciągu ostatniej dekady nastąpił gwałtowny rozwój bezzałogowych systemów powietrznych (ang. *Unmanned Aerial System* - UAS) we wszystkich sektorach: wojskowym, komercyjnym i cywilnym. Jednakże te postępy stwarzają również znaczące wyzwania w zakresie zabezpieczenia infrastruktury krytycznej, takiej jak energetyka, telekomunikacja i transport. W szczególności w ostatnich latach zaobserwowano znaczne zakłócenia na lotniskach spowodowane UASs, z których najpoważniejsze zostały wymienione w [1].

W ostatnich latach zaproponowano wiele metod detekcji i klasyfikacji sygnałów związanych z dronami w celu ochrony infrastruktury krytycznej. Rozwiązanie zaproponowane w [2] wprowadza podejście oparte na sieciach neuronowych do klasyfikacji sygnałów. Podejście to obejmuje następujące kroki: 1) tworzenie bazy danych sygnałów; 2) ręczne etykietowanie i adnotowanie bazy danych, co zazwyczaj jest procesem czasochłonnym; oraz 3) zastosowanie sieci neuronowej do rozpoznawania sygnałów. Alternatywnie, można wykorzystać ogólny model stworzony przez firmę DeepSig, który jest ciągle aktualizowany. Ten model potrafi rozróżniać różne sygnały, takie jak: LTE, Wi-Fi, Bluetooth. W publikacji [3] opisano implementację [2] na wydajnym serwerze HPE Edgeline EL8000, który obejmuje trening modelu i klasyfikację sygnałów.

Alternatywne podejście do wykrywania sygnałów związanych z transmisją danych w górę i w dół opiera się na analizie cyklostacjonarności [4, 5]. Jest to również metoda używana w analizie sygnałów FHSS (ang. *Frequency-Hopping Spread Spectrum*) [6]. W [7] zaproponowano metodę detekcji spektrum opartą na wartościach własnych macierzy kowariancji sygnałów odbieranych przez użytkowników wtórnych.

2. DETEKCJA SYGNAŁU

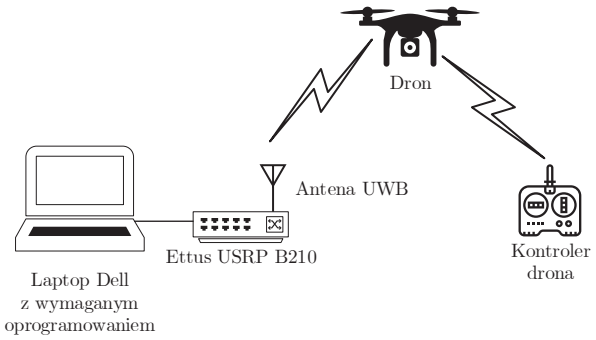
Głównym celem detekcji sygnału jest określenie czy dany nadajnik, w tym przypadku dron, transmituje sygnał. Matematycznie decyzję dotyczącą wykrywania sygnału $D(x)$ można traktować jako statystykę podwójnej hipotezy:

$$D(x) = \begin{cases} \mathcal{H}_0 & \text{jeżeli } r(x) = n(x) \\ \mathcal{H}_1 & \text{jeżeli } r(x) = s(x) + n(x), \end{cases} \tag{1}$$

gdzie $\mathcal{H}_0$ jest hipotezą, że odebrany sygnał to szum, natomiast $\mathcal{H}_1$ jest hipotezą, że odebrany sygnał jest sumą szumu oraz sygnału z nadajnika. Jakość detekcji jest opisana przez prawdopodobieństwo detekcji P_d czyli prawdopodobieństwo, że $D(x) = \mathcal{H}_1$ w przypadku, gdy nadajnik faktycznie transmituje sygnał oraz prawdopodobieństwo błędnej detekcji (ang. *miss-detection*) P_md tzn. prawdopodobieństwo, że $D(x) = \mathcal{H}_0$ w przypadku, gdy nadajnik jest aktywny a nie wykrył żadnego sygnału. Ponadto ważnym wskaźnikiem jest prawdopodobieństwo fałszywego alarmu P_f definiowane jako prawdopodobieństwo, że $D(x) = \mathcal{H}_1$ w przypadku, gdy nadajnik faktycznie nie transmituje.

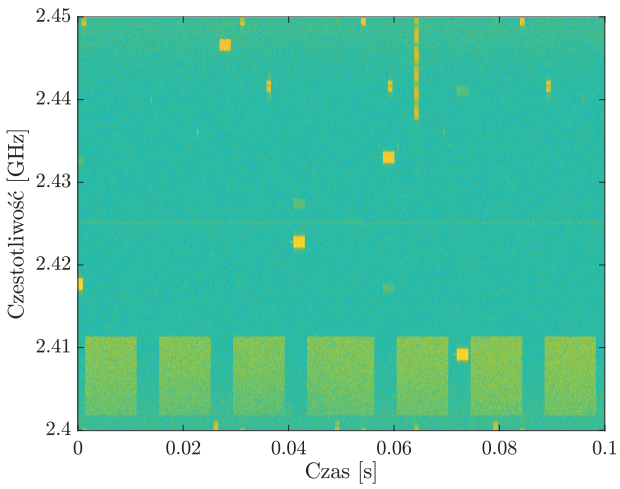
3. REJESTROWANIE SYGNAŁU

Głównym celem prezentowanych w artykule prac jest stworzenie algorytmu detekcji sygnałów transmitowanych przez drony komercyjne. W związku z tym w pierwszym etapie zarejestrowano sygnały w paśmie 2,4 GHz bez rozróżnienia czy są to sygnały sterujące w łączu w górę i w dół, czy transmisja obrazu. Do rejestracji sygnału użyto systemu przedstawionego na Rysunku 1.

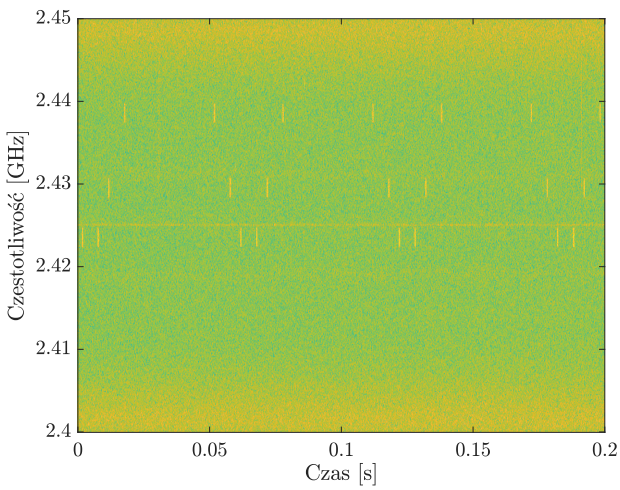


Rysunek 1: Schemat poglądowy stanowiska pomiarowego.

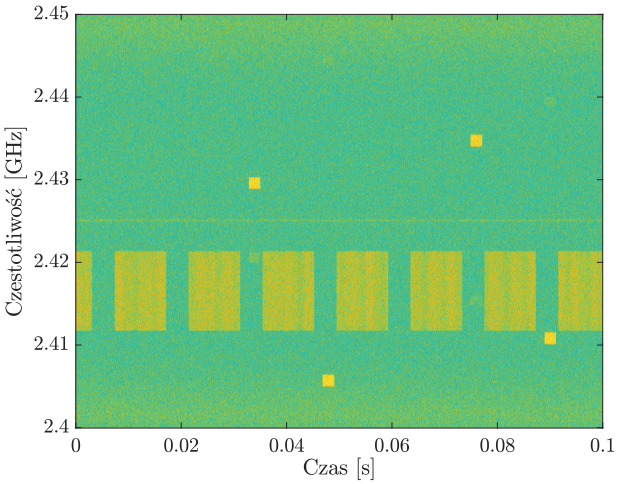
Częstotliwość próbkowania wynosiła 50 MHz a czas rejestrowanego sygnału to 2 sekundy. Sygnał był rejestrowany w miejscu oddalonym od innych nadajników w celu zarejestrowania tylko sygnału wykorzystywanego do komunikacji z dronem. Poniżej przedstawiono spektrogramy kilku zarejestrowanych sygnałów.



Rysunek 2: Spektrogram dla transmisji drona Autel Evo Nano+.



Rysunek 3: Spektrogram dla transmisji Dji MiniPro 3.



Rysunek 4: Spektrogram dla transmisji Dji Inspire.

Można zaobserwować, że poszczególne transmisje różnią się od siebie drastycznie, nawet jeżeli drony pochodzą od tego samego producenta (Rysunek 3 i Rysunek 4). Ponadto widoczne są transmisje różnego typu. W ogólności można wyróżnić transmisję OFDM (ang. *Orthogonal Frequency-Division Multiplexing*) wykorzystywana do transmisji video oraz FHSS, za pomocą której przesyłane są dane sterujące, które zajmują wąskie pasmo i trwają maksymalnie kilka milisekund. Dlatego też konieczne zaprojektowanie jest elastycznego algorytmu detekcji sygnału o niskiej złożoności, aby możliwe było zagłuszanie w czasie rzeczywistym.

4. ALGORYTM DETEKCJI

Zaproponowany rekurencyjny algorytm detekcji sygnału bazuje na jego analizie czasowo-częstotliwościowej reprezentowanej przez spektrogram. Problemem w detekcji bazującej na analizie czasowo-częstotliwościowej jest fakt, że parametry wyznaczonego spektrogramu mają istotny wpływ na jakość detekcji. Jest tak, ponieważ wysoka rozdzielczość w dziedzinie częstotliwości spowoduje obniżenie rozdzielczości w czasie i odwrotnie. Zatem wysoka jakość detekcji w dziedzinie częstotliwości może spowodować obniżenie jej jakości w czasie i vice versa. Rozwiązaniem tego problemu może być zwiększenie liczby próbek, które nakładają się między sąsiednimi segmentami, jednak skutkuje to znacznym zwiększeniem liczby punktów spektrogramu, a to może skutkować zwiększeniem czasu wyznaczania samego spektrogramu, ale także zwiększenia czasu detekcji. Dlatego też konieczny jest kompromis pomiędzy jakością detekcji a czasem detekcji.

Dla wyznaczonego spektrogramu kolejnym krokiem zaproponowanego algorytmu jest analiza zarejestrowanego sygnału w czasie. W tym celu dla danej chwili czasu zsumowano wszystkie próbki z dziedziny częstotliwości, otrzymując tym samym informację o tym w jakich chwilach czasu transmitowany jest sygnał. Matematycznie można tę operację przedstawić za pomocą następującego wzoru:

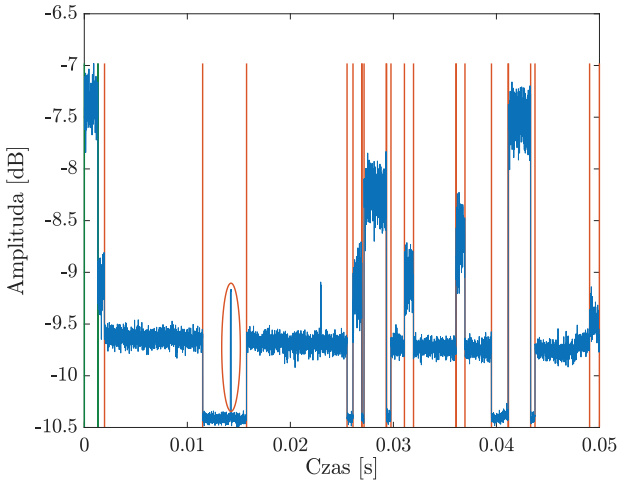
$$\forall n \quad \sum_{\omega \in \Omega} \left| \mathbf{STFT} \left\{ x\left[n\right] \right\} \left(\tau, \omega\right) \right|^2, \tag{2}$$

gdzie $\mathbf{STFT}\left\{x\left[n\right]\right\}\left(\tau,\omega\right)$ to krótkookresowa transforma-

ta Fouriera sygnału $x[n]$ i częstotliwości ω dana wzorem:

$$\mathbf{STFT}\{x[n]\}(\tau,\omega) \equiv X(\tau,\omega) = \sum_{n=-\infty}^{\infty} x[n]\, w[n-\tau]\, e^{-i\omega n}, \tag{3}$$

gdzie $w(\tau)$ to funkcja okna, którą w tym przypadku jest funkcja Hann. Dla tak wyznaczonego sygnału kolejnym krokiem jest wykrycie punktów, dla których następują zmiany (ang. *change point detection*). Wykorzystano w tym celu algorytm *bottom up segmentation* [8, 9] o złożoności obliczeniowej $\mathcal{O}(n \log n)$, gdzie n to liczba próbek. W celu przyspieszenia obliczeń, w zaproponowanym algorytmie analizowany sygnał jest dzielony na mniejsze fragmenty. Na Rysunku 5 przedstawiono przykładowe wyniki działania algorytmu *bottom up segmentation* dla sygnału po wykonaniu operacji (2) oraz spektrogramu z Rysunku 2. Można zaobserwować, że większość zmian w sygnale została poprawnie wykryta niemniej jednak są widoczne fragmenty, dla których następuje zmiana i nie została ona wykryta (przykład takiej sytuacji jest oznaczony elipsą). Wynika to z ustawień "czułości" algorytmu detekcji punktu zmiany, który z jednej strony może poprawić jakość detekcji sygnału, ale może też zwiększyć czas detekcji. Problem ten został rozwiązany przez rekurencyjne wywołanie zaproponowanej metody detekcji. Dzięki temu pierwsza (Rysunek 5) analiza w czasie nie jest ostateczna, a jest wejściem do kolejnego wywołania algorytmu.

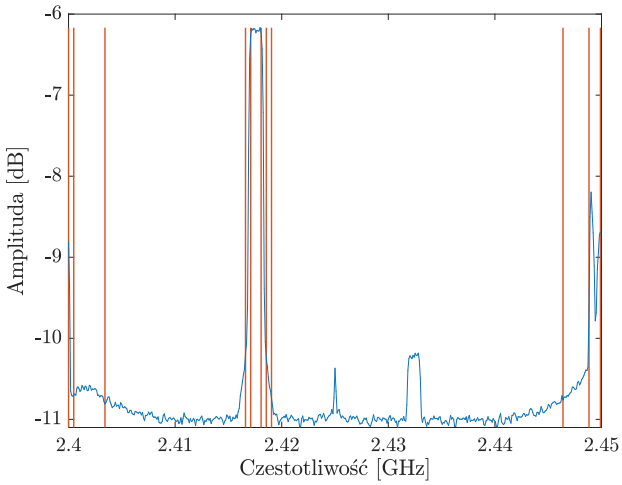


Rysunek 5: Rezultat operacji (2) oraz działania algorytmu *bottom up segmentation* dla Autel Evo Nano+.

Jak wspomniano, operacja (2) sumuje próbki ze wszystkich częstotliwości w danym czasie, a jak wiadomo w tym samym czasie może być transmitowanych kilka sygnałów na różnych częstotliwościach. Zatem w kolejnym kroku dla każdej wykrytej w czasie transmisji przeprowadzono analizę częstotliwością. W tym celu dla wyznaczonych wcześniej przedziałów czasowych zsumowano sygnał w dziedzinie czasu, co można przedstawić następującym wzorem:

$$\forall\omega \quad \sum_{n\in\mathcal{N}} |\mathbf{STFT}\{x[n]\}(\tau,\omega)|^2, \tag{4}$$

gdzie $\mathcal{N}$ to zbiór chwil czasu wykryty w poprzednim kroku. Następnie dla sygnału będącego wynikiem tej operacji zostają wykryte punkty zmiany za pomocą wcześniej wspominanego algorytmu *bottom up segmentation*. Poniżej, na Rysunku 6, przedstawiono wykryte punkty zmiany dla sygnału będącego wynikiem (4) dla chwil czasowych pomiędzy zielonymi liniami z Rysunku 5. Kolejny raz można zaobserwować, że nie wszystkie zmiany zostały wykryte. Powód jest ten sam, czyli "czułość" algorytmu detekcji punktów zmian. Oczywiście można zmienić parametr odpowiedzialny za "czułość", lecz znacznie zwiększa to czas detekcji i dużo lepszym rozwiązaniem jest rekurencyjne wywołanie algorytmu.



Rysunek 6: Rezultat operacji (4) oraz działania algorytmu *bottom up segmentation* dla Autel Evo Nano+.

Ostatnim etapem decydującym o tym, czy daną transmisję można uznać za wykrytą jest dalsza analiza w czasie, ale tylko dla konkretnych częstotliwości, czyli dla sygnału będącego wynikiem poniższej operacji:

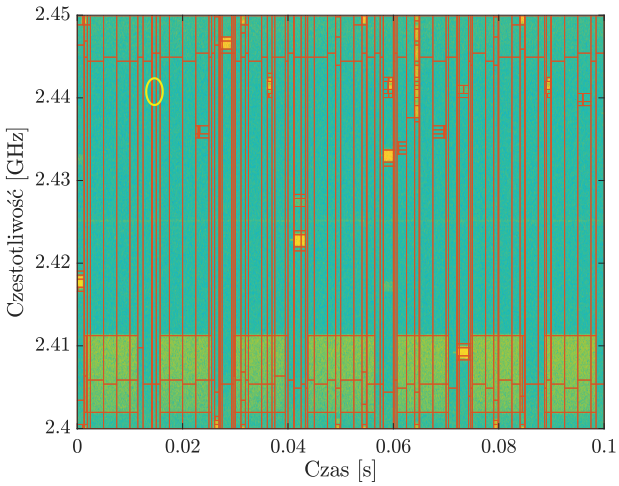
$$\forall n \in \mathcal{N} \quad \sum_{\omega\in\Psi} |\mathbf{STFT}\{x[n]\}(\tau,\omega)|^2, \tag{5}$$

gdzie Ψ to zbiór częstotliwości wykryty w poprzednim kroku. Jeżeli dla takiego sygnału nie wykryto nowych punktów zmiany oznacza to, że uznajemy transmisję za wykrytą. W przeciwnym przypadku oznacza to, że w danym paśmie i w danym czasie zostały wykryty sygnał, który wcześniej został pominięty. Dlatego też dla nowo wykrytego sygnału dokonywane jest ponowna analiza częstotliwościowa poprzez rekurencyjne wywołanie. Procedura ta jest powtarzana, aż do momentu, gdy nowy sygnał (punkt zmiany) dla danego pasma i czasu nie zostaje wykryty.

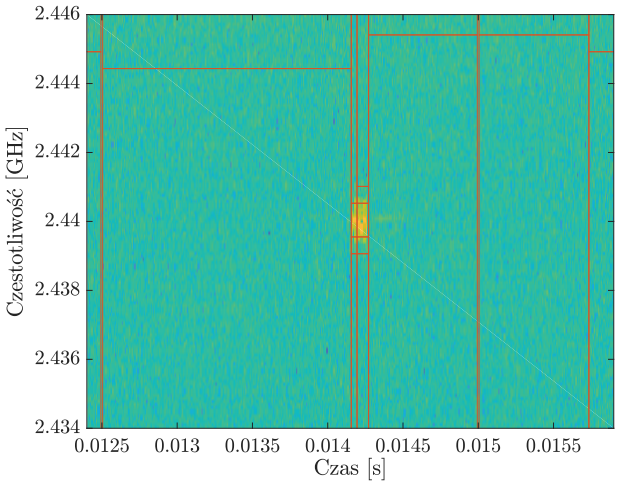
5. WYNIKI

W niniejszym rozdziale przedstawiono wyniki zaproponowanego rekurencyjnego algorytmu detekcji dla sygnałów zarejestrowanych zgodnie z procedurą i parametrami opisanymi w Rozdziale 3. Rysunek 7 przedstawia wynik działania algorytmu przed wykonaniem operacji progowania dla drona Autel Evo Nano+. Można zauważyć, że algorytm z dużą dokładnością rozdziela spektrogram na fragmenty zawierające transmisję, jak i te zawierające szum. Ponadto niektóre bloki są dzielone na mniejsze, ze względu na zmieniająca się moc. Widać to szczególnie na granicach pasma, gdzie występują wyższy poziom szumu wynikający z niedoskonałości urządzeń pomiarowych. Rysunek

8 przedstawia zbliżenie fragmentu z Rysunku 7 zaznaczonego żółtą elipsą, który również odpowiada niewykrytemu fragmentowi z Rysunku 5. Rysunek ten potwierdza, że rekurencyjne wywołanie algorytmu detekcji pozwala na zwiększenie skuteczności detekcji. Kolejne wyniki zostały wygenerowane dla drona DjiMini Pro3 i zostały przedstawione na Rysunku 9. Ponownie można zaobserwować wysoką skuteczność zaproponowanego algorytmu.

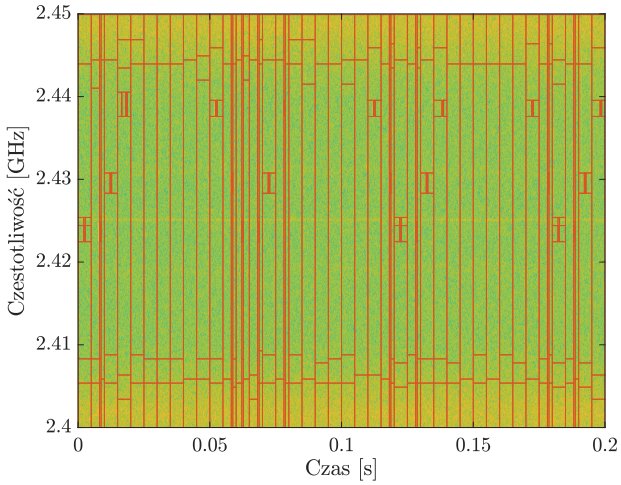


Rysunek 7: Wynik działania algorytmu dla drona Autel Evo Nano+.



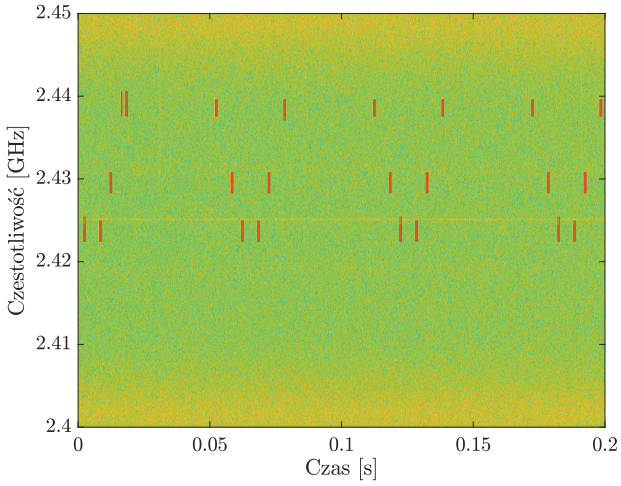
Rysunek 8: Zbliżony frgamentu z Rysunku 7 zaznaczony zieloną elpisą.

Podsumowując, zaproponowany rekurencyjny algorytm detekcji sygnału poprawnie rozdziela spektrogram na fragmenty zawierające transmisję, jak i te zawierające szum, zachowując przy tym dużą dokładność. Ponadto zaproponowany algorytm jest skalowalny w swej dokładności i możliwy do zrównoleglenia, ponieważ każdy wykryty w pierwszym kroku fragment transmisji może być przetwarzany niezależnie. Jak już wspomniano najbardziej czasochłonnym etapem zaproponowanego algorytmu jest wywoływanie metody znajdującą punkt zmiany, i to on w dużej mierze determinuje czas detekcji. W zaproponowanym algorytmie wykorzystano metodę *bottom up segmentation*, dla którego czas wykonania zaproponowanego algorytmu detekcji wynosi dwukrotność czasu trwania analizowanego sygnału.



Rysunek 9: Wynik działania algorytmu dla drona Dji MiniPro 3.

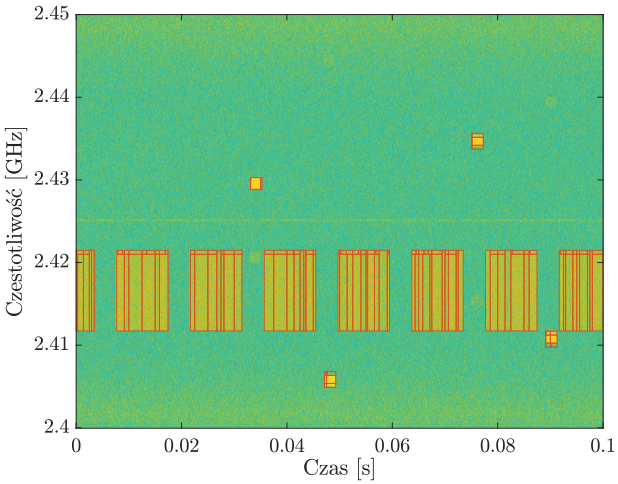
Jak wspomniano zaproponowany algorytm dzieli spektrogram na fragmenty zawierające transmisję, jak i te zawierające szum, zachowując przy tym dużą dokładność. Niemniej jednak w celu zagłuszania należy wyodrębnić tylko te fragmenty, dla których występuje transmisja sygnału. Metodą zaproponowaną w tym artykule jest klasyfikacja na bloki zawierające szum i transmisje poprzez progowanie na podstawie amplitudy. Jest to relatywnie prosta i szybka metoda, ale jej główną trudnością jest określenie wartości, od której blok jest uznawany za ten zawierający sygnał. Rysunek 10 oraz Rysunek 11 przedstawiają wyniki progowania na podstawie amplitudy dla Dji MiniPro 3. Można zaobserwować, że poprawnie dobrany próg odpowiednio usuwa fragmenty spektrogramu zawierające szum.



Rysunek 10: Wyniki dla progowania na podstawie amplitudy dla drona Dji MiniPro 3.

6. WNIOSKI

W artykule przedstawiono zaproponowany rekurencyjny algorytm detekcji sygnału. Zaproponowany algorytm bazuje na analizie czasowo-częstotliwościowej reprezentowanej przez spektrogram. Parametry wyznaczonego spektrogramu mają istotny na jakość detekcji, ponieważ wysoka rozdzielczość w dziedzinie częstotliwości spowoduje obniżenie rozdzielczości w czasie i odwrotnie. Zatem wysoka



Rysunek 11: Wyniki dla progowania na podstawie amplitudy dla drona Dji Inspire.

jakość detekcji w dziedzinie częstotliwości może spowodować obniżenie jej jakości w czasie i vice versa. W zaproponowanym algorytmie rozwiązano ten problem poprzez rekurencyjne wywołanie metody w myśl zasady "dziel i zwyciężaj". Zaproponowany algorytm został przetestowany dla zarejestrowanych rzeczywistych sygnałów dla szerokiej gamy dronów. Wyniki pokazują, że zaproponowany rekurencyjny algorytm detekcji sygnału poprawnie rozdziela spektrogram na fragmenty zawierające transmisję, jak i te zawierające szum, zachowując przy tym dużą dokładność. Ponadto zaproponowany algorytm jest skalowalny w swej dokładności, a zastosowanie odpowiedniego progowanie pozwala na wyodrębnienie użytecznego sygnału. Finalnie, mimo że w obecnej postaci detekcja zajmuje dwukrotność czasu trwania analizowanego sygnału, to istnieje możliwość zrównoleglenia obliczeń oraz zastosowanie procesorów graficznych GPU (ang. *Graphics Processing Unit*). Wstępne analizy pokazują, że możliwe jest zmniejszenie czasu detekcji, tak aby był on porównywalny bądź mniejszy niż czas trwania sygnału co pozwoli na zakłócanie w czasie rzeczywistym bazują na tej detekcji.

PODZIĘKOWANIA

Publikacja powstała w wyniku realizacji Projektu pt.: „Sieciocentryczny system monitorowania ruchu i ochrony przestrzeni powietrznej przed BSP dla prewencji w stanach zagrożenia publicznego, ochrony infrastruktury krytycznej i obiektów publicznych (SSAD)" realizowany w ramach Działania 1.1 Programu Operacyjnego Inteligentny Rozwój 2014-2020 współfinansowanego ze środków Europejskiego Funduszu Rozwoju Regionalnego.

LITERATURA

[1] Carolyn J. Swinney oraz John C. Woods. 2021. „The Effect of Real-World Interference on CNN Feature Extraction and Machine Learning Classification of Unmanned Aerial Systems". Aerospace 8, 7(179). https://doi.org/10.3390/aerospace8070179

[2] Deepsig. 2023. „Using OmniSIG SDK to Create a Drone Detection Model". https://www.deepsig.ai/white-papers/

[3] Deepsig. 2023. „Wireless Signal Identification and Analysis". https://www.deepsig.ai/white-papers/

[4] S. Horstmann, D. Ramírez oraz P. J. Schreier. 2020. „Two-Channel Passive Detection of Cyclostationary Signals". Proceedings of the IEEE Transactions on Signal Processing, 68:2340-2355, 2020, doi: 10.1109/TSP.2020.2981767

[5] P. Urriza, E, Rebeiz oraz D. Cabric. 2013. „Multiple antenna cyclostationary spectrum sensing based on the cyclic correlation significance test". IEEE Journal on Selected Areas in Communications, 31(11):2185-2195

[6] H. Fan, Y. Guo oraz Y. Xu. 2008, „A novel algorithm of blind detection of frequency hopping signal based on second-order cyclostationarity". Proceedings of the Congress on Image and Signal Processing, 5:399-402

[7] Y. Zeng, Y-C Liang. 2009. „Eigenvalue-based spectrum sensing algorithms for cognitive radio". IEEE transactions on communications, 57(6):1784-1793

[8] Keogh, E., Chu, S., Hart, D., and Pazzani, M. 2001. „An online algorithm for segmenting time series". Proceedings of the IEEE International Conference on Data Mining (ICDM), 289–296.

[9] Fryzlewicz, P. 2007. „Unbalanced Haar technique for nonparametric function estimation". Journal of the American Statistical Association, 102(480):1318–1327.

EFEKTYWNA METODA LOKALIZACJI URZĄDZEŃ NADAWCZYCH W ŚRODOWISKU Z PROPAGACJĄ WIELODROŻNĄ

EFECTIVE METHOD FOR LOCATING TRANSMITTER DEVICES IN MULTIPATH 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: 10.15199/59.2024.4.82

Streszczenie: **W artykule przedstawiono badania dotyczące lokalizacji urządzeń nadawczych w terenie zurbanizowanym przy użyciu zaawansowanych technik takich jak algorytm MUSIC, wsteczne śledzenie promieni i sztuczna inteligencja. Zastosowanie tych metod pozwoliło na znaczną poprawę dokładności lokalizacji, szczególnie przy użyciu sieci neuronowych do filtracji danych z BR a poprzez to na znaczące zwiększenie dokładności lokalizacji w terenie zurbanizowanym.**

Abstract: **This article presents research on the localization of transmitting devices in urban areas using advanced techniques such as MUSIC algorithm, reverse raytracing and artificial intelligence. The application of these methods significantly improved localization accuracy, especially when using neural networks for filtering raytracing data, and consequently increase accuracy in the urban environment.**

Słowa kluczowe: **algorytm MUSIC, lokalizacja, propagacja sygnału, raytracing, sieci neuronowe**

Keywords: **Localization, MUSIC algorithm, neural networks, raytracing, signal propagation**

1. WSTĘP

Lokalizacja urządzeń nadawczych stanowi kluczowy aspekt wielu dziedzin, w tym telekomunikacji, nawigacji, a także bezpieczeństwa publicznego [1]. W terenie zurbanizowanym gdzie gęstość zabudowy i inne przeszkody mogą znacząco zakłócać propagację sygnału, precyzyjne namierzanie urządzeń staje się szczególnym wyzwaniem. Tradycyjne metody lokalizacji, takie jak triangulacja oparta na pomiarach odległości, lub kątach, często nie są wystarczająco dokładne ze względu na zjawisko wielodrogowości. Ma to również znaczenie w kontekście działań militarnych [2], gdzie rozpoznanie radiowe jest pierwszym ogniwem walki, dostarczającym informacji o przeciwniku, w tym rozmieszczeniu jego ugrupowania.

Ponieważ coraz więcej operacji wojskowych przeprowadzanych jest w środowisku zurbanizowanym, powoduje to konieczność adaptacji sprzętu wojskowego do wymagań współczesnego pola walki. Precyzyjna lokalizacja nadajników w takich warunkach może znacząco wpłynąć na efektywność działań operacyjnych oraz bezpieczeństwo jednostek. Jedną z metod, które mogą skutecznie radzić sobie z tymi problemami, jest wykorzystanie zaawansowanych technik przetwarzania sygnałów oraz sztucznej inteligencji. W tym kontekście algorytm MUSIC (ang. Multiple Signal Classification) stanowi istotną metodę, która umożliwia jednoczesne namierzenie wielu ścieżek sygnału nawet w warunkach silnej wielodrogowości oraz niedostępności sygnału LOS (ang. Line of Sight).

Algorytm MUSIC opiera się na analizie przestrzennej sygnałów odbieranych przez szyk antenowy przy użyciu macierzy korelacji przekształconej do podprzestrzeni sygnałowej i szumowej [3]. W efekcie algorytm ten potrafi identyfikować kierunki nadejścia sygnałów oraz ich amplitudy, co można wykorzystać do określenia położenia nadajnika. Matematycznie algorytm MUSIC polega na dekompozycji przestrzeni sygnałowej na podprzestrzenie sygnałową i szumową przy użyciu analizy wartości osobliwych (SVD) macierzy korelacji sygnału:

$$R = A\Lambda A^H + \sigma I \qquad (1)$$

Gdzie R to macierz autokorelacji sygnału, A to macierz wektorów własnych sygnału, Λ to macierz wartości własnych, σ to moc szumu, a I to macierz jednostkowa [4]. Macierz wektorów własnych sygnału przechowuje informacje o kierunkach nadejścia sygnałów do szyku antenowego, natomiast macierz wartości własnych przechowuje informacje o sile tych sygnałów.

W opracowanym rozwiązaniu autorzy zaproponowali wykorzystanie algorytmu MUSIC do wyznaczenia ścieżek sygnałów odbitych [5]. Znajomość azymutów poszczególnych ścieżek można wykorzystać do określenia lokalizacji nadajnika z pomocą metody wstecznego śledzenia promieni (ang. backward raytracing - BR), które polega na analizie trajektorii odbitych promieni sygnału. BR symuluje propagację fal radiowych w odwrotnym kierunku od odbiornika do nadajnika, uwzględniając odbicia od przeszkód takich jak budynki czy góry. Dzięki temu możliwe jest dokładne określenie ścieżek sygnału i identyfikacja położenia nadajnika.

Matematycznie, BR polega na odwzorowaniu trajektorii odbitych fal i analizie punktów przecięcia tych trajektorii w przestrzeni. W tym procesie uwzględnia się zarówno kierunki nadejścia sygnałów zidentyfikowane przez algorytm MUSIC, jak i topologię środowiska zurbanizowanego.

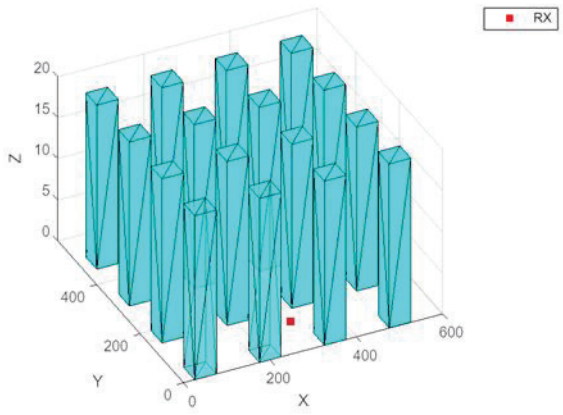
Analiza BR pozwala na znalezienie punktów przecięcia promieni, które określają przybliżoną lokalizację

nadajnika. W środowisku zurbanizowanym drogi sygnałów nie są liniami prostymi, lecz podlegają odbiciom i wielokrotnie wzajemnie się przecinają. W związku z tym, zastosowano sieci neuronowe, które na podstawie odebranych sygnałów są w stanie wskazać na obecność nadajnika w ramach niewielkiego obszaru. Pozwala to na odrzucenie punktów przecięcia się ścieżek sygnału znajdujących się poza wyznaczonym przez sieć sektorem, a w konsekwencji zwiększa to dokładność lokalizacji.

2. METODOLOGIA

2.1. Symulacje w środowisku MATLAB

Badanie rozpoczęto od opracowania symulacji w środowisku MATLAB. Wykorzystano model typu Manhattan, w którym umieszczono 16 wieżowców oraz urządzenie nadawcze i odbiorcze. Symulacje obejmowały wyznaczenie ścieżek sygnału dla stałego położenia odbiornika oraz różnych położeń nadajnika w celu zebrania danych treningowych dla sieci neuronowych. Zarówno nadajnik jak i odbiornik umieszczono na tej samej wysokości aby ograniczyć badanie do jednej płaszczyzny co znacząco upraszcza proces symulacji i uczenia. Opisaną sytuację ilustruje Rys. 1.



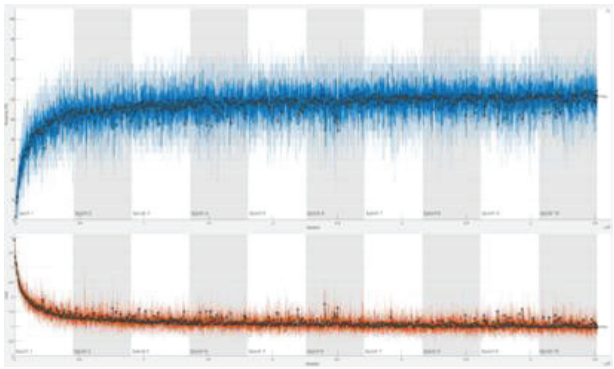
Rys. 1 Ilustracja wykorzystanego modelu symulującego środowisko miejskie

2.2. Przetwarzanie danych generowanych przez śledzenie promieni

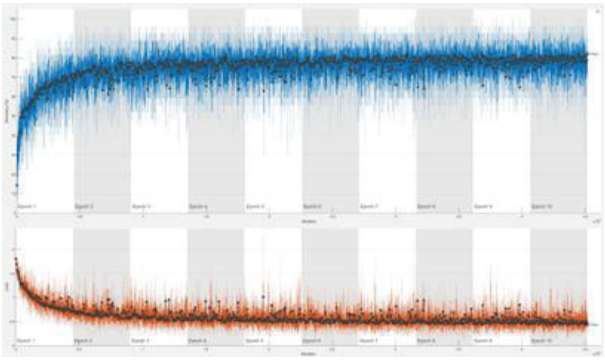
Dane ze śledzenia promieni pozyskane do treningu sieci neuronowych zostały poddane obróbce tak, aby dane jakie posłużą do treningu sieci były możliwe do zaobserwowania z punktu widzenia niezależnego odbiornika, tj. do treningu zostały wykorzystane dane o kącie nadejścia, zmianie fazy oraz mocy poszczególnych promieni a także czasie ich nadejścia względem pierwszego odebranego promienia. Dane te zostały następnie podzielone na zbiór treningowy, testowy i walidacyjny.

W ramach symulacji stworzono dwie sieci neuronowe. Jedną, która dzieliła badany obszar na 51 małych sektorów, oraz drugą, która dzieliła obszar na 6 większych sektorów. Zadaniem tych sieci było wskazanie na podstawie danych z odbiornika, w którym sektorze znajduje się nadajnik. Poddano je treningowi na przygotowanych wcześniej danych.

