МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ “ЛЬВІВСЬКА ПОЛІТЕХНІКА”

Кваліфікаційна наукова
праця на правах рукопису

РУДА ХРИСТИНА СТЕПАНІВНА

УДК 004.056.55

ДИСЕРТАЦІЯ

УДОСКОНАЛЕННЯ МЕТОДІВ ТА ЗАСОБІВ БІОМЕТРИЧНОЇ АВТЕНТИФІКАЦІЇ ЗА ГОЛОСОМ З ЗАСТОСУВАННЯМ ТЕХНОЛОГІЙ МАШИННОГО НАВЧАННЯ

125 Кібербезпека

12 “Інформаційні технології”

Подається на здобуття наукового ступеня доктора філософії
Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело _____ /Руда Христина Степанівна/

Науковий керівник: Микитин Галина Василівна, доктор технічних наук, професор

Львів – 2025

АНОТАЦІЯ

Руда Х.С. **Удосконалення методів та засобів біометричної автентифікації за голосом з застосуванням технологій машинного навчання.** – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю “125 – Кібербезпека. – Національний університет “Львівська політехніка” Міністерства освіти і науки України, Львів, 2025.

У сучасних умовах стрімкого розвитку інформаційних технологій та глобальної цифровізації питання надійної автентифікації особи набуває особливої актуальності. Одним із найбільш перспективних напрямів у цій сфері є біометрична автентифікація за голосом, що базується на унікальних акустичних характеристиках мовлення. Проте, з удосконаленням методів синтезу мовлення на основі штучного інтелекту, зокрема технологій клонування голосу, значно зростають ризики реалізації атак типу voice spoofing, що створює нові виклики для забезпечення безпеки біометричних систем.

У дисертаційній роботі обґрунтовано доцільність використання нейромережевих технологій машинного навчання для підвищення точності та стійкості голосової біометричної автентифікації. Запропоновано концептуальну модель системи, що реалізує поєднання класичних методів обробки мовного сигналу з сучасними нейронними архітектурами (ECAPA-TDNN, TitaNet, WavLM) і використовує векторні ембедінги для представлення голосових ознак. Проведено формалізацію основних етапів автентифікаційного процесу, зокрема відбору мовного сигналу, його попередньої обробки, екстракції ознак та верифікації особи на основі порівняння ембедінгів.

Розроблено та реалізовано експериментальну методику оцінювання ефективності системи в умовах spoofing-атак із використанням синтетичних голосових зразків, згенерованих за допомогою сучасних моделей голосового клонування (RVC, VITS, XTTS, ElevenLabs). Проведено аналіз стійкості до атак із різними типами текстового контенту та здійснено порівняння результативності

нейромережевих архітектур. Досліджено вплив масштабування системи та варіативності мовців на рівень точності верифікації.

Отримані результати засвідчують доцільність використання нейромережевих моделей для побудови високоточних, продуктивних та стійких до синтетичних атак систем голосової автентифікації.

Об’єктом дослідження є система біометричної автентифікації за голосом, яка забезпечує збір, обробку та аналіз мовних сигналів з метою верифікації особи, а також дозволяє оцінювати її стійкість до атак із використанням синтетичних голосів, згенерованих за допомогою нейромережевих технологій.

Предмет дослідження охоплює методи та засоби підвищення точності та стійкості голосових біометричних систем, зокрема нейромережеві моделі побудови ембедінгів, алгоритми порівняння мовних ознак та підходи до виявлення атак із використанням синтезованого мовлення.

У дослідженні застосовано методи: математичної статистики, машинного навчання, порівняльного аналізу, аналітичного й експериментального моделювання, оптимізації та цифрової обробки мовлення.

У першому розділі **«Аналіз особливостей процесу розпізнавання людини за голосом»** здійснено комплексний аналіз історичних передумов, теоретичних засад та практичних підходів до побудови систем біометричної автентифікації за голосом. Проведено класифікацію методів голосової автентифікації за принципами їх реалізації – від класичних і статистичних моделей до сучасних нейромережевих архітектур і ембедінг-орієнтованих підходів. Особливу увагу приділено порівнянню текстозалежних і текстонезалежних систем, виявлено їх переваги, обмеження та доцільність використання залежно від прикладного контексту. Окремим напрямом розглянуто сучасні загрози для голосової біометрії, зумовлені поширенням технологій клонування мовлення, зокрема Voice Conversion і Text-to-Speech Cloning. Деталізовано архітектурні особливості відповідних моделей та проаналізовано їхній потенціал для реалізації атак типу voice spoofing. Наголошено на необхідності впровадження антиспуфінгових механізмів, а також розглянуто

етичні виклики, пов’язані з використанням синтетичних голосів, що формує підґрунтя для подальшого дослідження стійкості систем до подібних загроз.

У другому розділі **«Концептуальна модель голосової автентифікації на основі нейромережевих трансформерів»** запропоновано концептуальну модель сучасної системи голосової біометрії, що інтегрує класичні та нейромережеві підходи до обробки мовного сигналу. Розглянуто ключові етапи процесу автентифікації: від відбору й попередньої обробки голосового зразка до формування голосового ембедінгу та здійснення верифікації шляхом порівняння тестового і реєстраційного шаблонів. Обґрунтовано ефективність використання глибоких моделей, зокрема ECAPA-TDNN, TitaNet і WavLM, у задачах побудови стійких до шуму і варіативності вхідних даних представлень користувача. Проведено порівняльний аналіз класичних (MFCC, LPC, i-vector) та сучасних нейромережевих методів екстракції ознак, показано їхні переваги й недоліки. Особливу увагу приділено проблемам безпеки, впровадженню стандартів ISO/IEC 30107 і C2PA, а також формуванню моделей, здатних протистояти атакам із застосуванням синтетичних голосів. Запропонована концептуальна модель забезпечує високий рівень точності, адаптивності й захищеності, що створює підґрунтя для її практичного застосування у критично важливих інформаційних системах.

У третьому розділі **«Імплементація вдосконалення системи голосової автентифікації»** здійснено практичну реалізацію та всебічний аналіз роботи вдосконаленої системи біометричної автентифікації за голосом. Основну увагу зосереджено на побудові архітектури системи на основі ембедінгів, де описано механізми реєстрації голосових профілів, налаштування порогових значень та процедуру верифікації користувачів. Запропоновано підхід до формування стійкого до варіативності мовлення голосового «відбитка» шляхом усереднення ембедінгів, що забезпечує точність і стабільність при зміні емоційного стану, темпу мовлення чи акустичних умов. Особливу увагу приділено метрикам подібності векторів ознак, проведено порівняння властивостей косинусної, евклідової, мангеттенської та Махаланобісової відстаней, із акцентом на доцільності

використання косинусної відстані як найбільш стійкої до акустичних флуктуацій. Деталізовано модулі обробки тестових зразків, екстракції ембедінгів за допомогою моделей ECAPA, TitaNet, WavLM, а також процедури прийняття рішень на основі обчисленої відстані. Окремий підрозділ присвячено формуванню аудіодатасету для експериментів та оцінювання ефективності системи. Запропоновано методику побудови тестових пар зразків для розрахунку FAR і FRR, проведено калібрування порогів для кожної моделі з урахуванням її метричних властивостей. Запропонована методологія дозволила забезпечити адаптивність системи до різних моделей ембедінгів і покращити загальну точність класифікації. Таким чином, розділ формує технічне підґрунтя для реалізації надійної системи голосової автентифікації, що поєднує сучасні нейромережеві методи та практичні механізми їх застосування у реальних сценаріях використання. Отримані результати засвідчують потенціал запропонованої архітектури до інтеграції у високонадійні інформаційні системи з підвищеними вимогами до безпеки.

У четвертому розділі **«Оцінювання стійкості системи до атак типу voice spoofing з використанням технологій клонування голосу»** представлено експериментальну методику перевірки захищеності біометричної системи автентифікації за голосом у сценаріях, що моделюють цілеспрямовані атаки із застосуванням синтетичного мовлення. Для дослідження було згенеровано масштабний корпус аудіозразків за допомогою моделей RVC, XTTS, ElevenLabs і Tortoise, що імітували голоси зареєстрованих користувачів у системі. Розглянуто два сценарії атаки: з використанням ідентичного текстового контенту (повторення реєстраційних фраз) та з новими фразами, не використаними під час навчання. Здійснено класифікацію результатів автентифікації на основі косинусної подібності ембедінгів, отриманих за допомогою моделей ECAPA, TitaNet, WavLM та інших. Продемонстровано залежність точності класифікації від типу клонувальної технології та сценарію текстового наповнення. Окрему увагу приділено впровадженню детекції синтетичного мовлення та оцінці ефективності когнітивних факторів як додаткового рівня захисту. Отримані результати підтверджують актуальність проблеми протидії генеративним атакам і слугують

основою для розроблення більш стійких до маніпуляцій архітектур біометричних систем автентифікації.

У **висновках** дисертаційної роботи викладено основні результати і рекомендації, які випливають з проведених досліджень, представлено та охарактеризовано кількісні оцінки показників ефективності в умовах використання запропонованих рішень.

У **додатках** до дисертації долучено акти впровадження результатів дисертаційної роботи, а також список наукових праць і апробацій автора за темою дисертації.

Ключові слова: кібербезпека, біометрична автентифікація, голосова біометрія, штучний інтелект, синтетичне мовлення, voice spoofing, клонування голосу, ембедінги, антиспуфінг, нейронні мережі, машинне навчання, верифікація особи, косинусна відстань, захист біометричних даних.

SUMMARY

Ruda K.S. ADVANCEMENT OF METHODS AND TOOLS FOR VOICE-BASED BIOMETRIC AUTHENTICATION USING MACHINE LEARNING TECHNOLOGIES. – Qualifying scientific work on the rights of the manuscript. Dissertation for obtaining the scientific degree of Doctor of Philosophy in the specialty 125 " Cyber Security". - Lviv Polytechnic National University, Lviv, 2025.