Obie sieci zostały zaimplementowane z wykorzystaniem architektury LSTM (Long Short-Term Memory), która jest rodzajem neuronowej sieci rekurencyjnej (RNN). W warstwie LSTM zastosowano funkcję sigmoidalną zaimplementowaną w bramkach odpowiadających za przepływ informacji z i do komórek pamięci. Funkcja tangensa hiperbolicznego zastosowana została natomiast do tworzenia nowych wartości komórek pamięci oraz przekształcania wyjścia komórki pamięci. Zastosowano również warstwę softmax, która korzystając z funkcji o tej samej nazwie przekształca wyjścia sieci na prawdopodobieństwo dla każdej klasy (sektora, w którym znajduje się źródło). Wykresy opisujące proces treningu ww. sieci zilustrowane są na Rys. 2 i Rys. 3.



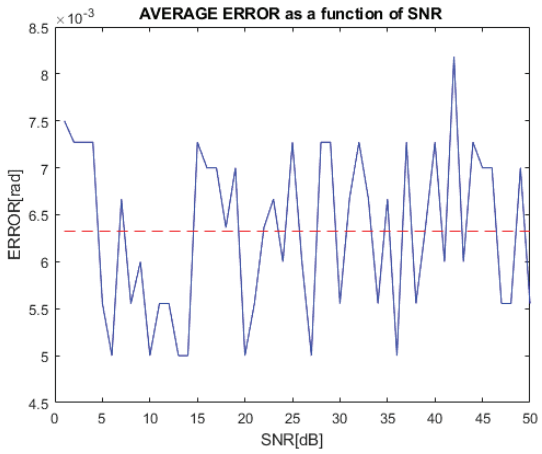
Rys. 2 Wykres przedstawiający dokładność (linia niebieska) i stratę (linia czerwona) w czasie treningu sieci dzielącej obszar na 51 sektorów



Rys. 3 Wykres przedstawiający dokładność (linia niebieska) i stratę (linia czerwona) w czasie treningu sieci dzielącej obszar na 6 sektorów

2.3. Realizacja BR

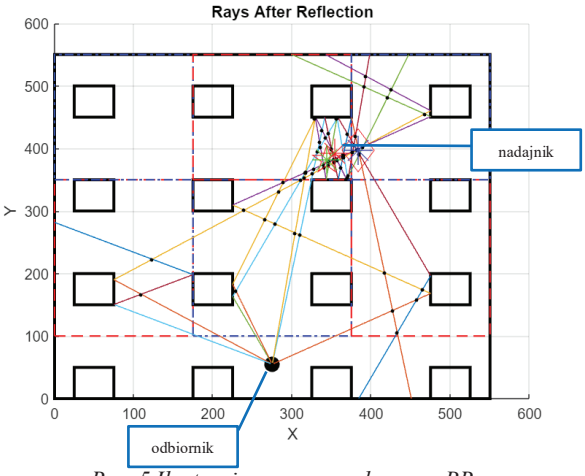
Dane ze śledzenia promieni służące do treningu sieci neuronowych podzielono na zbiory w zależności od ilości odebranych promieni. Odseparowano również pomiary, w których wystąpiły sygnały LOS, gdyż jest to niezwykle dogodny scenariusz w środowisku zurbanizowanym i zaburzało to poprawną interpretację wyników. Tak przygotowane zbiory wykorzystano do zasymulowania działania algorytmu MUSIC korygując uzyskaną w wyniku śledzenia promieni wartość o losowy błąd kątowy, wprowadzany przez algorytm MUSIC, który został zilustrowany na Rys 4.



Rys. 4 Wykres przedstawiający błąd kątowy namiaru sygnałów odbitych przy pomocy algorytmu MUSIC

Następnie wyznaczono drogi, którymi propagowały się poszczególne sygnały oraz znaleziono ich punkty przecięcia.

Punkty przecięcia następnie przefiltrowano poprzez odseparowanie punktów bardzo odległych od pozostałych oraz uśredniono wartość tak przefiltrowanych punktów przyjmując ją za lokalizację nadajnika, patrz Rys. 5.



Rys. 5 Ilustracja przeprowadzonego BR

W celu poprawy dokładności wskazania lokalizacji nadajnika i eliminacji błędnych punktów przecięcia się dróg poszczególnych promieni wykorzystano utworzone wcześniej sieci neuronowe, co pozwoliło odrzucić punkty znajdujące się poza wskazanym przez sieć sektorem.

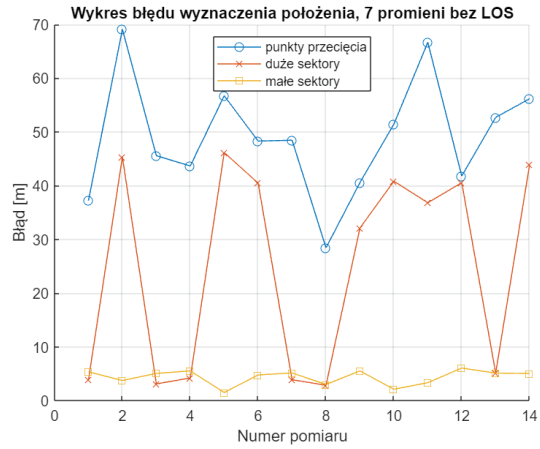
2.4. Porównanie wyników

Wyniki lokalizacji nadajnika zostały porównane dla przypadków z użyciem sieci neuronowych dla dużych i małych sektorów oraz bez ich użycia, gdzie wykorzystano punkty przecięcia promieni. Dokonano też analizy błędów lokalizacji w zależności od ilości odebranych promieni.

W celu ewaluacji skuteczności działania zaproponowanych sieci neuronowych, sprawdzono skuteczność obu sieci neuronowych w wyznaczaniu lokalizacji nadajnika oraz ich wpływ na zmniejszenie błędów w procesie lokalizacji.

3. WYNIKI

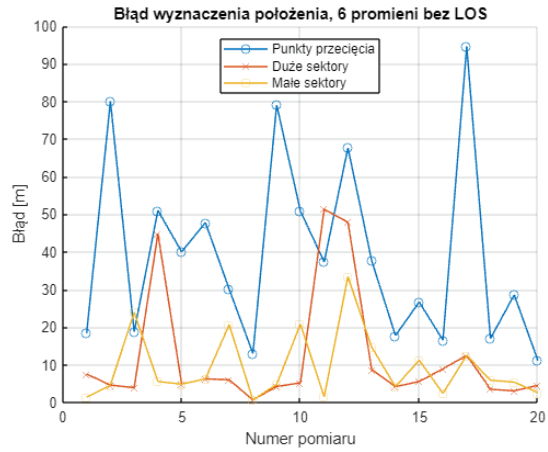
Zgodnie z danymi symulacyjnymi największą liczbą ścieżek sygnału jakie udało się odebrać było siedem ścieżek, więc badanie rozpoczęto od tego scenariusza. Jako miarę skuteczności zastosowanych technik przyjęto pomiar błędu liniowego wyznaczonego położenia stanowiącego odległość od rzeczywistego położenia nadajnika. Uzyskane wartości dla kolejnych pomiarów przedstawiono na poniższej charakterystyce (Rys. 6).



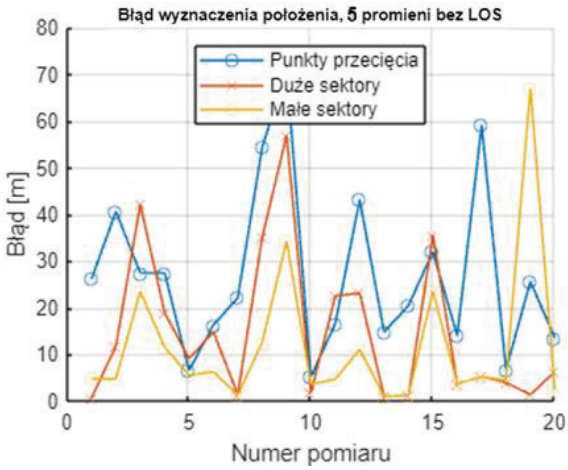
Rys. 6 Błąd wyznaczenia położenia różnymi metodami

Wraz ze spadkiem liczby odebranych promieni obserwuje się coraz mniejszą sprawność sieci przy jednoczesnej poprawie części wskazań uzyskanych bez ich wykorzystania. Wyniki pomiarów przeprowadzonych dla mniejszej liczby promieni przedstawione są na poniższych charakterystykach (Rys. 7 a-d).

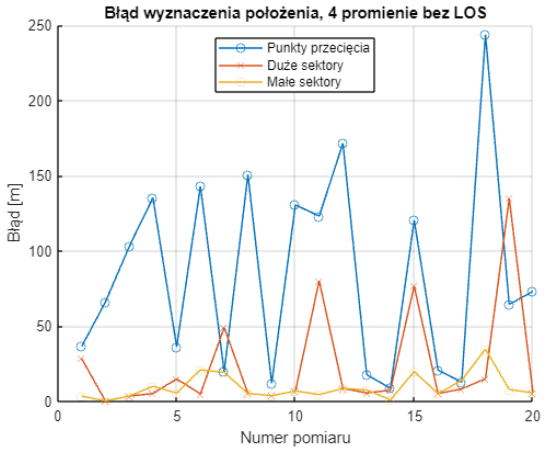
a)



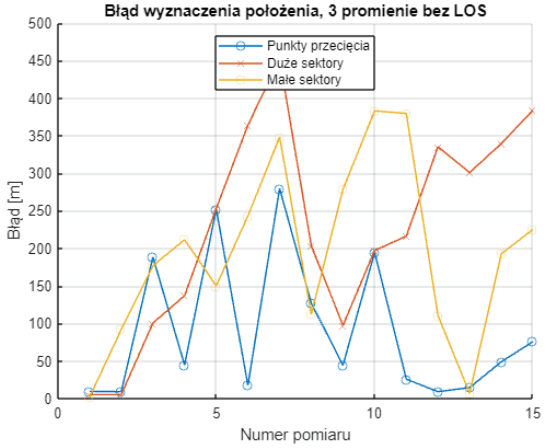
b)



c)



d)



Rys. 7 błędy wyznaczenia położenia dla różnych ilości odebranych promieni: a) 6 promieni b) 5 promieni c) 4 promieni d) 3 promieni

4. WNIOSKI

Sieci neuronowe dają tym lepsze i stabilniejsze wskazania im większą liczbą ścieżek sygnał dotarł do odbiornika. Dalszą poprawę lokalizacji można uzyskać poprzez eksperymentalny dobór hiper-parametrów sieci oraz dobór odpowiednich danych treningowych.

Warto też zauważyć, że wraz ze spadkiem liczby odebranych promieni coraz częściej sam BR zaczyna poprawnie wskazywać lokalizację nadajnika, co może być wywołane zmniejszoną liczbą punktów przecinania się promieni na drodze propagacji sygnału i w konsekwencji powstanie znacznie mniejszej liczby błędnych punktów przecięcia - nieodpowiadających lokalizacji nadajnika. Warto zauważyć, że przy prawidłowym działaniu sieci dla 7 odebranych ścieżek, dzięki zastosowaniu sieci neuronowych błąd namierzania udało się zmniejszyć 10 krotnie co jest znacznym udoskonaleniem.

5. DALSZE KIERUNKI BADAŃ

Kolejnym etapem badań będzie dobór hiper-parametrów sieci mający na celu zwiększenie jej dokładności, oraz próba statystycznego usunięcia błędów ich wskazań. Uwzględnienie dyfrakcji ścieżek sygnału również może umożliwić uzyskanie zwiększonej dokładności poprzez pojawienie się wielu nowych ścieżek o kącie elewacji na tyle małym, że można je wykorzystać nadal zawężając obliczenia do jednej płaszczyzny.

Istotnym wydaje się również rozszerzenie badań o propagację w przestrzeni trójwymiarowej, odwzorowującej prawdziwy obszar, w którym można by również przeprowadzić badania praktyczne celem weryfikacji uzyskanych wyników symulacyjnych.

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 5th 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.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

CZY KONIEC RADIOFONII UKF-FM JEST BLISKI?

IS THE END OF UKF-FM RADIO BROADCASTING NEAR?

Jacek Pieleszek[1]; Konrad Bródka[2]

[1] Urząd Komunikacji Elektronicznej, Wrocław, jacek.pieleszek@uke.gov.pl
[2] Urząd Komunikacji Elektronicznej, Warszawa, konrad.brodka@uke.gov.pl

DOI: 10.15199/59.2024.4.83

Streszczenie: **Niniejsze opracowanie przedstawia aktualną zajętość pasma UKF-FM w Polsce. Porusza aspekty doboru nowych częstotliwości do planu, możliwości rozwoju oraz przyszłość radiofonii w Polsce.**
Abstract: **This study presents the current occupancy of the VHF-FM band in Poland. It addresses aspects of adding new frequencies to the plan, development opportunities and the future of audio broadcasting in Poland.**

Słowa kluczowe: **UKF, FM, Radiofonia, Genewa 84**

Keywords: **VHF, FM, Broadcasting, Geneva 84.**

1. RADIOFONIA ANALOGOWA UKF-FM

Rok 1984 był punktem przełomowym dla dobrze znanej metody odbioru programów radiowych, tj. radia na falach ultrakrótkich (UKF-FM). Europa w tym czasie była podzielona na 2 strefy rozdzielone tzw. „żelazną kurtyną". Obowiązywała ona również w nadawaniu naziemnym. Kraje wschodnie Europy (w tym Polska) były zrzeszone w Międzynarodowej Organizacji Radia i Telewizji (OIRT), podczas gdy tzw. kraje zachodnie założyły istniejącą do dziś Europejską Unię Nadawców (EBU). Ważnym krokiem do ujednolicenia nadawania w Europie, Afryce i części Azji było zawarcie właśnie w 1984 r. nowego porozumienia i uzgodnienie planu regulującego zakres częstotliwości od 87,5 do 100 MHz wraz z rozszerzeniem go o dodatkowe 8 MHz tj. do 108 MHz [1]. W ramach tego porozumienia, zwanego Genewa 84, każdemu z krajów w nim uczestniczącym, przydzielono 6 pokryć ogólnokrajowych tworząc plan częstotliwościowy dla obiektów dużej mocy (powyżej 10 kW) znajdujących się poza miastami, jak i średniej mocy (10 kW i mniej) dla pokrycia miast. Dla Polski w ramach planu przewidziano 504 częstotliwości z czego 237 na obiektach dużej mocy i 267 częstotliwości dla miast.

W latach 90 XX w. doplanowano niemal 700 częstotliwości w celu uruchomienia emisji stacji radiowych i migracji pracujących stacji z pasma 65,5-74 MHz do pasma 87,5-108 MHz.

2. AKTUALNY STAN WYKORZYSTANIA WIDMA

Obecnie w kraju istnieje ok. 300 rozgłośni radiowych, które - na podstawie wydanych pozwoleń radiowych – prowadzą emisje z ponad 1200 nadawczych stacji radiofonicznych (których wyróżnikiem są: częstotliwość i lokalizacja). Do tego dochodzą testowe oraz liczne, związane z okazjonalnym przekazem informacji, stacje, z których prowadzone są emisje krótkoterminowe.

Zasięgi dostępnych emisji radiofonicznych znacząco się różnią. W kraju w paśmie UKF-FM funkcjonują 4 programy ogólnokrajowe Polskiego Radia S.A., z których tylko dwa (Program 1 i Program 3) charakteryzują się zasięgiem niemal ogólnopolskim. Do tego dochodzą programy Radia Maryja, Radia ZET i RMF FM, które pokrywają ponad 70% powierzchni Polski. Oprócz tego dostępne są programy o zasięgu regionalnym (np. regionalne rozgłośnie Polskie Radia S.A.), rozgłośnie dostępne w większości dużych miast, a także emisje o charakterze lokalnym, które można odebrać na niewielkim obszarze, zazwyczaj kilku gmin [2]. Jak wspomniano wyżej, nieodłącznym elementem radiodyfuzji naziemnej w Polsce są emisje okazjonalne, sprowadzające się do prowadzenia odsłuchów tzw. kin samochodowych, audiodeskrypcji wydarzeń, głównie sportowych, czy tłumaczeń symultanicznych na konferencjach lub innych wydarzeniach religijno-kulturowych. W zakresie 87,5-108 MHz prowadzone są również krótkoterminowe testy, które mają na celu przede wszystkim pomóc w ocenie zasadności przeniesienia emisji na inny obiekt.

Radio analogowe na falach ultrakrótkich jest nadal bardzo powszechnym środkiem przekazu w Polsce. Zgodnie z wynikami ankiety opublikowanej przez KIM, w roku 2023 w Polsce radia słuchało 18,1 mln osób, co stanowi 56,6% mieszkańców kraju w wieku 15+ [3].

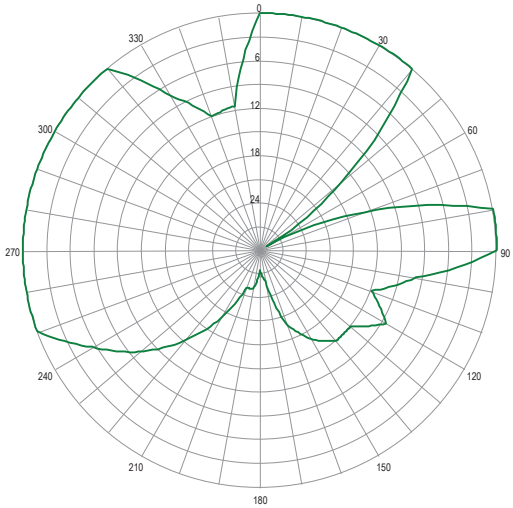
3. DOSTĘPNOŚĆ CZĘSTOTLIWOŚCI

Nowe, dobrane przez Prezesa UKE częstotliwości, po ich ew. uzgodnieniu międzynarodowym, przekazywane są one do dyspozycji Przewodniczącego KRRiT, który w oparciu o prowadzone postępowania z ogłoszeń o dostępności częstotliwości, rozdysponowuje je w formie wydawanych koncesji na rozpowszechnianie programu radio-

fonicznego z wykorzystaniem określonych częstotliwości[1]. Lista dostępnych częstotliwości sprowadza się do kilkudziesięciu pozycji, z czego wiele z nich zostało już ogłoszonych i toczą się postępowania koncesyjne.

W dużych, rzędu ok 100 tys. mieszkańców, miastach nie ma dostępnych częstotliwości. Ewentualne pojawienie się takiej jest wynikiem rezygnacji któregoś z dotychczasowych jej dysponentów (rozgłośni). Znalezienie w każdym z tych miast częstotliwości nawet pod tzw. emisje okazjonalne (np. kina samochodowe czy wydarzenia okolicznościowe) z mocą rzędu 1 W i zasięgu maksymalnie kilkuset metrów, jest wyjątkowo trudne. Największe problemy występują w Warszawie, Trójmieście, Poznaniu oraz w miastach aglomeracji Śląskiej.

Z uwagi na liczbę stacji w kraju (ale także za granicą), ich parametry i lokalizacje, dobór nowych częstotliwości jest utrudniony lub wręcz niemożliwy. Niemniej, podejmowane są przez pracowników Urzędu próby doboru, które rzadko kończą się powodzeniem. W sytuacji udanego doboru częstotliwości parametry techniczne gwarantują jedynie niewielki, rzędu ok. kilku kilometrów, średni promień zasięgu emisji, ograniczony przez zakłócenia od dużej liczby pracujących stacji w zakresie 87,5-108 MHz. Oprócz tego zachodzi niemal zawsze konieczność zastosowania skomplikowanego i kosztownego nadawczego systemu antenowego, zapewniającego duże ograniczenia mocy promieniowanej sygnału (wytłumienia) na wielu kierunkach.

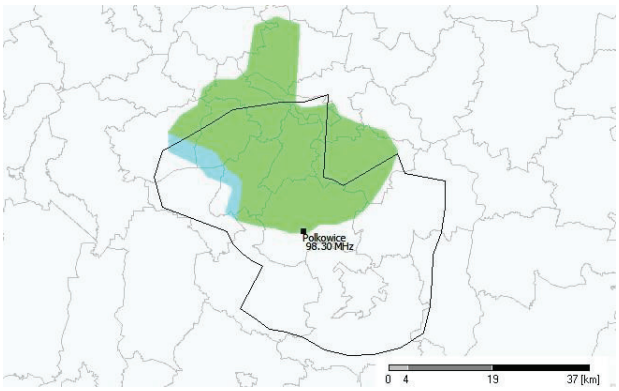


Rys. 1. Przykład skomplikowanej maski charakterystyki systemu antenowego dla stacji Bolechowo 95,9 MHz, dobranej w 2016 roku i przeniesionej o 14 km od miejsca pierwotnego doboru.

Ostatnie z tego typu prób pokazały, że nie ma możliwości doboru częstotliwości, która nie będzie miała dużych wytłumień (rzędu 15 dB i więcej) w kilku sektorach pomimo założonej przed doborem niewielkiej mocy promieniowanej (23 dBW – 200 W).

4. WYZWANIA I OGRANICZENIA

Największe problemy planowania radiofonii analogowej w paśmie UKF-FM w Polsce wiążą się przede wszystkim z zakłóceniami wewnątrzkrajowymi. Liczba i parametry stacji powodują, że dla większości stacji radiofonicznych zasięg emisji nie jest determinowany minimalnym poziomem sygnału możliwym do odbioru [54dBµV/m dla emisji stereo] a wysokim poziomem tzw. tła elektromagnetycznego, w związku z czym powszechne jest zjawisko, w którym stacje są słyszalne bez zakłóceń dopiero od poziomu sygnału rzędu 60-70dBµV/m, nawet w rejonach niezurbanizowanych[2]. Ochrona, wynikająca z rezerwacji częstotliwości i pozwoleń radiowych czy innych zezwoleń i decyzji zmieniających, dotychczasowych emisji powoduje, że nawet niewielka zmiana lokalizacji czy parametrów stacji nadawczej wymaga zastosowania większych lub/i dodatkowych wytłumień w masce charakterystyki promieniowania oraz skutkuje z reguły utratą zasięgu. Ponadto ciągłe zwiększanie mocy przez nadawców i operatorów powoduje komplikację charakterystyk promieniowania stacji nadawczych.



Rys. 2. Porównanie przewidywanego zasięgu użytecznego stacji na częstotliwości 98,3 MHz przed i po hipotetycznym przeniesieniu z obiektu w na obiekt w Jakubowie, w porównaniu z aktualnym chronionym zasięgiem użytecznym stacji nadawczej.

Dlatego, jak pokazały obliczenia prowadzone w ramach analiz na użytek własny Urzędu, zasadne wydaje się obniżenie mocy wszystkich nadajników. Skutkowałoby to wzrostem zasięgów wszystkich stacji nadawczych w Polsce[3].

Coraz częściej pojawiają się trudności w uzgodnieniach międzynarodowych stacji nadawczych. Część krajów, jak na przykład Litwa czy Niemcy, proponują nowe podejścia do analiz, bardziej restrykcyjne względem aktualnych, lub starają się chronić „widmo" w kraju, nie dopuszczając

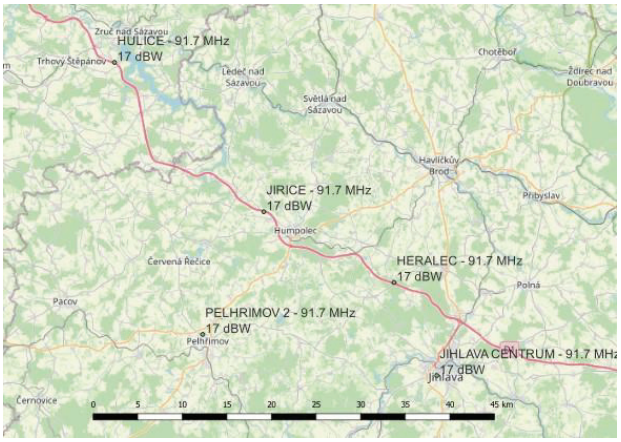
[1] W przypadku częstotliwości dedykowanych dla radiofonii publicznej, użytkowanie częstotliwości przez Polskie Radio S.A. i jej spółki regionalne, z wniosku o rezerwację częstotliwości, jest jedynie uzgadniane z Przewodniczącym KRRiT.

[2] Zależnie od zabudowy, zgodnie z rekomendacją ITU-R 412.9, wartości minimalnego użytecznego natężenia pola sygnału stereo powinny wynosić przynajmniej: 54dBuV/m (teren niezurbanizowany), 66dBuV/m (teren zurbanizowany), 74dBuV/m (gęsta zabudowa wielkomiejska).

[3] Analogiczne wnioski przedstawiono w artykule pt. „Implementacja parametru MPX dla polskich stacji radiofonicznych – umocowanie prawne oraz wpływ na warunki wykorzystania częstotliwości", KRRiT, Gdańsk, 14-16 maja 2012 r.

większego niż dotychczas poziomu sygnału z danej stacji na swoim terytorium.

Warto zwrócić uwagę również na temat powieleń częstotliwości, którego próbę forsowania można dostrzec w ostatnim czasie. O ile technika ta jest stosowana skutecznie w radiodyfuzji cyfrowej (sieci SFN), to w analogowej nie stanowi zadowalającego rozwiązania. Wyjątkiem potwierdzającym regułę są choćby Czechy, gdzie taka próba została jakiś czas temu podjęta.



Rys. 3. Przykład analogowych stacji czeskich znajdujących się blisko siebie pracujących na częstotliwości 91.7 MHz

Należy jednak podkreślić, że w Czechach powielono częstotliwości wzdłuż dróg szybkiego ruchu, położonych pomiędzy wysokimi wzniesieniami. Jak natomiast pokazują przykłady testów w innych krajach zasięg emisji przynajmniej dwóch wspólno-częstotliwościowych stacji jest mocno ograniczony przez rozległy obszar wzajemnych zakłóceń [4].

W ramach analogowej radiofonii jest uruchomionych około 20 sieci radiowych tj. programów posiadających więcej niż 5 emisji, jednak ich odbiór możliwy jest tylko w największych aglomeracjach i w związku z brakiem dostępnych częstotliwości nie ma możliwości ich rozszerzenia na cały kraj. Cierpi na tym różnorodność, ponieważ poza 4 stacjami o profilu ogólnym/uniwersalnym i 1 stacją społeczno-religijną nie ma możliwości odbioru innych na większości terytorium RP.

5. ALTERNATYWY

Mówiąc o przyszłości i rozwoju radiofonii należy patrzeć w kierunku rozwiązań cyfrowych. Jedną z alternatyw jest standard radiofonii cyfrowej DAB+. W Polsce, jak pokazują wyniki badania Radio Track, słuchalność DAB+ jest niska. na poziomie 0,3% w 2023 r., względem 0,2% 2022 r. Wynika to z wielu czynników, przede wszystkim konieczności wymiany odbiorników, do której słuchacze nie są przekonani[4][5]. Niemniej DAB+ to większa różnorodność programów dla słuchaczy, gdyż każdy multipleks może pomieścić od kilku do kilkunastu programów. Ponadto, standard ten zapewnia mniejsze koszty po stronie nadawcy w porównaniu z radiofonią analogową [5]. W DAB+ można budować sieci SFN, łatwiej pokrywając obszar, na którym ma być zapewniony odbiór. Ponadto DAB+ planowany jest i pracuje w zakresie częstotliwości 174-230 MHz, przez co może stanowić, przynajmniej na chwilę obecną, uzupełnienie oferty dostępnej na falach ultrakrótkich.

O ile wcześniej prowadzone były emisje testowe, to pierwsze uruchomienie nadajnika DAB+ w Polsce, na podstawie pozwolenia radiowego, nastąpiło w 2013 roku [6]. Zaplanowane są 3 Multipleksy i częstotliwości na użytek lokalny w dużych miastach.

Obecnie natomiast w standardzie DAB+ działa multipleks Polskiego Radia, (MUXR3), jeden z trzech zaplanowanych multipleksów o zasięgu ogólnokrajowym. W multipleksie oprócz programów dostępnych w UKF-FM (Jedynki, Dwójki, Trójki i PR 24) umieszczono dodatkowo ogólnopolskie programy: dla kierowców, dla dzieci, PR dla Ukrainy, Radio Poland i Chopin. Dodatkowo są umieszczone od 1 do 2 programów regionalnych rozgłośni. Pełne wdrożenie na zostać zakończone w październiku br. i według deklaracji ma objąć zasięgiem 93% populacji [7].

Jeden z pozostałych ogólnokrajowych multipleksów ma posłużyć odtworzeniu zasięgów obecnych koncesjonowanych stacji nadających w zakresie 87,5-108 MHz, oraz ma umożliwić wejście na rynek ogólnokrajowy nowym nadawcom [8]. Konkurs na MUXR1 powinien zostać rozpisany jeszcze w tym roku. Drugi z kolei multipleks może być zregionalizowany i w założeniu ma uzupełniać ofertę MUXR1 z tym, że będą na nim mogły nadawać podmioty regionalne i lokalne [9].

W 2019 r. Prezes UKE dodał do planu zagospodarowania częstotliwości dla zakresu 174-230 MHz 28 multipleksów lokalnych, z których obecnie 7 jest dostępnych dla słuchaczy[6]. Równolegle prowadzonych jest kilkanaście emisji testowych (eksperymentalnych).

Dalszy rozwój DAB+ w Polsce jest uzależniony od wielu czynników, m.in. strategii na szczeblu międzynarodowym czy krajowym, zainteresowania nadawców czy słuchaczy. Na przyszłość radia cyfrowego w Polsce ma również wpływ dyskutowany temat przeznaczenia całego zakresu 174-230 MHz dla radia cyfrowego, podobnie jak to ma miejsce w większości krajów ościennych. W Polsce, obok DAB+, dostępny jest w ww. zakresie multipleks naziemnej telewizji cyfrowej. Ewentualna zmiana przeznaczenia zakresu 174-230 MHz i optymalizacja jego wykorzystania wiązałaby się zatem z przeniesieniem MUX8 [10].

[4] W odróżnieniu od tzw. smartfonów, które średnio mieszkańcy Polski wymieniają co dwa lata.

[5] O ile koszt nowego odbiornika przenośnego/stacjonarnego z modułem DAB+ to koszt ok. 100 zł, to analogiczna wymiana w aucie to koszt już kilkuset złotych. Kolejna sprawa to wiek średni wiek aut w Polsce, sprawiający, że większość aut zarejestrowanych w Polsce nie posiada fabrycznie wbudowanego odbiornika DAB+.

[6] Częstochowa, Katowice, Poznań, Rzeszów, Tarnów, Toruń, Warszawa

Kolejnym alternatywnym standardem jest DRM, jednak jego popularyzacja przebiega na tyle wolno, że ciężko upatrywać w nim obecnie sukcesora emisji FM, choć sam standard przewiduje implementację w zakresie 87,5-108 MHz.

Radia można również słuchać odbierając emisje satelitarne, albo korzystając ze streamingu, przy czym to ostatnie rozwiązanie, choć oferuje ogromny zestaw programów radiowych znajduje się poza obszarem radiodyfuzji.

6. WNIOSKI

Obserwując obecne trendy i zapytania ze strony rynku można z pewnością stwierdzić, że radio jako sposób przekazu nie umiera. Jednak sposób w jaki przekaz jest dystrybuowany się zmienia i nie można przejść obok tego obojętnie. Popularyzacja aplikacji mobilnych do słuchania radia postępuje i w większości regionów jest ona jedyną szansą dla nadawców (wyłączając z tego rozgłośnie z największym zasięgiem w UKF-FM). Nadawcy wciąż pytają o wolne częstotliwości, których poza kilkoma wspomnianymi w akapicie o dostępności przypadkami nie ma i nie będzie z uwagi na bardzo duże wysycenie widma częstotliwości w zakresie 87,5 – 108 MHz. Dlatego sieć stacji radiofonii analogowej FM nie ma możliwości rozwoju, z uwagi na niemal całkowite wysycenie dostępnego widma częstotliwości. O ile optymalizacja wykorzystania zakresu 87,5-108 MHz jest teoretycznie możliwa, to proces ten wymagałby rewizji postanowień porozumienia Genewa 84, co wymagałoby zaangażowania również krajów ościennych. Biorąc pod uwagę aktualne małe zainteresowanie tym procesem, optymalizacja wydaje się niemożliwa.

Dlatego w obecnej formie słusznym wydaje się porzucenie doborów nowych częstotliwości na rzecz kolejnych emisji stałych, zostawiając jedynie możliwość przydziału niskomocowych częstotliwości dedykowanych okazjonalnemu przekazowi informacji.

LITERATURA

[1] International Telecommunication Union, „Final Acts of the Regional Administrative Conference for the Planning of VHF Sound Broadcasting (Regions 1 and Part of Region 3) (Regional Agreement relating to the Use of the Band 87.5 - 108 MHz for FM Sound Broadcasting (Region 1 and Part of Region 3)),” Genewa, 1984.

[2] J. Pieleszek, „Sieci radiofonii analogowej UKF,” Wrocław, 2022.

[3] „Badanie słuchalności radia – Audio Track. Wyniki za rok 2023 i porównanie z rokiem 2022 - Krajowy Instytut Mediów,” 12 03 2024. [Online]. Available: https://kim.gov.pl/2024/03/12/badanie-sluchalnosci-radia-audio-track-wyniki-za-rok-2023-i-porownanie-z-rokiem-2022/.

[4] P. Soellbeg, „Single Frequency Networks for FM Radio,” w *Audio Engineering Society 120th Convention*, Paryż, 2006.

[5] M. Lombardo, 07 2017. [Online]. Available: https://tech.ebu.ch/docs/techreview/EBU_Tech_Review_2017_Cost-benefit_analysis_of_FM_DAB_and_Broadband.pdf.

[6] G. Maj, „Polskie Radio 24,” 24 wrzesień 2013. [Online]. Available: https://polskieradio24.pl/artykul/939682,cyfryzacja-polskiego-radia.

[7] A. Gąbka, „Wirtualne Media,” 2023 grudzień 5. [Online]. Available: https://www.wirtualnemedia.pl/artykul/jak-odbierac-dab-multipleks-zasieg-polskie-radio.

[8] Krajowa Rada Radiofonii i Telewizji, „DAB+, wyniki konsultacji treści ogłoszeń Przewodniczącego KRRiT o możliwości uzyskania koncesji na nadawanie programu ogólnopolskiego,” 11 09 2020. [Online]. Available: https://www.gov.pl/web/krrit/wyniki-konsultacji-projektow-ogloszen-przewodniczacego-krrit-o-mozliwosci-uzyskania-koncesji-na-nadawanie-programu-w-ogolnopolskim-multipleksie-cyfrowym-dab.

[9] Krajowa Rada Radiofonii i Telewizji, „Zielona księga cyfryzacji radia w Polsce,” 31 03 2016. [Online]. Available: https://cyfrowapolska.org/wp-content/uploads/2017/09/zielona-ksiega-cyfryzacji-radia-w-polsce.pdf.

[10] Krajowa Rada Radiofonii i Telewizji, „Stanowisko KRRiT w sprawie zmiany zagospodarowania częstotliwości w paśmie 174-230 MHz.,” 29 11 2021. [Online]. Available: https://www.gov.pl/web/krrit/stanowisko-krrit-w-sprawie-zmiany-zagospodarowania-czestotliwosci-w-pasmie-174-230-mhz.

PORÓWNANIE CHARAKTERYSTYK ZANIKÓW W JEDNOCZĘSTOTLIWOŚCIOWEJ SIECI DAB+ SYMULOWANYCH I ZMIERZONYCH

Comparison of simulated and measured fading characteristics in a single-frequency DAB+ network

Igor Michalski, Ryszard J. Zieliński
[1] Instytut Łączności PIB, Wrocław, i.michalski@il-pib.pl
[2] Politechnika Wrocławska, Wrocław, Ryszard.zielinski@pwr.edu.pl

DOI: 10.15199/59.2024.4.84

Streszczenie: **W artykule zostaną pokazane wyniki pomiarów charakterystyk zaników w sieci jednocząstotliwościowej (SFN) DAB+ i ich porównanie z wynikami symulacji. Uzyskane wyniki pomiarów dowodzą rzeczywistego występowania wielu selektywnych zaników w paśmie sygnału DAB+. Wysoka zgodność wyników pomiarów charakterystyk zaników z wynikami symulacji, którą również zaprezentowano w artykule, wskazuje na możliwość przewidywania obszarów i charakterystyk zaników na podstawie badań symulacyjnych. Tym samym na ich podstawie możliwe będzie określenie miejsc występowania tego zjawiska i uwzględnienie go w procesie planowania sieci SFN.**

Abstract: **In this article the measurement results of fading characteristics of SFN DAB+ network and their comparison with simulation results will be shown. The obtained measurement results prove the real occurrence of many selective fading in the DAB + signal band. The high compliance of the measurement results of the fade characteristics with the simulation results, which were also presented in the article, indicates the possibility of predicting the areas and characteristics of fades on the basis of simulation tests. Thus, on their basis, it will be possible to determine the places where this phenomenon occurs and take it into account in the process of planning the SFN network.**

Słowa kluczowe: **sieć jednoczęstotliwościowa, radiofonia cyfrowa DAB+, analiza zaników, pomiary zaników.**

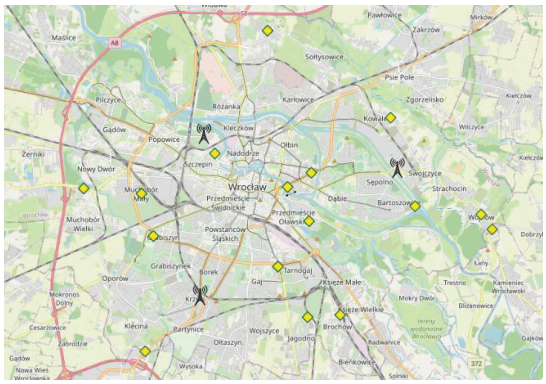
Keywords: **Single Frequency Network (SFN), Digital Audio Broadcasting (DAB+), fade analysis, fade measurements.**

1. WSTĘP

W wielu zaleceniach i publikacjach naukowych przedstawiono podstawy i zasady planowania sieci DAB+ w konfiguracji wieloczęstotliwościowej MFN (Multi Frequency Network) jak w również w konfiguracji jednoczęstotliwościowej SFN (Single Frequency Network), np. [1]. Nietrudno również znaleźć publikacje analizujące właściwości sieci SFN i wskazujące na dużą liczbę zalet wynikającą z ograniczonego zapotrzebowania energetycznego [5], tak ważnego w dobie kryzysu energetycznego (klimatycznego) jak również wynikającą ze znacznego poprawienia warunków odbioru sygnałów szczególnie w środowisku dużych miast, gdzie wielodrogowość silnie wpływa na warunki odbioru [7, 8]. Do głównych zalet zastosowania sieci SFN w emisji DAB+ można zaliczyć:

- poprawienie warunków odbioru wskutek odbioru replik nadanego sygnału, emitowanego przez nadajniki sieci SFN oraz powstałych na skutek zjawiska wielodrogowości, o ile obserwowane opóźnienia tych sygnałów mieszczą się w czasie trwania odstępu ochronnego.
- zmniejszenie obszarów znajdujących się w tzw. cieniu radiowym. Ma to szczególne znaczenie w środowiskach silnie zurbanizowanych. Ten sam efekt można uzyskać stosując sieć wieloczęstotliwościową MFN (*Multi Frequency Network*), ale wtedy trzeba zająć więcej zasobów widmowych.
- rozwiązanie bazujące na sieci nadajników SFN jest ze względu na zapotrzebowanie energetyczne bardziej ekologiczne. Wymaga ono mniejszej całkowitej mocy promieniowanej. Z tego względu generuje również mniejsze zakłócenia do innych sieci.
- sumowanie mocy sygnałów pochodzących od nadajników SFN. Efekt ten powinien być dobrze obserwowany wtedy, gdy sygnały nadawane są w polaryzacji pionowej a antena odbiorcza ma dookólna charakterystykę promieniowania.