In the context of rapid advancements in information technologies and global digitalization, the issue of reliable identity authentication is of critical importance. One of the most promising directions in this domain is voice-based biometric authentication, which leverages the unique acoustic characteristics of human speech. However, the development of AI-powered speech synthesis methods, particularly voice cloning technologies, significantly increases the risks of voice spoofing attacks, thereby introducing new challenges for the security of biometric systems.

This dissertation substantiates the feasibility of applying neural machine learning technologies to enhance the accuracy and robustness of voice biometric authentication systems. A conceptual system model is proposed that combines classical speech signal processing techniques with modern neural architectures (ECAPA-TDNN, TitaNet, WavLM), utilizing vector embeddings to represent voice features. The main stages of the authentication process are formalized, including speech signal acquisition, preprocessing, feature extraction, and identity verification through embedding comparison.

An experimental methodology for evaluating the system's performance under spoofing attacks using synthetic voice samples generated by state-of-the-art voice cloning models (RVC, VITS, XTTS, ElevenLabs) was developed and implemented. The system’s resilience was analyzed with respect to varying text content and different cloning technologies. Comparative assessments of neural architectures were performed, along with analysis of system scalability and speaker variability on verification accuracy.

The obtained results demonstrate the appropriateness of neural models for developing high-accuracy, high-performance voice authentication systems that are resilient to synthetic speech attacks.

The object of the study is a voice-based biometric authentication system that enables the acquisition, processing, and analysis of speech signals for identity verification, while also evaluating its resistance to attacks involving neural network-generated synthetic voices.

The subject of the study includes methods and tools for improving the accuracy and robustness of voice biometric systems, particularly neural embedding models, speech similarity algorithms, and approaches for detecting synthesized speech attacks.

The research employs methods of mathematical statistics, machine learning, comparative analysis, analytical and experimental modeling, optimization, and digital speech signal processing.

In Chapter 1, *"Analysis of Voice-Based Identity Recognition Systems"*, a comprehensive review is conducted on the historical development, theoretical foundations, and practical implementations of voice biometric authentication systems. The methods are classified by implementation principles, from classical and statistical models to modern neural and embedding-based approaches. Particular attention is given to comparing text-dependent and text-independent systems, highlighting their respective advantages, limitations, and applicability. The chapter also explores contemporary threats posed by the proliferation of speech cloning technologies, such as Voice Conversion and Text-to-Speech (TTS) Cloning. The architectural features of such models are detailed, and their potential for enabling voice spoofing attacks is analyzed. The need for anti-spoofing mechanisms is emphasized, alongside a discussion of the ethical implications of synthetic voice usage, forming a foundation for the system robustness investigations that follow.

Chapter 2, *"Conceptual Model of Voice Authentication Based on Neural Transformer Architectures"*, presents the conceptual model of an architecture of a modern voice biometric system that integrates classical and neural approaches to speech signal processing. The key stages of the authentication process are covered, from voice sample acquisition and preprocessing to embedding generation and template-based verification. The use of deep models–specifically ECAPA-TDNN, TitaNet, and WavLM–is justified for constructing speaker representations that are robust to noise and variability.

A comparative analysis of classical (MFCC, LPC, i-vector) and neural-based feature extraction methods is presented, outlining their respective strengths and weaknesses. Special attention is given to security issues, the integration of ISO/IEC 30107 and C2PA standards, and the development of models resistant to synthetic voice attacks. The proposed model ensures high accuracy, adaptability, and security, laying a solid foundation for deployment in mission-critical information systems.

Chapter 3, "Implementation of Improvements to the Voice Authentication System," presents the practical realization and detailed analysis of the enhanced voice-based biometric authentication system. The focus is placed on the system architecture built upon embedding-based representations, including modules for user registration, threshold calibration, and verification procedures. A robust method for constructing a user's voiceprint is proposed by averaging multiple embeddings, which ensures greater stability and accuracy in the presence of emotional, acoustic, or articulation variability. Particular emphasis is placed on similarity metrics used for comparing feature vectors. A comparative analysis of cosine, Euclidean (L2), Manhattan (L1), and Mahalanobis distances is conducted, demonstrating the advantages of cosine distance due to its invariance to scale and robustness to acoustic noise. The verification pipeline is described in detail, including the preprocessing of test samples, feature embedding extraction using ECAPA, TitaNet, and WavLM models, and the final decision-making process based on similarity scoring. A separate subsection is devoted to the construction of the experimental dataset used to evaluate the system's performance. A methodology for generating positive and negative test pairs is outlined to compute False Acceptance Rate (FAR) and False Rejection Rate (FRR), and individual threshold values are calibrated for each embedding model based on its statistical properties. This individualized calibration approach enhances the system's adaptability and accuracy. In summary, this chapter establishes the technical foundation for a resilient voice authentication system that combines state-of-the-art neural architectures with practical mechanisms for deployment. The results confirm the feasibility of embedding-based authentication in high-security applications and provide empirical validation of the proposed methods in real-world scenarios.

Chapter 4, *"Evaluation of System Resilience to Voice Spoofing Attacks Using Voice Cloning Technologies"*, presents an experimental framework for assessing the security of voice biometric authentication systems under targeted spoofing scenarios involving synthetic speech. A large-scale corpus of cloned audio samples was generated using four state-of-the-art voice cloning models–RVC, XTTS, ElevenLabs, and Tortoise–each designed to imitate the voices of enrolled users. The experiments covered two attack scenarios: one using phrases identical to those from the enrollment stage, and another with novel, previously unseen texts. Authentication decisions were made based on cosine similarity of embeddings extracted by models such as ECAPA, TitaNet, and WavLM. The results demonstrate the system's vulnerability depending on the cloning method and the linguistic context. Additional focus is placed on the integration of synthetic speech detection mechanisms and the role of cognitive factors as an auxiliary defense layer. The findings highlight the critical need for robust anti-spoofing strategies and contribute to the advancement of more resilient biometric system architectures capable of withstanding AI-driven threats.

The **conclusions** of the dissertation summarize the key findings and recommendations derived from the conducted research, and provide a detailed quantitative assessment of the system's performance indicators under the proposed solutions.

The **appendices** to the dissertation include implementation reports of the research results, as well as a list of the author's scientific publications and conference presentations related to the dissertation topic.

Keywords: cybersecurity, biometric authentication, voice biometrics, artificial intelligence, synthetic speech, voice spoofing, voice cloning, embeddings, anti-spoofing, neural networks, machine learning, speaker verification, cosine similarity, biometric data protection.

ПЕРЕЛІК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Наукові праці, в яких опубліковано наукові результати дисертації:

1. Дудикевич В. Б., Микитин Г. В., **Руда Х. С.** Застосування глибокого навчання для виявлення deepfake модифікацій біометричного зображення // Сучасна спеціальна техніка. – 2022. – № 1(68). – С. 13–22.

2. Dudykevych V., Yevseiev S., Mykytyn G., **Ruda K.,** Hulak H. Detecting deepfake modifications of biometric images using neural networks // CEUR Workshop Proceedings. – 2024. – Vol. 3654 : Cybersecurity providing in information and telecommunication systems 2024. Proceedings of the workshop cybersecurity providing in information and telecommunication systems (CPITS 2024), Kyiv, Ukraine, February 28, 2024 (online). – P. 391–397.

3. Zaiets I., Brydinskyi V., Sabodashko D., Khoma Y., **Ruda K**. Integrated system for speaker diarization and intruder detection using speaker embeddings // CEUR Workshop Proceedings. – 2024. – Vol. 3654 : Cybersecurity providing in information and telecommunication systems 2024. Proceedings of the workshop cybersecurity providing in information and telecommunication systems (CPITS 2024), Kyiv, Ukraine, February 28, 2024 (online). – P. 228–238.

4. Заєць І. С., Бридінський В. А., Сабодашко Д. В., **Руда Х. С.**, Хома Ю. В., Швед М. Є. Використання ембедінгів голосу в інтегрованих системах для діаризації мовців та виявлення зловмисників // Комп’ютерні системи та мережі. – 2024. – Вип. 6, № 1. – С. 54–66.

5. Микитин Г. В., **Руда Х. С**. Концептуальний підхід до виявлення deepfake-модифікацій біометричного зображення засобами нейронних мереж // Комп’ютерні системи та мережі. – 2024. – Вип. 6, № 1. – С. 124–132.

6. Микитин Г. В., **Руда Х. С.**, Яремчук Ю. Є. Методологія безпеки нейромережевих інформаційних технологій виявлення deepfake модифікацій біометричного зображення // Вісник Вінницького політехнічного інституту. – 2024. – № 1(172). – С. 74–80.

7. **Руда Х. С.,** Сабодашко Д. В., Микитин Г. В., Швед М. Є., Бордуляк С. М., Коршун Н. Порівняння методів цифрової обробки сигналів та моделей

глибинного навчання у голосовій аутентифікації // Кібербезпека: освіта, наука, техніка. – 2024. – № 5(25). – С. 140–160.

8. **Руда Х. С**. Дослідження масштабованості біометричних систем автентифікації на основі вбудовування голосу // Social Development and Security. – 2025. – Vol. 15, iss. 1. – P. 161–170.

Наукові праці, які засвідчують апробацію матеріалів дисертації:

9. **Руда Х. С**. Використання візуальних і неявних артефактів для виявлення deepfake-модифікацій у біометричних зображеннях // Сучасні методи, інформаційне, програмне та технічне забезпечення систем керування організаційно-технічними та технологічними комплексами : матеріали VIII Міжнародної науково-технічної Internet-конференції (Київ, 26 листопада 2021 року). – 2021. – С. 207–208.

10. Dudykevych V., Mykytyn H., **Ruda K.** The concept of a deepfake detection system of biometric image modifications based on neural networks // 2022 IEEE 3rd KhPI Week on Advanced Technology : conference proceedings, Kharkiv, 3–7 October 2022. – 2022. – P. 585–588.