Instalacja sieci SFN nie jest również pozbawiona pewnych niedogodności, w szczególności dla nadawców lokalnych koszt instalacji systemu stanowi barierę finansową. Z tego powodu w latach 2015 – 2017 realizowany był projekt pod nazwą „Sieć jednoczęstotliwościowa wykorzystująca platformę nadawczą DAB+ na potrzeby nadawców lokalnych w Polsce” – w skrócie LokalDAB finansowany przez Narodowe Centrum Badań i Rozwoju. Głównym celem programu była budowa SFN z wykorzystaniem niskokosztowych rozwiązań z zakresu budowy anten, systemów nadawczych, synchronizacji i zarządzania siecią oraz z wykorzystaniem uniwersalnych urządzeń programowalnych i oprogramowania typu „open source”. Sieć złożoną z trzech nadajników SFN (rys. 1) zbudowało konsorcjum składające się z Politechniki Wrocławskiej (PW), Instytutu Łączności PIB i Polskiego Radia Wrocław (PRW). Aby obniżyć koszty, instalacje nadawcze zlokalizowano na terenie należącym do Konsorcjantów. Sieć LokalDAB stała się doskonałym obiektem badawczym związanym z propagacją i zjawiskami towarzyszącymi odbiorowi sygnałów z kilku nadajników w środowisku dużego miasta. Ze względu na konieczność ciągłego nadawania programów radiowych do celów badawczych nie można było na czas badań wyłączać żadnego z trzech nadajników.



Rys. 1. Rozmieszczenie trzech nadajników we Wrocławiu w miejscach należących do Konsorcjantów (oznaczono symbolami nadajnika) oraz rozmieszczenie punktów pomiarowych (romby) w obszarach, w których potencjalnie zjawisko zaników może występować

Możliwości występowania zaników w jednoczęstotliwościowej sieci DAB+ była tematem kilku wcześniejszych publikacji. W artykule [9] przedstawiono warunki uzyskania poprawnego odbioru sygnału DAB+ w sieci jednoczęstotliwościowej, w [10] wyznaczono obszary potencjalnego występowania zaników na terenie miasta Wrocławia, pochodzących od sieci trzech nadajników DAB+ pracujących w kanale 11A (częstotliwość 216,928 MHz). Obszary te wytypowano na podstawie analizy natężenia pola e-m emitowanego przez te nadajniki obliczonego metodą ITU_R P.1546 [4]. Warunkiem występowania głębokich zaników był porównywalny poziom natężenia pola elektrycznego pochodzącego od dwóch lub trzech nadajników. W [11, 12] zaprezentowano wyniki obliczeń maksymalnych wartości zaników sygnału podczas jednoczesnego odbioru dwóch lub trzech sygnałów dochodzących do odbiornika we wcześniej wyznaczonych obszarach. W [13] wykazano selektywną charakterystykę zaników w paśmie kanału DAB+ o szerokości 1546 kHz i opracowaną metodę wyznaczania charakterystyki zaników na podstawie określenia różnicy dróg propagacji oraz poziomów poszczególnych sygnałów. W [14] zaprezentowano metodologię pomiarów natężeń pól wyznaczonych metodą ITU-R P.1546 we wcześniej wytypowanych miejscach oraz przedstawiono zastosowane procedury pomiarowe. Na podstawie tych wyników możliwa była analiza charakterystyk zaników wyznaczonych dla zmierzonych wartości natężenia pola elektrycznego i porównanie tak uzyskanych wyników ze zmierzonymi w warunkach polowych charakterystykach.

2. WYNIKI SYMULACJI I POMIARÓW CHARAKTERYSTYKI ZANIKÓW

Przebiegi charakterystyk zaników sygnału zostaną przedstawione zarówno dla symulacji uwzględniających zmierzone wartości natężenia pola elektrycznego jak i przebiegi zarejestrowane na analizatorze widma. W tabeli 1 przedstawiono natężenia pól zmierzone (będące podstawą analizy) i obliczone przy użyciu metody ITU-R P.1546. Ze względu na ograniczoną liczbę stron artykułu zaprezentowane zostaną jedynie wyniki z czterech punktów pomiarowych. W dwóch z nich można zauważyć duże podobieństwo przebiegu symulowanego poziomu sygnału w funkcji częstotliwości z przebiegiem zmierzonym. W dwóch pozostałych to podobieństwo jest znacznie mniejsze.

Tab. 1. Natężenie pola elektrycznego w dB[μV/m] obliczone/zmierzone

	Natężenie pola elektrycznego obliczone/zmierzone dB[μV/m]			
ulica	**Politechnika**	**Instytut Łączności**	**Polskie Radio**	**Wypadkowe**
Hallera	57,3/60,7	42,5/-	60,9/57,7	62,5/62,5
Armii Krajowej / Bardzka	54,1/58,1	54,6/55,9	55,5/57,6	59,5/62,1
Konduktor-ska	48,7/-	54,4/53,8	55,7/56,1	58,6/58,1
Szwajcarska	65,4/61,8	41,9/-	53,4/58,5	65,7/63,4

2.1. Metodologia obliczania charakterystyki zaniku

Wypadkowe natężenie pola elektrycznego $E_{Tot}(f_{cen_n})$ w dB[μV/m] w punkcie analizy położonym w odległości d_i od i-tego nadajnika na środkowej częstotliwości podnośnej ($f_{cen\text{-}n}$) pochodzące od dwóch lub trzech interferujących sygnałów (M=2 lub M=3) można wyznaczyć z zależności (1):

$$E_{Tot}(f_{cen_n}) = 20 \cdot \log\left(\left| \sum_{i=1}^{M} 10^{\frac{E_i(f_{cen_n}, d_i)}{20}} \cdot e^{j \cdot \phi_i(f_{cen_n}, d_i)} \right| \right) \quad (1)$$

W sygnale DAB+ liczba podnośnych K wynosi 1546, szerokość pasma B_{DAB} kanału DAB+ jest równa 1546 kHz. Częstotliwość środkowa sieci LokalDAB f_{cen_DAB} jest równa 216,928 MHz. Kanał ten zajmuje częstotliwości od f_p = 216,160 MHz do f_k = 217.696 MHz. Częstotliwość środkowa pierwszej podnośnej wynosi f_{cen_1} jest równa 216,1605 MHz. Częstotliwości środkowe pozostałych podnośnych można wyznaczyć z zależności (2):

$$f_{cen_n} = f_{cen_1} + (n-1) \cdot 0.001 \text{ [MHz]} \quad n=2, \ldots, 1536 \quad (2)$$

Jeżeli założymy, że w sygnale DAB+ wszystkie podnośne nadawane są z taką samą mocą, wtedy należy uwzględnić wprowadzenie współczynnika korekcyjnego mocy podnośnej Dp o wartości (3):

$$Dp = 10 \cdot \log(1536) = 31{,}86dB \quad (3)$$

Oznacza to, że każda podnośna z pasmem równym 1 kHz przenosi jedynie 1/1546 całkowitej mocy E_{iB} nadawanej przez i-ty nadajnik. Moc E_{iB} jest mierzona w całym 1546 kHz paśmie kanału DAB+. Dlatego też amplituda $E_i(f_{cen_n}, d_i)$ natężenia pola elektrycznego podnośnej jest wyznaczana z zależności (4):

$$E_i(f_{cen_n}, d_i) = E_{iB} - Dp \quad (4)$$

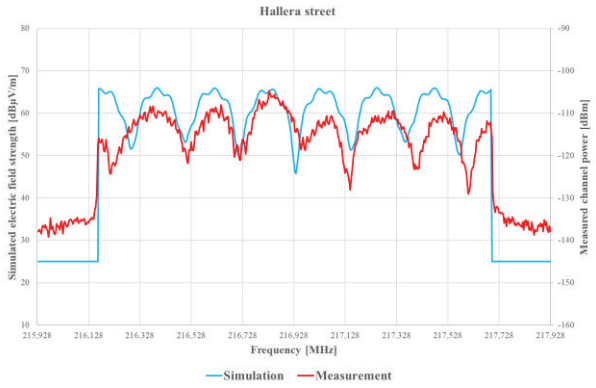
Faza każdej podnośnej interferującego sygnału ϕ_i (i = 1, 2 lub 3) jest wyznaczana ze wzoru (5):

$$\phi_i(f_{cen_n}, d_i) = -2 \cdot \frac{\pi \cdot f_{cen_n}}{300} \cdot d_i \quad (5)$$

w którym odległość d_i jest odległością analizowanego punktu od i-tego nadajnika.

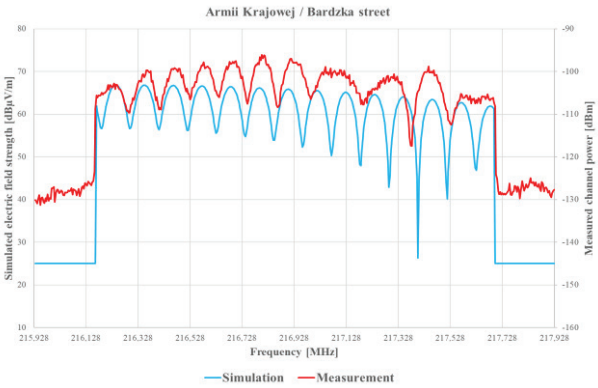
2.2. Porównanie wyników symulacji i pomiarów charakterystyki zaników

W tym rozdziale przedstawione zostaną charakterystyki zaników obliczone na podstawie zmierzonych wartości natężenia pola e-m (tab. 1) i ich porównanie ze zmierzonymi charakterystykami zaników (rys. 2-5).



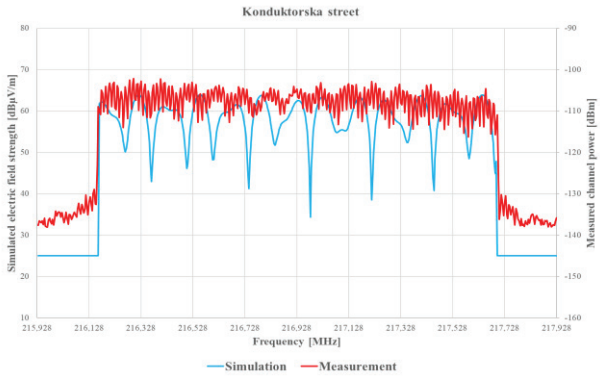
Rys. 2. Porównanie obliczonych (na podstawie pomiaru natężenia pola elektrycznego) i zmierzonych charakterystyk zaników na ul. Hallera. Głębokość zaników obliczona do 20 dB, zmierzona do 19 dB

W punkcie przy ul. Hallera (rys. 2) wystąpiła silna interferencja sygnałów od nadajników Politechniki Wrocławskiej i Polskiego Radia. Zmierzone natężenie pola e-m od tych nadajników różni się jedynie o 3 dB. Porównując charakterystykę zaników zmierzoną i obliczoną na podstawie zmierzonych natężeń pól widać bardzo duże podobieństwo. Zarówno liczba głębokich zaników jak i ich charakter są bardzo podobne. Również maksymalna zmierzona głębokość zaniku 19 dB jest bliska wartości obliczonej równej 20 dB. Uzyskane wyniki wskazują, że w tym punkcie dominującym mechanizmem interferencji jest interferencja sygnałów pochodząca z dwóch nadajników. Mały jest wpływ wielodrogowości spowodowanej przechodzeniem fali e-m przez obszar zurbanizowany. Wielodrogowość nie wpływa na charakterystykę zaniku.



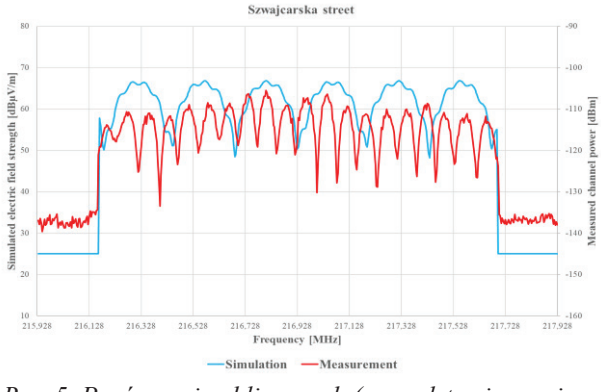
Rys. 3. Porównanie obliczonych (na podstawie pomiaru natężenia pola elektrycznego) i zmierzonych charakterystyk zaników na ul. Armii Krajowej. Głębokość zaników obliczona do 35 dB, zmierzona do 18 dB

W punkcie pomiarowym przy ul. Armii Krajowej (rys. 3) interferują ze sobą sygnały z trzech nadajników. Poziomy tych sygnałów różnią się o nie więcej niż 2,2 dB. W tym przypadku występuje 14 głębokich zaników w symulowanym przebiegu i 10 zaników w przebiegu zmierzonym. Występuje dość dobra zgodność w miejscach występowania zaników, również tych najgłębszych. Ich maksymalna głębokość wynosi 35 dB (symulacja) i 18 dB (pomiar). Poziom sygnału zmierzonego jest większy od poziomu sygnału symulowanego.



Rys. 4. Porównanie obliczonych (na podstawie pomiaru natężenia pola elektrycznego) i zmierzonych charakterystyk zaników na ul. Konduktorskiej. Głębokość zaników obliczona do 27 dB, zmierzona do 10 dB

W punkcie pomiarowym przy ul. Konduktorskiej (rys. 4) interferują ze sobą sygnały z nadajników Instytutu Łączności i Polskiego Radia. Różnica poziomów zmierzonych sygnałów wynosi 2,3 dB. Jednak w tym przypadku obraz interferencji zmierzonych i obliczonych znacznie się różni. Mamy 12 zaników w przebiegu obliczonym i 14 płytszych zaników w przebiegu zmierzonym. Maksymalne wartości zaników wynoszą odpowiednio 27 dB (obliczenia) i ok. 10 dB (pomiar). W tym przypadku znaczącą różnicę pomiędzy zanikami obliczonymi i zmierzonymi można tłumaczyć znaczącym wpływem wielodrogowości. Interferencja dwóch sygnałów z dwóch nadajników nie jest mechanizmem dominującym powstawania zaników.



Rys. 5. Porównanie obliczonych (na podstawie pomiaru natężenia pola elektrycznego) i zmierzonych charakterystyk zaników na ul Szwajcarskiej. Głębokość zaników obliczona do 10 dB, zmierzona do 22 dB

W punkcie pomiarowym przy ul. Szwajcarskiej (rys. 5) interferują ze sobą sygnały z nadajników Politechniki Wrocławskiej i Polskiego Radia. Zmierzona różnica poziomów sygnałów wynosi 3,3 dB (obliczona 12 dB). W

symulowanym przebiegu występuje 5 wyraźnych zaników. W przebiegu zmierzonym liczba zaników wynosi 18. W przebiegu symulowanym występują dodatkowe nierównomierności charakterystyki na częstotliwościach występowania zaników w przebiegu zmierzonym. Maksymalna głębokość zaniku wynosi 18 dB (symulacja) i 25 dB (pomiar). Poziom sygnału symulowanego jest wyższy niż sygnału zmierzonego.

3. PODSUMOWANIE

Przedstawione wyniki symulacji i pomiarów charakterystyki zaników spowodowanych interferencją sygnałów odbieranych z nadajników sieci jednoczęstotliwościowej umożliwiają wysuniecie szeregu wniosków.

Najważniejszym wnioskiem jest eksperymentalne potwierdzenie występowania tego zjawiska. W analizowanych punktach pomiarowych zaobserwowano zaniki, których liczba i głębokość może wpływać na jakość odbieranego sygnału. W badanych przypadkach we wszystkich punktach pomiarowych odbiór sygnału DAB+ był niezakłócony. Ale wynikało to ze stosunkowo dużego poziomu sygnału. Ograniczenia zasięgu sieci SFN na skutek występowania zaników można się spodziewać w miejscach, gdzie poziomy odbieranych sygnałów z nadajników zapewniają jeszcze poprawny odbiór ale są w pobliżu dolnej granicy. Wtedy zjawisko zaników może wpłynąć na ograniczenie pokrycia w obszarach, w których poziomy odbieranych sygnałów są porównywalne.

Zastosowanie metody propagacyjnej ITU-R P.1546 jest skuteczne do wyznaczania miejsc, w których spodziewamy się występowania głębokich interferencji sygnałów z sieci SFN. Jednakże symulacja charakterystyki zaników na podstawie symulowanych poziomów sygnałów może znacznie różnic się od charakterystyki uzyskanej z pomiarów. Metoda ITU-R P.1546 nie uwzględnia szeregu elementów wpływających na propagacje sygnału, szczególnie w środowisku dużego miasta. Przy czym uzyskane na podstawie symulacji wartości zaników mogą być większe lub mniejsze od wartości zmierzonych. W warunkach rzeczywistych w obszarze anteny odbiorczej mogą pojawić się dodatkowe sygnały wynikające z propagacji wielodrogowej, które w istotny sposób wpływają na liczbę i głębokość zaników.

Wyniki pomiarowe poziomu sygnału od nadajników sieci SFN nie są precyzyjne ze względu na dość ograniczony zysk kierunkowej anteny odbiorczej. Niestety w rozważanej sieci nie można było wyłączać nadajników. Zastosowanie anteny o większym zysku (kierunkowości) na częstotliwości 216,928 MHz związane jest z zastosowaniem anteny o większych wymiarach, co może utrudnić prowadzenie pomiarów.

LITERATURA

[1] Brugger Roland, Mayer Kerstin, kwiecień 2005, „RRC-06 — technical basis and planning configurations for T-DAB and DVB-T, EBU Technical Review.

[2] EBU, TECH 3391, 2018, “Guidelines for DAB network planning”, Genewa.

[3] EBU, „TR 24, 2013, „SFN Frequency planning and network implementation with regard to T-DAB and DVB-T,” Genewa.

[4] ITU-R P.1546-5, 2013, “Method for point-to-area predictions for terrestrial services in the frequency range 30 MHz to 3 000 MHz”.

[5] Mattson A, grudzień 2005, “Single Frequency Networks in DTV “,IEEE Transactions On Broadcasting, vol. 51, nr. 4, 413-422.

[6] Michalski Igor, 2020, praca dyplomowa, " Analiza rozkładu natężenia pola e-m i zaników od sieci SFN DAB+ we Wrocławiu*”, Politechnika Wrocławska*.

[7] Staniec Kamil, Kubal Sławomir, Kowal Michał, Piotrowski Piotr, 2020, “On the Influence of the Coding Rate and SFN Gain on DAB+ Coverage”, *Advances in Intelligent Systems and Computing*.

[8] Staniec Kamil, 2021, „Analysis of the Single Frequency Network Gain in Digital Audio Broadcasting Networks”, MDPI Sensors, 21(2), 569; https://doi.org/10.3390/s21020569.

[9] Zieliński Ryszard, 2017, “Warunki uzyskania poprawnego odbioru sygnału DAB+ w sieci jednoczęstotliwościowej”, *Przeglad Telekomunikacyjny i Wiadomosci Telekomunikacyjne,* nr. 6: 509-512.

[10] Zieliński Ryszard, 2018, „Analiza wielkości zaników w jednoczęstotliwościowej sieci DAB+ na przykładzie sieci we Wrocławiu”, *Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne*, nr. 6: 320-325.

[11] Zieliński Ryszard, 2018, “ Rozmieszczenie obszaru zaników w jednoczęstotliwościowej sieci SFN na przykładzie sieci DAB+”, *Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne*, nr. 8-9: 581-584.

[12] Zieliński Ryszard, 2019, “Analysis of the phenomenon of fades in the SFN DAB+ network with three transmitters on the example of the network in Wroclaw”, *Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne*, nr. 6: 386-391.

[13] Zieliński Ryszard, 2019, „Fade analysis in DAB+ SFN network in Wroclaw”, *Proc. of the 2019 International Symposium on Electromagnetic Compatibility (EMC Europe 2019)*, Barcelona, Hiszpania, 2– 6 wrzesień, (978-1-7281-0594-9/19/$31.00 © 2019 IEEE).

[14] Zieliński Ryszard, 2022, „Analiza porównawcza zmierzonych i obliczonych natężeń pól elektrycznych od nadajników SFN DAB+, *Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne,* nr 4: 612-615.

NAZIEMNA STACJA DO KOMUNIKACJI SATELITARNEJ W PASMACH VHF/UHF/S/X/C

EARTH STATION FOR SATELLITE COMMUNICATIONS IN THE VHF/UHF/S/X/C BANDS

Michał Sybis[1]; Wojciech Flakowski[2]; Marcin Rodziewicz[1]; Rafał Krenz[1]

[1] Politechnika Poznańska, Wydział Informatyki i Telekomunikacji, Instytut Radiokomunikacji, Poznań, {michal.sybis, marcin.rodziewicz, rafal.krenz}@put.poznan.pl
[2] Politechnika Poznańska, Wydział Informatyki i Telekomunikacji, Instytut Radiokomunikacji, Poznań,wojciech.flakowski@dostorate.put.poznan.pl

DOI: 10.15199/59.2024.4.85

Streszczenie: Artykuł przedstawia znaczenie stacji naziemnych w zapewnianiu skutecznej komunikacji z satelitami CubeSat, które stały się coraz bardziej powszechne w przestrzeni kosmicznej. Stacje naziemne pełnią kluczową rolę w ułatwianiu przesyłania danych, kontrolowania i monitorowania tych niewielkich satelitów, które często operują na niskich orbitach okołoziemskich. Artykuł podkreśla również znaczenie ciągłego rozwoju stacji naziemnych, aby sprostać dynamicznym potrzebom współczesnych misji kosmicznych wykorzystujących satelity CubeSat.

Abstract: The paper discusses the importance of ground stations in ensuring effective communication with CubeSat satellites, which have become increasingly common in space. Ground stations play a crucial role in facilitating data transmission, controlling, and monitoring these small satellites, which often operate in low Earth orbits. The article also emphasizes the significance of the continuous development of ground stations to meet the dynamic needs of contemporary CubeSat-based space missions.

Słowa kluczowe: stacja naziemna, komunikacja, satelita, Cubesat.

Keywords: Ground station, communication, satellite, CubeSat.

1. WSTĘP

W ostatnich dekadach obserwowany jest dynamiczny rozwój technologii kosmicznych, który umożliwia coraz szerszy dostęp do przestrzeni kosmicznej dla różnych podmiotów, w tym dla instytucji akademickich, małych przedsiębiorstw oraz entuzjastów kosmosu. Jednym z przejawów tej ewolucji są satelity typu CubeSat [2], które rewolucjonizują tradycyjne podejścia do projektowania, budowy i wykorzystania satelitów. Satelity CubeSat, należące do klasy tzw. nanosatelitow, to niewielkie sztuczne satelity o standardowych wymiarach, które charakteryzują się modułową konstrukcją i niskimi kosztami produkcji oraz umieszczenia na orbicie. Pomimo skromnych rozmiarów, satelity CubeSat posiadają wielorakie zastosowania w badaniach naukowych, telekomunikacji, obserwacji Ziemi czy eksploracji kosmosu. W niniejszej pracy przedstawiono podstawowe informacje związane z satelitami CubeSat, obejmujące ich historię, budowę, technologie wykorzystywane w ich konstrukcji, zastosowania oraz wyzwania z nimi związane. Wraz z postępem technologicznym i rosnącym zainteresowaniem eksploracją kosmosu, stacje naziemne stają się niezbędnym elementem infrastruktury wspierającej działania w kosmosie. Te zaawansowane systemy telekomunikacyjne odgrywają kluczową rolę w umożliwianiu efektywnej transmisji danych, kontroli oraz monitorowania satelitów, co ma zasadnicze znaczenie w realizacji różnorodnych misji kosmicznych.

W niniejszym artykule skupiono się na omówieniu funkcji, zasad działania oraz znaczenia stacji naziemnej pracującej na częstotliwościach VHF/UHF/C [3] oraz S/X [1], która powstała na Politechnice Poznańskiej w ramach realizacji projektu AEROSFERA2.0-POLYITAN. Analizie poddana została jej rola w zapewnianiu nieprzerwanej łączności z satelitami umieszczonymi na różnych orbitach, w tym zarówno w niskiej, jak i średniej orbicie okołoziemskiej.

Ponadto przedstawione zostały wyzwania techniczne związane z budową i utrzymaniem stacji naziemnych, uwzględniając specyficzne wymagania dotyczące transmisji danych na częstotliwościach VHF/UHF/C oraz S/X. Analizując najnowsze osiągnięcia w tej dziedzinie, zaproponowano także potencjalne rozwiązania, które mogą przyczynić się do dalszego rozwoju i doskonalenia tych kluczowych elementów infrastruktury kosmicznej. Artykuł został zorganizowany w następujący sposób. Rozdział 2 poświęcony jest idei satelitów CubeSat. W rozdziale 3 przedstawiono strukturę naziemnego systemu łączności satelitarnej, uwzględniając część przetwarzania danych satelitarnych, jak i segment kontrolny. Dodatkowo, w rozdziale 4, zaprezentowane zostały możliwości oraz przykładowe zastosowania omawianej stacji łączności satelitarnej.

2. SATELITY CUBESAT

Historia satelitów CubeSat rozpoczyna się w 1999 roku, kiedy Jordi Puig-Suari z California Polytechnic State University oraz Bob Twiggs z Stanford University stworzyli koncepcję małych, modułowych satelitów, które budowanych w oparciu o standaryzowane rozmiary i specyfikacje. Ich celem było umożliwienie uczelniom i innym instytucjom ograniczonym budżetowo dostępu do przestrzeni kosmicznej. Pierwszy satelita CubeSat, nazwany CP1, został wyniesiony na orbitę w 2003 roku w ramach misji japoń-

skiego satelity Nozomi.

CubeSat to niewielki sztuczny satelita, który charakteryzuje się standardowymi wymiarami i masą, co ułatwia ich produkcję i wyniesienie na orbitę. Nazwa "CubeSat" pochodzi od kształtu satelity, który jest zazwyczaj sześcianem. Satelity te są modułowe, co oznacza, że mogą być budowane z wielu identycznych części, zwanych modułami. Chociaż satelity CubeSat mogą mieć różne zastosowania, często są one używane w celach badawczych, do testowania nowych technologii kosmicznych, prowadzenia obserwacji Ziemi, monitorowania warunków pogodowych czy eksperymentów naukowych. Ze względu na swoją modułową konstrukcję, mogą być one dostosowywane do różnych potrzeb misji. Satelity CubeSat są wykorzystywane zarówno w misjach autonomicznych, jak i jako pomocnicze ładunki ładunków na większych satelitach. Ich rosnąca popularność w środowisku naukowym i komercyjnym sprawia, że stanowią one ważny element nowoczesnej eksploracji kosmicznej.

Istnieją trzy główne kategorie rozmiarów satelitów CubeSat, określane na podstawie liczby jednostek (U) CubeSat, gdzie jedna jednostka (1U) odpowiada sześcianowi o boku 10 centymetrów:

- 1U: Posiada wymiary 10 x 10 x 10 cm i wagę do kilku kilogramów. Jest to najmniejsza kategoria satelitów CubeSat, często wykorzystywana w celach edukacyjnych, do testowania nowych technologii lub jako dodatkowy ładunek na większych misjach.
- 2U: Składa się z dwóch jednostek (wymiary 10 x 10 x 20 cm). Te satelity CubeSat mogą pomieścić więcej instrumentów naukowych i czujników, co zwiększa ich możliwości badawcze.
- 3U: Jest to największa kategoria satelitów CubeSat, składająca się z trzech jednostek (wymiary 10 x 10 x 30 cm). Satelity CubeSat 3U mogą być wykorzystywane do bardziej zaawansowanych zadań, takich jak obserwacje Ziemi, eksperymenty naukowe lub testowanie zaawansowanych systemów komunikacyjnych.

Ponadto istnieją również niestandardowe konfiguracje satelitów CubeSat, takie jak np. 6U, 12U czy nawet większe, które mogą być dostosowane do specyficznych potrzeb misji. Jednak standardy od 1U do 3U są najczęściej stosowane ze względu na swoją uniwersalność i łatwość integracji z różnymi platformami startowymi.

Podstawowe założenia konstrukcyjne satelitów CubeSat są następujące:

- Mają małe wymiary, przez co znakomicie nadają się do zastosowań badawczych,
- Struktura - aluminiowy szkielet, do którego przymocowane są wszystkie elementy. Zapewnia właściwe ułożenie elementów w czasie startu rakiety nośnej, zabezpieczając podsystemy przed uszkodzeniem mechanicznym. Jednocześnie odgrywa ważną rolę w stabilizacji termicznej satelity. Jeśli jakiś podsystem wydziela dużo ciepła, jest ono przekazywane elementom struktury, cechującej się dużą pojemnością cieplną. Ze szkieletu nadmiar ciepła jest wypromieniowany w przestrzeń kosmiczną.
- CubeSat jest w całości zasilany energią słoneczną, zamienianą na prąd elektryczny w fotoogniwach, umieszczonych na ścianach satelity lub rozkładanych "skrzydłach". Fotoogniwa jednej ściany zapewnią prąd o mocy rzędu 2 W. Prąd zostaje częściowo wykorzystany do ładowania akumulatora litowo-jonowego, co zapewnia zasilanie w czasie, gdy satelita znajdzie się w cieniu Ziemi.
- Komunikacja - niezbędna do przesyłania przez satelitę danych na Ziemię oraz odbierania komend ze stacji naziemnej. System komunikacji składa się z dwóch modułów: komunikacyjnego oraz antenowego.
- Komputer pokładowy - odpowiada za zarządzanie pracą całego satelity. Poprzez system telekomunikacji odbiera z Ziemi komendy i przekazuje ich wykonanie do innych podsystemów. Jednocześnie zbiera informacje np. o temperaturze, buforuje je i przygotowuje do przesłania do naziemnej stacji odbiorczej w czasie sesji komunikacyjnej.
- Ładunek użyteczny - w zależności od przeznaczenia satelity może on zawierać zestaw dodatkowych modułów/czujników pozwalających na realizację określonych zadań badawczych.

3. STACJA NAZIEMNA

Zaprojektowana na Politechnice Poznańskiej i zlokalizowana na terenie Kampusu Kąkolewo stacja naziemna składa się z dwóch wież antenowych o wysokości 11m, wyposażonych w systemy antenowe umożliwiające komunikację w wybranych pasmach częstotliwości.

3.1 Wieża A

Wieża A została zaprezentowana na Rys. 1. Umieszczone na niej anteny zapewniają komunikację w pasmie VHF/UHF oraz pasmie C.

Rysunek 1: Wieża z systemami antenowymi na pasmo VHF/UHF oraz C

3.1.1 Pasmo VHF/UHF

Podsystem działający w pasmie VHF/UHF wykorzystywany jest przez wiele satelitów Cubesat, jako główny kanał komunikacji dla łącza telemetrycznego i sterującego.

Ponieważ zakres pasma C realizowany jest jedynie na łączu w dół, konieczne okazuje się zastosowanie innego łącza komunikacyjnego na łączu w górę, które byłoby przeznaczone do sterowania. W związku z tym podsystem transmisyjny dla pasma C ściśle współpracuje z innymi podsystemami komunikacyjnymi w celu przesyłania poleceń na łączu w górę.

W paśmie UHF/VHF wykorzystywane są dwie oddzielne anteny Yagi o wysokim zysku służące do ustanowienia dwukierunkowego łącza do komunikacji z satelitą CubeSat. Łącze w górę działa w paśmie 145 MHz i wykorzystuje antenę WX-220 firmy WiMo GmbH. Jest to antena Yagi składająca się z 2 x 10 elementów (polaryzacja kołowa) o długości 4,6 m i zysku 12 dBi. Łącze w dół działa w paśmie 432 MHz i wykorzystuje antenę WX-7036 firmy WiMo GmbH. Jest to antena Yagi o 2 x 18 elementach (polaryzacja kołowa), długości 3,4 m i zysku 14 dBi. Obie anteny Yagi są zamontowane na poprzeczce przymocowanej do precyzyjnego rotora azymutu i elewacji BIG-RAS/HR, który umożliwia śledzenie satelitów przemieszczających się po nieboskłonie.

3.1.2 Pasmo C

Antena używana w stacji naziemnej do komunikacji w pasmie C to siatkowa antena paraboliczna typu "prime focus" o średnicy 4,5 m, wyprodukowana przez firmę RF HAMDESIGN B.V. i dostarczana jako zestaw do samodzielnego montażu. Wspornik jest wykonany z aluminium, a siatka z ocynkowanej stali. Częstotliwość pracy anteny z siatką 2,8 mm wynosi 1 - 11 GHz. Zysk przy częstotliwości 5,8 GHz wynosi 43 dB, a kąt połowy mocy jest równy 1,3 stopnia.

Na potrzeby śledzenia satelitów używany jest rotor azymutu i elewacji. Precyzyjny rotor BIG-RAS/HR to rotor dwusilnikowy, który ma zainstalowane dwa czujniki położenia liczące 0,1 stopnia na impuls. Jest produkowany przez polską firmę SPID Elektronik i dostarczany razem ze sterownikiem MD-01 oraz zasilaczem PS-01 Dual Voltage Power Supply. Zakres obrotu dla AZ/EL wynosi 360/180 stopni z rozdzielczością 0,1 stopnia.

3.2 Wieża B

Antena wykorzystywana do łączności w pasmach S oraz X cechuje się modułową konstrukcją, składającą się z czterech połączonych ze sobą elementów, tworzących czaszę anteny parabolicznej. W przypadku systemu antenowego na pasmo S/X zapewniona jest dodatkowa ochrona przed niekorzystnymi czynnikami atmosferycznymi, takimi jak silny wiatr lub opady atmosferyczne w postaci kopuły wykonanej z włókien kompozytowych. Omawiany system antenowy może pracować w zakresach pasm 2,3 GHz (pasmo S) oraz 8 GHz (pasmo X) zapewniając tym samym efektywną transmisję w przypadku aplikacji satelitarnych o dużej przepływności danych (do 60 MBit/s), przy czym istnieje możliwość przełączania pasma, na którym w konkretnej chwili wymagana jest praca. W przypadku pasma X łącze realizowane jest tylko w dół (odbiór sygnału), natomiast układ radiowy pasma S wyposażony jest ponadto we wzmacniacz mocy cechujący się maksymalną mocą na poziomie 50 W wypromieniowanej fali nośnej.

Omawiany system antenowy cechuje się nieco odmiennymi możliwościami śledzenia satelitów w porównaniu z systemami antenowymi na pasma C oraz VHF/UHF, gdyż umożliwia ruch anteny w trzech płaszczyznach (azymut/elewacja/pochylenie), lepiej niwelując zjawisko martwych stref. Za sterowanie ruchem systemu antenowego odpowiada zintegrowana jednostka kontrolna, zarządzająca pracą serwo-motorów - ACU (ang. Allocated Control Unit). Wieża B została zaprezentowana na Rys. 2.

Rysunek 2: Wieża z systemem antenowym na pasma S/X wraz z kopułą

3.3 Centrum sterowania

Omawiana stacja naziemna składa się z trzech zasadniczych elementów. Pierwszy z nich stanowią opisane wcześniej systemy antenowe, realizujące łączność radiową z satelitami. Należy jednak zaznaczyć, że odebrany sygnał wymaga przetwarzania oraz stosownej konwersji z domeny analogowej do cyfrowej. Za cały proces odbioru oraz transmisji odpowiada Centrum Przetwarzania Danych, mieszczące się w kontenerze stacyjnym sąsiadującym z wieżami systemów antenowych. W skład Centrum Przetwarzania Danych wchodzą następujące grupy urządzeń:

- Moduły radia programowalnego - odpowiadają za obsługę torów radiowych dla pasm U-VHF/UHF/C, w tym konwersję sygnału z dziedziny analogowej do cyfrowej (w przypadku odbioru) lub cyfrowo-analogowej (w przypadku transmisji).
- Sterowniki rotorów anten - odpowiadają za sterowanie rotorami systemów antenowych dla pasm VHF/UHF/C, zgodnie z danymi sterującymi dostarczanymi poprzez interfejs sieciowy.
- Dedykowane modemy radiowe - realizują proces demodulacji oraz agregacji sygnałów transmitowanych z użyciem pasm S oraz X.
- Komputery dedykowane do pasm VHF/UHF/C - odpowiadają one za kontrolę funkcjonowania odbiorników radia programowalnego dla pasm VHF/UHF oraz agregację danych dla pasma C.
- Serwer przetwarzania danych - odpowiada za obróbkę odebranych danych przy użyciu wyspecyfikowanych skryptów przetwarzania danych w zależności od bieżącego typu danych.

- Serwer zarządzania - kontroluje pracę całego systemu satelitarnego, odpowiadając za planowanie obserwacji związanej z przelotami satelitów. Uczestniczy również w procesie dekodowania oraz demodulacji danych z pasm VHF/UHF/C, a także kontroluje pracę modemów dla pasm S/X oraz sterowników rotorów systemów antenowych.
- Macierz dyskowa - realizuje funkcję magazynowania odebranych danych, zarówno na etapie próbek sygnału przed demodulacją, jak i gotowych, zagregowanych danych satelitarnych.
- Komputery kontrolne - pełnią funkcje lokalnej kontroli systemu wewnątrz kontenera stacyjnego, umożliwiając lokalny dostęp do zasobów systemu poprzez stanowisko operatorskie.
- Przełączniki sieciowe - zapewniające połączenie wszystkich elementów systemu przy pomocy łączy optycznych z pozostałą infrastrukturą.

Oprócz Centrum Przetwarzania Danych oraz lokalnego dostępu, istnieje możliwość zdalnego zarządzania systemem oraz przeprowadzania obserwacji przy użyciu dedykowanego Centrum Zarządzania. Centrum to złożone jest z dedykowanych stanowisk operatorskich, połączonych z komputerami kontrolnymi znajdującymi się w osobnej lokalizacji w Instytucie Radiokomunikacji Politechniki Poznańskiej. Dzięki technice wirtualnych sieci lokalnych VLAN (ang. Virtual Local Area Network), możliwe jest ustanowienia połączenia z aparaturą Centrum Przetwarzania Danych i zdalne zarządzanie pracą systemu. Oprócz tego dostęp do niego możliwy jest z dowolnego miejsca na świecie przy użyciu wirtualnej sieci prywatnej - VPN (ang. Virtual Private Network). Zapewnia to możliwie największą elastyczność systemu i ciągły dostęp do usług łączności. W skład dedykowanego Centrum Zarządzania wchodzą:

- Dwa komputery w obudowach RACK 4U - umożliwiające zarządzanie systemem i wyposażone są w dedykowane układy graficzne. Dzięki nim dostęp do systemu możliwy jest przy użyciu interfejsu przeglądarki internetowej, umożliwiającej dostęp do zintegrowanych usług systemu.
- Dwa komputery w obudowach RACK 2U - zapewniające dostęp do dodatkowych programów sterujących poszczególnymi modułami Centrum Przetwarzania Danych oraz pełniące funkcje wspomagania dedykowanych komputerów kontrolnych.
- Przełącznik sieciowy - zapewniający szybkie oraz stabilne połączenie sieciowe komponentów Centrum Zarządzania z pozostałymi elementami systemu za pomocą łączy w standardzie Ethernet.