11. Микитин Г. В., **Руда Х. С**., Хома Ю. В. Categorization of deepfake methods: a comparison of single-shot and multi-shot transfer approaches // Захист інформації і безпека інформаційних систем : матеріали IX Міжнародної науково-технічної конференції (Львів, 25–26 травня 2023 р.). – 2023. – С. 109–110.

12. Сабодашко Д. В., **Руда Х. С**., Швед М. Є., Бордуляк С. М. Технологія ембедінгів голосу та діаризація мовців // XX International Scientific and Practical Conference «Scientific Research: Modern Challenges and Prospects» : collection of abstracts, April 24–26, 2024, Prague, Czech Republic. – 2024. – P. 72–74.

13. **Руда Х. С.,** Микитин Г. В. Методи детекції синтетичних аудіозаписів, створених системами клонування голосу // Scientific Research in the Age of Virtual Reality: Exploring New Frontiers : collection of abstracts of LII International Scientific and Practical Conference (December 18–20, 2024), Montreal, Canada. – 2024. – P. 132–136.

Наукові праці, які додатково відображають наукові результати дисертації.

14. Mykytyn H., **Ruda K.,** Shved M. The paradigm of building secure information technologies for detecting deepfake modifications of biometric images based on neural networks // *Informatyka techniczna і sztuczna inteligencja: колективна монографія*. – Chapter in a collective monograph. – Bielsko-Biała: Wydawnictwo naukowe Uniwersytetu Bielsko-Bialskiego, 2024. – P. 43–50.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

MFCC (*Mel-Frequency Cepstral Coefficients*) – коефіцієнти мел-частотного кепстрального перетворення.

FFT (*Fast Fourier Transform*) – швидке перетворення Фур'є.

LPC (*Linear Predictive Coding*) – лінійне передбачення (кодування).

CNN (*Convolutional Neural Network*) – згорткова нейронна мережа.

RNN (*Recurrent Neural Network*) – рекурентна нейронна мережа.

LSTM (*Long Short-Term Memory*) – довготривала короткочасна пам’ять.

TTS (*Text-to-Speech*) – перетворення тексту в мовлення.

VC (*Voice Conversion*) – перетворення голосу.

GMM (*Gaussian Mixture Model*) – модель змішування Гауссівських розподілів.

HMM (*Hidden Markov Model*) – прихована марківська модель.

FAR (*False Accept Rate*) – частка хибних допусків.

FRR (*False Reject Rate*) – частка хибних відмов.

EER (*Equal Error Rate*) – точка рівності помилок (коли FAR = FRR).

DCF (*Detection Cost Function*) – функція вартості детекції.

SVM (*Support Vector Machine*) – метод опорних векторів

АЦП (*аналогово-цифрове перетворення*) — перетворення аналогового аудіосигналу у цифровий формат.

VAD (*Voice Activity Detection*) — виявлення голосової активності; алгоритм для визначення ділянок мовлення у звуковому потоці.

ШІ (*штучний інтелект*) — галузь комп’ютерних наук, що розробляє системи, здатні до навчання, прогнозування та прийняття рішень.

ВСТУП

Актуальність. Сучасний цифровий ландшафт стрімко змінюється під впливом розвитку технологій штучного інтелекту, автоматизації, глобальної комунікації та зростаючого обсягу персональних даних, що передаються через мережі. Автентифікація користувача – ключовий компонент у будь-якій системі безпеки, яка функціонує в умовах недовіри: від електронного банкінгу до дистанційного доступу до критичних інфраструктур [1]. Зважаючи на слабкість традиційних методів автентифікації (паролі, PIN-коди, токени), біометричні системи стають основним інструментом встановлення цифрової ідентичності, особливо в секторах, де потрібно одночасно забезпечити зручність, захищеність і безперервність доступу [2].

Голосова біометрія, як різновид поведінкової автентифікації, має низку переваг порівняно з іншими біометричними ознаками: вона не потребує фізичного контакту, забезпечує автентифікацію на відстані, і є природною для користувача в комунікаційних середовищах (телефонія, віртуальні помічники, смарт-пристрої) [3]. За оцінками аналітичної компанії MarketsandMarkets, глобальний ринок голосової біометрії зростатиме на понад 20% щороку та перевищить $3,9 млрд до 2026 року [4].

Однак це зростання супроводжується фундаментальними загрозами. Прогрес у генеративному моделюванні мовлення, зокрема в таких напрямках як Text-to-Speech (TTS), Neural Voice Cloning та Zero-shot synthesis, призвів до появи нових класів атак на біометричні системи – presentation attacks, у яких зловмисник використовує не реальний голос, а штучно згенерований аудіосигнал [5]. Дослідження демонструють, що сучасні TTS-моделі (наприклад, VITS, YourTTS, Bark, Vall-E, NaturalSpeech 2) досягають такого рівня натуральності, що їхній вихід складно відрізнити навіть експертам [6].

Проблема посилюється через масову доступність відкритого коду (наприклад, RVC, OpenVoice, XTTS) і моделей на кшталт ElevenLabs, які доступні навіть для нетехнічних користувачів. Цей фактор створює критичний ризик для

будь-якої системи, що покладається на голос як фактор автентифікації, адже поріг входу для зловмисників суттєво знижується [7].

Низка міжнародних організацій, зокрема NIST та FIDO Alliance, наголошують на потребі переосмислення стратегії розробки біометричних систем: сьогодні вони мають бути не лише точними, а й стійкими до штучного обману [8, 9]. У відповідь на ці виклики було започатковано низку ініціатив: наприклад, конкурси ASVspoof Challenge [10], проєкти EU Horizon (TReSPAsS, PriMa) [11] та документи ISO/IEC 30107-3:2023 і ISO/IEC 30136:2018, які встановлюють вимоги до оцінювання стійкості до атак та захисту біометричних шаблонів [12], [13].

Ще одним важливим фактором є зміна парадигми побудови біометричних систем. Якщо раніше основна увага зосереджувалася на екстракції ознак за допомогою статистичних методів, то сьогодні домінують глибокі нейронні архітектури: ECAPA-TDNN, Res2Net, TitaNet, WavLM, Whisper, BigVGAN. Вони не лише підвищують точність верифікації, але й відкривають можливості для побудови універсальних систем, здатних до мультимовного, текстонезалежного розпізнавання у складних акустичних умовах [14]

У цьому контексті постає комплексне науково-практичне завдання підвищення ефективності та безпеки систем біометричної автентифікації за голосом шляхом використання нейромережевих технологій, спрямоване на розроблення, реалізацію та верифікацію архітектурних і алгоритмічних рішень, здатних забезпечити:

1. високу точність автентифікації користувачів у реальних умовах експлуатації, з урахуванням варіативного мовлення, акустичного шуму, міжсесійних змін голосу та природної мовленнєвої спонтанності;
2. стійкість до атак типу voice spoofing, зокрема до генеративного синтетичного мовлення, створеного сучасними моделями клонування голосу (RVC, Tacotron, VITS, XTTS, ElevenLabs тощо);
3. практичну придатність до інтеграції у прикладні сценарії – від автономних мобільних застосунків до масштабованих хмарних API-рішень для автентифікації, з урахуванням обмежень продуктивності та реального часу;

4. відповідність міжнародним стандартам у сфері біометрії та кібербезпеки.

Розв’язання цього науково-практичного завдання, має не лише теоретичне, але й практичне значення для проєктування сучасних голосових біометричних систем, здатних працювати в умовах ризиків генеративного шахрайства. Це відкриває шлях до інтеграції біометричної автентифікації у масштабовані, багатокомпонентні середовища – банківські сервіси, eGovernment, системи eHealth, call-центри, хмарні платформи безпеки та військові IT-комплекси.

Зв’язок роботи з науковими програмами, планами, темами.

Дисертаційні дослідження виконано відповідно до наукового напряму кафедри захисту інформації Національного університету “Львівська політехніка”.

Основні положення та результати дисертаційної роботи впроваджені у навчальний процес кафедри Захист інформації Національного університету “Львівська політехніка” з вивчення дисципліни “Гарантоздатність автоматизованих систем” для студентів напрямку підготовки “125 – Кібербезпека та захист інформації”, освітньої програми “Системи технічного захисту інформації, автоматизація її обробки”

Окремі частини роботи виконано в межах держбюджетної науково-дослідної роботи: «Дослідження стійкості біометричних систем автентифікації до атак із застосуванням технології клонування голосу на основі глибинних нейронних мереж» (№ держреєстрації 0124U000407).

А також в діяльності комерційного підприємства UNIDATALAB LTD”.

Мета роботи. Метою дисертаційної роботи є удосконалення методів та засобів біометричної автентифікації за голосом шляхом побудови високоточних систем із використанням нейромережевих технологій машинного навчання та підвищення їх стійкості до атак із застосуванням синтетичних голосових зразків.

Завдання. Дисертаційна робота присвячена вирішенню актуального науково-практичного завдання з підвищення ефективності та безпеки систем біометричної автентифікації за голосом шляхом використання нейромережевих технологій, що забезпечують високу точність верифікації особи та стійкість до атак

із застосуванням клонованих голосів. Для успішного досягнення мети даної роботи необхідно виконати наступні завдання:

1. Провести комплексний аналіз сучасних методів голосової автентифікації, класифікувати їх за принципами роботи (класичні, статистичні, нейромережеві), дослідити особливості текстозалежних і текстонезалежних підходів та оцінити їх ефективність і стійкість до зовнішніх впливів.

2. Дослідити новітні загрози для систем голосової біометрії, що пов’язані з технологіями клонування голосу, провести огляд сучасних генеративних моделей (RVC, Tacotron, VITS тощо) і сформулювати ключові виклики для підвищення безпеки автентифікаційних систем.

3. Розробити концептуальну модель системи голосової автентифікації, що інтегрує структурно-функціональні блоки на основі нейромережевих технологій (CNN, LSTM, Transformers), визначити ключові етапи обробки мовного сигналу та забезпечити відповідність розробленої моделі міжнародним стандартам у сфері біометричної безпеки.

4. Провести порівняльний аналіз класичних та нейромережевих методів виділення та обробки голосових ознак (MFCC, LPC, Titanet, ECAPA-TDNN, WavLM), визначити їх ефективність у задачах побудови голосових ембедінгів і оцінити переваги трансформерних архітектур у забезпеченні високої точності та стійкості систем автентифікації.

5. Розробити архітектурну модель системи голосової автентифікації на основі ембедінгів, реалізувати модулі реєстрації, калібрування та порівняння голосових зразків, а також визначити оптимальні методи оцінювання схожості векторів ознак для забезпечення високої точності верифікації.

6. Провести експериментальне оцінювання точності, продуктивності та масштабованості системи голосової автентифікації з використанням нейромережевих моделей (TitaNet, ECAPA-TDNN, WavLM), дослідити вплив збільшення кількості зареєстрованих користувачів на ефективність системи та сформулювати практичні рекомендації для оптимізації порогових значень і стійкості до акустичних впливів.

7. Розробити експериментальну методику тестування стійкості системи голосової автентифікації до атак з використанням сучасних технологій клонування голосу (RVC, ElevenLabs, XTTS, Tortoise), створити відповідний тестовий датасет і провести порівняльний аналіз ефективності різних моделей ембедінгів у сценаріях атак із ідентичним і новим текстовим контентом.

8. Розробити, реалізувати та експериментально оцінити методи детекції синтетичного мовлення як засіб підвищення захищеності голосових біометричних систем; здійснити порівняльний аналіз ефективності фірмового рішення (ElevenLabs) та авторської моделі на основі MFCC-ознак і SVM-класифікатора; виявити їхні функціональні обмеження та переваги; сформулювати рекомендації щодо інтеграції універсальних антиспуфінгових модулів у загальну архітектуру системи автентифікації.

Об’єктом дослідження є система біометричної автентифікації за голосом, що виконує збір, обробку та аналіз мовних сигналів для верифікації особи, а також забезпечує дослідження стійкості до атак із використанням синтетичних голосових зразків на основі нейромережевих моделей і методів штучного інтелекту.

Предметом дослідження є методи та засоби підвищення точності й стійкості систем біометричної автентифікації за голосом, зокрема нейромережеві моделі побудови голосових ембедінгів, алгоритми порівняння мовних зразків та підходи до захисту від атак із використанням синтетичних голосів.

Методи дослідження. Виконання завдань наукової роботи здійснювалося з використанням таких методів: математичної статистики, порівняльного аналізу, машинного навчання, експериментального та аналітичного моделювання, оптимізації, а також обробки мовних сигналів.

Наукова новизна роботи полягає в тому, що:

1. Отримала подальший розвиток методологія класифікації та аналізу сучасних методів голосової автентифікації, що передбачає структуризацію систем за принципами дії (класичні, статистичні, нейромережеві), а також комплексне порівняння текстозалежних і текстонезалежних підходів. Це забезпечило

обґрунтовану основу для вибору архітектур, орієнтованих на стійкість і точність у реальних умовах застосування.

2. Вперше розроблено концептуальну модель системи голосової автентифікації, яка об’єднує структурно-функціональні блоки (реєстрація, обробка, верифікація) на основі нейромережевих технологій (CNN, LSTM, Transformers) з урахуванням міжнародних стандартів у сфері біометричної безпеки. Це створює технологічне підґрунтя для впровадження біометричних рішень у системах підвищеної надійності.

3. Вдосконалено архітектуру системи голосової автентифікації на основі ембедінгів, що включає реалізацію модулів порівняння, налаштування та реєстрації голосових зразків із використанням обґрунтованих метричних функцій. Це дозволяє досягти стабільної точності верифікаційних рішень за умов міжсесійних варіацій та масштабованих багатокористувацьких сценаріїв.

4. Вперше запропоновано та імплементовано експериментальну методику оцінювання стійкості до атак із застосуванням синтетично згенерованого мовлення, зокрема на основі генеративних моделей (RVC, ElevenLabs, XTTS, Tortoise). Створений тестовий корпус дозволив провести порівняльне оцінювання ефективності моделей ембедінгів у spoofing-сценаріях з різним лінгвістичним наповненням.

5. Вперше реалізовано та зіставлено підходи до детекції синтетичного мовлення на основі фірмового рішення ElevenLabs і авторської моделі MFCC+SVM. Отримані результати лягли в основу практичних рекомендацій щодо впровадження універсальних антиспуфінгових механізмів у архітектуру голосових біометричних систем.

Практичне значення одержаних результатів полягає у можливості їх безпосереднього застосування для підвищення точності та стійкості систем біометричної автентифікації за голосом, а також для удосконалення механізмів виявлення та запобігання spoofing-атакам у державних і комерційних інформаційно-комунікаційних системах**.**

1. Розроблена архітектура системи голосової автентифікації на основі ембедінгів забезпечила відповідність принципам надійності, масштабованості та обмеження доступу до біометричних шаблонів, визначеним у ISO/IEC 24745:2022 та ISO/IEC 19795-1:2021. Впровадження архітектури в експериментальному середовищі дозволило досягти точності автентифікації до 96%.
2. Розроблена методика тестування стійкості системи до spoofing-атак із використанням синтетичних голосів, згенерованих моделями RVC, XTTS, ElevenLabs та Tortoise, дозволяє верифікувати ефективність системи згідно з вимогами ISO/IEC 30107-3:2023 щодо тестування Presentation Attack Detection (PAD). Виявлено, що всі протестовані моделі демонструють вразливість до атак з застосуванням клонування голосу, що стало підставою для рекомендації впровадження детекторів синтетичного мовлення як обов’язкового компонента систем біометричної голосової автентифікації.
3. Імплементована модель виявлення синтетичного мовлення на основі MFCC+SVM забезпечує точність виявлення атак до 89.75%, при цьому точність класифікації справжніх голосів залишається низькою (25.96%), що свідчить про потребу у додатковому фільтруванні рішень. Модель відповідає вимогам ISO/IEC 30107-1:2023 до PAD-механізмів і рекомендована до застосування як додатковий рівень захисту в багатофакторних біометричних системах, що працюють із критичними даними.
4. Результати порівняльного тестування моделей TitaNet, ECAPA-TDNN, WavLM та PyAnnote можуть слугувати основою для обґрунтованого вибору архітектури в задачах біометричної автентифікації, орієнтованих на великомасштабні користувацькі бази. Зокрема, модель TitaNet продемонструвала найвищу стійкість до масштабування: при збільшенні кількості зареєстрованих користувачів від 10 до 70 середня точність знизилася лише на 3,55%, що вказує на ефективне збереження продуктивності в умовах зростання навантаження. Процедура оцінювання масштабованості здійснювалася відповідно до методологічних положень стандарту ISO/IEC 19795-2:2007, який регламентує

сценарне тестування з варіативною чисельністю суб’єктів як одного з ключових параметрів експериментального дизайну.

5. Оцінка ресурсної ефективності моделей TitaNet, ECAPA-TDNN та WavLM вказала на їхню придатність для розгортання в обмежених обчислювальних середовищах. Зокрема, модель WavLM-base-plus-sv забезпечує високоточну автентифікацію (93%) при середньому часі порівняння ембедінгів лише 4.34 мс, що задовольняє вимоги до латентності в edge-системах та пристроях реального часу. Отримані результати відповідають критеріям ефективності, визначеним у ISO/IEC 29167-1:2014, та засвідчують можливість застосування таких моделей у вбудованих біометричних рішеннях, мобільних пристроях і системах фізичного контролю доступу.

Наукові та практичні результати виконаних досліджень використані у навчальному процесі кафедри захисту інформації Національного університету “Львівська політехніка”, зокрема для студентів спеціальності “125 – Кібербезпека та захист інформації” в курсі лекцій з дисципліни “Гарантоздатність автоматизованих систем”.

Розроблені підходи до побудови архітектури систем впроваджено у професійну діяльність компанії UNIDATALAB LTD у контексті виконання прикладних завдань у рамках комерційних проєктів.

Особистий внесок. Важливі наукові результати, викладені у цій дисертаційній роботі, були здобуті автором самостійно. У публікаціях, підготовлених у співавторстві, провідна роль у формуванні ідей, реалізації експериментів та інтерпретації результатів належить здобувачеві.

Зокрема, у публікаціях [1, 2, 5, 10] автором було розроблено архітектуру та концептуальні засади побудови систем аналізу біометричних даних із використанням нейромережевих моделей. У працях [3, 4, 12] реалізовано модулі для виділення голосових ембедінгів, інтегрованих у багатокомпонентні системи для діаризації мовців, а також проведено експериментальну перевірку точності таких систем. У публікаціях [6, 11, 14] автором здійснено аналіз безпекових аспектів використання біометричних даних та класифікацію методів їх генерації,

що мають потенційну загрозу цілісності систем автентифікації. У роботах [7, 13] виконано порівняльний аналіз методів цифрової обробки мовного сигналу та побудовано підхід до оцінювання достовірності голосових зразків у контексті автентифікації користувачів. У праці [8] здобувач провів експериментальне дослідження масштабованості біометричних систем із різною кількістю зареєстрованих користувачів, з аналізом впливу розмірів вибірки на точність класифікації. У публікації [9] реалізовано підхід до підвищення якості ознак біометричних зображень шляхом попередньої обробки вхідних даних, з урахуванням структурних і візуальних артефактів.

Таким чином, здобувачем забезпечено системний внесок у формування методичних, алгоритмічних та прикладних засад побудови сучасних біометричних систем автентифікації з урахуванням актуальних загроз, вимог продуктивності та точності.