4. MOŻLIWOŚCI STACJI NAZIEMNEJ

Możliwości jakie oferuje stacja naziemna można zaklasyfikować do trzech kategorii:

- Transmisja satelitarna w pasmach VHF/UHF/C oraz S/X. Jest to podstawowa funkcjonalność stacji na potrzeby której stacja została zaprojektowana. Stacja ma możliwość nawiązania łączności z satelitami na niskich orbitach (ang. Low Earth Orbit, LEO) w pasmach VHF/UHF/C z wykorzystaniem układów nadawczo-odbiorczych w technologii SDR. Warto zauważyć, że takie rozwiązanie nie ogranicza komunikacji do z góry założonego standardu, ale pozwala na zdefiniowanie standardu transmisji zgodnie z wymaganiami użytkownika.

  W ramach pasm S/X stacja ma możliwość nawiązania łączności z satelitami na niskich orbitach (LEO) z przepływnością do 60Mbits/s w standardach CCSDS oraz DVB-S2/DVB-S2X.
- Pomiary sygnałów i systemów radiokomunikacyjnych z wykorzystaniem wektorowego generatora i analizatora sygnałów. Stacja naziemna umożliwia pomiar i badanie parametrów sygnałów dla systemów pracujących na częstotliwościach do 30 GHz. Na wyposażeniu stacji znajdują się urządzenia firmy Rohde&Schwarz: wektorowy analizator sygnałów FSV 3030 oraz generator wektorowy SMM100A.
- Przetwarzanie danych z akceleracją GPU. Przetwarzanie danych odbieranych z satelitów CubeSat może wymagać znacznych mocy obliczeniowych, dotyczy to m.in. przetwarzania zdjęć/obrazów. Dlatego stacja naziemna została wyposażona w serwer przetwarzania bazujący na procesorach GPU, pozwalający na wykorzystanie ponad 80 000 rdzeni CUDA, 2600 rdzeni Tensor oraz 256GB RAM do przyspieszenia obliczeń.

5. PODSUMOWANIE

Stacje naziemne są niezbędnym punktem infrastruktury w badaniach kosmicznych, obsługując komunikację z satelitami. Umożliwia to eksplorację kosmosu i pozwala na zbieranie cennych danych wykorzystywanych w badaniach naukowych. Współczesne stacje naziemne charakteryzują się wykorzystaniem zaawansowanych technologii, wykorzystaniem dużych anten parabolicznych i zastosowaniem zaawansowanych systemów odbioru i transmisji sygnałów.

Warto podkreślić, iż zaprezentowana stacja naziemna jest pierwszym tego rodzaju obiektem w kraju, należącym do uczelni wyższej. Będzie ona wykorzystywana zarówno do prowadzenia badań naukowych jak i działalności komercyjnej poprzez udostępnianie czasu antenowego jednostkom z poza Politechniki Poznańskiej.

Projekt finansowany w ramach RPWP.01.01.00-IZ.00-30-001/19 dla Działania 1.1 „Wsparcie infrastruktury B+R w sektorze nauki” Wielkopolskiego Regionalnego Programu Operacyjnego na lata 2014-2020

LITERATURA

[1] H. H. Abdullah, A. Elboushi, A. E. Gohar and E. A. Abdallah, "An Improved S-Band CubeSat Communication Subsystem Design and Implementation," in IEEE Access, vol. 9, pp. 45123-45136, 2021, doi: 10.1109/ACCESS.2021.3066464.

[2] J. Puig-Suari, C. Turner and W. Ahlgren, "Development of the standard CubeSat deployer and a CubeSat class PicoSatellite," 2001 IEEE Aerospace Conference Proceedings, Big Sky, MT, USA, 2001, pp. 347-353 vol.1, doi: 10.1109/AERO.2001.931726.

[3] R. Yadav, V. S. Pandey, S. Kumar and S. Gotra, "CubeSat Inter-Satellite Link-based Log Periodic Antenna for C-band Applications," 2022 IEEE Students Conference on Engineering and Systems (SCES), Prayagraj, India, 2022, pp. 1-6, doi: 10.1109/SCES55490.2022.9887711.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

Modelowanie nieliniowej charakterystyki szerokopasmowych wzmacniaczy radiowych o zmiennym napięciu zasilania

Modeling Nonlinear Characteristics of Wideband Radio Frequency Amplifiers with Variable Supply Voltage

Kornelia Kostrzewska[1]; Paweł Kryszkiewicz[2]

[1] Politechnika Poznańska, Poznań, kornelia.kostrzewska@student.put.poznan.pl
[2] Politechnika Poznańska, Poznań, pawel.kryszkiewicz@put.poznan.pl

DOI: 10.15199/59.2024.4.86

Streszczenie: Praca ma na celu zaproponowanie nowego modelu dla nieliniowej charakterystyki wzmacniacza radiowego ze zmiennym napięciem zasilania pracującym w szerokim zakresie częstotliwości. Przedstawiona została propozycja rozszerzonego modelu Rappa. Zaproponowany model zweryfikowano na podstawie pomiarów charakterystyk trzech różnych wzmacniaczy. Model ten może być wykorzystany do projektowania systemów 6G "świadomych" niedoskonałości układów wejściowo-wyjściowych. [1]

Abstract: The work aims to propose a new nonlinear characteristics model for a wideband radio amplifier of variable supply voltage. An extended Rapp model proposal is presented. The proposed model has been verified by measurements of three different amplifiers. This model can be used to design frontend-aware 6G systems.

Słowa kluczowe: Modelowanie wzmacniacza, Model Rappa, Śledzenie obwiedni, Wzmacniacz mocy

Keywords: Amplifier modeling, Rapp model, Envelope tracking, Power amplifier

1. WSTĘP

Prężny rozwój komunikacji bezprzewodowej jest ściśle związany z odpowiednim dostosowaniem systemów do obsługi coraz większej liczby urządzeń, przy zachowaniu odpowiedniej jakości proponowanych rozwiązań. Jedną z najnowszych odpowiedzi na te rosnące wymagania jest technologia 5G, która przewyższa swoich poprzedników na wielu płaszczyznach, zapewniając nawet kilkadziesiąt razy większą przepustowość komórkową oraz efektywność widmową [21]. Jednym z powodów tak dobrych własności tego systemu jest wykorzystanie złożonych schematów modulacji i technik wielodostępu, np. OFDM (ang. *Orthogonal Frequency Division Multiplexing*) [5, 22]. Zwiększają one przepustowość przy ograniczonym paśmie częstotliwości. Jednak sygnały te charakteryzują się wysokim stosunkiem mocy szczytowej do mocy średniej (ang. *Peak-to-Average Power Ratio*, PAPR) i stawiają duże wymagania co do liniowości układu nadajnika w celu utrzymania zniekształceń sygnału na niskim poziomie.

Kluczowym, w tej kwestii, elementem w torze nadawczym jest wzmacniacz mocy (ang. *Power Amplifier*, PA). Ich wykorzystanie jest związane ze znacznym zużyciem energii, a wprowadzane przez niego zniekształcenia nielinowe mogą przyczyniać się do zniekształceń sygnału w całym pasmie. Konieczna jest zatem wiedza o podstawowych parametrach PA np. jego liniowości. Celem jest zidentyfikowanie najkorzystniejszego punktu pracy będącego kompromisem pomiędzy efektywnością energetyczną, a efektywnością widmową.

W tym celu wykorzystuje się różne podejścia do architektury układów wzmacniaczy [7], które mają na celu osiągnięcie jak najbardziej liniowego wzmocnienia sygnału przy jednoczesnej minimalizacji zużycia energii. Śledzenie obwiedni (ang. *Envelope Tracking*, ET) wyróżnia się jako jedno z najbardziej popularnych i bardzo skutecznych podejść [8]. Podstawowym założeniem jest zmienianie amplitudy napięcia zasilającego wzmacniacz, do chwilowej obwiedni sygnału podawanego na tor radiowy wzmacniacza.

Z perspektywy projektowania algorytmów nadawczo-odbiorczych lub optymalizacji parametrów transmisyjnych konieczna jest znajomość charakterystyki nieliniowej wzmacniacza. Najczęściej odbywa się to poprzez modelowanie behawioralne [8]. Modele behawioralne nie wymagają głębokiej wiedzy na temat budowy wewnętrznej wzmacniacza np. układu tranzystorów. Skupiają się one na sygnale podawanym na wejście i uzyskanym na wyjściu wzmacniacza w wyniku pomiarów lub symulacji obwodów elektronicznych [4, 5].

Istnieje wiele publikacji naukowych [4, 7, 10, 21], które szeroko omawiają podstawowe modele behawioralne. Niektóre modelują tylko efekty bezpamięciowe (Soft limiter, Rapp, Saleh). Inne uwzględniają również „pamięć" wzmacniacza (Voltera, Wiener, Hammerstein). Wybór w dużej mierze zależy od badanego w danym przypadku wzmacniacza oraz wymaganej dokładności.

Jako że technika ET upowszechnia się pozwalając osiągać wysoką jakość nadawanego sygnału przy wysokiej efektywności energetycznej konieczne jest opracowanie nowych modeli behawioralnych uwzględniających zmienne napięcie zasilania jak również szerokopasmowy charakter transmisji oczekiwany od systemów 6G. W pracach [1, 2] zaprezentowane są propozycje rozszerzonego modelu Saleha. Autorzy dodają niezależne współczynniki, które poprzez funkcję wielomianową uzależniają charakterystykę nieliniową od chwilowego napięcia zasilania.

[1]Praca powstała w ramach projektu OPUS finansowanego przez Narodowe Centrum Nauki nr 2021/41/B/ST7/00136.

Z drugiej strony, w [13, 14] punktem wyjścia jest tradycyjny pamięciowy model wielomianowy, w którym współczynniki są uzależniane od obwiedni sygnału na wejściu wzmacniacza i wejścia modulatora zasilania. Przeanalizowano także prace badawcze koncentrujące się na rozszerzeniu modelu Rappa [20, 23]. Podobnie jak poprzednio napięcie zasilania jest włączone do modelu podstawowego jako dodatkowa zmienna, która wpływa na współczynniki definiujące nieliniowość wzmacniacza.

Wszystkie powyższe prace skupiają się na włączeniu do modelu napięcia zasilania PA. Jednak, z uwagi na to, że współczesne systemy telekomunikacyjne działają w szerokim zakresie częstotliwości, włączenie częstotliwości nośnej jako parametru do analizy sygnału może korzystnie wpłynąć na efektywność widmową, a także energetyczną systemu. W żadnym z wymienionych powyżej artykułów nie rozważano tego problemu.

Ta praca przedstawia nowy model behawioralny oparty na tradycyjnym modelu Rappa, który został rozszerzony o wpływ zmiennego napięcia zasilania wzmacniacza, a także częstotliwości nośnej sygnału nadawanego. Model Rappa został wybrany z uwagi na swoją prostotę, wyrażoną niewielką liczbą parametrów, oraz dużą popularność, wynikająca z dobrego dopasowania do współczesnych wzmacniaczy mocy wykorzystujących tranzystory, [6]. Został też zaproponowany do analizy wpływu wzmacniacza na sygnał systemu 5G [18].

W rozdziale 2 zaprezentowany został wykorzystany system pomiarowy oraz wyniki pomiarów. W rozdziale 3 przedstawiony jest proponowany model charakterystyki wzmacniacza. Analiza poprawności zaproponowanego modelu zamieszczona została w rozdziale 4, a wnioski znajdują się w rozdziale 5.

2. SYSTEM POMIAROWY I WYNIKI POMIARÓW

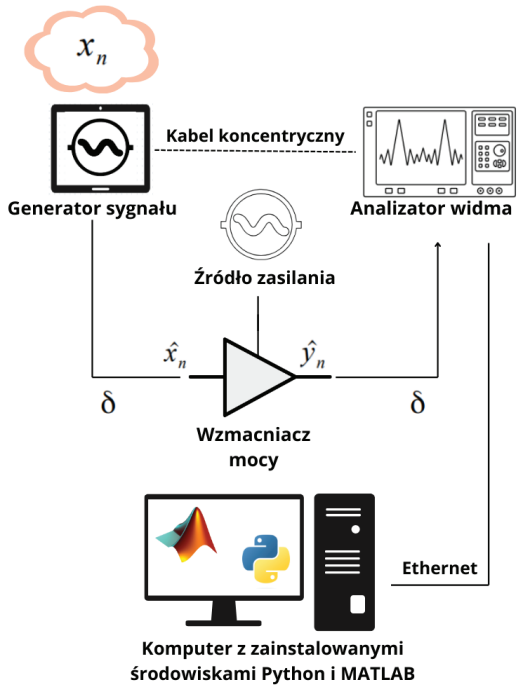
Schemat pomiarowy przedstawiono na Rys. 1 i składa się z następujących elementów:

- generator sygnału: Rohde & Schwarz SMBV 100A,
- analizator widma: Rohde & Schwarz FSL 6 [19],
- badane wzmacniacze mocy od Mini-Circuits: ZFL-2000+ [15], ZX60-5916 [17], ZX60-2534 [16],
- źródło zasilania: NN M10-QP-305E,
- komputer z zainstalowanymi środowiskami MATLAB i Python, połączony za pomocą kabla Ethernetowego.

Głównym celem tego zestawu jest zebranie próbek po ich zniekształceniu przez wzmacniacz mocy.

Generator sygnału wysyła cyklicznie 10 symboli OFDM, o zadanej częstotliwości, które są podawane na wejście testowanego wzmacniacza. Jeden symbol składa się z 4096 podnośnych, spośród których zajętych jest 590 podnośnych z zakresu od -300 do -1 oraz od 10 do 300.

Wzmacniacz zasilany jest napięciem stałym z zasilacza. Dostępny układ pomiarowy nie pozwalał na dynamiczną zmianę napięcia zasilania jak w przypadku architektury ET. Napięcie zasilania zmieniane jest z pewnym krokiem zależnym od modelu wzmacniacza (co opisano w rozdziale 2.1) w całym dostępnym dla danego modelu przedziale ręcznie. Podobnie zmieniano częstotliwość nośną nadawanego sygnału. Zabieg ten ma na celu uzyskanie pełnej charakterystyki zachowania wzmacniacza.



Rysunek 1: Schemat pomiarowy

Sygnał z wyjścia wzmacniacza jest odbierany przez analizator widma. Pobrane próbki mają formę zespolonych próbek IQ, które umożliwiają analizę sygnału w paśmie podstawowym. Podejście to jest powszechne w wielu badaniach dotyczących analizy lub modelowania sygnałów [3, 11, 10]. W odebranym wektorze próbek IQ zapisany jest jeden pełny okres sygnału nadawanego. Aby uniknąć zniekształceń nieliniowych na wejściu analizatora widma, konieczne było ustawienie silnego tłumienia na wejściu analizatora, tzn. około $50dB$.

Warto również wspomnieć, że oscylatory lokalne nadajnika i odbiornika zostały zsynchronizowane poprzez kabel koncentryczny.

Ostatecznie odebrany wektor próbek IQ trafia do komputera podłączanego do analizatora za pomocą kabla Ethernetowego.

Przygotowanie próbek obejmowało uwzględnienie tłumienia kabli łączących urządzenia oraz tzw. 'wirtualnego wzmocnienia' w przypadku sygnału wejściowego, wynikającego z faktu, że próbki cyfrowe sygnału nadawanego są „wzmacniane" poprzez generator sygnałowy do określonej mocy. Dokładny sposób pomiaru tłumienia oraz przeskalowanie próbek sygnału wejściowego i wyjściowego przedstawiono w [9].

W następnym kroku próbki poddano korekcji przesunięcia w dziedzinie czasu oraz w dziedzinie częstotliwości. Wykorzystana procedura korekcji została przedstawiona w [9]. W efekcie dysponujemy dwoma sygnałami: wejściowym $\hat{x}$ oraz wyjściowym y_n które są zależne poprzez funkcję $g(\ )$ stanowiąca charakterystykę wzmacniacza:

$$y_n = g\left(\hat{x}\right). \tag{1}$$

Możliwe jest więc dopasowanie modelu behawioralnego, jak pokazano w [9].

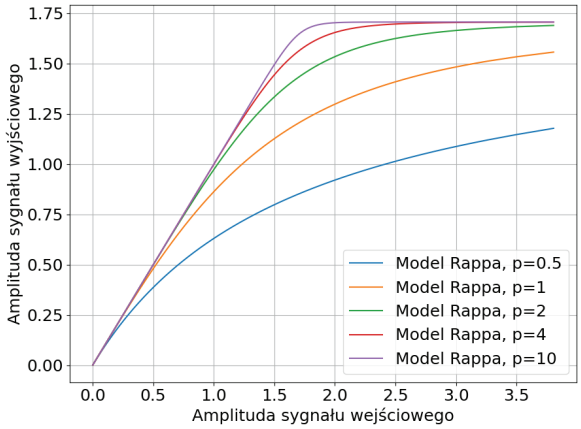
Dla każdego zestawu parametrów wejściowych dopasowano do próbek wyjściowych model Rappa. W tym modelu sygnał (zespolony) na wyjściu wzmacniacza jest

opisywany jako:

$$y_n = G\hat{x_n}\left(1+\left(\frac{|\hat{x_n}|}{V_{sat}}\right)^{2p}\right)^{-\frac{1}{2p}}, \qquad (2)$$

gdzie p to współczynnik gładkości (liczba rzeczywista dodatnia), V_{sat} napięcie nasycenia na wejściu wzmacniacza, G to wzmocnienie liniowe amplitudy sygnału, a $\hat{x_n}$ odnosi się do sygnału podawanego na wejście wzmacniacza.

Rys. 2 przedstawia charakterystykę AM/AM wzmacniacza tzn. amplitudę sygnału wyjściowego w funkcji amplitudy sygnału wejściowego, modelowanego wzorem (2) dla różnych wartości p. Napięcie nasycenia V_{sat} oraz wzmocnienie G zostały przyjęte jako stałe. Zauważalne jest, że im większa wartość parametru p, tym w większym zakresie amplitud wzmacniacz działa liniowo, z wyraźnym obcięciem wyższych amplitud.



Rysunek 2: Charakterystyka AM/AM dla różnych wartości p, $V_{sat} = 1.7V$, $G = 1$ w skali liniowej (V).

2.1 Zakresy parametrów wejściowych wzmacniaczy

Dla każdego modelu badanego wzmacniacza, na podstawie danych przedstawionych w jego nocie katalogowej, dobrano zakres napięcia zasilającego wzmacniacz (Tabela 1) oraz zakres częstotliwości nośnej (Tabela 2).

Tabela 1: Zakres napięcia zasilania i testowane wartości dla wzmacniaczy mocy

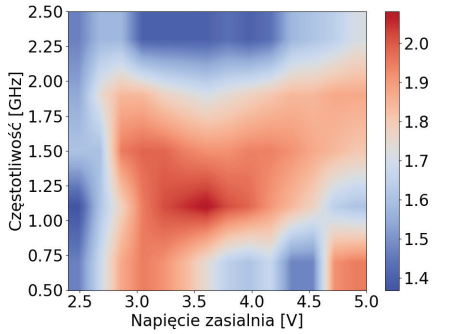
Wzmacniacz	*Katalogowe wartości [V]*	*Testowane wartości [V]*
ZFL-2000+	max 17	8.0,9.5,. . . , 14.5,15.0
ZX60-5916	2.8, 5.0	2.4,2.6,. . . , 4.8,5.0
ZX60-2534	2.8, 5.0	2.4,2.6,. . . , 4.8,5.0

Tabela 2: Zakres częstotliwości nośnej i testowane wartości dla wzmacniaczy

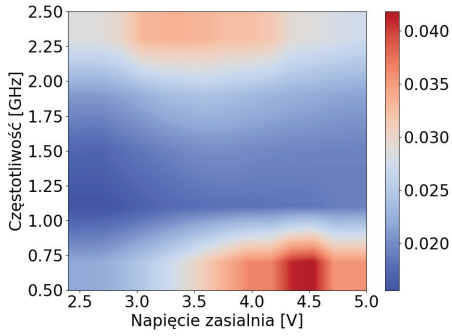
Wzmacniacz	*Katalogowe wartości [MHz]*	*Testowane wartości [GHz]*
ZFL-2000+	10 – 2000	0.5,1,. . . ,2
ZX60-5916	1500 — 6000	2,2.5,. . . ,5.5
ZX60-2534	500 — 2500	0.5,1,. . . ,2.5

2.2 Analiza wyników pomiarów

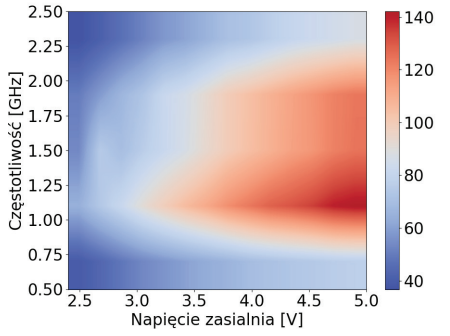
Pomiary oraz korekcje przeprowadzono dla każdego modelu w całym zakresie jego parametrów. Następnie wyznaczono współczynniki modelu Rappa, wykorzystując do tego algorytm przedstawiony w [9]. Wszystkie otrzymane wyniki zostały udostępnione w [12].



(a) Parametr p



(b) Parametr V_{sat}



(c) Parametr G

Rysunek 3: Mapy cieplne dla modelu ZX60-2534

Rys. 3 przedstawia otrzymane wartości współczyn-

ników Rappa dla wzmacniacza ZX60-2534 oraz ich zmienność w funkcji parametrów wejściowych za pomocą map cieplnych.

Widoczne jest, że wzmocnienie rośnie wraz z napięciem zasilania. Tempo tego wzrostu jest jednak istotnie różne w zależności od częstotliwości nośnej. Należy pamiętać, że jest to wzmocnienie napięcia w skali liniowej. Analizując zmienność współczynnika gładkości p zauważalna jest znacząca zmienność wartości tego parametru. Pomimo istotnych różnic między wzmacniaczami (wyniki dla pozostałych dwóch wzmacniaczy mogą być znalezione w [9]), maksymalne wartości p dla wszystkich przypadków nie przekraczają $p = 3,0$. Oznacza to, że badane wzmacniacze wykazują silnie nieliniowe zachowanie. Natomiast napięcie nasycenia V_{sat} zmienia się wraz z parametrami wejściowymi co najmniej dwukrotnie dla każdego badanego modelu wzmacniacza. Ogólnie można powiedzieć, że jego zmienność stanowi funkcję dwóch badanych parametrów wejściowych. Podsumowując, każdy z obserwowanych parametrów zmienia się znacząco zarówno w funkcji częstotliwości jak i napięcia zasilania wzmacniacza co oznacza konieczność ich zamodelowania. Wykresy nie wykazują jednak gwałtownych zmian (nieciągłości), co jest podstawą do zaproponowanego rozszerzenia modelu Rappa, w którym wartości współczynników wyrażone są przez funkcje dwóch zmiennych wejściowych, co pokazano w rozdziale 3.

3. PROPONOWANY MODEL CHARAKTERYSTYKI WZMACNIACZA

Przedstawiony w tej pracy model został zainspirowany rozszerzeniem modelu Rappa, przedstawionym w [20]. Jego autorzy opisali współczynniki modelu G, p oraz V_{sat} jako funkcje zależne od obwiedni napięcia zasilania PA V_{sup}:

$$y_n = G(V_{sup})\hat{x}_n \left(1+\left(\frac{|\hat{x}_n|}{V_{sat}(V_{sup})}\right)^{2p(V_{sup})}\right)^{-\frac{1}{2p(V_{sup})}} \tag{3}$$

Nasza propozycja to uwzględnienie dodatkowo wartości częstotliwości nośnej w funkcjach opisujących wartości współczynników modelu Rappa tzn.

$$y_n = G(V_{sup}, f)\hat{x_n}\left(1+\left(\frac{|\hat{x_n}|}{V_{sat}(V_{sup}, f)}\right)^{2p(V_{sup},f)}\right)^{-\frac{1}{2p(V_{sup},f)}} \tag{4}$$

Pierwszym krokiem było przeprowadzenie pomiarów 3 różnych wzmacniaczy dla określonych wartości napięcia zasilania wzmacniacza V_{sup} oraz częstotliwości nośnej f i wyznaczenie dla każdej badanej pary wartości współczynników modelu Rappa: p, V_{sat} oraz G. Następnie na tej podstawie dopasowano funkcje opisujące te parametry. W tym celu wykorzystanie jedno z rozszerzeń środowiska MATLAB - MATLAB Curve Fitting Tool.

W przypadku każdego współczynnika przetestowano kilka funkcji szukając takiej o najlepszym dopasowaniu dla wzmacniacza ZX60-2534.

Pierwsze podejście polegało na dopasowaniu powierzchni zdefiniowanej przez wielomian drugiego stopnia zarówno w funkcji f jak i V_{sup}, niezależnie dla współczynników p, V_{sat} oraz G,

$$\begin{aligned} poly22 = {} & p00 + p10 \cdot V_{sup} + p01 \cdot f + p20 \cdot V_{sup}^2 + \\ & + p11 \cdot V_{sup} \cdot f + p02 \cdot f^2. \end{aligned} \tag{5}$$

Rozważono także wielomian trzeciego stopnia:

$$\begin{aligned} poly33 = {} & p00 + p10 \cdot V_{sup} + p01 \cdot f + p20 \cdot V_{sup}^2 + \\ & + p11 \cdot V_{sup} \cdot f + p02 \cdot f^2 + p30 \cdot V_{sup}^3 + \\ & + p21 \cdot V_{sup}^2 \cdot f + p12 \cdot V_{sup} \cdot f^2 + p03 \cdot f^3. \end{aligned} \tag{6}$$

Następne podejście polegało na dopasowaniu dwóch niezależnych wielomianów, uwzględniających zależność od V_{sup} oraz f, i wyrażanie całkowitej zależności jako iloczyn równań opisujących te charakterystyki tzn.

$$f(V_{sup}, f) = g(V_{sup}) \cdot h(f). \tag{7}$$

Dla parametrów V_{sat} oraz G funkcje te mają więc postaci:

$$f_{V_{sat}}(V_s up, f) = (log(V_{sup}) + a) \cdot (b \cdot f^3 + c \cdot f + d), \tag{8}$$

oraz

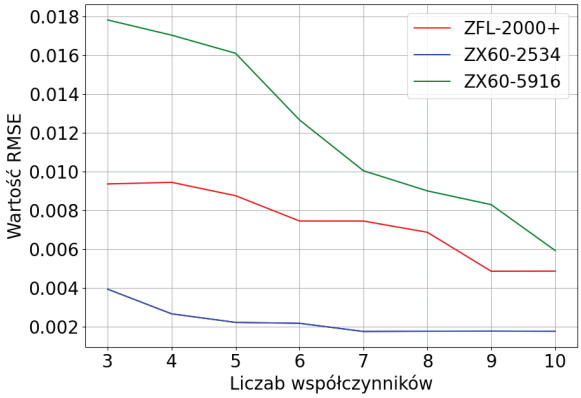
$$f_G(V_s up, f) = (log(V_{sup}) + a) \cdot (b \cdot f^3 + c \cdot f^2 + d \cdot f + e). \tag{9}$$

W przypadku parametru p nie udało się dobrać optymalnych funkcji opisujących zmienność oddzielnie dla parametrów V i f. Złożoność zmienności tego parametru jest na tyle wysoka, że nie można jej precyzyjnie opisać oddzielnie dla zmiennych wejściowych, ponieważ wykazuje ona silną zależność zarówno od napięcia, jak i częstotliwości.

W ostatnim podejściu skupiono się na stopniowym eliminowaniu kolejnych współczynników wielomianu dla optymalizacji dopasowania. Proces ten rozpoczął się od zdefiniowania funkcji złożonej ze wszystkich możliwych monomianów do stopnia N składających się z kombinacji zmiennych V_{sup} oraz f. N dobrano tak wysokie, żeby wartość pierwiastka z błędu średniokwadratowego (ang. *Root Mean Square Error*, RMSE) osiągnęła minimum tzn. dalsze zwiększanie stopnia wielomianu nie wpłynęło na obniżenie RMSE. Następnie współczynniki były usuwane jeden po drugim, obserwując, jak zmieniały się wartości RMSE. Najmniej istotny człon wielomianu tzn. ten, którego usunięcie najmniej podniosło RMSE, był trwale usuwany z funkcji. Te operacje były iteracyjnie powtarzane, dopóki wartości RMSE pozostawały na podobnym poziomie.

Przykładowe wyprowadzenie wzoru wykorzystując ostatnie podejście przedstawiono dla współczynnika V_{sat} (analogicznie wyglądało ono dla pozostałych współczynników). W tym przypadku danymi, na których dokonywano redukcji współczynników były te otrzymane z pomiaru wzmacniacza ZX60-2534, a wielomianem początkowym był (6) o 10 współczynnikach. Na Rys. 4 przedstawiono wartości RMSE dla otrzymanych równań. Równania te następnie zaaplikowano do danych z pozostałych dwóch wzmacniaczy. Otrzymane wartości RMSE również przedstawiono na Rys.4. Można zauważyć, że we wszystkich przypadkach wartość RMSE zwiększa się w miarę usuwania kolejnych wyrazów wielomianu, a dodatkowo jest różna pomiędzy wzmacniaczami. Celem było więc wybranie wielomianu o małej liczbie współczynników, który dobrze dopasowuje się dla wszystkich rozważanych wzmacniaczy. W tym przypadku wybór padł na wielomian opisany czterema współczynnikami,

$$poly_{V4} = p10 \cdot V_{sup} + p11 \cdot V_{sup} \cdot f + p02 \cdot f^2 + p12 \cdot V_{sup} \cdot f^2. \tag{10}$$



Rysunek 4: RMSE jako funkcja liczby współczynników dla parametru V_{sat}

Aby uzyskać wzór najlepiej opisujący zmienność danego parametru modelu Rappa, porównano równania uzyskane z różnych podejść, uwzględniając wartości RMSE. Postać tych funkcji (np. zbiór monominanów) wyznaczono na podstawie danych ze wzmacniacza dla którego zmienność danego parametru (np. p) była największa. Następnie, równania te zastosowano do pozostałych dwóch wzmacniaczy, co umożliwiło porównanie dokładności dopasowania między nimi.

Przykładowe porównanie modelowania współczynnika V_{sat} przedstawiono w Tabeli 3, oraz na Rys. 5 dla wzmacniacza ZFL-2000+ transmitującego sygnał na nośnej 1.5 GHz. Widoczne jest, że wartości RMSE dla wzmacniacza ZX60-2534 są najniższe. Wynika to z faktu, że dane pomiarowe z tego wzmacniacza zostały wykorzystane do wyznaczania postacji funkcji opisującej parametr V_{sat}. Funkcje $poly22$ (5) oraz $f(V_{sup}, f)$ (7) dla każdego wzmacniacza osiągają największe wartości RMSE, natomiast najlepsze dopasowanie wykazuje równanie $poly33$ (6), ale kosztem wykorzystania największej liczby współczynników. Równanie $poly_{V4}$ (10) wykorzystuje jedynie cztery współczynniki i osiąga mniejsze wartości RMSE, niż równanie $f(V_{sup}, f)$ (7), które wykorzystuje tą samą liczbę współczynników, dla dwóch z trzech badanych wzmacniaczy. Widać również, że krzywa wyznaczona przez równanie $poly_{V4}$ (10) dobrze pokrywa się z danymi pomiarowymi, dlatego zostało ono wybrane jako najlepiej opisujące zmienność parametru:

$$V_{sat}(V_{sup}, f) = p10 \cdot V_{sup} + p11 \cdot V_{sup} \cdot f + p02 \cdot f^2 + p12 \cdot V_{sup} \cdot f^2. \quad (11)$$

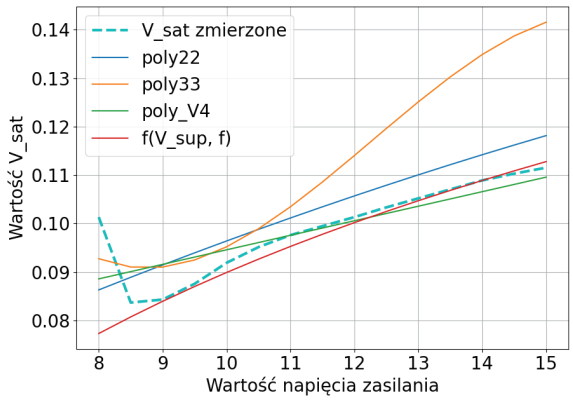
Analogicznie przeprowadzono modelowanie parametrów p oraz G. Uzyskano następujące funkcje:

$$G(V_{sup}, f) = (log(V_{sup}) + a) \cdot (b \cdot f^3 + c \cdot f^2 + d \cdot f + e), \quad (12)$$

$$p(V_{sup}, f) = p10 \cdot V_{sup} + p01 \cdot f + p20 \cdot V_{sup}^2 + p02 \cdot f^2. \quad (13)$$

Tabela 3: Modelowanie V_{sat}

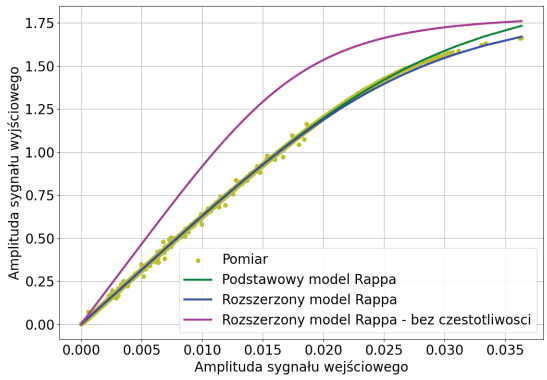
Wzmacniacz	*Równanie*	*Liczba wsp.*	*RMSE*
ZFL-2000+	$poly22$	6	0.007
	$poly33$	10	0.004
	$poly_{V4}$	4	0.009
	$f(V_{sup}, f)$	4	0.014
ZX60-2534	$poly22$	6	0.002
	$poly33$	10	0.001
	$poly_{V4}$	4	0.002
	$f(V_{sup}, f)$	4	0.003
ZX60-5916	$poly22$	6	0.011
	$poly33$	10	0.005
	$poly_{V4}$	4	0.017
	$f(V_{sup}, f)$	4	0.014



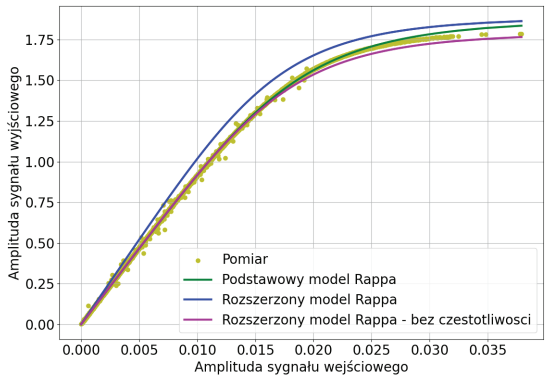
Rysunek 5: Porównanie modelowania V_{sat} dla częstotliwości 1.5GHz

4. ANALIZA JAKOŚCI DOPASOWANIA PROPONOWANEGO MODELU

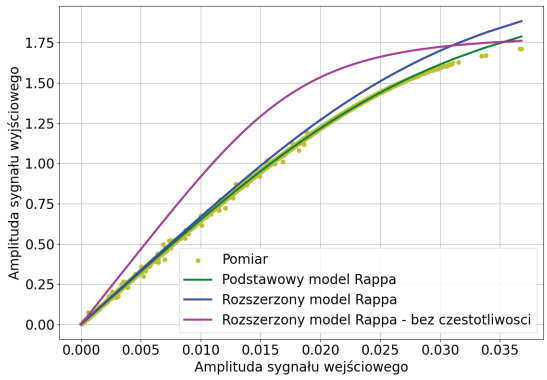
Dla każdego z badanych wzmacniaczy i uzyskanych współczynników p, V_{sat}, G dopasowano funkcje dwuwymiarowe (11), (12) i (13). Jakość tego dopasowania można ocenić poprzez analizę charakterystyki AM/AM wzmacniacza używając współczynników modelu Rappa: podstawowego (tzn. skalarów pochodzących z dopasowania dla danej częstotliwości i napięcia zasilania), rozszerzonego (tzn. zaproponowanego w tym artykule) oraz rozszerzonego bez częstotliwości (tzn. zaproponowanego w [20]). Rysunek 6 przedstawia przykładowe porównanie wykreślone dla wzmacniacza ZX60-2534 dla napięcia zasilania $V_{sup} = 3,6V$ oraz różnych częstotliwości. Tu warto zaznaczyć, ze model przedstawiony w [20] nie uwzględnia zmiennej częstotliwości, współczynniki dla tego modelu w przypadku wzmacniacza ZX60-2534 zostały wyznaczone dla częstotliwości $f = 1.5GHz$. Dla pozostałych wzmacniaczy również wybrano częstotliwości środkowe do wyznaczenia współczynników.



(a) Częstotliwość $f = 0.5GHz$



(b) Częstotliwość $f = 1.5GHz$

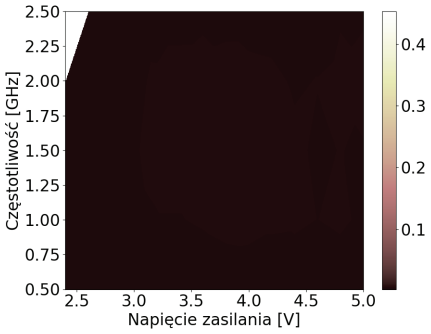


(c) Częstotliwość $f = 2.5GHz$

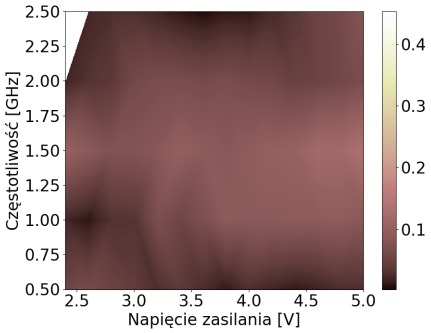
Rysunek 6: Porównanie charakterystyk AM/AM (w V) modeli Rappa dla wzmacniacza ZX60-2534

Z analizy wykresu wynika, że zaproponowany rozszerzony model Rappa aproksymuje charakterystyki wzmacniacza w ich liniowym zakresie pracy praktycznie tak dokładnie jak model podstawowy. W każdym przypadku obserwuje się bardziej znaczące różnice podczas modelowania próbek o najwyższych amplitudach. Co ważne, dokładność przybliżenia jest dobra niezależnie od częstotliwości nośnej sygnału nadanego. Tej uniwersalności nie zapewnia model zaproponowany w [20]. Dla częstotliwości $1.5GHz$, dla której zostały wyznaczone współczynniki Rappa, przybliżenie wykazują prawie zerowy błąd. Jednak w przypadku pozostałych częstotliwości tzn. 0,5 GHz oraz 2,5GHz, występują bardzo duże odchylenia.

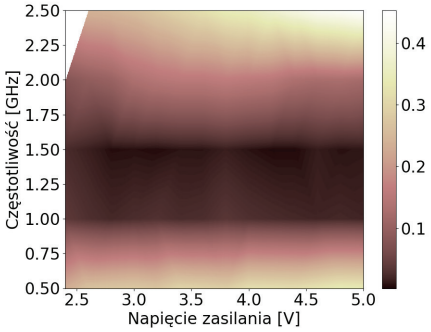
Pokazuje to znormalizowany (względem wartości zmierzonej charakterystyki AM/AM) pierwiastek ze średniego błędu kwadratowego (ang. *Normalized Root Mean Squared Error*, NRMSE) dopasowania krzywej AM/AM pokazany w funkcji napięcia zasilania i częstotliwości dla wzmacniacza ZX60-2534 na Rys. 7 c). Osiągane są wartości do 0.4 (co można interpretować jako 40% błąd w estymowanej wartości wyjściowej charakterystyki AM/AM wzmacniacza). Dla porównania NRMSE dla proponowanego modelu (Rys. 7 b)) w całym zakresie działania wzmacniacza rozkłada się równomiernie i nie przekracza 12%. Zgodnie z oczekiwaniami, model podstawowy dysponujący niezależnym zestawem parametrów G, V_{sat}, p, dla każdego punktu wykresu daje najdokładniejsze wyniki tzn. NRMSE nie przekracza 0.6%.