Апробація результатів. Ключові наукові результати, отримані у процесі виконання дисертаційної роботи, були апробовані на провідних міжнародних та вітчизняних науково-технічних конференціях. Зокрема, положення дисертації було представлено на VIII Міжнародній науково-технічній Internet-конференції «Сучасні методи, інформаційне, програмне та технічне забезпечення систем керування організаційно-технічними та технологічними комплексами» (Київ, 2021); на 2022 IEEE 3rd KhPI Week on Advanced Technology (Харків, 2022); на IX Міжнародній науково-технічній конференції «Захист інформації і безпека інформаційних систем» (Львів, 2023); на XX Міжнародній науково-практичній конференції «Scientific Research: Modern Challenges and Prospects» (Прага, Чехія, 2024); а також на LII Міжнародній науково-практичній конференції «Scientific Research in the Age of Virtual Reality: Exploring New Frontiers» (Монреаль, Канада, 2024). Окремі результати дисертації були також представлені на фахових наукових семінарах кафедри захисту інформації Національного університету «Львівська політехніка» у 2021–2025 роках. Участь у зазначених заходах сприяла фаховій перевірці результатів дослідження, отриманню експертних зауважень і врахуванню

їх у процесі вдосконалення теоретичних і прикладних положень дисертаційної роботи.

Публікації. Основні результати дослідження викладено у чотирнадцяти наукових публікаціях, а саме: у восьми статтях (із них шість– у фахових наукових виданнях України та дві – у періодичному виданні закордоном), одному розділі монографії, опублікованому в іноземному виданні і п’яти тезах виступів на науково-практичних заходах.

Структура та обсяг дисертації. Дисертація складається зі вступу, чотирьох розділів, що охоплюють 23 підрозділів, висновків, списку використаних джерел і додатків. Загальний обсяг дисертації становить 162 сторінок,з яких 142 – основний текст, 13 – список використаних джерел (100 найменувань), 7 – додатки

РОЗДІЛ 1. АНАЛІЗ ОСОБЛИВОСТЕЙ ПРОЦЕСУ РОЗПІЗНАВАННЯ ЛЮДИНИ ЗА ГОЛОСОМ

У цьому розділі представлено історичний розвиток технологій автентифікації голосової біометричної автентифікації, розглянуто основні підходи до мовця та проаналізовано найбільш поширені методи, які знаходяться в основі сучасних систем біометричної автентифікації. Основну увагу приділено порівнянню текстозалежних і текстонезалежних підходів, а також характеристиці класичних, статистичних та нейромережевих методів обробки голосових даних.

1.1. Історія, розвиток і сучасний стан голосової автентифікації

Голосова автентифікація, як метод біометричної автентифікації, має тривалу історію розвитку, яка тісно пов'язана з еволюцією технологій розпізнавання голосу. Початок її формування припадає на першу половину XX століття. У 1940-х роках винайдення звукового спектрографа (сонографа) дозволило науковцям вперше детально візуалізувати акустичні характеристики мовлення та аналізувати спектральні відбитки голосу людини. Це відкриття, описане в класичній роботі «Visible Speech» Поттера, Коппа і Ґріна [15], стало фундаментом для подальших досліджень і розвитку голосової автентифікації.

Наступний важливий етап розпочався у 1960-1970-х роках із появою перших автоматизованих систем розпізнавання голосу, які стали можливими завдяки прогресу комп'ютерних технологій і розвитку алгоритмічного забезпечення. Вагомим досягненням цього періоду були роботи Лі, Рабінера та їх колег, зокрема фундаментальна праця «Digital Processing of Speech Signals», в якій було розроблено і популяризовано методи, засновані на аналізі лінійного передбачення [16]. Ці методи суттєво вплинули на точність і надійність систем розпізнавання мовлення і стали стандартом для наступних поколінь систем голосової авторизації.

У 1980-х роках інтенсивно досліджувались проблеми залежності голосових систем від акустичних умов, зокрема фонових шумів та варіативності голосу користувача. Значний внесок зробили роботи Фуруї [17], які визначили основні

методики покращення надійності голосових систем шляхом нормалізації акустичних ознак.

У 1990-х роках голосові технології почали активно інтегруватися у комерційні рішення, насамперед у вигляді систем інтерактивної голосової відповіді. Такі системи, що автоматизували процес обслуговування клієнтів у телефонних центрах, значно скорочуючи витрати підприємств, описував Фуруї у праці 2005 року Fifty years of progress in speech and speaker recognition [18]. Крім того, важливим напрямком стали дослідження безпеки та захисту голосових даних, що детально описано у роботах з безпеки біометричних систем [19].

Із початку 2000-х років і до сьогодні, завдяки стрімкому розвитку штучного інтелекту та методів глибокого навчання, голосова автентифікація стала значно точнішою та надійнішою. Зокрема, дослідження Гінтона, Денґа та інших учених, як-от у праці "Deep Neural Networks for Acoustic Modeling in Speech Recognition" [20], довели, що використання глибоких нейронних мереж суттєво знижує рівень помилок і забезпечити високу ефективність навіть у складних акустичних умовах. Цей напрямок продовжує активно розвиватись, зокрема з акцентом на використання рекурентних нейронних мереж і трансформерних моделей для подальшого вдосконалення систем голосової автентифікації.

Голосова авторизація поступово впроваджується в різні галузі, зокрема у фінансовий сектор, мобільні технології та системи безпеки. Проте, незважаючи на зростаючий інтерес до голосової біометрії, доступна статистика щодо її поширення є фрагментарною, що ускладнює точну оцінку її впливу на ринок і рівень впровадження в різних країнах та сферах діяльності. Це може бути пов'язано з тим, що багато компаній інтегрують голосову авторизацію як частину ширших рішень з біометричної автентифікації або систем розпізнавання мовлення, що ускладнює виділення окремих даних для цього сегменту.

Компанія Statista у своєму огляді на стан ринку голосової автентифікації вказує, що на даний момент обсяг ринку технологій розпізнавання мовлення досягне 8,58 мільярда доларів США у 2025 році [21]. Очікується, що ринок зростатиме із середньорічним темпом (CAGR) 13,08% у період 2025–2030 років,

досягнувши обсягу 15,87 мільярда доларів США до 2030 року. У глобальному масштабі найбільший ринок припадатиме на США, де його обсяг у 2025 році складе 2,29 мільярда доларів США. Прогнозоване зростання ринку розпізнавання голосу зображене на рис. 1.1.

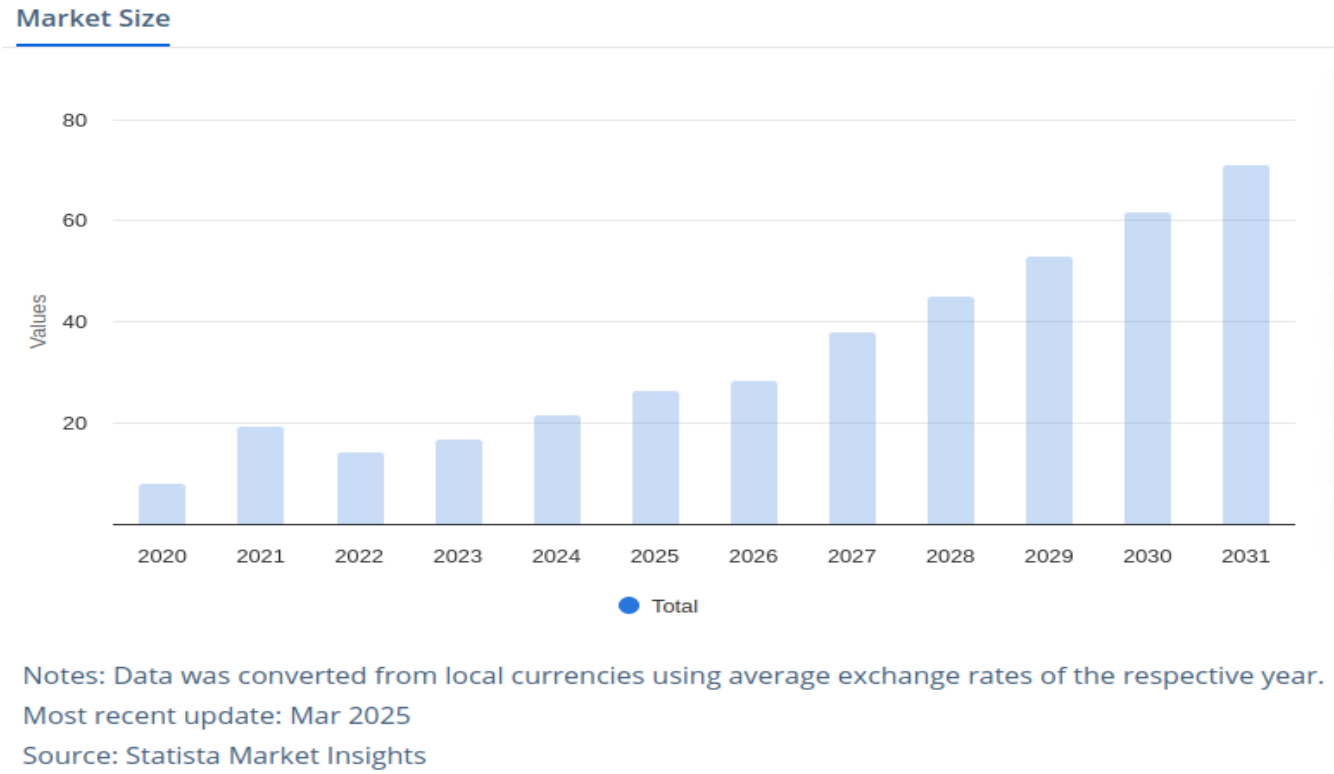

Рис. 1.1. Зростання ринку розпізнавання голосу[21]

Ринок розпізнавання мовлення у сфері штучного інтелекту демонструє стійке зростання на глобальному рівні. Основними драйверами цього процесу є активне впровадження цифрових технологій, підвищена увага до охорони здоров'я, а також зростаюча популярність онлайн-медичних послуг. Розвиток даного ринку обумовлений низкою чинників, зокрема вдосконаленням алгоритмів обробки природної мови (Natural Language Processing, NLP) та широким використанням глибокого навчання, що забезпечує вищу точність і надійність розпізнавання голосу.