(a) NRMSE dla modelu podstawowego



(b) NRMSE dla modelu rozszerzonego



(c) NRMSE dla modelu rozszerzonego bez częstotliwości

Rysunek 7: NRMSE dla danych z modelu ZX60-2534

Tabela 4 przedstawiają średnie wartości NRMSE wyliczone na podstawie pełnych danych dla każdego badanego wzmacniacza. Jak widać w przypadku wzmacniacza ZX60-2534 proponowany model wykazuje ponad dwu-

krotnie mniejsze wartości NRMSE niż model nieuwzględniający częstotliwości. Jednak dla pozostałych wzmacniaczy wartości NRMSE są wyższe. Szczególnie w przypadku modelu ZFL-2000+ błąd otrzymany z zaproponowanego modelu jest znacznie wyższy, co może sugerować inną budowę tego wzmacniacza. Trzeba podkreślić, że postaci funkcji aproksymujących współczynniki (np. p) wzmacniaczy tj. (11), (12) i (13) były dobierane dla wzmacniacza ZX60-2534. Uzasadnia to bardzo dobre zamodelowanie jego własności przy pomocy niewielkiej liczby współczynników. W przypadku innych wzmacniaczy, choć współczynniki funkcji były dobierane niezależnie struktura tych funkcji była nie zmieniana. Zasadne wydaje się dostosowanie struktury funkcji dla konkretnych urządzeń.

Tabela 4: NRMSE dla różnych modeli Rappa

	Model podstawowy
ZFL-2000+	0.017825
ZX60-2534	0.003980
ZX60-5916	0.006975
	Model rozszerzony
ZFL-2000+	0.476953
ZX60-2534	0.055058
ZX60-5916	0.103444
	Model rozszerzony bez częstotliwości
ZFL-2000+	0.118918
ZX60-2534	0.162267
ZX60-5916	0.102723

5. WNIOSKI

W artykule została przedstawiona propozycja rozszerzonego modelu Rappa poprzez wyrażenie jego parametrów jako funkcji wielomianowych parametrów wejściowych: napięcia zasilania wzmacniacza i częstotliwości nośnej przesyłanego sygnału. Wykorzystanie takiego podejścia może otworzyć nowe możliwości modelowania sygnału na wyjściu wzmacniaczy ze zmiennym napięciem zasilania, na przykład dla stacji bazowych używanych w systemach komórkowych następnej generacji.

Zaproponowany model zapewnia wystarczającą dla wielu celów dokładność opisu charakterystyki nieliniowej przy użyciu niewielkiej liczby współczynników koniecznych do estymacji dla konkretnego wzmacniacza.

Choć zaproponowany model dobrze oddaje charakterystykę nieliniową dla wzmacniacza dla którego dobierana była funkcja opisująca parametry modelu Rappa, badania wykazały, że jego skuteczność znacząco spada w przypadku innych wzmacniaczy. Wyniki te sugerują, że model nie jest wystarczająco uniwersalny i konieczne jest dostosowanie funkcji opisującej parametry wzmacniacza dla każdego wzmacniacza indywidualnie. Sugerowane są zatem dalsze badania dla znalezienia bardziej uniwersalnego modelu.

LITERATURA

[1] Al-Kanan, Haider, Li, Fu, Tafuri, Felice Francesco, 2017, "Extended saleh model for behavioral modeling of envelope tracking power amplifiers". *2017 IEEE 18th Wireless and Microwave Technology Conference (WAMICON)*, 1–4, IEEE.

[2] Al-kanan, Haider, Tafuri, Felice, Li, Fu, 2018, "Hysteresis nonlinearity modeling and linearization approach for envelope tracking power amplifiers in wireless systems". *Microelectronics journal*, 82: 101–107.

[3] Boumaiza, Slim, Helaoui, Mohamed, Hammi, Oualid, Liu, Taijun, Ghannouchi, Fadhel M, 2007, "Systematic and adaptive characterization approach for behavior modeling and correction of dynamic nonlinear transmitters". *IEEE Transactions on Instrumentation and Measurement*, 56 (6): 2203–2211.

[4] Ghannouchi, Fadhel M., Hammi, Oualid, 2009, "Behavioral modeling and predistortion". *IEEE Microwave Magazine*, 10 (7): 52–64.

[5] Glock, Stefan, Rascher, Jochen, Sogl, Bernhard, Ussmueller, Thomas, Mueller, Jan-Erik, Weigel, Robert, 2015, "A memoryless semi-physical power amplifier behavioral model based on the correlation between am–am and am–pm distortions". *IEEE Transactions on Microwave Theory and Techniques*, 63 (6): 1826–1835.

[6] Glock, Stefan, Rascher, Jochen, Sogl, Bernhard, Ussmueller, Thomas, Mueller, Jan-Erik, Weigel, Robert, 2015, "A memoryless semi-physical power amplifier behavioral model based on the correlation between am–am and am–pm distortions". *IEEE Transactions on Microwave Theory and Techniques*, 63 (6): 1826–1835.

[7] Joung, Jingon, Ho, Chin Keong, Adachi, Koichi, Sun, Sumei, 2014, "A survey on power-amplifier-centric techniques for spectrum-and energy-efficient wireless communications". *IEEE Communications Surveys & Tutorials*, 17 (1): 315–333.

[8] Kim, Dongsu, Kang, Daehyun, Choi, Jinsung, Kim, Jooseung, Cho, Yunsung, Kim, Bumman, 2011, "Optimization for envelope shaped operation of envelope tracking power amplifier". *IEEE transactions on microwave theory and techniques*, 59 (7): 1787–1795.

[9] Kostrzewska, Kornelia, Kryszkiewicz, Pawel, 2024, "Power amplifier modeling framework for front-end-aware next-generation wireless networks". *Electronics*, 13 (9): 1643.

[10] Kryszkiewicz, Pawel, 2018, "Amplifier-coupled tone reservation for minimization of ofdm nonlinear distortion". *IEEE Transactions on Vehicular Technology*, 67 (5): 4316–4324.

[11] Kryszkiewicz, Pawel, Kliks, Adrian, Bogucka, Hanna, 2015, "Obtaining low out-of-band emission level of an nc-ofdm waveform in the sdr platform". *2015 International Symposium on Wireless Communication Systems (ISWCS)*, 66–70, IEEE.

[12] Kryszkiewicz, Pawel, Kostrzewska, Kornelia, 2024, "Measurements of nonlinearity characteristics and power consumption of 3 power amplifiers (ZFL-2000+,ZX60-2534,ZX60-5916)".

[13] Li, Delong, Yu, Hui, 2016, "A new model for envelope tracking power amplifier modeling and digital predistortion". *2016 8th International Conference on Wireless Communications & Signal Processing (WCSP)*, 1–5, IEEE.

[14] Mengozzi, Mattia, Angelotti, Alberto Maria, Gibiino, Gian Piero, Florian, Corrado, Santarelli, Alberto, 2021, "Joint dual-input digital predistortion of supply-modulated rf pa by surrogate-based multi-objective optimization". *IEEE Transactions on Microwave Theory and Techniques*, 70 (1): 35–49.

[15] Mini-Circuits, ZFL-2000+, "Power amplifier's datasheet". "https://www.minicircuits.com/pdfs/ZFL-2000+.pdf".

[16] Mini-Circuits, ZX60-2534, "Power amplifier's datasheet". "https://www.minicircuits.com/pdfs/ZX60-2534MA+.pdf".

[17] Mini-Circuits, ZX60-5916, "Power amplifier's datasheet". "https://www.minicircuits.com/pdfs/ZX60-5916MA+.pdf".

[18] Nokia, 2016, "Realistic power amplifier model for the New Radio evaluation". 3GPP doc. R4-163314.

[19] Rohde & Schwarz, R&S FSL6, "Manual for the spectrum analyzer". "https://scdn.rohde-schwarz.com/ur/pws/dl_downloads/dl_common_library/dl_manuals/gb_1/f/sfl_1/FSL_OperatingManual_en_12.pdf".

[20] Tafuri, Felice Francesco, Sira, Daniel, Nielsen, Troels Studsgaard, Jensen, Ole Kiel, Mikkelsen, Jan Hvolgaard, Larsen, Torben, 2015, "Memory models for behavioral modeling and digital predistortion of envelope tracking power amplifiers". *Microprocessors and Microsystems*, 39 (8): 879–888.

[21] Thota, Sravanti, Kamatham, Yedukondalu, Paidimarry, Chandra Sekhar, 2020, "Analysis of hybrid papr reduction methods of ofdm signal for hpa models in wireless communications". *IEEE Access*, 8: 22780–22791.

[22] Thota, Sravanti, Kamatham, Yedukondalu, Paidimarry, Chandra Sekhar, 2020, "Analysis of hybrid papr reduction methods of ofdm signal for hpa models in wireless communications". *IEEE Access*, 8: 22780–22791.

[23] Wang, Feipeng, Yang, Annie Hueiching, Kimball, Donald F, Larson, Lawrence E, Asbeck, Peter M, 2005, "Design of wide-bandwidth envelope-tracking power amplifiers for ofdm applications". *IEEE Transactions on Microwave theory and techniques*, 53 (4): 1244–1255.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

LUSTRZANY KODEK VIDEO BAZUJĄCY NA GŁĘBOKIEJ SIECI NEURONOWEJ

MIRRORED VIDEO CODEC BASED ON DEEP NEURAL NETWORK

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: 10.15199/59.2024.4.87

Streszczenie: **Zagadnienie stabilnego przekazywania informacji, a zwłaszcza stabilnej transmisji informacji w formie video w warunkach braku stabilnej prędkości przesyłania danych dla specyficznych zastosowań jest kwestią istotną. Celem niniejszego artykułu jest przedstawienie efektywnej metody modyfikacji kodeka video zapewniającej zwiększenie jakości video dla wysokich wartości kompresji wprowadzanej przez kodek video poprzez jego modyfikacje z wykorzystaniem jego lustrzanego odbicia bazując na głębokiej sieci neuronowej.**

Abstract: **The issue of stable information transmission, and especially stable information transmission in the form of video in the absence of a stable data transmission speed for specific applications, is an important issue. The aim of this article is to present an effective method of modifying the video codec ensuring an increase in video quality for high values of compression introduced by the video codec by modifying it using its mirror image based on a deep neural network.**

Słowa kluczowe: **kodek video, obraz, video, głęboka sieć neuronowa, kompresja.**

Keywords: **video codec, image, video, deep neural network, compression.**

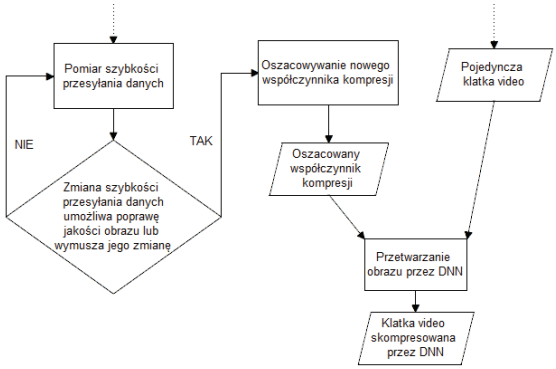
1. WSTĘP

Współcześnie przesyłanie informacji często następuje w formie video. Zachowanie stabilności transmisji z możliwie jak największą jakością video w warunkach zmiennej prędkości przesyłania informacji stanowi jedną z istotnych kwestii dla zagadnienia przesyłania informacji nie tylko w sieci Internet, ale również innych profesjonalnych zastosowaniach. Dla potrzeb rozpoznania obrazowego, nawigacji czy też wykonywania operacji medycznych na odległość istnieje potrzeba stabilnego przesyłania obrazu w czasie możliwie zbliżonego do rzeczywistego przy zachowaniu możliwie najwyższej jakości video celem zachowania jak największej liczby szczegółów. Degradacja obrazu wprowadzana przez kompresję kodeka video wraz z wzrostem współczynnika kompresji prowadzi do zaniku szczegółowości obrazu jak i zniekształceń elementów charakterystycznych. Degradacja obrazu może prowadzić do podejmowania niepewnych lub niewłaściwych decyzji, a nawet do opóźnienia lub braku podjęcia decyzji ze względu na niepewność lub brak danych do ich podjęcia [11].

W celu zapewnienia stabilności transmisji najczęściej stosowana jest zmiana współczynnika kompresji video lub odpowiednia priorytetyzacja przesyłanej informacji w sieci [1, 2]. Dla specyficznych potrzeb istnieją również dedykowane modyfikacje kodeka video [3, 12]. W literaturze istnieją metody poprawy już skompresowanego video takie jak: video restoration [10, 13], video enhancement [4, 8] oraz video super-resolution [9]. Wyżej wymienione metody bazują na poprawie już zdegradowanego obrazu. Proponowana w tym artykule metoda ma na celu zwiększenie jakości obrazu (ilości oraz szczegółowości detali w obrazie) przed zastosowaniem opcjonalnego procesu poprawy jakości video po stronie odbiorczej wpływając również na finalną efektywność metod poprawiających jakość zdegradowanego obrazu. Poprawa jakości strumieniowanego video po stronie odbiorczej wiąże się z dodatkowym opóźnieniem czasowym potrzebnym na dokonanie stosownych obliczeń zwiększając opóźnienie przekazywanego strumienia video, a tym samym opóźniając przekazywaną finalnie informację. Podając przykład nawigacji drona, w którym dodatkowe opóźniania w transmisji rzędu sekund wymuszają wprowadzenie opóźnienia reakcji zautomatyzowanego systemu jak i nawigatora, co w konsekwencji może prowadzić do zderzenia drona z dynamicznie pojawiającą się przeszkodą. Proponowana metoda nie tylko umożliwia potencjalną poprawę jakości działania metod poprawiających jakość zdegradowanego obrazu, ale również może być stosowana, bez z nich z zmodyfikowaną charakterystyką degradacji wprowadzoną przez głęboką sieć neuronową (DNN - ang. Deep Neural Network).

2. KONCEPCJA METODY

Proponowana metoda zachowania jakości video przedstawiona jest na schemacie ideowym (Rys. 1). W założeniach metody zakładane jest istnienie systemu mierzącego prędkość przesyłania informacji. W przypadku zmian prędkości powodujących opóźnienia w przesyłaniu informacji następuje zmiana współczynnika jakości video zwiększając stopień kompresji obrazu. W przeciwnym przypadku poprawy szybkości przesyłania informacji następuje zmniejszenie współczynnika kompresji.



Rys. 1. Schemat ideowy proponowanej metody

Po podjęciu decyzji o zmianie współczynnika kompresji następuje szacowanie nowego współczynnika kompresji na bazie zmian szybkości przesyłania danych. Następnie nowo określony współczynnik kompresji podawany jest na wejście DNN wraz z kolejną klatką video, która to ma zostać poddana kompresji. Wynikiem działania DNN jest obraz skompresowany z użyciem nowej charakterystyki kompresji opartej na lustrzanym odbiciu kodeka video.

Na potrzeby badań zastosowano modyfikację szeroko stosowanego kodeka video HEVC (High-Efficiency Video Coding). Proponowana metoda może zostać zastosowana z użyciem innych kodeków video zawracając uwagę na zakres wartości współczynnika kompresji wynikający z specyfikacji danego kodeka video. Współczynnik kompresji dla HEVC określany jest w zakresie wartości $\langle 0; 51\rangle$ oraz opisywany jest za pomocą współczynnika QP (ang. Quantization Parameter) lub CRF (ang. Constant Rate Factor). Dla współczynnika kompresji QP wartość kompresji posiada stałą określoną wartość, natomiast dla współczynnika kompresji CRF współczynnik kompresji może posiadać mniejszą bądź większą wartość niż początkowo określoną w zależności od kompresowanego obrazu oraz dynamiki jego zmian wraz z kolejnymi klatkami. Domyślna wartość współczynnika kompresji dla HEVC wynosi 28.

2.1. Proces uczenia sztucznej sieci neuronowej

Proces uczenia realizowany był na bazie algorytmu wstecznej propagacji błędów. W celu realizacji procesu uczenia lustrzanego kodeka video przyjęto następujące zależności funkcji celu:

$$MSE_{HevcDiff} = \frac{\left((1.00 \cdot MSE_{HevcNormal}(QP)) + (0.25 \cdot MSE_{HevcMirror}(QP))\right)}{2} \quad (1)$$

$$MSE_{HevcNormal}(QP) = MSE\begin{pmatrix} Image_{Hevc}(QP), \\ Image_{DNN}\left(\frac{QP}{51}\right) \end{pmatrix} \quad (2)$$

$$MSE_{HevcMirror}(QP) = MSE\begin{pmatrix} Image_{Hevc}(QP), \\ Image_{DNN}\left(\frac{(51-QP)}{51}\right) \end{pmatrix} \quad (3)$$

$$L_{DNNLoss} = \left(1.75 \cdot MSE_{HevcDiff}\right) \quad (4)$$

$$\nabla L_{DNN} = \frac{\partial L_{DNNLoss}}{\partial Var_{DNN}} \quad (5)$$

Gdzie MSE to błąd średniokwadratowy (ang. Mean Square Error); $MSE_{HevcDiff}$ to błąd średniokwadratowy obrazu przetworzonego przez DNN w porównaniu do obrazu poddanego kompresji HEVC parametryzowany przez współczynnik kompresji QP z uwzględnieniem lustrzanego odbicia kompresji kodeka video HEVC. Lustrzane odbicie kodeka video jest odbijane częściowo $0.25 \cdot MSE_{HevcMirror}$; QP (Quantization Parameter) to parametr determinujący stopień kompresji HEVC w zakresie wartości $QP \in \langle 0\,;51\rangle$; $Image_{Hevc}$ to obraz uzyskany w wyniku kompresji HEVC; $MSE_{HevcNormal}$ to błąd średniokwadratowy obrazu przetworzonego przez DNN w porównaniu do obrazu poddanego kompresji HEVC parametryzowany przez współczynnik kompresji QP; $Image_{DNN}$ to obraz uzyskany w wyniku kompresji DNN; $MSE_{HevcMirror}$ to błąd średniokwadratowy obrazu przetworzonego przez DNN w porównaniu do obrazu poddanego kompresji HEVC parametryzowany przez współczynnik kompresji QP odwrócony w sposób lustrzany poprzez odjęcie od maksymalnej wartości $QP_{Max} = 51$ podanej wartości QP; ∇L_{DNN} to gradient funkcji celu DNN; $L_{DNNLoss}$ to błąd DNN; Var_{DNN} to tensor wag modelu DNN.

Zbiór uczący składał się z 4000 obrazów z zbiorów danych Places [14] oraz Flickr-Faces-HQ (FFHQ) [5]. Obrazy z zbioru uczącego zostały przekształcone do rozdzielczości 128 × 128 stanowiącej rozmiar wejściowy obrazu dla DNN. Zbiór walidacyjny składał się z klatek materiału video [7].

2.2. Architektura sztucznej sieci neuronowej

Tabele 1–2 przedstawiają strukturę DNN. Tabela 1 przedstawia strukturę logicznej warstwy podstawowej (LWP). Natomiast Tabela 2 przedstawia strukturę DNN. Przedstawione w tabelach podstawowe typy warstw są zdefiniowane w bibliotece uczenia maszynowego TensorFlow. Zastosowany optimizer jest oparty na algorytmie Adam o współczynniku 10^{-4}. Funkcje aktywacji dla poszczególnych warstw podstawowych stanowi LeakyReLU z przypisaną wartością aktywacji 10^{-2}.

Tab. 1. Logiczna Warstwa Podstawowa – LWP

Numer warstwy	*Typ warstwy*
1	Conv2D(Wejście)
2	Conv2D(1)
3	Conv2D(2)
4	BatchNormalization(3)

Tab. 2. Struktura modelu DNN

Numer warstwy	*Typ warstwy*
1	LWP(Obraz wejściowy)
2	LWP(1)
3	Multiply(2, QP)
4	Concatenate((2, 3), axis = 3)
5	BatchNormalization(4)
6	LWP(5)
7	LWP(5)
8	LWP(5)
9	Concatenate((6, 7, 8), axis = 3)
10	LWP(9)
11	LWP(10)
12	Conv2D(11)

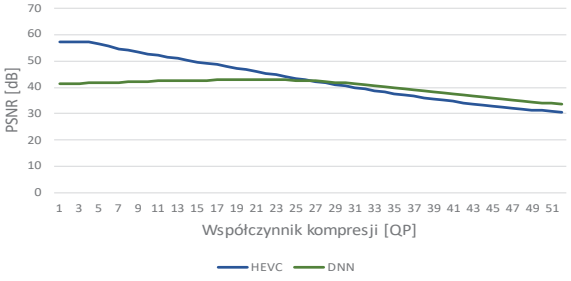
Struktura DNN, która ze względu na zastosowane warstwy stanowi sieć splotową posiada 5491047 parametrów. Do zapisu DNN na nośniku danych wymagane jest 67 MB. Liczba zastosowanych filtrów wynosi 156, a rozmiar jądra dla warstw splotowych wynosi 3 × 3.

W Tabeli nr 2 *Obraz wejściowy* jak i wynik działania DNN stanową obraz znormalizowany w zakresie wartości $\langle 0; 1 \rangle$ o rozdzielczości 128 × 128. QP stanowi wartość współczynnika kompresji z zakresu $\langle 0; 51 \rangle$ znormalizowaną w zakresie $\langle 0; 1 \rangle$.

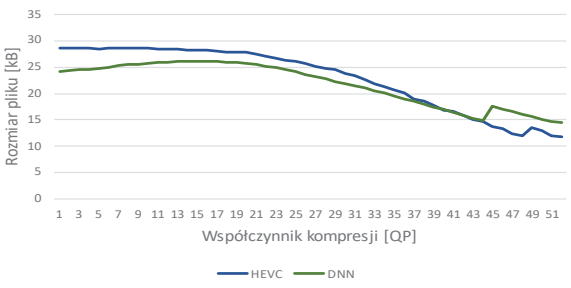
3. WYNIKI BADAŃ

Badania przeprowadzono na systemie operacyjnym Linux (Ubuntu 23.10) z wykorzystaniem biblioteki uczenia maszynowego TensorFlow w wersji 2.15.1. Do przetwarzania wideo wykorzystano bibliotekę obsługi wideo FFmpeg (wersja 7:6.0-6ubuntu1). Kody źródłowe zostały napisane w Python w wersji 3.11.6. Charakterystyki (Rys. 2–3) przedstawiają porównanie działania DNN z lustrzanym odbiciem kodeka video w porównaniu do natywnego kodeka HEVC. Proponowana metoda nie modyfikuje wyłączenie poszczególnych elementów składowych dedykowanego kodeka video celem poprawy działania kompresji. Ze względu na kompleksowość zmian oraz adaptowalność metody do dowolnego kodeka video jak również brak podobnych metod w dostępnej literaturze efektywność metody zostanie przedstawiona poprzez bezpośrednie porównanie do działania wybranego kodeka video.

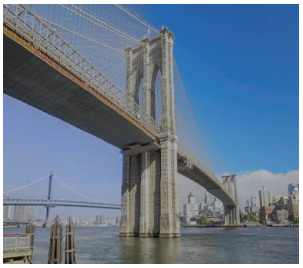
Rysunki 4–5 przedstawiają zobrazowanie wyników działania metody. Zbiór walidacyjny składał się z klatek materiału video o rozdzielczości 384 × 384 [7].



Rys. 2. Charakterystyka uśrednionych wartości podobieństwa od oryginalnego (niezmodyfikowanego) obrazu dla kompresji HEVC oraz kompresji DNN



Rys. 3. Charakterystyka uśrednionych wartości potrzebnej pamięci do zapisu obrazu dla kompresji HEVC oraz kompresji DNN



Rys. 4. Podgląd przykładowej oryginalnej (niemodyfikowanej) klatki video

Rys. 5. Podgląd działania kompresji (W kolumnach od lewej kompresja HEVC, kompresja DNN. W wierszach kolejno od góry zobrazowania kompresji dla wartości QP: 40, 51).

Miara podobieństwa określana jest metryką PSNR (ang. Peak Signal-to-Noise Ratio). Wielkość potrzebnej pamięci do zapisu jest mierzona poprzez zapis klatek video na dysku w formacie pliku JPEG.

Analizując charakterystykę z Rysunku 2 widoczna jest zmiana charakterystyki kompresji DNN w porównaniu do kompresji HEVC. Kompresja HEVC posiada wyższą wartość PSNR dla niskich wartości kompresji. W tym miejscu warto zauważyć, że standardowa wartość współczynnika kompresji dla HEVC wynosi 28 oraz, że ze względu na duży rozmiar pliku wynikowego bardzo niskie wartości kompresji nie są często stosowane. W dalszej części charakterystyki widoczny jest generalny trend zwiększenia podobieństwa obrazu przetworzonego przez DNN w porównaniu do kodeka video HEVC począwszy od wartości współczynnika kompresji wynoszącej 26.

Rysunek 3 przedstawia zmiany wielkości pliku wynikowego dla pojedynczej klatki video. Ze względu na funkcję celu (1) omówioną w sekcji 2.1 charakterystyka rozmiaru pliku wyjściowego w funkcji współczynnika kompresji różni się od charakterystyki HEVC. DNN dla niskich wartości kompresji charakteryzuje się mniejszym rozmiarem pliku wynikowego. Dla wysokich wartości współczynnika kompresji HEVC posiada mniejsze wartości potrzebnej pamięci do zapisu jednakże kosztem większej degradacji obrazu wynikowego (Rys. 2 i 5).

4. PODSUMOWANIE

Dobór odpowiedniej jakości strumieniowania video celem zapewnienia stabilności rozumianej jako brak dodatkowych opóźnień w transmisji ze względu na zbyt wysoką jakość video względem posiadanej prędkości transmisji danych stanowi istotne zagadnienie dla dedykowanych zastosowań. Przedstawiona metoda lustrzanego kodeka video bazującego na głębokiej sieci neuronowej przedstawia metodę modyfikacji dowolnego kodeka video celem zmiany jego charakterystyki kompresji.

Zmniejszeniu podlega jakość obrazu (PSNR) dla niskich wartości współczynnika kompresji, natomiast jakość podlega zwiększeniu dla wyższych wartości kompresji. Zmianą podlega również charakterystyka zmiany wielkości pliku wynikowego zmniejszając rozmiar pliku wynikowego dla niskich wartości kompresji oraz analogicznie zwiększając dla wyższych wartości kompresji w porównaniu do HEVC co jest związane z zwiększeniem jakości dla tych wartości kompresji przez DNN. Należy zauważyć, że przy niewielkim wzroście rozmiaru plików wynikowych dla kompresji DNN uzyskiwane są wyżej jakościowo obrazy zachowujące więcej detali w obrazie oraz wprowadzające efekt rozmycia (ang. Blurr) zamiast degradacji kształtów w porównaniu do kompresji HEVC. Fakt ten ułatwia potencjalne odzyskanie jakości obrazu ze względu na opisywane w literaturze efektywne metody usuwania rozmycia (ang. Deblurring) [6].

Potencjalnym miejscem aplikacji proponowanej metody są urządzenia przekazujące obraz w warunkach zakłóceń pasywnych jak i aktywnych w tym celowych. Przykładami tego typu urządzeń są drony znajdujące się na lądzie, wodzie w powietrzu lub podwodą.

Dalsze badania powinny uwzględniać inne mnożniki wagi lustrzanego odbicia oraz łączenie charakterystyk różnych kodeków video.

LITERATURA

[1] Antoniou Zinonas, Panayides Andreas, Pantzaris Marios, Constantinides Anthony, Pattichis Constantinos, Pattichis Marios. 2018. „Real-Time Adaptation to Time-Varying Constraints for Medical Video Communications". *IEEE Journal of Biomedical and Health Informatics* 22, 4, 1177–1188. DOI:https://doi.org/10.1109/JBHI.2017.2726180.

[2] Balagalla Umaya Bhashini, Sivanatham Savansi, Munasinghe Kusala, Subasinghe Akila, Chamitha de Alwis, Wijewardhana Uditha, and Dharmaweera M. Nishan. 2019. „Efficient Medical Video Streaming by Pre-Processing and Network Traffic Prioritization in Real-Time". In 2019 International Conference on Advanced Technologies for Communications (ATC), 123–128. DOI:https://doi.org/10.1109/ATC.2019.8924511.

[3] Bouaafia Soulef, Khemiri Randa, Messaoud Seifeddine, Ben Ahmed Olfa, Sayadi Fatma Ezahra. 2022. „Deep learning-based video quality enhancement for the new versatile video coding". *Neural computing & applications* 2022;34(17):14135-14149. DOI:https://doi.org/10.1007/s00521-021-06491-9.

[4] Chen Shengjie and Ye Mao. 2021. „Two-stage Multi-frame Cooperative Quality Enhancement on Compressed Video". In 2021 11th International Conference on Intelligent Control and Information Processing (ICICIP), 94–99. DOI:https://doi.org/10.1109/ICICIP53388.2021.9642200.

[5] Karras Tero, Laine Samuli, Aila Timo. 2019. „A Style-Based Generator Architecture for Generative Adversarial Networks". DOI:https://doi.org/10.48550/arXiv.1812.04948.

[6] Kong Lingshun, Dong Jiangxin, Li Mingqiang, Ge Jianjun, Pan Jinshan. 2022. „Efficient Frequency Domain-based Transformers for High-Quality Image Deblurring". DOI:https://doi.org/10.48550/arXiv.2211.12250.

[7] „LG: Spain and Patagonia". 2019. Available online: https://4kmedia.org/lg-spain-and-patagonia-uhd-4k-demo/ (accessed on 29 April 2024).

[8] Liu Ying, Wong Alexander, Fieguth Paul. 2011. „A structure-guided conditional sampling model for video resolution enhancement". In 2011 18th IEEE International Conference on Image Processing, 1169–1172. DOI:https://doi.org/10.1109/ICIP.2011.6115637.

[9] Lu Mingxuan, Peng Zhang. 2024. „Real-World Video Super-Resolution with a Degradation-Adaptive Model" Sensors 24, no. 7: 2211. DOI:https://doi.org/10.3390/s24072211.

[10] Mathew Liz Maria, Suma R, Kizhakkethottam Jubilant J. 2015. „A survey on different video restoration techniques". In 2015 International Conference on Soft-Computing and Networks Security (ICSNS). DOI:https://doi.org/10.1109/ICSNS.2015.7292442.

[11] O'Byrne Michael, Vibhoothi Michael, Sugrue Mark, Kokaram Anil. 2022. „Impact of Video Compression on the Performance of Object Detection Systems for Surveillance Applications". DOI:https://doi.org/10.48550/arXiv.2211.05805.

[12] Tian Kuan, Guan Yonghang, Xiang Jinxi, Zhang Jun, Han Xiao, Yang Wei. 2023. „Towards Real-Time Neural Video Codec for Cross-Platform Application Using Calibration Information". In Proceedings of the 31st ACM International Conference on Multimedia (MM '23), Association for Computing Machinery, Ottawa ON, Canada, , 7961–7970. DOI:https://doi.org/10.1145/3581783.3611955.

[13] Yu Bing, Ding Youdong, Huang Xi, Wu Bing. 2016. „Corrupted old film sequences restoration using improved PatchMatch and low-rank matrix recovery". In 2016 International Conference on Audio, Language and Image Processing (ICALIP), 295–300. DOI:https://doi.org/10.1109/ICALIP.2016.7846589.

[14] Zhou Bolei, Lapedriza Agata, Xiao Jianxiong, Torralba Antonio, Oliva Aude. 2014. „Learning Deep Features for Scene Recognition using Places Database". *In Advances in Neural Information Processing Systems*. Curran Associates. ISBN: 9781510800410.

PREDYKCJA MIĘDZYWIDOKOWA DLA KOŁOWEGO ROZMIESZCZENIA KAMER

INTER-VIEW PREDICTION FOR CIRCULAR CAMERA ARRANGEMENT

Jarosław Samelak

Politechnika Poznańska, Instytut Telekomunikacji Multimedialnej, jaroslaw.samelak@gmail.com

DOI: 10.15199/59.2024.4.88

Streszczenie: **Niniejsza praca przedstawia opis badań nad nową metodą predykcji międzywidokowej, dedykowanej systemom kamer rozmieszczonych w przybliżeniu na okręgu. W pierwszej części przedstawiono autorską metodę rektyfikacji wielowidokowych sekwencji wizyjnych o kołowym rozmieszczeniu kamer. W drugiej części dostosowano technikę 3D-HEVC do szybkiej i efektywnej kompresji zrektyfikowanych sekwencji. Ocena eksperymentalna wykazała, że zaproponowana metoda jest zarówno szybsza, jak i bardziej efektywna niż 3D-HEVC bez autorskich modyfikacji.**

Abstract: **The paper presents the author's research on novel inter-view prediction technique, dedicated to systems with cameras distributed roughly on a circle. First, the author proposes a novel rectification method of multiview video acquired by nearly-circular camera arrangements. Then, the author modifies the inter-view prediction of 3D-HEVC for fast and efficient compression of rectified video. According to the results of experimental evaluation, the proposal is both faster and more efficient than the state-of-the-art 3D-HEVC.**

Słowa kluczowe: **3D-HEVC, kompresja wizji, predykcja międzywidokowa, rektyfikacja wizji, wizja wielowidokowa.**

Keywords: **3D-HEVC, video compression, inter-view prediction, video rectification, multiview video.**

1. WSTĘP

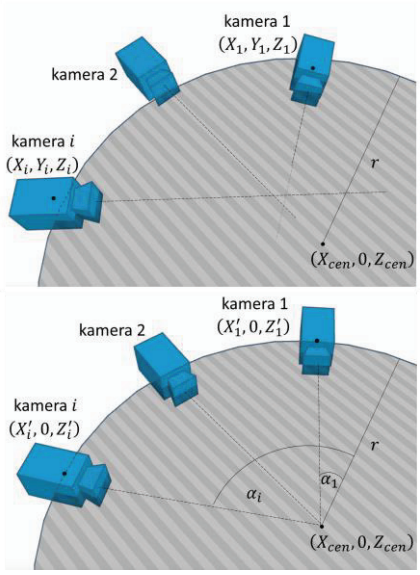
W ostatnich latach rośnie zainteresowanie nowymi technikami multimedialnymi, takimi jak rzeczywistość wirtualna i rozszerzona, czy też wizja wszechogarniająca [1]. Rozwój tego typu technik przyczynia się do wzrostu znaczenia kompresji wizji wielowidokowej, rejestrowanej za pomocą systemów kamer rozmieszczonych wokół sceny, często na łuku lub okręgu. Niestety, zarówno MV-HEVC, jak i 3D-HEVC, które są aktualnie dedykowanymi kodekami dla kompresji wizji wielowidokowej, są przystosowane do efektywnej kompresji wizji rejestrowanej za pomocą kamer gęsto rozmieszczonych na linii [6]. W przypadku innego rozmieszczenia kamer, efektywność kompresji MV-HEVC oraz 3D-HEVC znacząco spada ze względu na niską skuteczność realizowanej przez nie predykcji międzywidokowej [5]. Poprawę efektywności kompresji dla dowolnie rozmieszczonych kamer można uzyskać wykonując w predykcji międzywidokowej rzutowanie punktów obrazu pomiędzy płaszczyznami kamer [4-5]. Niestety, takie rozwiązanie wiąże się z istotnym wydłużeniem czasu kodowania oraz koniecznością przesyłania do dekodera wszystkich parametrów kamer. W niniejszym artykule przedstawiono rozwiązanie pośrednie, umożliwiające szybką oraz efektywną kompresję wizji rejestrowanej za pomocą kamer rozmieszczonych w przybliżeniu na okręgu.

Predykcja międzywidokowa w 3D-HEVC zakłada, że kamery rejestrujące obraz zostały rozmieszczone idealnie na linii, a osie optyczne kamer są względem siebie równoległe [6]. W praktyce jest to niemożliwe do wykonania ze względu na ograniczone możliwości pozycjonowania oraz różnice pomiędzy parametrami wewnętrznymi kamer. Z tego względu, przed kompresją, zazwyczaj wykonywana jest rektyfikacja wizji wielowidokowej, której celem jest kompensacja niedokładności rozmieszczenia kamer oraz eliminacja różnic w ich właściwościach [2]. Wynikiem rektyfikacji są sekwencje wizyjne skorygowane w taki sposób, jak gdyby były zarejestrowane przez idealnie rozmieszczone, identyczne kamery. Należy podkreślić, że rektyfikacja nie zmienia rzeczywistych pozycji kamer, a jedynie przekształca zarejestrowane przez nie obrazy. Rektyfikacja przed kodowaniem 3D-HEVC zapewnia bardziej dokładną predykcję międzywidokową, a przez to także lepszą efektywność kompresji [6].

W przypadku systemów, w których kamery są rozmieszczone na okręgu wokół rejestrowanej sceny, precyzyjne ustawienie kamer stanowi jeszcze większe wyzwanie. W niniejszym artykule przedstawiono oryginalną procedurę rektyfikacji dla systemów kamer rozmieszczonych w przybliżeniu na okręgu, zwaną dalej rektyfikacją kołową. Rektyfikacja kołowa ma na celu kompensację niedokładności w rozmieszczeniu kamer na okręgu oraz różnic pomiędzy parametrami kamer. W drugiej części artykułu przedstawiono modyfikację predykcji międzywidokowej w kodeku 3D-HEVC, której celem jest efektywna i jednocześnie szybka kompresja wizji wielowidokowej zrektyfikowanej kołowo.

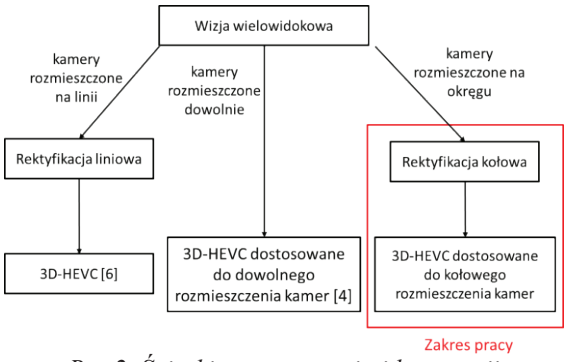
2. CEL I ZAKRES BADAŃ

Pierwszym celem opisywanych badań było opracowanie i implementacja rektyfikacji kołowej, która jest procesem analogicznym do rektyfikacji liniowej stosowanej przed kodowaniem 3D-HEVC. Efektem rektyfikacji kołowej jest takie przekształcenie sekwencji wizyjnych, jak gdyby zostały zarejestrowane za pomocą kamer rozlokowanych idealnie na okręgu, z osiami optycznymi leżącymi na jednej płaszczyźnie i skierowanymi w stronę środka okręgu (Rys. 1). Takie rozmieszczenie dostarcza znacznie więcej informacji o scenie, niż ustawienie kamer na linii, co jest kluczowe dla wielu zastosowań.



Rys. 1. Wizualizacja rozmieszczenia kamer przed (góra) i po (dół) rektyfikacji kołowej

Drugim celem badań było dostosowanie techniki 3D-HEVC do efektywnej kompresji sekwencji zrektyfikowanych kołowo. Proponowane podejście stanowi alternatywną ścieżkę przetwarzania wizji wielowidokowej, w porównaniu do dwóch rozwiązań już opisanych w literaturze (Rys. 2) [4][6]. Połączenie rektyfikacji kołowej z dedykowanym kodekiem ma na celu osiągnięcie kompromisu pomiędzy szybkością i prostotą 3D-HEVC, a zaletami wynikającymi z rozmieszczenia kamer wokół sceny.



Rys.2. Ścieżki przetwarzania i kompresji wizji wielowidokowej

3. REKTYFIKACJA KOŁOWA

W tym rozdziale opisano proponowany proces rektyfikacji kołowej, w tym wyprowadzenie parametrów okręgu, modyfikację parametrów kamer oraz przekształcenie sekwencji testowych.

3.1. Parametry kamer

W artykule zakłada się, że wszystkie parametry kamer są znane i reprezentowane przez macierz parametrów wewnętrznych [3×3] $\mathbb{K}$ (1), macierz obrotu [3×3] $\mathbb{R}$ oraz wektor translacji [3×1] $\mathbb{T}$.

$$\mathbb{K} = \begin{bmatrix} f_x & c & o_x \\ 0 & f_y & o_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{1}$$

gdzie: f_x, f_y – długości ogniskowych, o_x, o_y – współrzędne punktu głównego, c – współczynnik skośności.

Wspomniane wyżej parametry można wykorzystać do obliczenia macierzy projekcji [4×4] $\mathbb{P}$ dla każdej kamery, za pomocą równania (2):

$$\mathbb{P} = \begin{bmatrix} \mathbb{K} & 0 \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} \mathbb{R} & \mathbb{T} \\ 0^T & 1 \end{bmatrix}. \tag{2}$$

Następnie, położenia odpowiadających sobie punktów w widokach A i B wyznacza się za pomocą równania (3).

$$\begin{bmatrix} z_B \cdot x_B \\ z_B \cdot y_B \\ z_B \\ 1 \end{bmatrix} = \mathbb{P}_B \cdot \mathbb{P}_A^{-1} \begin{bmatrix} z_A \cdot x_A \\ z_A \cdot y_A \\ z_A \\ 1 \end{bmatrix}, \tag{3}$$

gdzie:
- (x_A, y_A), (x_B, y_B) - współrzędne odpowiadających sobie punktów w widokach A i B,
- z_A, z_B – wartości głębi odpowiadających sobie punktów obrazu w widokach A i B,
- $\mathbb{P}_A$, $\mathbb{P}_B$ – macierze projekcji widoków A i B.

Autor zakłada również, że kamery znajdują się w przybliżeniu na okręgu wokół sceny. W przeciwnym razie rektyfikacja mogłaby wprowadzić nieakceptowalne zniekształcenia sekwencji wizyjnych. Systemy wielokamerowe o takim rozmieszczeniu kamer zostały już przygotowane i wykorzystane do nagrywania sekwencji testowych.

3.2. Wyznaczenie parametrów okręgu oraz skorygowanych pozycji kamer

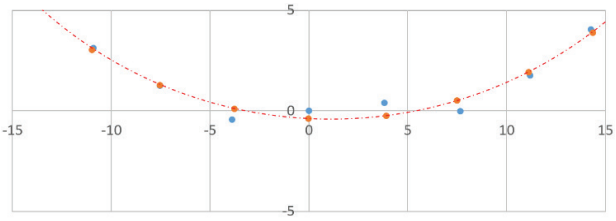
Pierwszym etapem rektyfikacji kołowej jest znalezienie parametrów okręgu na podstawie oryginalnych pozycji kamer. Do wyznaczenia tych parametrów wykorzystano równanie okręgu oraz metodę regresji nieliniowej, zgodnie z równaniami (4) i (5). Minimalizując błąd S wyznacza się pozycję środka okręgu $(x_{cen}, 0, z_{cen})$ oraz jego promień r.

$$(x_i - x_{cen})^2 + (z_i - z_{cen})^2 = r^2 \tag{4}$$

$$S = \sum_{i=1}^{N} (\sqrt{(x_i - x_{cen})^2 + (z_i - z_{cen})^2} - r)^2 \tag{5}$$

W proponowanej metodzie współrzędne wertykalne zostają zignorowane ($y_{cen} = 0$, $y_i = 0$), ponieważ przyjęto, że po rektyfikacji wszystkie kamery są umieszczone na tej samej wysokości.

Po wyznaczeniu parametrów okręgu, kolejnym krokiem jest znalezienie dla każdej kamery jej zmodyfikowanej pozycji $(x_i', 0, z_i')$ na okręgu, najbliższej oryginalnej lokalizacji. Rysunek 3 przedstawia zarówno oryginalne, jak i nowe pozycje kamer dla jednej z testowych sekwencji.



Rys. 3. Oryginalne (niebieski) oraz skorygowane (pomarańczowy) pozycje kamer dla sekwencji testowej Breakdancers, rzut z góry

3.3. Macierze obrotu kamer idealnie rozmieszczonych na okręgu

Jak przedstawiono w rozdziale 2, proponowana metoda rektyfikacji kołowej zakłada nie tylko korekcję położenia kamer do idealnego okręgu, ale także precyzyjne skierowanie ich osi optycznych do punktu środkowego tego okręgu, który wyznaczono w poprzednim podrozdziale. Aby to osiągnąć, konieczna jest modyfikacja macierzy obrotu.

Macierz obrotu $\mathbb{R}$ reprezentuje łączny obrót kamery wokół trzech ortogonalnych osi. W rozpatrywanym przypadku kamer rozmieszczonych idealnie na okręgu zakłada się, że osie optyczne znajdują się na wspólnej płaszczyźnie równoległej do ziemi. Taki obrót kamery można przedstawić za pomocą następującej macierzy (6):

$$\mathbb{R}_i' = \begin{bmatrix} \cos\alpha_i & 0 & \sin\alpha_i \\ 0 & 1 & 0 \\ -\sin\alpha_i & 0 & \cos\alpha_i \end{bmatrix}, \quad (6)$$

gdzie α_i oznacza kąt pomiędzy pozycją *i*-tej kamery a wybranym promieniem okręgu, stąd:

$$\cos\alpha_i = \frac{z_i' - z_{cen}}{r}, \qquad \sin\alpha_i = \frac{x_i' - x_{cen}}{r}. \quad (7)$$

Do wyznaczenia macierzy obrotu każdej z kamer idealnie rozmieszczonych na okręgu wystarczą zatem jedynie parametry okręgu (położenie środka okręgu i jego promień) oraz pozycja kamery na okręgu.

3.4. Parametry wewnętrzne kamer idealnie rozmieszczonych na okręgu

W rozdziałach 3.2 oraz 3.3 przedstawiono proces wyznaczania parametrów zewnętrznych kamer umieszczonych na idealnym okręgu. W tym podrozdziale autor przedstawia metodykę wyznaczania parametrów wewnętrznych.

W pierwszej kolejności przyjęto, że współczynnik skośności wszystkich kamer po rektyfikacji powinien wynosić *c=0*, podobnie jak ma to miejsce podczas rektyfikacji liniowej przed kodowaniem 3D-HEVC. Następnie, ogniskowe f_y oraz składowe o_y punktów głównych są uśredniane i dla każdej kamery przypisywana jest ich wartość średnia. Dla porównania, w 3D-HEVC wartości f_y i o_y w ogóle nie są używane [6].

Ostatnim parametrem wewnętrznym jest składowa o_x punktów głównych. Dla każdej kamery jest ona wyznaczana niezależnie na podstawie składowej o_x kamery źródłowej oraz macierzy obrotu przed i po korekcji. Ma to na celu kompensację przesunięcia pola widzenia kamery wynikającego ze zmiany macierzy obrotu.

4. MODYFIKACJA PREDYKCJI MIĘDZYWIDOKOWEJ 3D-HEVC

Jak wcześniej wspomniano, predykcja międzywidokowa w 3D-HEVC jest znacznie uproszczona, ponieważ zakłada liniową rektyfikację sekwencji. Z kolei predykcja oparta na pełnym rzutowaniu nie zakłada żadnego specyficznego rozmieszczenia kamer, ale wymaga skomplikowanych operacji na macierzach i jest znacznie wolniejsza. Proponowane podejście wykorzystujące rektyfikację kołową stanowi kompromis pomiędzy istniejącymi rozwiązaniami. Z jednej strony, kamery mogą być umieszczone na okręgu, co prowadzi do lepszego pokrycia rejestrowanej sceny. Z drugiej strony, zakładając że sekwencje są zrektyfikowane kołowo, możliwe jest dostosowanie predykcji międzywidokowej do takiego rozmieszczenia kamer i przez to redukcji złożoności obliczeniowej, przy zachowaniu wysokiej efektywności kompresji. Po zastosowaniu parametrów kamer po rektyfikacji, wyznaczonych w rozdziale 3, do równania (3), opracowano uproszczone wzory rzutowania punktów obrazu pomiędzy widokami *A* i *B*:

$$z_B = (x_A - o_{xA})\frac{z_A}{f_x}\sin\Delta\alpha + (z_A - r)\cos\Delta\alpha + r, \quad (8)$$

$$y_B = o_y + \frac{z_A}{z_B}(y_A - o_y), \quad (9)$$

$$x_B = o_{xB} + \frac{1}{z_B}[(x_A - o_{xA})z_A\cos\Delta\alpha - (z_A - r)f_x\sin\Delta\alpha] \quad (10)$$

gdzie $\Delta\alpha = \alpha_B - \alpha_A$.

Powyższe wzory umożliwiają predykcję położenia punktu w widoku *B* na podstawie położenia odpowiadającego mu punktu w widoku *A*, przy założeniu rektyfikacji kołowej obu widoków. Autor dokonał szeregu modyfikacji w oprogramowaniu modelowym kodeka 3D-HEVC, których celem była implementacja powyższych wzorów jako podstawy predykcji międzywidokowej. Należy zaznaczyć, że wektor różnicy współrzędnych odpowiadających sobie punktów w dwóch widokach ma dwie składowe ($x_B - x_A$, $y_B - y_A$), w przeciwieństwie do jednej składowej w 3D-HEVC, co wymusza zmiany również innych narzędzi kodowania wykorzystujących predykcję międzywidokową, m.in. Międzywidokowej Predykcji Ruchu (ang. *Inter-view Motion Prediction*).

Kolejną zmianą w stosunku do 3D-HEVC jest konieczność transmisji dodatkowych parametrów kamer, tj. uśrednionej składowej o_y punktów głównych, promienia okręgu *r* oraz kąta *α* dla każdej kamery (w zastępstwie pozycji kamery na linii w 3D-HEVC). Liczba wymaganych parametrów jest zatem nieznacznie większa niż w 3D-HEVC i znacząco mniejsza w porównaniu z kodekiem wspierającym dowolne rozmieszczenie kamer.

5. OCENA EKSPERYMENTALNA

Celem przeprowadzonych eksperymentów była ocena efektywności kompresji oraz czasu kodowania 3D-HEVC dostosowanego do kamer rozmieszczonych na okręgu. Rezultaty porównano z wynikami dla oryginalnego kodera 3D-HEVC oraz kodera 3D-HEVC dostosowanego do dowolnego rozmieszczenia kamer. Dodatkowo wyodrębniono i porównano czasy predykcji międzywidokowej obu zmodyfikowanych koderów. Porównanie efektywności kompresji wykonano poprzez pomiar uśrednionej zmiany przepływności bitowej dla składowej luminancji, wykorzystując metrykę Bjøntegaarda. Eksperymenty przeprowadzono dla 7 widoków z 4 powszechnie stosowanych wielowidokowych sekwencji testowych. Kodowanie zostało wykonane na 100 ramkach oraz dla 4 wartości parametru kwantyzacji QP={25, 30, 35, 40}. Sekwencje testowe zostały uprzednio zrektyfikowane kołowo. Konfiguracje koderów były identyczne i zgodne z zaleceniami zawartymi w CTC (ang. *Common Test*

Conditions), z wyjątkiem parametrów kamer, które dostosowano do wymagań każdego z koderów. Wszystkie kodery bazują na publicznie dostępnym modelu testowym 3D-HEVC, oznaczonym jako HTM-13.0 [3].

Tab. 1. Procentowa zmiana prędkości bitowej (ΔBD) oraz czasu kodowania (Δt) proponowanego kodera, w odniesieniu do oryginalnego 3D-HEVC oraz 3D-HEVC dostosowanego do dowolnego rozmieszczenia kamer

Sekwencja	*vs 3D-HEVC*		*vs 3D-HEVC dla dowolnego rozmieszczenia kamer*	
	ΔBD	*Δt*	*ΔBD*	*Δt*
Ballet	-8.14	-1.89	-0.23	-10.21
Breakdancers	-8.22	-8.50	0.82	-7.46
BBBFlowers	-3.15	-1.78	-0.54	-7.66
PoznanBlocks	-4.61	-4.50	-0.23	-9.89
Średnia	**-6.03**	**-4.17**	**-0.05**	**-8.81**

Tab. 2. Procentowa zmiana czasu predykcji międzywidokowej (Δt) proponowanego kodera, w odniesieniu do 3D-HEVC dostosowanego do dowolnego rozmieszczenia kamer

Sekwencja	*Δt*
Ballet	-97.96%
Breakdancers	-97.79%
BBBFlowers	-97.77%
PoznanBlocks	-97.43%
Średnia	**-97.74%**

Tabela 1 pokazuje zmianę przepływności bitowej oraz czasu kodowania dla 3D-HEVC dostosowanego do kompresji wizji zrektyfikowanej kołowo, w porównaniu z niezmodyfikowanym 3D-HEVC i zmodyfikowanym 3D-HEVC dla dowolnego rozmieszczenia kamer. Autorska propozycja zmniejsza przepływność bitową średnio o 6% w porównaniu z oryginalną techniką 3D-HEVC, co wynika z niedokładności predykcji międzywidokowej w 3D-HEVC dla układów kamer inne niż liniowe. W porównaniu z koderem wspierającym dowolne rozmieszczenie kamer, rozwiązanie zoptymalizowane dla okręgu daje nieco lepsze wyniki. Różnica wynika z mniejszej liczby parametrów kamer przesyłanych w strumieniu bitowym oraz prostszej predykcji międzywidokowej, co skutkuje m.in. zmniejszoną liczbą błędów zaokrągleń.

Porównując czas kompresji można zauważyć, że proponowany koder jest do 10% szybszy niż koder dedykowany dla dowolnego rozmieszczenia kamer, a sama predykcja międzywidokowa jest aż 44 razy szybsza (Tabela 2). Ten wynik osiągnięto dzięki znacznie prostszym równaniom projekcji, zoptymalizowanym dla sekwencji zrektyfikowanych kołowo. Co zaskakujące, proponowany koder okazał się również o około 4% szybszy niż oryginalny 3D-HEVC, mimo że predykcja międzywidokowa tego pierwszego jest bardziej złożona. Wynika to z faktu, że dokładniejsza predykcja międzywidokowa pozwala na szybsze zlokalizowanie odpowiadających sobie obszarów w dwóch widokach.

6. PODSUMOWANIE

W niniejszej pracy zaproponowano oryginalne metody rektyfikacji i predykcji międzywidokowej, dedykowane dla sekwencji rejestrowanych wieloma kamerami i rozmieszczonymi w przybliżeniu na okręgu. Autor opracował proces korekcji parametrów kamer oraz rektyfikacji rejestrowanych przez nie widoków. Dzięki rektyfikacji kołowej możliwe było dostosowanie kodeka 3D-HEVC do efektywnej kompresji tego typu sekwencji. Autor wyprowadził uproszczone równania rzutowania punktów obrazu pomiędzy widokami i zaimplementował je w oprogramowaniu modelowym 3D-HEVC. W rezultacie otrzymano nowy kodek ze zmodyfikowaną predykcją międzywidokową i narzędziami kodowania, które ją wykorzystują.

Eksperymentalna ocena wykazała, że zmodyfikowany przez autora kodek 3D-HEVC poprawia efektywność kompresji sekwencji kołowych średnio o 6%, oraz szybkość kodowania średnio o 4%, w porównaniu ze standardowym kodekiem 3D-HEVC. Zaproponowany kodek okazał się również znacznie szybszy od 3D-HEVC dostosowanego do dowolnego rozmieszczenia kamer, oraz nieco bardziej efektywny ze względu na mniejszą liczbę transmitowanych parametrów kamer.

Podsumowując, zaproponowana metoda jest atrakcyjną alternatywą dla pozostałych dwóch rozwiązań, ponieważ umożliwia efektywną kompresję wizji rejestrowanej przez kamery rozmieszczone dookoła sceny, a przy tym nie wydłuża czasu kodowania tego typu sekwencji.

LITERATURA

[1] Boyce Jill, Renaud Doré, Adrian Dziembowski, Julien Fleureau, Joel Jung, Bart Kroon, Basel Salahieh, Vinod Kumar Malamal Vadakital, Lu Yu. 2021. „MPEG Immersive Video Coding Standard". *Proceedings of the IEEE* 109, no. 9 (2021): 1521–1536.

[2] Ho Yo-Sung. 2012. „Video Processing Techniques for 3D Television". *2012 Visual Communications and Image Processing*, 1–13.

[3] Model testowy kodeka 3D-HEVC. Dostęp 10.06.2024 r. https://hevc.hhi.fraunhofer.de/svn/svn_3DVCSoftware/tags/HTM-13.0

[4] Samelak Jarosław, Jakub Stankowski, Marek Domański. 2016. „Adaptation of the 3D-HEVC coding tools to arbitrary locations of cameras", *International Conference on Signals and Electronic Systems (ICSES)*, 107–112.

[5] Stankowski Jakub, Łukasz Kowalski, Jarosław Samelak, Marek Domański, Tomasz Grajek, Krzysztof Wegner. 2015. „3D-HEVC extension for circular camera arrangements". *2015 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, 1–4.

[6] Tech Gerhard, Ying Chen, Karsten Müller, Jens-Rainer Ohm, Anthony Vetro, Ye-Kui Wang. 2016. „Overview of the Multiview and 3D Extensions of High Efficiency Video Coding". *IEEE Transactions on Circuits and Systems for Video Technology* 26, no. 1 (2016): 35–49.

Wykrywanie trudnych do kompresji sekwencji wideo na podstawie obiektywnych metryk jakości

Detection of Hard-To-Compress Video Sequences Based on Objective Quality Metrics

Filip Korus[1]; Dawid Juszka[2]; Michał Grega[3]; Mikołaj Leszczuk[4]; Michał Suchoń[5]

[1] AGH Akademia Górniczo-Hutnicza, Kraków,
fkorus@student.agh.edu.pl
[2] AGH Akademia Górniczo-Hutnicza, Kraków,
dawid.juszka@agh.edu.pl
[3] AGH Akademia Górniczo-Hutnicza, Kraków,
michal.grega@agh.edu.pl
[4] AGH Akademia Górniczo-Hutnicza, Kraków,
mikolaj.leszczuk@agh.edu.pl
[5] tvhive sp. z o.o., Kraków,
msuchon@quickstream.tech

DOI: 10.15199/59.2024.4.89

Streszczenie: Serwisy strumieniowania wideo odpowiadają za 82% całego ruchu IP (dane na rok 2022 według Cisco). Netflix, TikTok, YouTube czy Twitch, zyskują popularność, przyciągając zarówno uwagę fanów filmów i seriali, jak i miłośników gier komputerowych. Jednak, kompresja materiałów wideo generowanych podczas wydarzeń e-sportowych staje się wyzwaniem, nawet dla zaawansowanych kodeków. Badania oparte na metrykach jakości wideo ujawniają te trudności, co pozwoliło opracować model uczenia maszynowego do oceny trudności kompresji.

Abstract: Video streaming services account for 82% of all IP traffic (as of 2022, according to Cisco). Netflix, TikTok, YouTube and Twitch are gaining popularity, attracting the attention of both fans of films and series, as well as computer game enthusiasts. However, compressing video content generated during e-sports events becomes a challenge, even for advanced video codecs. Research based on video quality metrics reveals these difficulties. As a result, a machine learning model was developed to assess compression difficulty.

Słowa kluczowe: gry komputerowe, metryki jakości wideo, model uczenia maszynowego, QoE

Keywords: gameplay, machine learning model, QoE, video quality metrics

1. WSTĘP

Według Cisco, pomiędzy 2017 a 2022 rokiem, ilość przenoszonego ruchu IP zwiększyła się trzykrotnie, z tego 82% przypada na strumieniowanie treści wideo [2]. Platformy takie jak Netflix, YouTube czy Twitch diametralnie zmieniły sposób konsumpcji materiałów wideo, zapewniając dostęp do ogromnej różnorodności treści na żądanie. Nową kategorią treści, która istotnie zwiększa swój udział w ruchu IP jest strumieniowanie tzw. gameplay, czyli rozgrywek użytkowników gier komputerowych oraz strumieniowanie wydarzeń e-sportowych.

Podczas widowisk e-sportowych profesjonalni gracze rywalizują między sobą, a ich umiejętności w prowadzeniu rozgrywki oraz determinacja przyciągają miliony widzów na całym świecie. Wielkie imprezy e-sportowe stały się równie ważnymi wydarzeniami, co tradycyjne mecze sportowe, gromadząc fanów zarówno w czasie transmisji na żywo, jak i tych, którzy oglądają retransmisje.

Jednakże, pomimo ogromnej popularności i zapotrzebowania na treści wideo związane z e-sportem, aktualne pozostaje wyzwanie techniczne związane z ich transmisją na żywo. Materiały wideo z gier komputerowych są szczególnie trudne do skompresowania [10], co może prowadzić do utraty jakości lub opóźnień podczas transmisji strumieniowej. Nawet zaawansowane kodeki kompresji, takie jak H.265 i H.264, w niektórych przypadkach nie przetwarzają tych materiałów na tyle dobrze, aby osiągnęły one wysoką jakość. Tę anomalię można wykryć dzięki metrykom jakości wideo, takim jak Video Multimethod Assessment Fusion (VMAF), które zostały użyte do przeprowadzenia badań nad skutecznością kompresji wideo treści rozrywkowych.

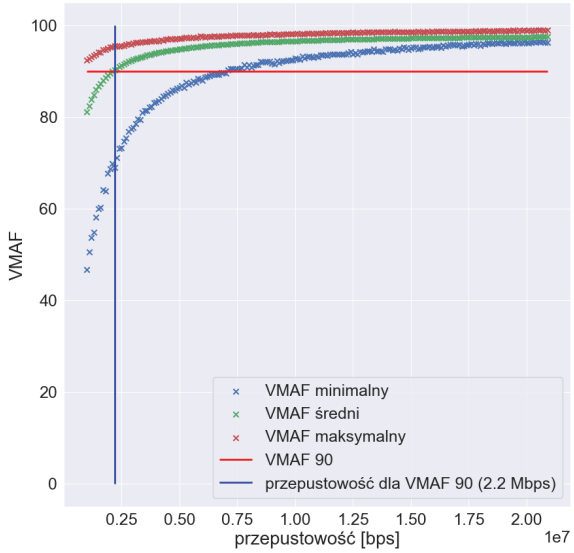
W tym celu został opracowany model uczenia maszynowego, który analizując metryki ocenia, czy dane ujęcie jest trudne do skompresowania. Dokładność poprawnej klasyfikacji przez model wynosi 0.92.

Zaproponowane rozwiązanie wpisuje się w zakres prac prowadzonych przez grupę No Reference Metrics (NORM) - Video Quality Experts Group (VQEG), której celem jest połączenie wielu wskaźników No Reference (NR) w jeden na tyle dokładny model, aby można było go uwzględnić w zaleceniu ITU.

2. EKSPERYMENT

Eksperyment polega na skompresowaniu kodekiem H.264 wszystkich filmów z bazy 210 razy z różnymi przepływnościami - od 100 kbps do 21000 kbps (21 Mbps) z krokiem co 100 kbps. Następnie na każdym wariancie przepływ-

nościowym skompresowanego ujęcia liczony jest minimalny, maksymalny oraz średni poziom VMAF. Efektem tej operacji jest wykreślenie krzywych zależności VMAF od zadanej przepustowości dla poszczególnych filmów i znalezienie przepływności, dla której wartość średniego VMAF wynosi 90, co zostało zobrazowane na Rysunku 1.



Rysunek 1: VMAF w funkcji przepustowości dla filmu GTAV_plane1.

2.1 Baza ujęć

Do przeprowadzenia eksperymentu została utworzona baza sekwencji wideo, czyli krótkich filmów zawierających urywki Netflixowych seriali, nagrania z popularnych gier, a także sekwencje udostępnione przez organizacje CableLabs oraz NERC-DTV. Przykładowe klatki z filmów znajdują się na Rysunku 2.

Sekwencje zawierają zarówno ujęcia statyczne, na których nieruchoma kamera jest skierowana w stronę nagrywanych obiektów, jak i dynamiczne, gdzie kamera się porusza. Na ujęciach znajdują się różne ilości ruchomych obiektów oraz szczegółów, które potencjalnie będą wymagały zastosowania wyższej przepustowości przy kompresji, aby utrzymać jakość zadowalającą użytkownika.

Rysunek 2: Przykładowe klatki filmów z bazy.

2.2 Metryki jakości

Jako obiektywną metrykę oceny jakości, na podstawie której można stwierdzić czy dane ujęcie zostało skompresowane zachowując pożądaną jakość, wykorzystano Video Multimethod Assessment Fusion (VMAF). Metryka ta jest powszechnie stosowana przy optymalizacji przepustowości transmisji wideo ze względu na zaimplementowaną korelację z ocenami jakości postrzeganej przez człowieka [3] [8]. Wartości VMAF zawierają się w przedziale 0-100, natomiast poziomem powszechnie uważanym za subiektywny brak straty jakości sekwencji jest wartość 90 [7].

Do oceny złożoności sekwencji w kontekście kodowania zostały użyte obiektywne metryki: Spatial Information (SI), Temporal Information (TI) oraz criticality (o) [4].

SI mierzy jak dużo szczegółów przestrzennych znajduje się w danym ujęciu. Jest obliczane w kilku krokach; najpierw piksele danej klatki (F_n) są przepuszczane przez filtr Sobela ($Sobel()$), następnie liczone jest odchylenie standardowe po tych pikselach. Z powstałej listy wartości znajdowana jest maksymalna wartość - SI: [9]

$$SI = max_{time}\{std_{space}[Sobel(F_n)]\} \tag{1}$$

Zatem im więcej szczegółów znajduje się na klatce, tym SI będzie miało większą wartość.

TI mierzy jak dużo ruchu znajduje się w danym filmie. Jest zdefiniowane jako maksymalne odchylenie standardowe różnic wartości pikseli następujących po sobie klatek: [9]

$$TI = max_{time}\{std_{space}[F_n - F_{n-1}]\} \tag{2}$$

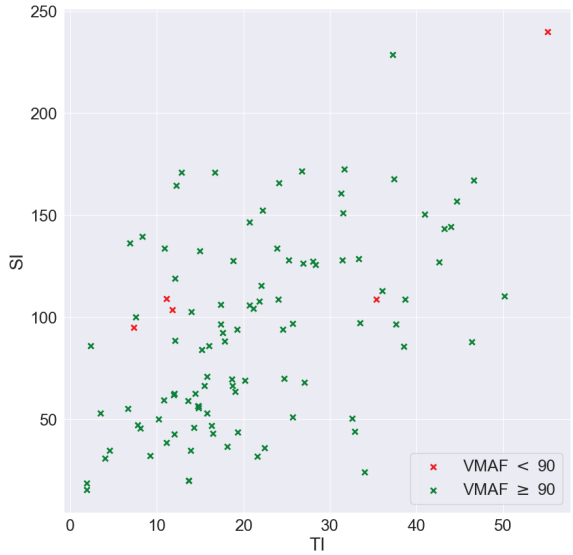
Zatem im bardziej sąsiadujące ze sobą klatki różnią się od siebie, tym większa będzie wartość TI.

Criticality jest definiowane jako logarytm dziesiętny ze średniej w czasie wartości iloczynu SI oraz TI obliczanych dla danej klatki (F_n): [6]

$$o = log_{10}\{mean_{time}[SI(F_n) * TI(F_n)]\} \tag{3}$$

Określa jak bardzo dana scena jest wymagająca podczas kodowania. [6]

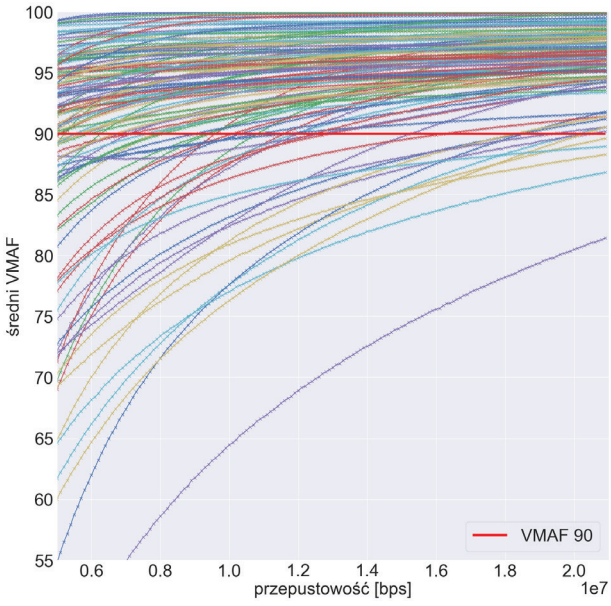
Filmy trudne do kompresji mają wartość SI już w okolicach 100, a TI około 10. Jednak istnieją ujęcia posiadające SI większe niż 150, dla których kompresja nie wymaga użycia wysokiej przepływności w celu zachowania jakości. Zobrazowane jest to na Rysunku 3.



Rysunek 3: SI w funkcji TI dla sekwencji z bazy z podziałem na grupy filmów, które: osiągnęły VMAF 90, nie osiągnęły VMAF 90.

2.3 Wyszukiwanie sekwencji trudnych do kompresji

Po nałożeniu na jeden wykres krzywych średnich poziomów VMAF dla całej bazy filmów można zauważyć, że zgodnie z oczekiwaniami, różne rodzaje treści uzykują znacząco inne wyniki VMAF dla tych samych wartości przepływności. Niektóre ujęcia osiągają VMAF na poziomie 90 dla niskich przepływności, inne dopiero dla zdecydowanie wyższych przepływności. Jednak kilka sekwencji nigdy nie osiąga tego pułapu – jest to zilustrowane na Rysunku 4.



Rysunek 4: Średnie wartości VMAF w funkcji przepustowości dla filmów z bazy. Jedna krzywa odpowiada jednej sekwencji.

Sekwencje, dla których metryka VMAF nigdy nie osiąga poziomu 90 mają wspólny mianownik. Są to przede wszystkim nagrania tzw. gameplay, czyli zarejestrowanych rozgrywek gier komputerowych. Występuje w nich duża ilość szczegółów, ogień, iskry, deszcz, promienie słoneczne, poruszająca się tafla wody lub szybko przemieszczająca się kamera itp. Charakterystyczną cechą gameplayów jest także przesunięcie kamery (ang. pan) związane z perspektywą "z oczu gracza" lub z umieszczeniem kamery kontrolowanej przez gracza nad polem gry. Sceny zawierające tego typu elementy charakteryzują się wyższymi (w porównaniu do scen osiągających VMAF większą lub równą 90) wartościami metryk SI, TI oraz criticality.

Według artykułu [4]:

- trudne do kodowania sekwencje mają wartość $criticality \geq 3.5$,
- łatwe do kodowania sekwencje mają wartość $criticality \leq 2.5$,
- sekwencja zawierająca dużo detali przestrzennych ma wartość $SI \geq 200$,
- sekwencja zawierająca dużo ruchu ma wartość $TI \geq 60$.

Są to zakresy wartości zdefiniowane przed ogłoszeniem metryki VMAF. Niestety, te zakresy nie są odzwierciedlone w przypadku eksperymentów z sekwencjami z przedmiotowej bazy, które nie osiągnęły wartości VMAF równej 90. Pokazuje to Tabela 1. Tylko jedna z tych sekwencji posiada wartość dla wskazania criticality na poziomie określonym w artykule [4]. Określony w [4] próg TI dla żadnej z tych sekwencji nie został przekroczony, a zalecana wartość SI została przekroczona tylko dla jednego filmu. Wynika z tego, że zalecane w artykule [4] progi omawianych metryk są zbyt wygórowane i należy postulować by zostały zrewidowane i ustalone na znacznie niższym poziomie.

Tabela 1: Parametry TI, SI, criticality dla filmów, które nie osiągnęły VMAF 90.

Nazwa filmu	*TI*	*SI*	*o*
clabs_Wipe_forest	7.36	95.10	2.61
Fallout4_venery	55.13	240.00	3.70
Fountains	11.10	109.02	3.06
RUST_settlement	35.34	108.72	3.03
WITCHER3_wolves2	11.79	103.56	2.92

2.4 Klasyfikator trudności kodowania scen

Mając na uwadze omówione wcześniej zachowanie metryk w zależności od zawartości ujęcia, został stworzony model uczenia maszynowego służący do binarnej klasyfikacji filmów ze względu na trudność kodowania.

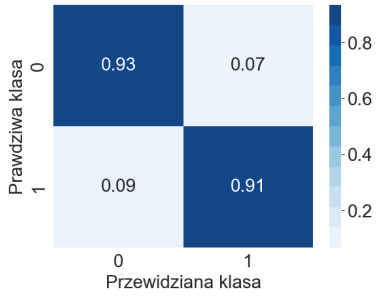
Jako klasyfikator wykorzystany został Random Forest Classifier z biblioteki $scikit-learn$. Las losowy charakteryzuje się losowym wyborem podzbiorów danych treningowych. Następnie na każdym podzbiorze budowane są osobne drzewa decyzyjne. Podczas tego procesu losowo wybiera się podzbiór cech do rozważenia podczas poszukiwania najlepszego podziału w każdym węźle drzewa. Ma to na celu uniknięcie zbyt dużego dopasowania do konkretnych cech i zmniejszenie ryzyka nadmiernego dopasowania modelu do danych treningowych. Dzięki tym cechom każde drzewo w lesie jest inne, co prowadzi do większej różnorodności i generalizacji modelu. Po zbudowaniu wszystkich drzew decyzyjnych, binarne przyporządkowanie próbki do danej klasy polega na przeprowadzeniu osobnej klasyfikacji przez każde drzewo w lesie, a następnie na przyporządkowaniu klasy na podstawie głosowania większościowego.

Cechami wejściowymi modelu są metryki SI, TI, o oraz średnie wartości SI, TI w czasie. Natomiast klasą wyjściową modelu jest binarny wskaźnik mówiący o tym, czy próg VMAF 90 został osiągnięty.

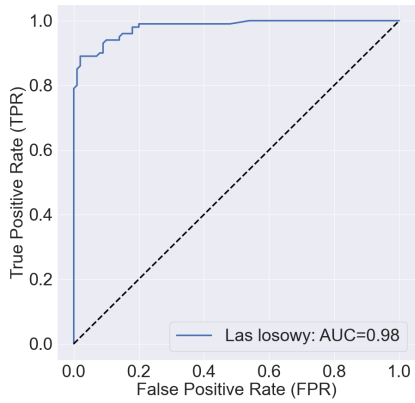
Modele uczenia maszynowego osiągają najlepsze wyniki, gdy zbiory danych uczących model są zbalansowane, czyli gdy ilość próbek w obu klasach (w przypadku klasyfikacji binarnej) jest identyczna lub zbliżona. W przypadku tego modelu stanowi to wyzwanie, ponieważ ilość próbek w klasie ujęć, które osiągnęły VMAF 90 jest bardzo mała. Ilość danych w najliczniejszej klasie jest niemal 20 razy większa niż w klasie najmniej licznej. Gdyby wytrenować model na takich danych zachowywałby się on bardzo stronniczo i częściej przewidywał liczniejszą klasę.

W celu rozwiązania tego problemu została wykorzystana technika nadpróbkowania danych, a właściwie Synthetic Minority Oversampling Technique (SMOTE). SMOTE tworzy syntetyczne dane treningowe dla mniejszościowej klasy metodą interpolacji liniowej. Po wykonaniu nadpróbkowania obie klasy stały się jednakowo liczne i nadają się do użycia ich w procesie trenowania.

W celu ewaluacji modelu została użyta walidacja krzyżowa. Polega ona na dzieleniu zbioru danych na k równych części, trenując i testując model k razy na różnych podzbiorach. Poprzez iteracyjne wykorzystanie danych treningowych i testowych, pozwala na dokładniejszą ocenę modelu pod kątem przyzwyczajenia się go do danych uczących. Po uśrednieniu wyników z każdej iteracji dokładność modelu wynosi 0.92 (średnia z wartości na przekątnej macierzy pomyłek - Rysunek 5), natomiast powierzchnia pod krzywą ROC 0.98 (Rysunek 6).

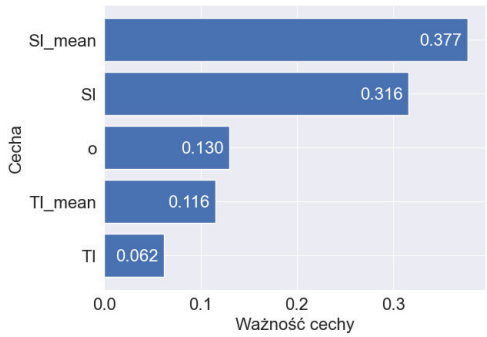


Rysunek 5: Znormalizowana macierz pomyłek.



Rysunek 6: Krzywa ROC modelu.

Rysunek 7 pokazuje, które zmienne wejściowe są najbardziej istotne dla predykcji modelu. Jak widać, największy wpływ na klasyfikację mają parametry SI oraz średnia z SI dla całego filmu. Oznacza to, że filmy zawierające dużą ilość detali przestrzennych, z większym prawdopodobieństwem będą klasyfikowane jako trudne do kompresji.



Rysunek 7: Ważność cech modelu.

3. PODSUMOWANIE

Stworzony model pomaga klasyfikować filmy w kontekście trudności kodowania stratnego. Może on zostać wykorzystany przy doborze odpowiednich sekwencji do bazy filmów.

W przyszłości można by rozwinąć ten pomysł i stworzyć narzędzie, które wskazywałoby ujęcia (fragmenty filmu) wymagające użycia wyższej przepustowości przy kodowaniu. Na podstawie takich danych można by czasowo zmieniać parametry kodeka w celu utrzymania stałej jakości.

4. PODZIĘKOWANIA

Prezentowane wyniki badań pochodzą z projektu badawczo-naukowego pt. „Technologia do bezpiecznego i niezawodnego dostarczania profesjonalnych przekazów kontrybucyjnych audio/wideo na żywo przy zachowaniu minimalnego możliwego opóźnienia.” w ramach Poddziałania 1.1.1 Programu Operacyjnego Inteligentny Rozwój 2014-2020 współfinansowanego ze środków Narodowego Centrum Badań i Rozwoju.

LITERATURA

[1] Dey Ishani, Pratap Vibha. 2023. „A Comparative Study of SMOTE, Borderline-SMOTE, and ADASYN Oversampling Techniques using Different Classifiers”. 2023 3rd International Conference on Smart Data Intelligence (ICSMDI): 294-302.

[2] Cisco. 2020. „Cisco Annual Internet Report (2018–2023)”. White Paper.