У межах глобального ринку штучного інтелекту, технології розпізнавання мовлення інтегруються в численні галузі, включаючи охорону здоров'я, фінансовий сектор і роздрібну торгівлю. Цей процес підвищує функціональність цифрових сервісів і оптимізує взаємодію з користувачем. Зокрема, інтеграція NLP-технологій сприяє більш ефективному виконанню голосових команд, що

забезпечує зручність використання як побутових пристроїв, так і мобільних застосунків.

Серед провідних технологічних компаній Google, Apple, Amazon і Meta активно впроваджують голосові інтерфейси у свої продукти. Наприклад, Google застосовує технології розпізнавання мовлення у Google Assistant, що дає змогу здійснювати голосове керування пристроями. Хоча конкретна статистика щодо масштабів впровадження голосової автентифікації не розкривається, відомо, що Google постійно вдосконалює свої сервіси з метою підвищення рівня безпеки та зручності користувачів [22].

Apple інтегрувала голосовий асистент Siri у власні пристрої, що забезпечує голосову взаємодію з операційною системою [23]. Хоча Siri на даному етапі не виконує функцію автентифікації, компанія активно досліджує можливості використання біометричної автентифікації, включаючи аналіз голосу, для посилення захисту даних [24].

Amazon запровадила голосову авторизацію у продуктах Echo, обладнаних голосовим асистентом Alexa [25]. Це дає можливість користувачам здійснювати транзакції, керувати розумними пристроями вдома та виконувати інші дії за допомогою голосу. Meta (раніше Facebook) також досліджує потенціал голосових технологій, однак публічна інформація щодо реалізації голосової автентифікації залишається обмеженою [26].

На корпоративному рівні рішення голосової автентифікації поєднують штучний інтелект і біометричні технології для забезпечення надійного розпізнавання користувачів. Сучасні системи дозволяють автоматизувати процес автентифікації, мінімізуючи потребу в паролях або PIN-кодах, що суттєво знижує ризики несанкціонованого доступу.

Однією з провідних технологій у цій сфері є голосова біометрія (Voice Biometrics), яка аналізує унікальні параметри мовлення, такі як тембр, ритм, інтонація і швидкість. Серед відомих рішень – Nuance Gatekeeper і Verint Voice Biometric Authentication, які активно застосовуються у фінансових установах та

call-центрах. Такі системи забезпечують високий рівень точності в режимі реального часу, навіть за умов фонових шумів або багатоголосих середовищ.

Крім того, хмарні платформи, зокрема Google Cloud Speech-to-Text та Microsoft Azure Speech, пропонують підприємствам інструменти для голосової автентифікації й інтеграції у корпоративні системи управління доступом, комунікації та безпеки.

Водночас одним із ключових викликів залишаються атаки типу voice spoofing (імітація або підробка голосу). Для підвищення стійкості систем до таких загроз застосовуються методи денойзингу, перевірки діапазону, а також багаторівнева автентифікація, що поєднує голосову біометрію з іншими параметрами, зокрема відбитками пальців або розпізнаванням обличчя.

Таким чином, розвиток систем голосової автентифікації демонструє поступове вдосконалення технологій, що створює основу для використання сучасних нейромережевих підходів у забезпеченні надійної та зручної автентифікації користувачів, що буде розглянуто в наступних розділах.

1.2. Порівняльна характеристика підходів і методів побудови систем розпізнавання голосу.

У сучасних системах голосової автентифікації застосовуються два принципово різні підходи: текстозалежний (text-dependent) та текстонезалежний (text-independent). Вибір між ними визначається специфікою застосування, вимогами до рівня безпеки та точності автентифікації [27].

Текстозалежні методи передбачають, що користувач вимовляє заздалегідь визначену або динамічно згенеровану фразу, яка використовується як ключовий елемент автентифікації. Система може функціонувати на основі фіксованих паролів або випадкових фраз, що оновлюються перед кожною сесією. Такий підхід забезпечує підвищену стійкість до атак типу «відтворення голосу» (playback attack), оскільки навіть за наявності запису попередньої автентифікаційної сесії ймовірність відтворення коректної фрази без її актуального знання є вкрай

низькою. Крім того, текстозалежні системи, як правило, потребують менше навчальних даних, оскільки працюють з обмеженим набором голосових команд.

Водночас такі системи мають певні обмеження щодо зручності використання. Зокрема, вони вимагають від користувача точної вимови заданої фрази, що може бути ускладнено у випадках фонового шуму, фізіологічних змін голосу (наприклад, під час захворювання) або емоційної нестабільності. Це може впливати на стабільність автентифікації та викликати помилки навіть за наявності коректного користувача.

Текстонезалежні методи, на відміну від текстозалежних, не потребують вимови фіксованого набору слів чи фраз. Вони ґрунтуються на аналізі індивідуальних акустичних характеристик мовлення, незалежно від лексичного змісту вимовленого. Такий підхід забезпечує вищу гнучкість і зручність використання, дозволяючи здійснювати автентифікацію у процесі природної мовної взаємодії – зокрема, під час розмови з голосовими асистентами, телефонного обслуговування клієнтів або інших діалогових сценаріїв. Відсутність обмежень щодо тексту розширює можливості інтеграції таких систем у реальні комунікаційні середовища, що робить їх привабливими для практичного застосування.

Разом із тим, реалізація текстонезалежних систем є значно складнішою порівняно з текстозалежними. Основними викликами є необхідність опрацювання значно ширшої фонетичної та акустичної варіативності, зниження чутливості до змін мовленнєвого контексту, а також потреба у великомасштабних корпусах, що охоплюють різноманітні умови запису та мовців. Крім того, такі системи виявляють підвищену чутливість до атак із використанням синтезованого мовлення, зокрема генерованого за допомогою нейронних мереж. Для підвищення стійкості до таких загроз системи доповнюються механізмами виявлення штучних голосів, що ґрунтуються на аналізі спектральних, фазових та часових характеристик мовного сигналу.

У статті «A Survey on Text-Dependent and Text-Independent Speaker Verification» [28] представлено ґрунтовний огляд сучасних підходів до верифікації

мовця, з особливою увагою до текстонезалежних методів. Такі підходи забезпечують автентифікацію в умовах природної комунікації – наприклад, під час взаємодії з віртуальним асистентом або телефонної розмови з оператором банку. Гнучкість та невимушеність процесу автентифікації зумовлюють високу привабливість текстонезалежних систем для практичного застосування.

Сучасні системи голосової автентифікації спираються на широкий спектр методологій, що еволюціонували від класичних алгоритмів обробки мовного сигналу до складних архітектур глибокого навчання, здатних формувати високороздільні векторні репрезентації, відомі як ембедінги. З огляду на складність реалізації, стійкість до завад, вимоги до обчислювальних ресурсів і гнучкість інтеграції в реальні прикладні сценарії, ці методи доцільно класифікувати на чотири ключові групи: класичні алгоритми, статистичні моделі, нейромережеві підходи глибокого навчання та методи, що грунтуються на генерації ембедінгів [30].

Класичні методи. Класичні підходи, зокрема Mel-Frequency Cepstral Coefficients (MFCC) та Linear Predictive Coding (LPC), вважаються базовими у галузі цифрової обробки мовлення. MFCC забезпечує перетворення аудіосигналу на компактне представлення з урахуванням нелінійної чутливості людського слуху до частот, що робить цей підхід ефективним для виокремлення спектральних ознак мовця [31]. LPC, своєю чергою, моделює голосовий тракт людини через систему лінійних предикторів, що дозволяє прогнозувати наступні зразки мовлення на основі попередніх [32].

Незважаючи на ефективність у простих умовах, обидва методи мають обмежену стійкість до фонових шумів, акустичних спотворень та фізіологічних змін голосу, а отже, не завжди придатні для реального середовища [33, 34].

Статистичні моделі. До статистичних підходів належать Gaussian Mixture Models (GMM) та Hidden Markov Models (HMM). GMM апроксимує розподіли голосових ознак за допомогою суміші гауссових функцій, що забезпечує високу гнучкість у моделюванні міжмовцевої варіативності [35]. HMM, у свою чергу,

враховує часову послідовність мовного сигналу, моделюючи його як послідовність латентних станів з ймовірнісними переходами [36].

Незважаючи на успішне використання в ранніх комерційних системах, ці моделі демонструють обмежену стійкість до атак типу spoofing (використання записів або синтетичного мовлення), а також потребують попередньої інженерії ознак, часто базованої на класичних перетвореннях [37].

Методи глибокого навчання. Поява глибоких нейронних мереж істотно трансформувала підходи до голосової автентифікації, зробивши можливим автоматичне вилучення релевантних ознак із мовного сигналу. Згорткові нейронні мережі (CNN) показали високу ефективність при обробці спектрограм, дозволяючи виявляти локальні частотно-часові патерни [38]. Рекурентні мережі (RNN, LSTM, GRU) дозволяють моделювати довготривалі залежності в мовленні, що особливо корисно для збереження тембрального та ритмічного контексту [39].

Окрему нішу в цій категорії займають трансформерні архітектури, серед яких варто виділити Wav2Vec 2.0, HuBERT, Whisper та WavLM. Ці моделі поєднують механізми самозвернення (*self-attention*) з попереднім самонавчанням (*self-supervised pretraining*), що дає змогу їм ефективно працювати на неанотованих даних та забезпечувати універсальні векторні представлення мовного сигналу [40].

Разом із тим, реалізація нейромережевих систем супроводжується високими обчислювальними витратами, як під час навчання, так і в процесі інференсу, що потребує спеціалізованої інфраструктури [41].

Підходи на основі ембедінгів. Новітнім і найбільш перспективним напрямом розвитку є підходи, що ґрунтуються на генерації ембедінгів – багатовимірних векторних представлень голосу, що відображають унікальні біометричні особливості мовця. Моделі цього класу (наприклад, x-vector, ECAPA-TDNN, TitaNet, pyannote.audio) дозволяють отримувати фіксовані за розмірністю представлення незалежно від тривалості вхідного сигналу [42].