[3] Mercat Alexandre, Mäkinen Arttu, Sainio Joose, Lemmetti Ari, Viitanen Marko, Vanne Jarno. 2021. „Comparative Rate-Distortion-Complexity Analysis of VVC and HEVC Video Codecs”. IEEE Access 9: 67813-67828.

[4] Pinson Margaret, Boyd Karen Sue, Hooker Jessica, Muntean Kristina. 2013. „HOW TO CHOOSE VIDEO SEQUENCES FOR VIDEO QUALITY ASSESSMENT”. Proceedings of the Seventh International Workshop on Video Processing and Quality Metrics for Consumer Electronics.

[5] Robitza Werner, Rao Ramachandra Rao Rakesh, Göring Steve, Raake Alexander. 2021. „Impact of Spatial and Temporal Information on Video Quality and Compressibility”.

[6] Fenimore, Charles, John Libert, and Stephen Wolf. 1998. „Perceptual Effects of Noise in Digital Video Compression.” 140th SMPTE Technical Conference.

[7] Uddin Syed, Leszczuk Mikołaj, Grega Michał. 2022. „Preliminary Study on Video Codec Optimization Using VMAF”. Intelligent Information and Database Systems: 469–480.

[8] Li Zhi, Aaron Anne, Katsavounidis Ioannis, Moorthy Anush, Manohara Megha. 2016. „Toward a practical perceptual video quality metric”. The Netflix Tech Blog 6 (2): 2.

[9] ITU-T Recommendation P.910. 1994. „Two criteria for video test scene selection”.

[10] Barman Nabajeet, Martini Maria. 2022. „User Generated HDR Gaming Video Streaming: Dataset, Codec Comparison, and Challenges”. IEEE Transactions on Circuits and Systems for Video Technology 32 (3): 1236-1249.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

Automatyczne generowanie środowisk testowych w systemie zarządzania zmianą w sieciach

Automatic generation of test environments in the network change management system

inż. Wiktor Tekiela[1]; dr hab. inż. Jerzy Domżał[2];

[1] Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, Kraków, wiktort@student.agh.edu.pl
[2] Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, Kraków, jerzy.domzal@agh.edu.pl

DOI: 10.15199/59.2024.4.90

Streszczenie: Utrzymanie wysokiej dostępności sieci jest wyzwaniem dla wielu dostawców usług telekomunikacyjnych oraz chmurowych. System proponowany w artykule pozwala uporządkować proces rekonfiguracji oraz całkowicie wyeliminować przerwy techniczne związane z rekonfiguracją infrastruktury.

Abstract: Maintaining high network availability is a challenge for many telecommunications and cloud service providers. The system proposed in the article allows to organize the reconfiguration process and completely eliminate technical interruptions related to infrastructure reconfiguration.

Słowa kluczowe: Automatyzacja infrastruktury, DevOps, Środowiska testowe, Zarządzanie zmianami.

Keywords: Infrastructure automation, DevOps, Test sandboxes, Change management.

1. WSTĘP

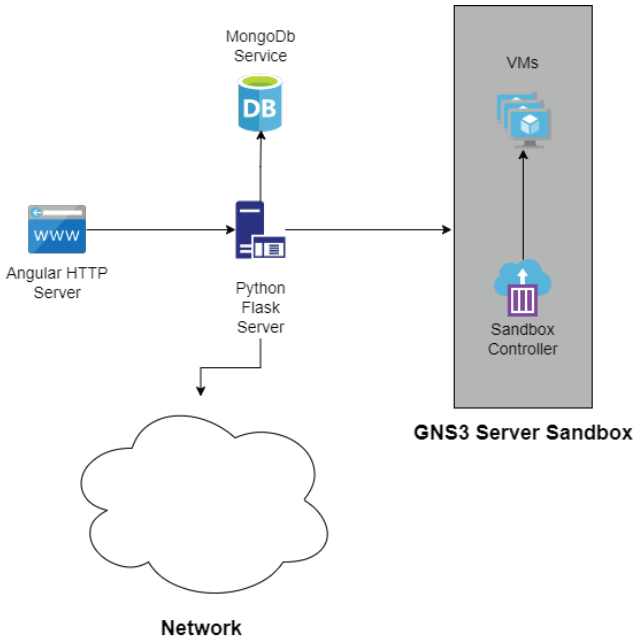
Proces zarządzania zmianą w sieciach ma znaczący wpływ na utrzymanie w nich wysokiej dostępności. Odpowiednie uporządkowanie wszystkich kroków zmniejsza prawdopodobieństwo wystąpienia błędu, zmniejsza czas potrzebny na przeprowadzenie całego procesu oraz niweluje lub całkowicie eliminuje czas niedostępności sieci. Systemy "five nines", czyli o dostępności 99.999%, pozwalają tylko na 5 minut 26 sekund czasu niedostępności w przeciągu roku. Przeprowadzenie testów jednostkowych na urządzeniach warstwy drugiej oraz trzeciej, minimalizujących ryzyko wystąpienia błędu oraz dających pewność, że sieć po wprowadzeniu nowej konfiguracji będzie posiadać zadane parametry, potrzebuje czasu dostępowego do urządzenia niezależnie od tego, jaka implementacja systemu zarządzania zmianą jest używana. Podstawowe testy w małej sieci mogą zajmować nawet 140 sekund [1], co eliminuje możliwość przeprowadzania pełnego procesu rekonfiguracji w sieciach z wysoką dostępnością, co z kolei zwiększa ryzyko wystąpienia błędu. Rozwiązaniem problemu jest symulacja sieci przy użyciu dodatkowych zasobów. Jedną z możliwości, która dostarczy najbardziej wiarygodne rezultaty, jest posiadanie laboratorium z odwzorowaną produkcyjną siecią, która pozwoli na wgranie konfiguracji, przeprowadzanie testów jednostkowych, a następnie wgranie zmian w docelowej sieci bez konieczności wprowadzania przerwy technicznej. Rozwiązanie charakteryzuje się wysokim kosztem (wymagany jest duplikat wszystkich urządzeń sieciowych) oraz skrajnie nieefektywnym wykorzystaniem zasobów (urządzenia przystosowane do ciągłej pracy używane są tylko podczas wprowadzania zmian). Znacznie bardziej efektywnym rozwiązaniem jest symulowanie sieci przy użyciu oprogramowania do wirtualizacji [2]. Podejście takie pozwala na alokację zasobów tylko wtedy kiedy testy jednostkowe muszą zostać przeprowadzane a wyniki otrzymane w tym podejściu nie będą się różnić, jeśli administrator nie zdecyduje się na przeprowadzenie testów obciążeniowych [10]. Podejście to charakteryzuje się również znaczną redukcją kosztów oraz elastycznością, gdyż samo środowisko testowe może się znajdować na dowolnym serwerze w tym na przykład w infrastrukturze chmurowej na serwerze z zagnieżdżoną wirtualizacją. Pomimo wielu plusów, użycie wirtualizacji i symulowanie sieci w zwirtualizowanym środowisku nie jest prostym procesem i wymaga szerokiej wiedzy administratora [9]. Ponadto systemy operacyjne nie zawsze bazują na Linuxie, jak na przykład Cisco IOS, co może powodować trudności przy wirtualizacji takiego urządzenia. Administratorzy sieci na dzień dzisiejszy do wirtualizacji mają do dyspozycji oprogramowanie takie jak Ansible czy Chef. Programy te umożliwiają:

1. zdefiniowanie urządzeń sieciowych w plikach YAML,
2. zdefiniowanie testów oraz kolejności ich przeprowadzania,
3. wgrywanie wcześniej przygotowanej konfiguracji, również w większości przypadków w pliku YAML,

co pozwala na przejście kompletnego procesu wdrożenia zmiany. Cały proces definicji tak zwanych "cookbooków" jest czasochłonny i żaden dostawca oprogramowania nie zapewnia podłączenia do środowiska testowego, które byłoby generyczne. Aby rozwiązać ten problem, przygotowany został system automatyzujący wszystkie wymienione wcześniej kroki [1], który przeprowadza testy jednostkowe w wyizolowanym, zwirtualizowanym środowisku testo-

wym, które sam przygotowuje w zautomatyzowany sposób.

2. STRUKTURA SYSTEMU



Rysunek 1: Diagram zależności pomiędzy komponentami systemu.

System przedstawiony na rysunku 1 został rozbity na kilka logicznych komponentów, które mogą zostać wdrożone rozdzielnie na różnych maszynach, co pozwala na lepszą skalowalność w środowiskach chmurowych.

2.1 Angular HTTP Server

Serwer HTTP umożliwia obsługę systemu przy użyciu przeglądarki internetowej. Sama aplikacja została przygotowana przy użyciu biblioteki Angular oraz języka TypeScript. Interfejs został opracowany przy użyciu komponentów Bootstrap i umożliwia on na kontrolę wersji, zarządzanie testami, definicjami urządzeń oraz samym procesem wdrażania zmiany [1]. Interfejs użytkownika komunikuje się bezpośrednio z serwerem/kontrolerem sieci.

2.2 Python Flask Server

Centralny punkt systemu odpowiedzialny jest za zarządzanie wszystkimi pozostałymi komponentami systemu. Serwer napisany w języku Python przy użyciu biblioteki Flask i pełni rolę serwera web dla części HTTP. Część odpowiedzialna za komunikację z urządzeniami sieciowymi została opracowana przy użyciu biblioteki Cisco PyATS [3] oraz Netmiko. Serwer wykonuje również operacje na danych znajdujących się w serwisie bazodanowym NoSQL, do którego powinien mieć bezpośredni dostęp. Ostatnim komponentem serwera jest kontroler środowiska testowego, który został wykonany przy użyciu biblioteki gns3fy [4].

2.3 MongoDb Service

MongoDb Service to serwis z implementacją bazy danych NoSQL MongoDb, która jest wykorzystywana do przechowywania poświadczeń użytkowników, definicji urządzeń sieciowych, metadanych systemowych, konfiguracji środowiska testowego oraz historycznych konfiguracji używanych w systemie kontroli wersji.

2.4 GNS3 Server Sandbox

GNS3 Server Sandbox to środowisko testowe, którego bazą jest GNS3 server. Środowisko zarządzane jest przez wyspecjalizowaną część serwera Python. Jest ono bazą do wirtualizacji wielu typów urządzeń sieciowych z różnymi dystrybucjami systemów operacyjnych tych urządzeń. Samo środowisko testowe składa się z kontrolera, który na rysunku 1 jeden reprezentuje "Sandbox Controller" oraz symulowanych urządzeń sieciowych, które są oznaczone na rysunku 1 jako "VMs".

3. GENERYCZNE ŚRODOWISKO TESTOWE

Platforma GNS3 Server, na której bazuje środowisko testowe, umożliwia import szablonów urządzeń. Wiele z nich jest otwartoźródłowych i darmowych. Na rynku dostępne są alternatywy do GNS3 Server [8] takie jak EVE-NG, jednak wszystkie działają na tej samej zasadzie i wybór platformy nie wpływa znacząco na wynik końcowy.

3.1 Wirtualizacja i szablony urządzeń sieciowych

Do każdego urządzenia sieciowego wymagane jest wybranie odpowiedniego typu wirtualizacji oraz szablonu. Wiele urządzeń producentów takich jak Cisco, Arista, Juniper czy VyOS dostarcza swoje szablony. W przypadku braku takiego szablonu możliwe jest zaimportowanie dockerowego obrazu systemu lub opracowanie go samemu przy użyciu systemu z urządzenia. Platforma może zostać dostosowana do urządzeń sieciowych różnych producentów dzięki implementacji wirtualizacji QEMU, emulacji Dynamips oraz obrazów Docker.

3.2 Kontroler

Elementem nieodłącznym jest załączony do pracy obraz docker z kontrolerem środowiska [11]. Jako kontroler użyty został system Linux Ubuntu z podstawowymi pakietami sieciowymi. Częścią kontrolera jest pakiet z interpreterem Python 3.10 oraz biblioteką PyATS wraz z dependencjami. Miejsce podłączenia kontrolera jest ustalane w metadanych w systemie, co pozwala na odwzorowanie dokładnej lokalizacji kontrolera, tak jak ma to miejsce w rzeczywistości. Podejście to umożliwia lepszą kontrolę pakietów, na przykład pól TTL. Kontroler jest zdalnie zarządzany przez serwer podczas przeprowadzania testów.

4. AUTOMATYZACJA OBSŁUGI ŚRODOWISKA TESTOWEGO

Aby uniknąć ręcznej konfiguracji serwera GNS3 oraz ręcznej definicji urządzeń wymagane jest zdobycie informacji

na temat interfejsów i połączeń pomiędzy ruterami. Środowisko ponadto powinno efektywnie wykorzystywać dostępne zasoby, dlatego powinno być tworzone na potrzeby testów, a następnie usuwane po ich przeprowadzeniu.

4.1 Wykrywanie połączeń

Do zautomatyzowanego ustalenia połączeń pomiędzy ruterami przygotowany został algorytm, który przetwarza aktualną konfigurację urządzeń w sieci. Serwer łączy się za pomocą REST API z każdym z ruterów (w przypadku implementacji urządzeń nieobsługujących REST API możliwe jest wykorzystanie konsoli SSH), a następnie pobiera konfigurację interfejsów sieciowych urządzenia. Z każdego interfejsu pobierany jest adres IP wraz z maską sieciową, po pobraniu danych wszystkie adresy są grupowane z uwzględnieniem podsieci. Każdy interfejs ma specjalnie utworzony skrót (ang. hash), którego źródłem jest nazwa hosta rutera oraz numer interfejsu. Tak przygotowany skrót przypisywany jest do grupy konkretnego adresu IP.

4.2 Tworzenie zwirtualizowanych urządzeń sieciowych

Administrator przed przystąpieniem do testów, jest zobligowany do przygotowania definicji urządzeń. W skład definicji urządzenia wchodzą: nazwa użytkownika, hasło dostępowe do konsoli klucz API, nazwa hosta, typ urządzenia. Pierwsze cztery parametry służą do połączenia z urządzeniem sieciowym w celu pobrania wymaganych danych i wprowadzaniu zmian. Piąty parametr definiuje typ systemu operacyjnego urządzenia, a co za tym idzie konkretną implementację. W skład implementacji wchodzi klasa w języku Python odpowiedzialna za połączenie z urządzeniem (wiele z nich dostępnych jest w bibliotece Netmiko) oraz szablon urządzenia w GNS3 server. Tak przygotowane wcześniej dane pozwalają na utworzenie wirtualnych odpowiedników fizycznych urządzeń w środowisku testowym.

4.3 Tworzenie sieci w środowisku wirtualnym

Każda sieć reprezentowana jest przez przełącznik sieciowy w środowisku wykonawczym. Do tak przygotowanych przełączników podpinane są kolejno urządzenia sieciowe na podstawie rozkodowanego skrótu. Podejście to pozwala na połączenie dokładnie tych samych interfejsów do konkretnej sieci, co w rzeczywistym środowisku. Ponadto sam system nie ogranicza się tylko do połączeń punkt-punkt. Do poprawnego działania systemu wymagane jest zdefiniowanie wszystkich urządzeń warstwy trzeciej, natomiast sugerowana jest również definicja urządzeń warstwy drugiej, aby osiągnąć bardziej wiarygodne rezultaty testów jednostkowych.

4.4 Konfiguracja zwirtualizowanych urządzeń

Urządzenia po utworzeniu, uruchomieniu i poprawnym podłączeniu do wirtualnej sieci są konfigurowane za pomocą dostarczonej przez GNS3 server konsoli Telnet. Konsola ta jest osiągalna przez serwer zarządzający, gdyż GNS3 otwiera połączenie do niej spoza środowiska wydzielonego. Konfiguracja bazowa pobierana jest z urządzeń docelowych przez serwer zarządzający (przy pomocy SSH/REST API), a następnie wgrywana jest do urządzeń wirtualnych. Po jej wgraniu wgrywane są zmiany zdefiniowane przez użytkownika.

Gdy serwer zarządzający wykona wszystkie wymienione wyżej kroki, kończy przygotowanie środowiska testowego tworzeniem kontrolera oraz podpięciem go w środowisku wirtualnym do punktu zdefiniowanego w metadanych (punktem powinien być interfejs rutera lub przełącznika).

5. PRZEPROWADZANIE TESTÓW JEDNOSTKOWYCH W ŚRODOWISKU ZWIRTUALIZOWANYM

Do przeprowadzenia testów jednostkowych, używany jest kontroler w środowisku testowym. Serwer zarządzający łączy się z nim, podobnie jak z urządzeniami sieciowymi, przez konsolę Telnet oraz interfejs udostępniony przez dockera. Testy jednostkowe przygotowane przez administratora przy użyciu biblioteki AETest są kopiowane do kontrolera. Tak wgrane testy są przeprowadzane w kontrolerze. Serwer zarządzający sprawdza stan testów raz na pięć sekund. Po ukończeniu, rezultaty zapisywane są w kontrolerze w pliku JSON, który następnie odczytywany jest przez serwer zarządzający.

6. ZARZĄDZANIE ŚRODOWISKIEM

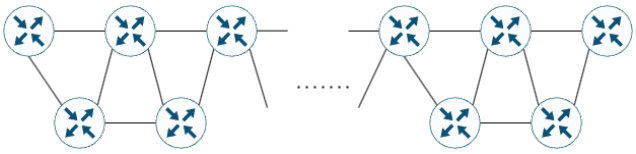
Środowisko testowe tworzone jest za każdym razem od podstaw, przed etapem przeprowadzania testów jednostkowych oraz jest usuwane zaraz po jego zakończeniu. Podejście to sprawdza się w stabilnych sieciach, gdzie nad zmianami nie pracuje wielu administratorów w oddzielnych zespołach. System został opracowany z myślą o środowisku chmurowym. Przykładową implementacją może być implementacja przy użyciu rozwiązań Amazon Web Services. W przypadku takiej implementacji każdy komponent systemu powinien zostać wdrożony jako oddzielna maszyna wirtualna EC2 z dowolnym systemem Linux wspieranym przez GNS3 server. Jednak znacznie bardziej elastyczną i prostszą do wdrożenia opcją będzie wdrożenie systemu w środowisku mieszanym, gdzie wszystkie komponenty są hostowane w środowisku on premise, a GNS3 server jest wdrożony jako maszyna wirtualna typu bastion host w środowisku chmurowym. To zapewni odpowiednią skalowalność i stabilność całego systemu oraz zmniejszy czas potrzebny na przeprowadzanie procesu wdrożenia.

7. TESTOWA IMPLEMENTACJA

System został skonfigurowany na potrzeby urządzeń z systemem VyOS. W celu wirtualizacji urządzeń tego typu użyty został dostarczony przez producenta szablon, który korzysta z wirtualizacji typu QEMU [7].

7.1 Infrastruktura testowa

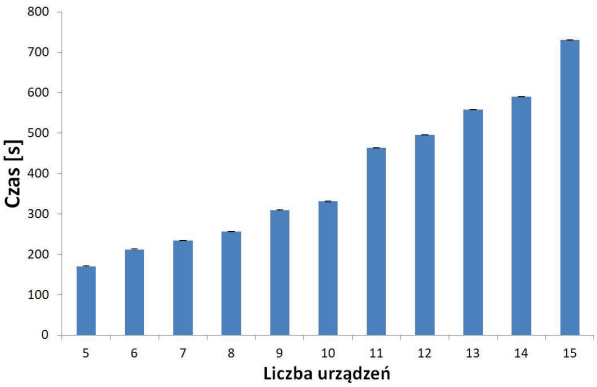
Wzór testowej sieci przedstawiono na rysunku numer 2. Na każdym urządzeniu uruchomiony został klient SSH, serwer HTTPS do integracji REST API oraz procesy sieciowe, w tym proces rutingu dynamicznego OSPF.



Rysunek 2: Diagram testowej sieci.

8. BADANIA WYDAJNOŚCI

Do przeprowadzania badania użyto N ruterów, gdzie N należy do zbioru od 5 do 15. Wynik testu to uśredniony czas potrzebny na przeprowadzenie pełnego procesu tworzenia środowiska testowego. Platforma testowa, na której przeprowadzono badanie to Intel® Core™ i5-7200U (posiada 4 wątki logiczne) wraz z 16 gigabajtami pamięci operacyjnej. Wszystkie komponenty były hostowane na jednym systemie jako procesy. Użyto systemu Linux Ubuntu w wersji 20.04.



Rysunek 3: Wykres zależności czasu potrzebnego na generowanie środowiska testowego od liczby urządzeń sieciowych w sieci.

8.1 Podsumowanie wyników

Wraz ze zwiększającą się liczbą urządzeń do zwirtualizowania, czas potrzebny na uruchomienie środowiska testowego rośnie nieliniowo. Pomiędzy liczbą urządzeń równą 10 a 11 widać znaczny wzrost potrzebnego czasu. Wynika on z dostępnej mocy obliczeniowej procesora. Przy 10 urządzeniach możliwe jest równoległe uruchamianie urządzeń natomiast przy 11 urządzeniach i więcej część urządzeń musi oczekiwać na zwolnienie zasobów. Kolejny nagły wzrost potrzebnego czasu ma miejsce przy 15 urządzeniach. W tym miejscu systemowi kończy się dostępna pamięć operacyjna i zaczyna on korzystać z przestrzeni wymiany. Dalsze dodawanie urządzeń do sieci przy tych zasobach będzie skutkowało wysoką niestabilnością systemu. Pomimo ograniczonych zasobów, system jest w stanie zasymulować sieć LAN, w skład której wchodzi 15 ruterów. Rutery VyOS charakteryzują się również wysokim zapotrzebowaniem na pamięć operacyjną (każdy ruter wymaga 2048 MB pamięci) w porównaniu do innych tego typu urządzeń stosowanych w sieciach LAN. Dla porównania ruter Cisco przy użyciu emulatora Dynamips może już działać przy użyciu 256 MB pamięci RAM.

9. PODSUMOWANIE

Przedstawiony w artykule system do automatycznego generowania środowiska testowego pozwala na przeniesienie kluczowych testów sieci do zwirtualizowanego środowiska co zwiększa bezpieczeństwo i zmniejsza ryzyko wystąpienia błędu spowodowanego przez zmianę. System dzięki modularności jest łatwo skalowalny, a jego implementacja w średniej wielkości sieciach lokalnych (do 30 urządzeń sieciowych) nie wiąże się z dużymi kosztami, dzięki optymalnemu wykorzystaniu zasobów. Sam kontroler został przygotowany również z myślą o wykorzystaniu w innych systemach takich jak Ansible czy Chef. Poprzez dostęp REST API lub użycie klasy SandboxController [5], możliwe jest wygenerowanie środowiska testowego, nie używając do tego aplikacji przeglądarkowej. System ze względu na komunikację przy użyciu HTTPS z serwerem GNS3, implementuje wszystkie funkcjonalności, jakie oferuje ta aplikacja. System z powodzeniem może zostać wykorzystany do wdrażania i rozwoju sieci SDN wraz z komponentami NFV [6].

LITERATURA

[1] Tekiela, Wiktor. (2023). Automatyzowanie procesów zarządzania zmianami w sieciach. PRZEGLĄD TELEKOMUNIKACYJNY - WIADOMOŚCI TELEKOMUNIKACYJNE. 1. 393-396. 10.15199/59.2023.4.89.

[2] Zhang, H., Wang, Y., Qiu, X., Li, W., & Zhong, Q. (2015). Network operation simulation platform for network virtualization environment. 2015 17th Asia-Pacific Network Operations and Management Symposium (APNOMS). doi:10.1109/apnoms.2015.7275351

[3] Cisco, "Cisco pyATS: Network Test & Automation Solution" https://developer.cisco.com/docs/pyats/

[4] David Flores, "Gns3fy overview", https://davidban77.github.io/gns3fy/

[5] Wiktor Tekiela, Klasa w języku Python odpowiedzialna za zarządzanie środowiskiem testowym, https://github.com/abcdrefg/ NetworkDeploymentTool/blob/dev/ backend/SandboxController.py

[6] Kalliola, A., Lal, S., Ahola, K., Oliver, I., Miche, Y., & Holtmanns, S. (2017). Testbed for security orchestration in a network function virtualization environment. 2017 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN). doi:10.1109/nfv-sdn.2017.8169857

[7] Carvalho, H., Nelissen, G., & Zaykov, P. (2020). mcQEMU: Time-Accurate Simulation of Multicore platforms using QEMU. 2020 23rd Euromicro Conference on Digital System Design (DSD). doi:10.1109/dsd51259.2020.00024

[8] Bhola, Abhishek & Jain, Arpit & Lakshmi, Bhavani & Lakshmi, Tulasi & Hari, Chandana. (2022). A Wide Area Network Design and Architecture using Cisco Packet Tracer. 1646-1652. 10.1109/IC3I56241.2022.10073328.

[9] Emiliano, R., & Antunes, M. (2015). Automatic network configuration in virtualized environment using GNS3. 2015 10th International Conference on Computer Science & Education (ICCSE). doi:10.1109/iccse.2015.7250212

[10] CV, R. K., & Goyal, H. (2019). IPv4 to IPv6 Migration and Performance Analysis using GNS3 and Wireshark. 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN). doi:10.1109/vitecon.2019.8899746

[11] Wiktor Tekiela, Obraz Docker kontrolera środowiska testowego, https://aghedupl-my.sharepoint.com/:u:/g/personal/wiktort_student_agh_edu_pl/EQv0glVkIRJImVfJhxUoYN4BV9XNtgjK5p2QHrkZs9uKNQ?e=zTbxnG

RÓWNOWAŻENIE OBCIĄŻEŃ I OPÓŹNIENIA W ARCHITEKTURZE CHMUR BRZEGOWYCH DLA OBLICZEŃ MOBILNYCH

LOAD AND LATENCY BALANCING IN EDGE-CLOUD ARCHITECTURE FOR MOBILE COMPUTING

Jerzy Martyna
Wydział Matematyki i Informatyki UJ
ul. Prof. S. Łojasiewicza 6, 30-348 Kraków
e-mail: jerzy.martyna@uj.edu.pl

DOI: 10.15199/59.2024.4.91

Streszczenie: **Obliczenia mobilne z użyciem chmur brzegowych są jedną z kluczowych technologii sieci komórkowych. Odpowiednie rozmieszczenie chmur brzegowych w sposób istotny wpływa na zrównoważenie ich obciążenia oraz zmniejszenie czasu opóźnienia transmisji. W artykule przedstawiono metodę, która może w istotny sposób poprawić efektywność działania sieci komórkowej. Eksperyment symulacyjny potwierdza, że zastosowanie tej metody zapewnia wymagania stawiane przez użytkowników.**

Abstract: **Mobile computing using edge clouds is one of the key technologies for cellular networks. The appropriate placement of edge clouds has a significant impact balancing their load and reducing transmission delay time. The paper presents a method that can significantly improve efficiency of the mobile network. The simulation experiment confirms that the application this method ensures the requirements set by users**

Słowa kluczowe: **Chmura brzegowa, mobilne przetwarzanie brzegowe, równoważenie obciążenia, zerowe opóźnienie.**

Keywords: **Edge-cloud, load balancing, mobile edge-computing, zero latency.**

1. WPROWADZENIE

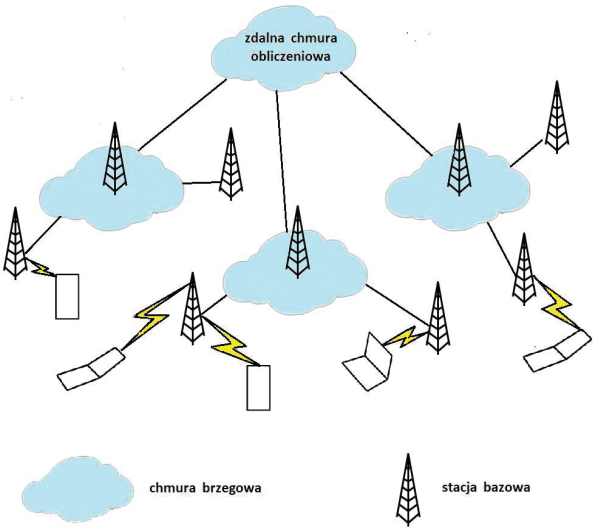
Szereg nowych aplikacji mobilnych, np. do rzeczywistości rozszerzonej AR (ang. *augmented reality*) i rzeczywistości wirtualnej VR (ang. *virtual reality*), do rozpoznawania twarzy itp., wymaga dostępu do systemów komputerowych o dużej mocy obliczeniowej. Jedną z technologii, która może być zastosowana dla rozwiązania tego problemu, są obliczenia mobilne z użyciem chmur brzegowych (ang. *mobile edge-computing*, MEC) [1], [2]. Ich zastosowanie jest alternatywą dla centralnej chmury obliczeniowej [3] oraz sieci mgłowych [4].

Dostawcy usług telekomunikacyjnych wykazują stale rosnące zainteresowanie wdrożeniami chmur brzegowych, co potwierdzają dokumenty i standardy, m.in. [5], [6]. Rozpowszechnienie się technologii MEC wiąże się z odpowiednią lokalizacją chmur brzegowych w strukturze sieci komórkowej 4G/5G. Ich stosowna lokalizacja pozwala na zrównowazenia średniego obciążenia chmur brzegowych oraz minimalizację średniego czasu opóźnienia transmisji. Umożliwia to zwiększenie wydajności obliczeniowej chmur brzegowych oraz poprawia akceptację usług mobilnych, jak również pozwala na długotrwałe utrzymanie parametrów jakości obsługi QoS (ang. *Quality of Service*).

Współczesne badania dotyczące obliczeń mobilnych z użyciem chmur brzegowych koncentrują się m.in. na optymalizacji obliczeń w tego rodzaju systemach [7] oraz dostarczaniu usług o pożądanych parametrach QoS [8]. W wielu pracach zastosowano teorię gier dla rozwiązania problemu alokacji zasobów w obliczeniach mobilnych z użyciem chmur brzegowych. Między innymi Zhang i in. [9] zaproponowali grę koalicyjną dla ograniczenia kosztów komunikacji dla uczynienia ich bardziej opłacalnymi. Alokację zasobów w chmurach brzegowych wraz z uwzględnieniem zużycia energii elektrycznej w oparciu o teorię gier przedstawiono w pracy [10].

Fizyczna lokalizacja chmur brzegowych w sieci komórkowej ma kluczowe znaczenie dla zapewnienia ich wysokiej wydajności. Z kolei nieefektywne rozmieszczenie chmur brzegowych powoduje drastyczne zwiększenie średniego opóźnienia i nierówne ich obciążenie. Oznacza to, że niektóre chmury brzegowe są przeciążone, podczas gdy inne są wykorzystywane w niewystarczającym stopniu [11]. W pracach poświęconych tej tematyce brano pod uwagę m.in. klastrowanie połączone z optymalizacją [12], algorytmy heurystyczne [13], metodę k-średnich [14], teorię grafów [15].

Głównym celem tego artykułu jest przedstawienie metody obliczenia dla zrównoważenia obciążenia i opóźnienia chmur brzegowych w strukturze sieci komórkowej. Dla przyjętego modelu określono wskaźniki obciążenia i opóźnienia chmur brzegowych. Wyznaczono granice efektywności tego systemu. Oryginalnym rozwiązaniem jest optymalizacja efektywności takiego systemu, co może być przydatne przy projektowaniu rozmieszczenia chmur brzegowych w sieciach komórkowych.



Rys. 1. Model systemu komórkowego z chmurami brzegowymi oraz stacjami bazowymi.

W rozdziale 2 przedstawiono model systemu komórkowego z zastosowanymi chmurami brzegowymi. Rozdział 3 podaje główne wskaźniki efektywności takiego systemu, którymi tutaj są wskaźniki nierównowagi obciążenia oraz opóźnienia transmisji. W rozdziale 4 zaprezentowano metodę optymalizacji efektywności tego systemu. Rozdział 5 przedstawia wyniki badań symulacyjnych. Podsumowanie pracy przedstawiono w końcowym rozdziale pracy.

2. MODEL SYSTEMU

Niech będzie dany system komórkowy z chmurami brzegowymi oraz zbiorem stacji bazowych. Każda ze stacji bazowych ma stałe połączenie z najbliższą dla niej chmurą brzegową. Z kolei chmury brzegowe są połączone ze zdalną chmurą obliczeniową. Każdy mobilny użytkownik (ang. *mobile user* - MU), przebywając na danym obszarze systemu komórkowego, może łączyć się z najbliższą mu stacją bazową (patrz rys. 1).

Niech $P(x)$ będzie mocą transmisji pojedynczego użytkownika MU znajdującego się w lokalizacji x. Współczynnik SNR (ang. *signal to noise ratio*) dla tego użytkownika wynosi:

$$SNR_{m_j}(x) = \frac{P(x)g_{m_j}(x)}{\sigma^2}, \quad 1 \le m_j \le M_j, \quad 1 \le j \le J \tag{1}$$

gdzie $g_{m_j}(x)$ jest zyskiem propagacji kanału dla lokalizacji x, σ^2 jest mocą szumu, m_j jest indeksem stacji bazowej, M_j jest liczbą stacji bazowych, j jest indeksem chmury brzegowej, J jest liczbą chmur brzegowych.

Szybkość przesyłania danych (ang. *data rate*) dla lokalizacji x w odniesieniu do m_j-tej stacji bazowej wynosi [16]:

$$r_{m_j}(x) = W_{m_j} \log(1 + SNR_{m_j}(x)), \quad 1 \le m_j \le M_j, \quad 1 \le j \le J \tag{2}$$

gdzie W_{m_j} jest szerokością pasma m_j-tej stacji bazowej.

Zakłada się, że długości wszystkich danych w sieci komórkowej są o rozkładzie wykładniczym o wartości średniej $l(x)$. Przyjmuje się też, że przepływ danych z lokalizacji x jest o rozkładzie Poissona z wartością średnią $\lambda(x)$. Wówczas średnie obciążenie pochodzące z lokalizacji x, będących w zasięgu stacji bazowej m_j, można przedstawić [17]:

$$\rho_{m_j}(x) = \frac{\lambda(x)l(x)}{r_{m_j}}, \quad 1 \le m_j \le M_j, \quad 1 \le j \le J \tag{3}$$

Średnie obciążenie m_j-tej stacji bazowej może być uzyskane dzięki zsumowaniu obciążeń pochodzących ze wszystkich lokalizacji x w zasięgu tej stacji bazowej, a mianowicie:

$$\rho_{m_j} = \sum_{x \in \mathcal{A}_{m_j}} \rho_{m_j}(x), \quad 1 \le m_j \le M_j, \quad 1 \le j \le J \tag{4}$$

gdzie $\mathcal{A}_{m_j}$ jest zasięgiem radiowym stacji bazowej m_j.

Średnie obciążenie j-tej chmury brzegowej wynosi

$$L_j = \sum_{m_j=1}^{M_j} \rho_{m_j}, \quad 1 \le j \le J \tag{5}$$

Przyjmując, że modelem pojedynczej stacji bazowej jest system $M/M/1$, można podać wartość średniego czasu dostarczenia danych [18] z lokalizacji x do stacji bazowej m_j, a mianowicie

$$t_{m_j}(x) = \frac{l(x)}{r_{m_j}(x)(1 - \rho_{m_j})}, \quad 1 \le m_j \le M_j, \quad 1 \le j \le J \tag{6}$$

Średni czas oczekiwania dla przepływu danych z lokalizacji x wynosi:

$$w_{m_j}(x) = t_{m_j}(x) - s_{m_j}(x), \quad 1 \le m_j \le M_j, \quad 1 \le j \le J \tag{7}$$

gdzie $s_{m_j}(x)$ jest średnim czasem obsługi przepływów danych, który wynosi $\frac{l(x)}{r_{m_j}(x)}$.

Współczynnik opóźnienia dla przepływów danych jest ilorazem średniego czasu oczekiwania oraz czasu obsługi tych danych w m_j-tej stacji bazowej, co można przedstawić następująco:

$$\mu_{m_j}(x) = \frac{w_{m_j}(x)}{s_{m_j}(x)} = \frac{\rho_{m_j}(x)}{1 - \rho_{m_j}(x)}, \quad 1 \le m_j \le M_j, \quad 1 \le j \le J \tag{8}$$

Przyjmując, że wszystkie urządzenia mobilne mają ten sam współczynnik opóźnienia we wszystkich lokalizacjach x, można uogólnić wzór (8), a mianowicie

$$\mu_{m_j} = \frac{w_{m_j}}{s_{m_j}} = \frac{\rho_{m_j}}{1 - \rho_{m_j}}, \quad 1 \le m_j \le M_j, \quad 1 \le j \le J \tag{9}$$

Zakłada się, że C_{m_j} będzie mocą obliczeniową (ang. *computing capacity*) m_j-tej stacji bazowej wyrażaną w cyklach CPU na sekundę. Niech $\nu(x)$ jest średnim rozmiarem (ang. *average computing size*) przepływów danych wyrażaną w liczbie cykli CPU. Wówczas średnie obciążenie stacji bazowej uwzględniające moc obliczeniową stacji bazowej m_j wynosi:

$$\hat{\rho}_{m_j}(x) = \frac{\lambda(x)\nu(x)}{C_{m_j}}, \;\; 1 \le m_j \le M_j, \;\; 1 \le j \le J \quad (10)$$

Sumując średnie obciążenia wszystkich stacji bazowych związanych z chmurą brzegową j oraz zakładając, że wszystkie urządzenia mobilne mają taki sam współczynnik opóźnienia, można podać wartość średniego obciążenia uwzględniającą moce obliczeniowe stacji bazowych, a mianowicie

$$\hat{\rho}_j = \sum_{m_j=1}^{M_j} \sum_{m_j \in \mathcal{A}_{m_j}} \hat{\rho}_{m_j}(x), \quad j \le j \le J \quad (11)$$

3. WSKAŹNIKI EFEKTYWNOŚCI SYSTEMU KOMÓRKOWEGO Z CHMURAMI BRZEGOWYMI

Biorąc pod uwagę wszystkie chmury brzegowe sieci komórkowej można podać wskaźnik równowagi chmur brzegowych, który definiuje się następująco:

$$\Omega_J = \frac{1}{J} \sum_{j=1}^{J} (L_j - \overline{L}_j)^2, \quad (12)$$

gdzie $\overline{L}_j$ jest średnim obciążeniem j-tej chmury brzegowej, które wynosi

$$\overline{L}_j = \frac{1}{J} \sum_{m_j=1}^{M_j} L_{m_j}, \;\; 1 \le j \le J \quad (13)$$

Wskaźnik równowagi pomiędzy opóźnieniami transmisji wszystkich chmur brzegowych wynosi:

$$\Psi_J = \frac{1}{J} \sum_{j=1}^{J} (\mu_j - \hat{\mu}_j)^2 \quad (14)$$

Wskaźnik powyższy wskazuje, jak bardzo różnią się opóźnienia transmisji dla chmur brzegowych po uwzględnieniu różnych mocy obliczeniowych wszystkich stacji bazowych.

Podstawową miarą efektywności systemu jest minimalizacja wskaźnika równowagi oraz równowagi opóźnień. Dla każdego z tych wskaźników można podać warunki ograniczające, a mianowicie

$$0 \le \Omega_J \le \Omega_J^{(max)} \quad (15)$$

$$0 \le \Psi_J \le \Psi_J^{(max)} \quad (16)$$

gdzie

$$\Omega_J^{(max)} = \frac{1}{J} \sum_{j=1}^{J} L_j^2 \quad (17)$$

$$\Psi_J^{(max)} = \frac{1}{J} \sum_{j=1}^{J} \mu_j^2 \quad (18)$$

Można podać warunek ograniczający efektywność systemu

$$\Omega_J^{(max)} + \Psi_J^{(max)} = \frac{1}{J} \sum_{j=1}^{J} (L_j^2 + \mu_j^2) \quad (19)$$

4. OPTYMALIZACJA EFEKTYWNOŚCI SYSTEMU KOMÓRKOWEGO Z CHMURAMI BRZEGOWYMI

Dla optymalizacji efektywności systemu komórkowego z chmurami brzegowymi wykorzystano optymalizację bez użycia pochodnych (ang. *derivative-free optimization*) [19], [20], [21]. Metody te dla znalezienia minimum funkcji wykorzystują w każdej iteracji informacje jedynie o wartościach minimalizowanej funkcji. Minimalizowaną funkcją jest tutaj

$$F(\Omega_J, \Psi_J) = (\Omega_J(\Omega_J, \Psi_J) - K)^2 + (\Psi_J(\Omega_J, \Psi_J) - T)^2 \quad (20)$$

gdzie K oraz T są znanymi wartościami maksymalnymi odpowiednio dla wskaźnika nierównowagi obciążenia oraz opóźnienia.

Użyta metoda polega utworzeniu lokalnego modelu kwadratowego dla punktu $p^k = (\Omega_J^k, \Psi_J^k)$, $k = 0, 1, 2, \ldots$, a mianowicie

$$\eta_k(p^k + d) = F(p^k) + (y^k)^T d + \frac{1}{2} d^T \mathcal{H}^k d \quad (21)$$

W powyższym równaniu zakłada się, że $p^{k+1} = p^k + d^k$. Dla małych wartości ϵ, $\epsilon > 0$, uzyskuje się:

$$g^k = (g_1, g_2)^T, \;\; g_i = \frac{F(p^k + \epsilon e_i) - F(p^k - \epsilon e_i)}{2\epsilon^2} \quad (22)$$

gdzie $e_1 = (1, 0)^T$ oraz $e_2 = (0, 1)^T$ są przybliżonymi gradientami. Macierz hesjanu $\mathcal{H}^{\mathbf{k}}$ jest dana jako

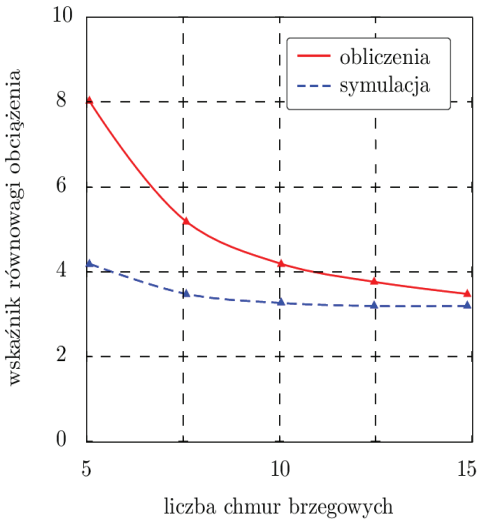
$$\mathcal{H}^k = \begin{pmatrix} \vartheta_{11} & \vartheta_{12} \\ \vartheta_{21} & \vartheta_{22} \end{pmatrix} \quad (23)$$

gdzie ϑ_{ij} wynosi

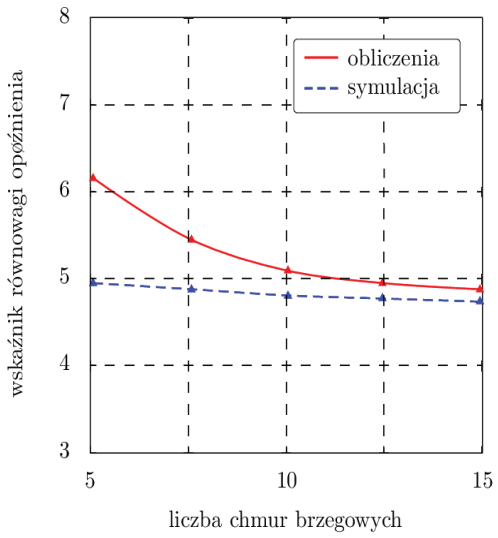
$$\vartheta_{ij} = \frac{F(p^k + \epsilon e_i + \epsilon e_j) - F(p^k) - F(p^k + \epsilon e_j) + F(p^k)}{\epsilon^2} \quad (24)$$

5. WYNIKI BADAŃ SYMULACYJNYCH

W modelu symulacyjnym założono, że sieć fizyczną stanowi sieć 5G na obszarze 2000 × 1500 m^2. Przyjmuje się, że każdy z mobilnych użytkowników może łączyć się z najbliższą dla niego stacją bazową. Wszystkie stacje bazowe są połączone z najbliższymi dla nich chmurami brzegowymi. Przyjęto, że liczba chmur brzegowych może zmieniać się w granicach od 5 do 15, podczas gdy liczba stacji bazowych zmienia się od 10 do 100.



Rys. 2. Wskaźnik równowagi obciążenia chmur brzegowych w funkcji liczby chmur.



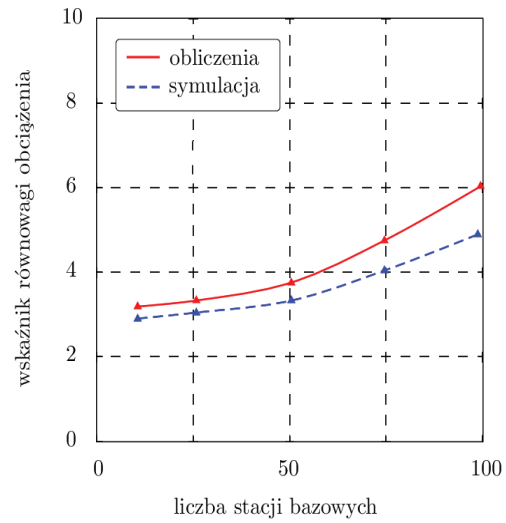
Rys. 3. Wskaźnik równowagi opóźnienia chmur brzegowych w funkcji liczby chmur brzegowych.

W przyjętym modelu symulacyjnym przyjęto, że istnieje 7500 lokalizacji, z których każda reprezentuje kwadrat o rozmiarze 20 m $\times$ 20 m. Przepływ danych z każdej lokalizacji jest procesem Poissona z intensywnością zgłoszeń równą 0.5 przepływów/sekundę. Rozmiar danych każdego z przepływów jest o rozkładzie wykładniczym z wartością średnią 0.1 Mbit/s. Każdy z przepływów danych wymaga średnio 4000 cykli CPU. Przyjęto, że moc urządzeń mobilnych wynosi 100 mW, natomiast szerokość pasma częstotliwości każdej stacji bazowej wynosi 10 MHz. Założony poziom szumów wynosi tutaj -104 dBm.

Na rys. 2 przedstawiono wskaźnik równowagi obciążenia chmur brzegowych w funkcji liczby chmur brzegowych uzyskany dzięki symulacji oraz przy użyciu przedstawionej metody obliczeniowej. Przyjęto, że liczba stacji bazowych jest stała i wynosi 50. Wykres wskazuje, że wraz ze zwiększającą się liczbą chmur brzegowych wyniki symulacji są zbliżone do rezultatów obliczeniowych.

Rysunek 3 zaprezentowano wskaźnik równowagi opóźnień chmur brzegowych w funkcji liczby chmur brzegowych. Przyjęto tutaj, że jest tutaj użytych 50 stacji bazowych. Można zauważyć, że wraz ze wzrostem liczby chmur brzegowych wyniki uzyskane dzięki obliczeniom są zbliżone rezultatów symulacyjnych.

Na rys. 4 przedstawiono wskaźnik równowagi obciążenia chmur brzegowych w funkcji liczby stacji bazowych. dla ustalonej liczby chmur brzegowych, która jest równa 10. Wykres potwierdza, że wraz ze wzrostem liczby stacji bazowych rośnie wskaźnik równowagi obciążenia, co potwierdzają obliczenia i symulacja.



Rys. 4. Wskaźnik równowagi obciążenia chmur brzegowych w funkcji liczby stacji bazowych.

6. WNIOSKI

W artykule przedstawiono metodę obliczenia obciążeń i opóźnień dla chmur brzegowych w systemach komórkowych. Dzięki niej jest możliwe poprawienie działania tych urządzeń, a tym samym usunięcie ewentualnych przeciążeń lub usunięcie ich niskiego wykorzystywania. Zezwala to także na minimalizację opóźnień, co ma znaczenie w przypadku transmisji z określonymi parametrami QoS. Badania symulacyjne potwierdziły, że metoda ta może być przydatna dla wyrównywania obciążeń i opóźnień chmur brzegowych, co może być pomocne przy projektowaniu i eksploatacji tych systemów.

LITERATURA

[1] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, D. Sabella, On Multi-Access Edge Computing: A Survey of the Emerging 5G Network Edge Cloud Architecture and Orchestration, IEEE Comm. Surveys and Tutorials, vol. 19, no. 3, pp. 1657 - 1681, 2017.

[2] P. Mach, Z. Becvar, Mobile Edge Computing: A Survey on Architecture and Computation Offloading, 2017, [https://arxiv.org/abs/1702.05309]

[3] S. Rizvi, S., J. Ryoo, J., Y. Liu, Y.,D. Zazworsky, D.,A. Cappeta, A centralized trust model approach for cloud computing. In IEEE 2014 23rd Wireless and Optical Communication Conference (WOCC), pp. 1 - 6, 2014.

[4] M. Chiang, Fog Networking: An Overview on Research Opportunities, Princeton University, 2015, [https://www.princeton.edu/ chiangm/FogResearchOverview.pdf]

[5] Y. Chan Hu i in., Mobile Edge Computing. A Key Technology Towards 5G, ETSI White Paper No. 11, 2015.

[6] M. Patel i in., Mobile Edge Computing (MEC), Introductory Technical White Paper, ostatnia aktualizacja 2016.

[7] F. Wang, J. Xu, X. Wang, S. Cui, Joint Offloading and Computing Optimization in Wireless Powered Mobile-Edge Computing Systems, in IEEE Trans. on Wireless Comm., vol. 17, no. 3, pp. 1784 - 1797, 2018.

[8] G. Wang, F. Xu, C. Zhao, QoS-enabled resource allocation algorithm in internet of vehicles with mobile edge computing. IET Comm., vol. 14 no. 14, 2326 - 2333, 2020.

[9] T. Zhang T, W. Chen, F. Yang, Data offloading in mobile edge computing: a coalitional game based pricing approach. IEEE Access 2018, vol. 6, pp. 2760 - 2767, 2018.

[10] W. Binwei, J. Zeng, Iu ge, Y. Tang, X. Su, Game-Theoretical Approach for Energy-Efficient Resource Allocation in MEC Network, 2019 IEEE Int. Conf. on Comm. (ICC), 2019.

[11] D. Bhatta, L. Mashayekhy, Generalized cost-aware cloudlet placement for vehicular edge computing systems, in: 2019 IEEE International Conference on Cloud Computing Technology and Science, CloudCom, pp. 159 - 166, 2019.

[12] N. Mohan, A. Zavodovski, P. Zhou, J. Kangasharju, Anveshak: Placing edge servers in the wild, in: Proceedings of the 2018 Workshop on Mobile Edge Communications, ACM, United States, pp. 7 - 12, 2018.

[13] L. Ma, J. Wu, L. Chen, Z. Liu, Fast algorithms for capacitated cloudlet placements, in: IEEE 21st International Conference on Computer Supported Cooperative Work in Design, pp. 439 - 444, 2017.

[14] J. Liu, U. Paul, S. Troia, O. Falowo, G. Maier, K-means based spatial base station clustering for facility location problem in 5G, in: J. Lewis, Z. Ndlela (Eds.), Proceedings of Southern Africa Telecommunication Networks and Applications Conference, SATNAC, pp. 406–409, 2018.

[15] F. Zeng, Y. Ren, X. Deng, W. Li, Cost-effective edge server placement in wireless metropolitan area networks, Sensors 19, vol.1, 2018.

[16] T. Han, N. Ansari, A traffic load balancing framework for software-defined radio access networks powered by hybrid energy sources, IEEE/ACM Trans. on Networking, vol. 24, no. 2, pp. 1038 - 1051, 2016.

[17] N. Sapountzis, T. Spyropoulos, N. Nikaein, U. Salim, User association in hetnets: Impact of traffic differentiation and backhaul limitations, IEEE/ACM Trans. on Networking, vol. 25, no. 6, pp. 3396 - 3410, 2017.

[18] L. Kleinrock, Queueing Systems. Computer Applications, vol. 2, John Wiley and Sons, 1976.

[19] W. Spendley, G.R. Hext, F.R. Himsworth, Sequential application for simplex designs in optimisation and evolutionary operation. Technometrics 4, pp. 441–461, 1962.

[20] M. J.D. Powell, An efficient method for finding the minimum of a function of several variables without calculation derivatives, The Computer Journal, vol. 7, pp. 155 - 162, 1964.

[21] J. A. Nelder, R. Mead, R.: A simplex method for function minimization. Comput. J. 7, 308–313, 1965.

Model oparty na danych do długoterminowego prognozowania przepustowości i opóźnień plastrów sieci 5G z kontekstem konfiguracji komórki

Data-driven model for long-term forecasting of 5G throughput and delay per network slice with the context of cell configuration

Dominik Dulas[1,2]; Justyna Witulska[1,2]; Agnieszka Wyłomańska[2]; Krzysztof Walkowiak[2]

[1] Nokia Solutions and Networks, Wroclaw, Poland, email: name.surname@nokia.com
[2] Wroclaw University of Science and Technology, Wroclaw, Poland, email: name.surname@pwr.edu.pl

The research leading to these results has received partial funding from the Polish Ministry of Education and Science via the Industrial PhD program.

DOI: 10.15199/59.2024.4.92

Streszczenie: Podstawowym wkładem tej pracy jest ocena wielowymiarowego modelu VARMAX, wykorzystującego rzeczywiste dane o ruchu i środowisku sieci w połączeniu na etapie uczenia z danymi o konfiguracji komórek, w celu dokładnego prognozowania wskaźników telekomunikacyjnych, które są kluczowe przy wymiarowaniu sieci 5G z zastosowaniem plastrowania. Wyniki pokazują, że technika ta jest skuteczna w długoterminowym przewidywaniu opóźnień i przepustowości, które znacząco wpływają na jakość usługi w plastrze, w horyzoncie 3 miesięcy.

Abstract: The primary innovation and contribution of this study is the evaluation of a multidimensional VARMAX model, utilizing real traffic and environmental network data combined with cell configuration during the training phase, to accurately forecast telecommunication metrics, which are crucial in the dimensioning of 5G networks with network slicing. The results show that this technique is effective in predicting delay and throughput, which significantly influence the slice quality of service, over a long-term horizon of approximately 3 months.

Słowa kluczowe: jakość usług, planowanie wydajności, plastrowanie sieci, prognozowanie

Keywords: capacity planning, forecasting, network slicing, quality of service

1. INTRODUCTION

As networks expand, it becomes necessary for Communication Service Providers (CSPs) to consistently assess and keep track of their capacity. There should be a decrease in this capacity, they engage in a process of network re-planning that includes capacity dimensioning [1]. This dimensioning process is typically performed using proprietary tools from vendors, which rely on the inherent capabilities of their products and their internal expertise. It is also common to perform this process without any advanced techniques, instead using spreadsheets [2]. Furthermore, this method is based on conventional linear or queueing models [3], which are calibrated and validated through network simulations or laboratory tests. The influence of product enhancements, such as new or expanded features, is incorporated into the model as an additional linear block or coefficient. However, this strategy can lead to imprecise outcomes with intricate deployments due to the multicollinearity effect [2]. Moreover, the introduction of Network Slicing (NS) in 5G network deployments, which involves segmenting the existing network into layers with distinct Quality of Service (QoS) requirements, further complicates the issue. The application of sophisticated statistical or machine learning models that utilize data can address this challenge, as these methods allow the specific modeling of individual cells or groups of cells that share similar configurations or performance characteristics, customized for their unique features [4], [5]. Models that incorporate multiple variables consider a range of factors to predict traffic growth for each cell. This approach leads to precise and varied resource planning for networks, catering to specific local demands rather than adopting a one-size-fits-all strategy.

The traffic model is a crucial component for dimensioning. Contemporary research indicates that artificial intelligence techniques such as deep neural networks, long short-term memory, and recurrent neural networks have the ability to precisely predict traffic load specific to services [4]. The use of data allows for the application of forecasting techniques to accurately predict future resource needs. This issue can be divided into short- and long-term forecasts. Short-term forecasts improve operational management and can be beneficial in network management, where continuous predictions are utilized to optimize specific processes (e.g. allocation of capacity per slice). Short-duration forecasting is particularly valuable in environments subject to high variability. The illustrative example of a short-term prediction for 5G data is a power forecast for 5G photovoltaic base stations [6]. In this case, the goal is to achieve more precise energy allocation and management. In [6], the authors explain that short-term energy fluctuations are influenced by low latency, high data transmission rates, and the variety of connected devices specific to 5G. The advantage of long-term predictions is the ability to plan future strategies over a longer time horizon. However, in contrast, the limitation of long-term predictions that can be observed in systems is the chaotic

dynamics of these systems [7].

This study presents the results of the evaluation of multivariate models to predict long-term throughput and delay within the 5G NS dimensioning framework. The predictive model is developed and assessed using actual traffic and environmental data. Its application in the dimensioning process eliminates the requirement for a predefined traffic model [8]. Due to the prediction of both key metrics, which are crucial from a QoS standpoint, it becomes feasible to determine when each metric will reach its capacity threshold (such as link bandwidth or delay) and by what extent it will surpass this limit. This subsequent data can be utilized for the redimensioning of BTS to enhance its capacity. Furthermore, the study illustrates how the incorporation of cell configuration data can generalize the training of models from a per-cell basis to a per-band basis.

The paper is organized as follows. Section 2 describes the data and methods that have been used in the investigation. Section 3 presents the approach and precision of the predictions. Section 4 concludes this study.

2. MATERIALS AND METHODS

2.1 Data set

The research used a data set consisting of time series averages taken every hour from 33 operational 5G Base Transceiver Stations (BTS) in a live network environment. Data were collected from each BTS and each cell configured within these stations over the course of eight months in 2023 and 2024. Following the preparation and cleaning of the data, the models were trained and tested using data from 158 and 92 days, respectively. Each cell is characterized by a variety of configuration attributes (such as cell duplex mode, channel bandwidth, etc.) and Key Performance Indicators (KPIs) derived from counters. The counters provide a low-level description of events in a mobile network [9] (for example, the mean delay in the downlink radio link control in gNodeB per slice). Subscribers in this network cluster have been divided into three groups:

- Slice A - mobile subscribers with high priority,
- Slice B - mobile subscribers with low priority,
- Slice C - fixed wireless access subscribers with lowest priority.

2.2 Short-term forecasting

In our previous research published in [10], we used statistical models and neural networks to predict short-term delay and performance for 5G networks with different network slices using a common model. There, the forecasting problem was part of a data-driven framework for the dimensioning and planning of sliced 5G networks. In addition to prediction, the models were also used for "what-if" analysis. As we investigate the applicability of these methods for both short-term and long-term dimensioning processes, it is beneficial to discuss the results of our previous studies.

Tab. 1: The mean of nMAE corresponding to the short-term forecasts made for 66 cells.

Variable	A	B	C
Delay	0.080	0.077	0.089
Throughput	0.010	0.011	0.032

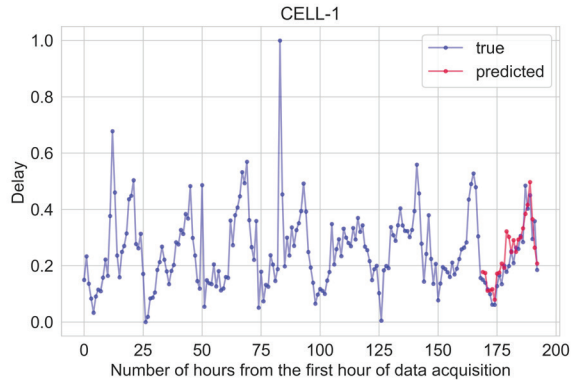


Fig. 1: The short-term forecast for normalized delay for Slice C using VARMAX(1,0) for illustrative cell.

In [10], the data was aggregated over an hour interval and the prediction applied to the next 24 hours. The illustrative results of the prediction for the selected cell and for Slice C are visible in Fig. 1. In the Tab. 1, the mean of the normalized mean absolute error (nMAE) calculated for 66 cells is provided. It corresponds to throughput and delay per network slice. MAE is normalized by dividing by range. As can be seen, the nMAE values are low for each network slice and each variable.

2.3 Feature selection

The features used in modeling are shown in Fig. 2. The full name, description, and unit for each variable is provided in Tab. 3.

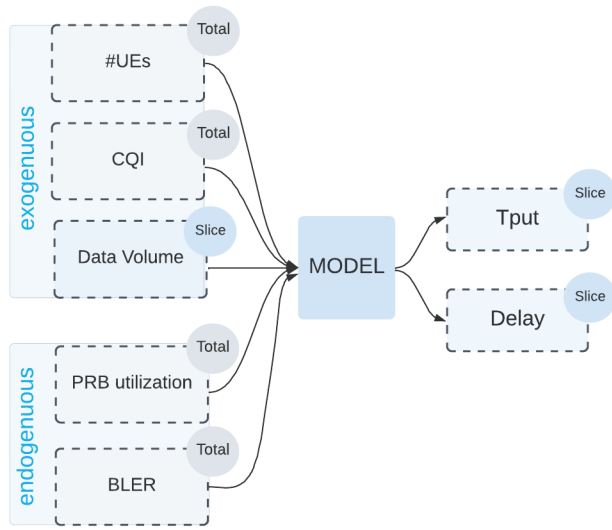


Fig. 2: Model scheme for throughput and delay forecasting.

The KPIs presented were selected according to the best knowledge of the telecommunication experts and based on previous research.
The variables can be divided in two areas:

- traffic conditions: #UEs (number of User Equipments), DV (Data Volume), PRB (Physical Resource Block) utilization,
- environmental conditions: CQI (Channel Quality Indicator) and BLER (Block Error Rate).

The metrics marked in Fig. 2 as *Total* are calculated for all slices and the metrics marked as *Slice* are calculated for each slice separately. Thanks to them, it is possible to create multivariate models that include traffic load and radio environment metrics [11] that have a direct effect on throughput and delay.

2.4 Selection of busy hour

For predictions that span several months, examining hourly patterns may introduce extraneous details. This study aims to estimate the maximum future capacity requirements. Consequently, we have focused on analyzing the delay and throughput during peak network activity periods (over 24 hours). Hence, we identified the Busy Hour (BH) for each day, discarding data from less active hours for both training and evaluating of our model. The BH was determined on the basis of the maximum PRB utilization, which indicates the maximum load on the radio interface. Furthermore, we examine the effects of choosing every BH from each day in the data set compared to selecting only those BHs that exhibit the highest load.

2.5 Forecasting method

In [10], it has been demonstrated that VARMAX serves as an effective approach for short-term prediction, and this study further explores its long-term prediction capabilities. A time-series $\{\boldsymbol{X_t}\}$ is a m-variate ARMA(p,q) process (called also vector ARMA, VARMA) if it is formulated in the following way:

$$\Phi(B)\boldsymbol{X_t} = \Theta(B)\boldsymbol{Z_t}, \tag{1}$$

where $\{X_t\}$ is a stationary solution of difference equations (1), where

- $\Phi(z) := I - \Phi_1 z - ... - \Phi_p z^p$, where $\Phi_1, ..., \Phi_p$ are $m \times m$ matrices,
- $\Theta(z) := I - \Theta_1 z - ... - \Theta_q z^q$, where $\Theta_1, ..., \Theta_q$ are $m \times m$ matrices.

Moreover:

- I is $m \times m$ identity matrix,
- B is the backward shift operator (for more details, see Sec. 1.4 in [12], p. 19),
- $\{\boldsymbol{Z_t}\}$ is multivariate white noise sequence.

For more details on the VARMA model, the definition of multivariate white noise, and methods for estimating parameters, see [12]. In this research, the VARMAX model has been used, that is, the vector ARMA model with additional exogenous components. The tools used enable fit of the multivariate model using the Maximum Likelihood Estimate (MLE) under the assumption that $\{\boldsymbol{Z_t}\}$ has multivariate Gaussian distribution (for more details, see [13, 14]).

Data preprocessing starts with seasonal decomposition to achieve stationarity. The MLE-based procedure selects the model coefficients for the appropriate order of the models (p, q). Different pairs (p, q) have been evaluated and the one that minimizes the information criteria has been selected. The information criteria considered are as follows: Akaike Information Criterion, Bayesian Information Criterion, Hannan-Quinn Information Criterion [15]. Finally, the VARMAX(p, q) model is fitted for the selected p and q.

2.6 Information about configuration

Training of the model with historical data for each cell separately will create a model that will accurately predict its future values. However, when there is a short history of data or the traffic within cell is not saturated, then such data might be insufficient, and the prediction error is higher. Therefore, we have investigated the approach to generalize the model by training on a larger set of cells. The frequency band and bandwidth were chosen as a criterion to group cells for a unified model, and these criteria characterize the overall conditions of the radio interface regarding propagation features and spectrum availability. Configuration combinations available in the data set are shown in Tab. 2.

Tab. 2: Frequency band and channel bandwidth configurations in data set.

	Frequency band [Mhz]	**Channel bandwidth [Mhz]**
Band-1	2500(B41)	80, 100
Band-2	600(B71)	15, 20

3. RESULTS

The research on methods used in short-term predictions from [10] has been continued to evaluate its applicability to long-term forecasting relevant for many dimensioning

Tab. 3: The variables used in the modeling.

Abbreviation	**Full name**	**Unit**	**Description**
#UEs	Number of user equipments	#	The average number of user equipments which have buffered data in downlink direction
BLER	Block Error Rate	%	A ratio of the number of erroneous blocks to the total number of transmitted blocks
PRB utilization	5G PRB utilization for PDSCH	%	Utilization of PRBs for physical downlink shared channel (PDSCH)
CQI	Channel Quality Indicator	#	This indicates the average level of modulation and coding the UE could operate
DV	Data Volume	kbit	Amount of data send per particular network slice
-	Delay	microsecond	Calculated as time difference between the reception of the RLC SDU from PDCP layer and when first RLC SDU is sent over the air interface
TPut	Throughput	kbit/ms	Average downlink throughput volume at PDCP SDU level for a given network slice

use cases (e.g., estimation of time to reach capacity limit). Our goal was to propose a high-quality solution that will have a prediction error similar to that of short-term forecasts.

3.1 Model per cell

As explained in Sec. 2.4, for long-term forecasting, we focus on BH. An example of a delay and throughput prediction for Slice C is shown in Fig. 3 and Fig. 4 (the same cell as in Fig. 1). It can be seen that the model predictions are correct, even though the signal characteristics have changed completely (e.g. by smoothing seasonal trends during aggregation to BH).

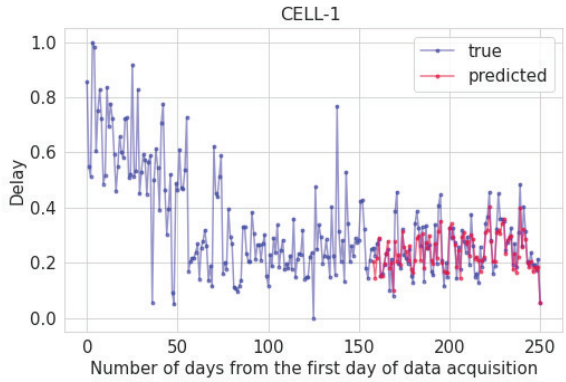


Fig. 3: The long-term forecast for normalized delay for Slice C using VARMAX(1,0) for illustrative cell.

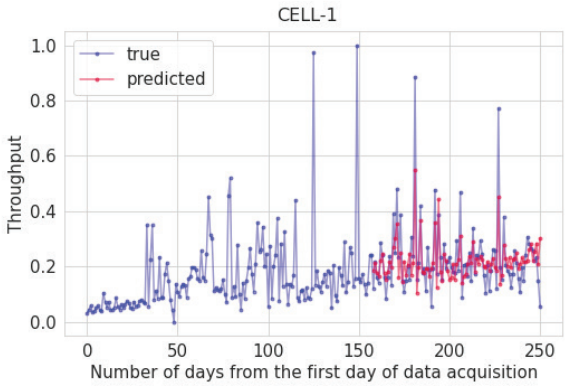


Fig. 4: The long-term forecast for normalized throughput for Slice C using VARMAX(1,0) for illustrative cell.

Tab. 4 presents the mean normalized Mean Absolute Error across predictions for the 185 cells included in the analysis, as well as for cells under high load. High-loaded cells (HL) have high PRB utilization (that is, 0.95 quantile with PRB utilization $> 80\%$). These cells have been analyzed separately because they show the behavior of BTS during the busiest times, which is particularly important for dimensioning process. The high precision of the results for each slice demonstrates that this technique is effective in predicting delay and throughput, which significantly influence the slice QoS, over a long-term horizon of approximately 3 months.

Tab. 4: The mean of nMAE (± standard deviation) corresponding to the long-term forecasts made for all and high-loaded cells.

Variable	Cells	A	B	C
Delay	**All**	0.100 ± 0.043	0.094 ± 0.043	0.010 ± 0.046
	HL	0.098 ± 0.054	0.087 ± 0.040	0.087 ± 0.047
Throughput	**All**	0.109 ± 0.045	0.112 ± 0.041	0.097 ± 0.036
	HL	0.127 ± 0.053	0.116 ± 0.041	0.073 ± 0.027

3.2 Model per frequency band

Moving from the model created for the single slice in the cell (related work: [10]) we have reached the point where we start to analyze the behavior of the network for aggregated cells. We look at aggregated cells for a single band, which allows us to gain new knowledge about the future values of KPIs within a frequency band. The modeling for the individual band is made as follows:

- Take the cells that correspond to the particular band.
- For each exogenous and endogenous variable, calculate the average value over time.
- Fit the model to the trajectory obtained.

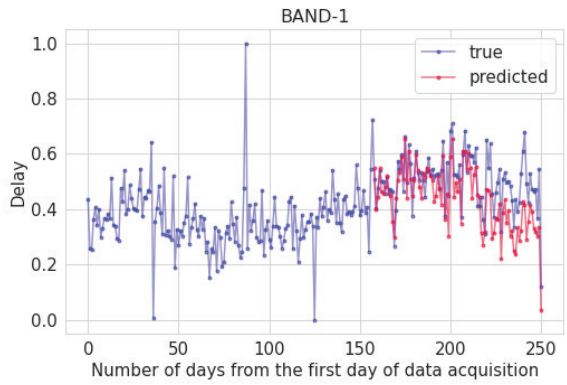


Fig. 5: The long-term forecast for normalized delay for Slice C using VARMAX(1,0) for band-1.

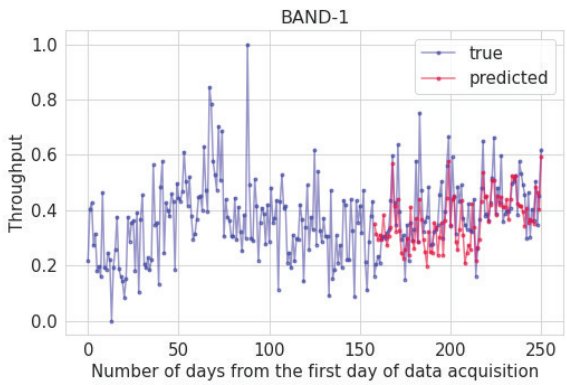


Fig. 6: The long-term forecast for normalized throughput for Slice C using VARMAX(1,0) for band-1.

As mentioned earlier, cells with high PRB utilization are especially interesting. Forecasts for illustrative bands (for which only highly loaded cells were taken into modeling) are visible in Fig. 5 – Fig. 8. It can be seen that the

predictive model is effective in both cases, despite different data characteristics. It is worth adding that band 1 (Fig. 5, Fig. 6) corresponds to the cell in Fig. 1 and Fig. 3.

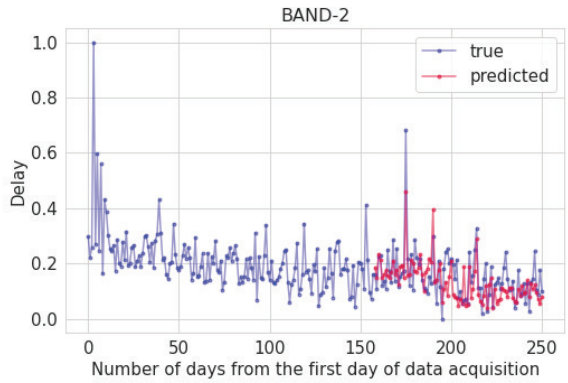


Fig. 7: The long-term forecast for normalized delay for Slice C using VARMAX(1,0) for band-2.

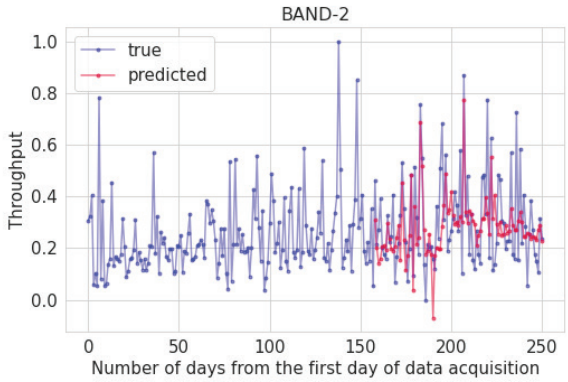


Fig. 8: The long-term forecast for normalized throughput for Slice C using VARMAX(1,0) for band-2.

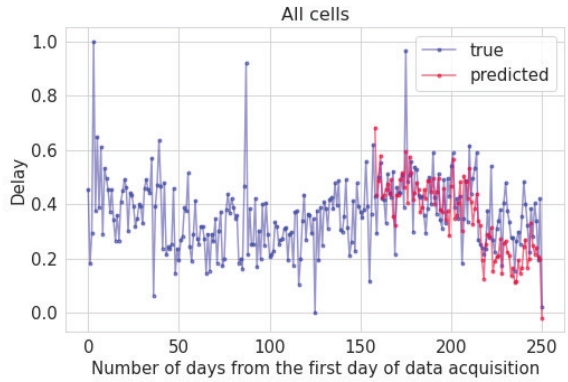


Fig. 9: The long-term forecast for normalized delay for Slice C using VARMAX(1,0) for all high-loaded cells in the network.

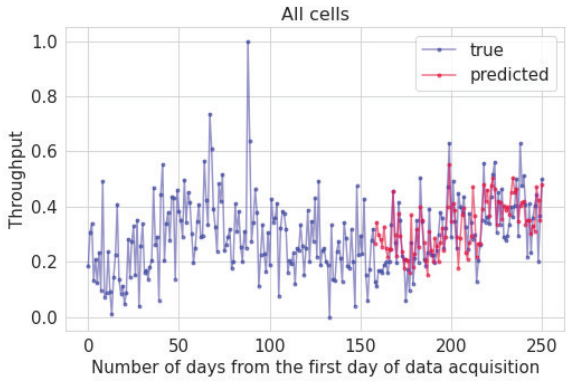


Fig. 10: The long-term forecast for normalized throughput for Slice C using VARMAX(1,0) for all high-loaded cells in the network.

Tab. 5 shows the nMAE values corresponding to individual slices and bands. It can be seen that the error values differ slightly between bands. However, the nMAEs are still small and in most cases lower than in the model per cell and for all cells. It is also worth adding that the delays errors are smaller because the KPI determining throughput is more noisy, which affects the quality of model fitting.

Tab. 5: The nMAE corresponding to the long-term forecasts made for selected bands.

Variable	Cells	A	B	C
Delay	**Band-1**	0.097	0.085	0.075
	Band-2	0.088	0.053	0.050
	All	0.121	0.073	0.088
Throughput	**Band-1**	0.209	0.141	0.073
	Band-2	0.122	0.093	0.114
	All	0.116	0.104	0.071

4. CONCLUSIONS

In this study, we demonstrate the utility of the multidimensional VARMAX model in forecasting telecommunication data at different levels of aggregation. Enhancements in previous models for short-term forecasting have been shown to extend their use to forecasts up to three months. Additionally, expanding the model to train on all cells within band data, rather than just individual cell data, has improved prediction accuracy. The results suggest viable applications in the planning and scaling of 5G networks. Empirical tests using real commercial network data have confirmed the practicality and reproducibility of predictive models and methodologies developed for the long-term planning of network slicing.

REFERENCES

[1] M. U. Khan, A. García-Armada, J. J. Escudero-Garzás, Service-based network dimensioning for 5g networks assisted by real data, IEEE Access 8 (2020) 129193–129212.

[2] M. Gutierrez, S. P. Puthenpurakel, S. Vasudevan, Network dimensioning: Next-gen ai-driven solutions - ericsson (Mar 2022).
URL https://www.ericsson.com/en/blog/2022/3/next-gen-ai-driven-network-dimensioning-solution

[3] W. Wu, C. Zhou, M. Li, H. Wu, H. Zhou, N. Zhang, X. S. Shen, W. Zhuang, Ai-native network slicing for 6g networks, IEEE Wireless Communications 29 (1) (2022) 96–103.

[4] A. A. Barakabitze, A. Ahmad, R. Mijumbi, A. Hines, 5g network slicing using sdn and nfv: A survey of taxonomy, architectures and future challenges 167 (C) (2020).

[5] M. Panek, I. Jabłoński, M. Woźniak, Modeling configuration-performance relation in a mobile network: a data-driven approach, IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (2–5 September 2024, Valencia, Spain (Accepted)).

[6] J. Huang, W. Guo, R. Wei, M. Yan, Y. Hu, T. Qin, Short-term power forecasting method for 5g photovoltaic base stations on non-sunny days based on sdn-integrated ingo-bp and rgan, IET Renewable Power Generation (2024).

[7] Y. Neuman, Y. Cohen, B. Tamir, Short-term prediction through ordinal patterns, Royal Society Open Science 8 (1) (2021) 201011.

[8] X. Shen, J. Gao, W. Wu, K. Lyu, M. Li, W. Zhuang, X. Li, J. Rao, Ai-assisted network-slicing based next-generation wireless networks, IEEE Open Journal of Vehicular Technology 1 (2020) 45–66.

[9] M. Panek, A. Pomykała, I. Jabłoński, M. Woźniak, 5g/5g+ network management employing ai-based continuous deployment, Applied Soft Computing 134 (2023) 109984.

[10] D. Dulas, J. Witulska, A. Wyłomańska, I. Jabłoński, K. Walkowiak, Data-driven model for sliced 5g network dimensioning and planning, featured with forecast and" what-if" analysis, IEEE Access 12 (2024) 50067 – 50082.

[11] S. Roy, D. Applegate, Z. Ge, A. Mahimkar, S. Pathak, S. Puthenpura, Quantifying the service performance impact of self-organizing network actions, in: 2016 12th International Conference on Network and Service Management (CNSM), 2016, pp. 37–45.

[12] P. J. Brockwell, Davis, Time series: theory and methods, Springer Series in Statistics, 1991.

[13] H. Lütkepohl, New introduction to multiple time series analysis, Springer Science & Business Media, 2005.

[14] Documentation statsmodels 0.15.0, https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.varmax.VARMAX.html, [Accessed 22-05-2024].

[15] M. Chapman, A Meta-Analysis of Metrics for Change Point Detection Algorithms, Spring, 2017.