На етапі верифікації ембедінги порівнюються за допомогою метрик відстані, найпоширенішими з яких є косинусна подібність та Probabilistic Linear Discriminant Analysis (PLDA) [43]. Такі системи демонструють високу масштабованість,

адаптивність до нових користувачів та підтримують механізми transfer learning, що сприяє швидкій адаптації моделі до нових акустичних умов або мовців [44].

Зважаючи на наведені особливості кожного підходу, доцільно узагальнити ключові характеристики методів голосової автентифікації в порівняльній таблиці. Такий формат наочно демонструє переваги та обмеження різних класів методів у контексті їх практичного застосування. В таблиці 1.1 подано порівняльну характеристику методів голосової автентифікації.

Таблиця 1.1.

Порівняльна характеристика методів голосової автентифікації за основними параметрами ефективності

Метод	Переваги	Недоліки
Класичні (MFCC, LPC)	Простота реалізації Низькі обчислювальні витрати Швидка обробка	Низька стійкість до шуму Погано обробляють часову динаміку Обмежена точність
Статистичні (GMM, HMM)	Моделюють розподіли й часову структуру Краща адаптація до варіативності мовлення	Вразливість до spoof-атак Потребують ручної інженерії ознак
Глибоке навчання (CNN, RNN, Wav2Vec, WavLM)	Висока точність Автоматичне вилучення ознак Стійкість до шуму	Висока складність Значні обчислювальні ресурси Довгий час тренування

Ембедінги (x-vector, ECAPA, TitaNet)	Масштабованість Підтримка швидкої верифікації Сумісність із метричними методами	Залежність від якості ембедінгів Необхідність донавчання для нових середовищ

Узагальнення існуючих підходів до голосової автентифікації свідчить про поступову еволюцію від класичних алгоритмів цифрової обробки мовного сигналу до складних моделей глибокого навчання та ембедінг-орієнтованих систем. Кожен із розглянутих методологічних напрямів характеризується власними перевагами та обмеженнями: класичні та статистичні моделі відзначаються простотою реалізації, однак демонструють обмежену стійкість до завад та атак, тоді як нейронні мережі й методи на основі ембедінгів забезпечують високий рівень точності та адаптивності за рахунок значних обчислювальних витрат і складності впровадження. Ці висновки підкреслюють необхідність комплексного підходу до вибору методів залежно від конкретних умов експлуатації та вимог до системи.

Водночас із розвитком технологій голосової автентифікації зростає і рівень загроз, пов’язаних із використанням синтетичних голосів, що імітують мовлення легітимних користувачів. У цьому контексті важливим аспектом забезпечення безпеки є аналіз сучасних систем клонування голосу та їхнього потенційного впливу на надійність біометричних систем. Наступний підрозділ присвячений розгляду архітектур і принципів роботи таких систем, а також оцінюванню їхньої ролі в контексті атак типу spoofing.

1.3. Аналіз загроз і вразливостей системи голосової автентифікації.

Сучасні досягнення в галузі штучного інтелекту (ШІ) сприяли стрімкому розвитку технологій синтезу та обробки мовлення, зокрема в контексті біометричної автентифікації. Голос, як унікальний біометричний ідентифікатор, забезпечує зручний та ефективний спосіб підтвердження особистості. Проте, розвиток технологій водночас породжує нові загрози – однією з яких є клонування голосу за допомогою ШІ.

Клонування голосу означає створення штучних аудіозаписів, які імітують голос конкретної особи з високою точністю, використовуючи обмежений обсяг початкових голосових даних. Сучасні алгоритми ШІ дозволяють досягти надзвичайного рівня реалістичності, що відкриває нові можливості, але водночас створює значні ризики. Технології клонування голосу знаходять застосування у створенні персоналізованих голосових асистентів, в кінематографі, відеоіграх, навчальних платформах, маркетингу та рекламі. Водночас їхня здатність відтворювати голос практично ідентично до оригіналу ставить під загрозу системи інформаційної безпеки, зокрема в контексті шахрайства та атак на голосову автентифікацію.

Сьогодні можна виокремити два провідні підходи до реалізації систем синтезу голосу на основі ШІ: перетворення голосу (Voice Conversion, VC) та технології клонування голосу з тексту (Voice Cloning або Text-to-Speech Cloning) [45].

Перетворення голосу (Voice Conversion, VC). Перетворення голосу передбачає зміну голосових характеристик мовця без втручання у зміст висловлювання. На відміну від TTS, цей підхід модифікує вже існуючий аудіосигнал, адаптуючи його до тембру й особливостей іншої особи. Сучасні реалізації VC базуються на глибокому навчанні та включають такі підходи, як Retrieval Voice Conversion (RVC), CycleGAN, AutoVC тощо [46].

VC є важливим напрямом досліджень у сфері ШІ та обробки мовлення. Найбільш перспективною серед новітніх технологій вважається Retrieval Voice Conversion (RVC), яка поєднує в собі можливості генеративних змагальних мереж (GAN), автоенкодерів та інших рекурсивних нейронних моделей.

RVC – це нейромережева архітектура, призначена для високоякісної трансформації голосових характеристик одного мовця в голосові особливості іншого. На відміну від традиційних статистичних підходів, RVC використовує можливості глибокого навчання для автоматизованого вивчення прихованих характеристик голосу.

Перший етап включає екстракцію ключових акустичних ознак: спектральні параметри, фундаментальну частоту (f0), мел-кепстральні коефіцієнти (MFCC) тощо. Далі за допомогою автоенкодера відбувається формування латентного простору (latent representation) – стислого та інформативного представлення мовлення, що дає змогу зменшити потребу в обсязі навчальних даних і підвищити точність трансформації [47].

На наступному етапі використовується генеративна змагальна мережа: генератор створює аудіосигнали на основі трансформованих латентних векторів, тоді як дискримінатор оцінює реалістичність згенерованих сигналів. Завдяки цьому змагальному процесу модель поступово навчається відтворювати максимально природні голоси. Завершальним кроком є реконструкція аудіосигналу за допомогою вокодерів – таких як HiFi-GAN, WaveNet, WaveGlow тощо.

Основною перевагою RVC є можливість ефективного навчання на обмежених наборах даних, що дає змогу швидко та точно здійснювати перетворення голосу. Ця технологія забезпечує високу природність мовлення та зводить до мінімуму виникнення артефактів, таких як металевий відтінок чи спотворення тембру. Завдяки цим характеристикам вона знаходить широке застосування у створенні персоналізованих голосових асистентів, дубляжі мультимедійного контенту, розробленні персоналізованої реклами, а також у наданні голосової підтримки людям із мовними порушеннями.

Найновіші дослідження у сфері RVC зосереджені на підвищенні реалістичності звучання, зменшенні обчислювальних затрат, створенні систем для реального часу (real-time RVC), використанні варіаційних автоенкодерів (VAE) та розвитку мультиспікерних моделей [48, 49].

Клонування голосу (Voice Cloning або Text-to-Speech Cloning). На відміну від VC, технологія voice cloning не модифікує існуючий аудіосигнал, а синтезує мовлення з нуля, генеруючи аудіо на основі тексту. Модель ШІ, пройшовши навчання на прикладах мовлення конкретної особи, відтворює нове мовлення, що відповідає голосовим характеристикам цієї особи [50].

Популярні моделі, що реалізують даний підхід, включають Tacotron, FastSpeech, VITS.

Моделі voice cloning зазвичай реалізуються як енкодер-декодерні архітектури з механізмами уваги (attention) або трансформерними структурами. Першим етапом є побудова прихованих голосових репрезентацій, що ґрунтуються на невеликому наборі аудіозразків цільового мовця. Після цього здійснюється генерація спектрограми, яка згодом перетворюється на аудіосигнал за допомогою нейромережевих вокодерів (наприклад, WaveNet, HiFi-GAN або WaveGlow) [51, 52].

Tacotron та його вдосконалена версія Tacotron 2 демонструють здатність генерувати мовлення високої якості з природною інтонацією та гнучкістю синтезу. Технологія клонування голосу відкриває перспективи для створення цифрових копій голосів, впровадження персоналізованих освітніх рішень, а також використання у сфері мультимедійного й кінематографічного контенту. Завдяки трансферному навчанню і мультиспікерним моделям, сучасні алгоритми потребують всього кілька хвилин аудіо для навчання, що робить технологію доступною та масштабованою.

У контексті дослідження загроз для систем біометричної автентифікації за голосом особливу увагу заслуговують технології синтезу мовлення, зокрема системи тексту-в-мовлення (TTS). Ці системи відіграють ключову роль у створенні високоякісних клонованих голосів, які можуть бути використані для реалізації атак на голосові біометричні системи. Сучасні TTS-моделі, зокрема ті, що базуються на глибокому навчанні, демонструють здатність генерувати мовлення з індивідуальними характеристиками конкретного мовця, що суттєво ускладнює задачу розпізнавання підробки. Тому вивчення принципів роботи, можливостей та обмежень TTS-систем є невід'ємною частиною комплексного аналізу стійкості голосової автентифікації до атак із використанням синтетичного мовлення.

У сучасних дослідженнях та розробках у галузі синтезу мовлення на основі тексту (Text-to-Speech, TTS) значну увагу привертають три проекти: VITS, Coqui

TTS та PFlow-TTS. Кожен із них пропонує унікальні підходи та рішення для генерації високоякісного мовлення.

Проєкт VITS, розроблений Jaehyeon Kim (jaywalnut310), є сучасною енд-то-енд моделлю синтезу мовлення, яка поєднує переваги умовних варіаційних автоенкодерів (Conditional Variational Autoencoders, CVAE), генеративних змагальних мереж (Generative Adversarial Networks, GAN) та нормалізуючих потоків (Normalizing Flows). Такий підхід дає змогу одночасно моделювати текстову та голосову інформацію в єдиній архітектурі. Особливістю VITS є те, що на відміну від класичних двоетапних моделей TTS (окремий синтез спектрограм і вокодер), тут процес синтезу повністю інтегрований. Це забезпечує високу природність синтезованого мовлення і знижує кількість артефактів. Зокрема, застосування GAN дає змогу отримати мовлення з природними варіаціями і реалістичністю, тоді як нормалізуючі потоки забезпечують ефективне моделювання складних залежностей між текстом та мовленням. VITS демонструє високі результати в завданнях мультиспікерного синтезу, а також може бути ефективно адаптована до малих наборів даних. Це робить її привабливою для практичного використання у персоналізованих додатках синтезу мовлення [53].

Проєкт Coqui TTS є відкритою платформою для глибокого навчання, орієнтованою на мультиспікерний та багатомовний синтез мовлення. Цей інструментарій побудований на основі популярних та перевірених нейромережевих архітектур, таких як Tacotron, Glow-TTS, VITS, FastSpeech 2, а також сучасних нейромережевих вокодерів (наприклад, HiFi-GAN). Головна особливість Coqui TTS – це гнучкість і модульність платформи, яка дає змогу дослідникам і розробникам легко налаштовувати і розгортати різні конфігурації моделей, залежно від поставлених завдань та наявних ресурсів. Платформа також містить велику кількість попередньо навчених моделей для понад тисячу мов, що суттєво знижує поріг входження для розробників, які не мають великих наборів навчальних даних або ресурсів для тривалого навчання. Coqui TTS активно використовується в освітніх цілях, дослідженнях, розробці персоналізованих голосових асистентів, а також у комерційних продуктах. Завдяки широкій спільноті розробників цей

проєкт постійно отримує оновлення та підтримку нових алгоритмів, архітектур і мов [54].

Проєкт PFlow-TTS є неофіційною реалізацією моделі P-Flow, яка базується на підході NVIDIA, спрямованому на швидкий та якісний синтез мовлення у так званому режимі zero-shot (без попереднього тривалого навчання для конкретного мовця). Ключовим інноваційним аспектом цього підходу є використання технології "Speech Prompting", що дає змогу системі адаптуватися до голосу нового мовця за дуже коротким аудіозразком (prompt). В основі архітектури PFlow-TTS лежить комбінація текстового енкодера і генеративної моделі на основі потоків (Normalizing Flows). Модель навчається таким чином, щоб ефективно представляти голосові особливості мовців у прихованому просторі і швидко адаптуватися до нових мовних підказок. Використання нормалізуючих потоків дає змогу суттєво покращити якість і швидкість генерації мовлення, порівняно з традиційними TTS-архітектурами. PFlow-TTS підходить для сценаріїв, коли потрібно швидко створювати персоналізоване мовлення з мінімальною кількістю доступних даних, наприклад, для персоналізованих голосових асистентів, автоматичного дубляжу або інтерактивних систем [55].

Описані проєкти є сучасними зразками ефективних рішень для синтезу мовлення та демонструють тенденцію переходу до інтегрованих енд-то-енд архітектур, що покращує природність та якість генерованого аудіо. VITS представляє комбіноване рішення, що максимально використовує переваги генеративних моделей та варіаційних автоенкодерів, тоді як Coqui TTS робить акцент на гнучкості, масштабованості та доступності для широкого кола розробників. Водночас PFlow-TTS пропонує сучасний підхід, що мінімізує витрати на адаптацію до нових мовців. Усі три підходи активно розвиваються, представляючи значний інтерес для дослідників та розробників у галузі синтезу мовлення, що підтверджує їх широке використання як в академічних дослідженнях, так і у практичних розробках.

Виклики та етичні аспекти застосування систем клонування голосу. Етичні аспекти застосування технологій клонування голосу на основі штучного інтелекту

набувають дедалі більшого значення в контексті зростання їх точності, доступності та широкого поширення. Однією з ключових моральних дилем є використання синтезованих голосів для імітації мовлення конкретних осіб без їхньої згоди. Це може призводити до серйозних порушень права на приватність та особисту недоторканність, зокрема у випадках, коли клонований голос використовується для створення фальшивих заяв, дискредитаційних матеріалів або маніпулятивних повідомлень. Така практика створює передумови для глибоких суспільних і політичних загроз, включаючи поширення дезінформації, втручання у виборчі процеси, фабрикацію доказів у судових справах або реалізацію фінансових шахрайств через обман голосової автентифікації. Окремо постає питання захисту цифрової ідентичності, адже в умовах зростаючої здатності систем ШІ до імітації голосу розмиття межі між реальним і штучним мовленням ускладнює процес розпізнавання автентичної комунікації. Виникає також ризик експлуатації технологій voice cloning у вразливих соціальних групах – наприклад, у дітей, літніх людей або осіб із психічними розладами, які можуть бути легко введені в оману за допомогою синтетичного голосу знайомої чи авторитетної для них особи. Крім того, використання таких технологій у художньому чи рекламному контексті без чіткого інформування споживачів про синтетичну природу мовлення може порушувати принципи чесності та прозорості. Таким чином, розвиток і застосування систем клонування голосу має супроводжуватись не лише технічними запобіжниками, як-от маркування синтетичного мовлення чи впровадження механізмів аудіо-верифікації, а й широким етичним осмисленням – із залученням розробників, правників, етичних комітетів і суспільства загалом для формування стандартів відповідального використання таких технологій.

Сучасні технології клонування голосу є надзвичайно потужними інструментами, які відкривають широкі можливості для персоналізації, автоматизації мовленнєвої взаємодії та інноваційних застосунків у сфері мультимедіа та комунікації. Водночас зростає потреба у вивченні їхніх безпекових аспектів, етичних викликів, а також у розробленні захисних технологій, здатних

забезпечити достовірну автентифікацію голосу в умовах зростаючої загрози зловживань.

1.4. Формулювання задач дисертаційного дослідження

У контексті реалізації поставленої мети та відповідно до визначених наукових і прикладних проблем, сформульованих у межах попередніх підрозділів, доцільним є виокремлення комплексу задач, вирішення яких забезпечує системний підхід до створення ефективної та захищеної системи біометричної автентифікації за голосом. Враховуючи багатовекторність дослідження, завдання охоплюють питання архітектурного проєктування, алгоритмічного забезпечення, моделювання загроз та експериментального тестування.

З метою забезпечення комплексної функціональності системи необхідно розробити концептуальну модель архітектури біометричної автентифікації за голосом, яка б інтегрувала структурно-функціональні блоки на основі глибоких нейронних мереж (CNN, LSTM, Transformers), підтримувала модульність, відповідала стандартам інформаційної безпеки та була придатною до подальшої експериментальної реалізації.

Для вибору оптимального підходу до екстракції ознак мовлення доцільно провести порівняльний аналіз методів побудови систем біометричної автентифікації за голосом, з урахуванням їхньої здатності забезпечувати стійкість ознак до завад і високу дискримінативність у задачах верифікації.

Необхідно реалізувати архітектурну модель системи автентифікації, що включає модулі збору голосових зразків, побудови ембедінгів, калібрування порогів і прийняття рішення про відповідність, а також дослідити ефективність різних методів обчислення відстаней між векторними представленнями мовлення.

На основі отриманих експериментальних результатів необхідно здійснити оцінювання масштабованості, продуктивності та точності системи біометричної автентифікації, дослідити вплив збільшення кількості зареєстрованих користувачів на верифікаційну ефективність, а також сформулювати обґрунтовані рекомендації

щодо оптимізації порогових значень та вибору архітектурних конфігурацій для різних умов експлуатації.

З метою оцінювання стійкості системи до атак на основі клонованого мовлення доцільно розробити експериментальну методику моделювання таких атак, створити відповідний тестовий набір синтетичних голосів, згенерованих за допомогою сучасних TTS і VC-моделей, та провести верифікацію ефективності системи в умовах повторного та нового текстового контенту.

Для підвищення рівня захищеності автентифікаційної системи необхідно розробити і протестувати модуль детекції синтетичного мовлення, реалізований на основі SVM-класифікатора та спектральних ознак провести порівняльний аналіз із комерційними рішеннями, виявити їх обмеження та сформулювати рекомендації щодо впровадження антиспуфінгових механізмів в архітектуру системи.

Таким чином, послідовне вирішення зазначених завдань дозволить створити високоточну, стійку до атак та нормативно узгоджену систему біометричної автентифікації за голосом, що відповідає сучасним вимогам кібербезпеки та може бути адаптована до практичних умов експлуатації.

1.5. Висновки до розділу 1

Розділ 1 робить акцент на комплексному огляді історичних, теоретичних і практичних аспектів голосової автентифікації. Основні висновки можна сформулювати так:

1. Детально проаналізовано історію розвитку голосової автентифікації, яка пройшла етапи від первинних методів спектрографії до інтеграції сучасних нейромережевих технологій.
2. Розглянуто і проведено порівняльний аналіз текстозалежних та текстонезалежних підходів до голосової автентифікації. Виявлено, що текстозалежні системи забезпечують високу точність при коротких зразках мовлення і мають підвищену стійкість до атак відтворення, але менш зручні у використанні через залежність від конкретних фраз. Текстонезалежні системи, хоча й складніші у впровадженні та більш ресурсомісткі,

дозволяють інтегрувати автентифікацію у природну комунікацію, що розширює можливості їхнього застосування у реальних сценаріях.

3. Проаналізовано класифікацію методів голосової автентифікації за принципами роботи: від класичних і статистичних до сучасних нейромережевих та ембедінг-орієнтованих підходів. Узагальнено переваги та обмеження кожного класу методів. Зокрема, встановлено, що ембедінг-орієнтовані моделі поєднують можливість масштабованості, адаптивність до нових користувачів і сумісність із сучасними метричними методами (наприклад, косинусна відстань, PLDA), що робить їх найперспективнішими для впровадження у реальні системи безпеки.
4. Проведено аналітичний огляд новітніх технологій клонування голосу, включаючи моделі Voice Conversion і Text-to-Speech Cloning. Розкрито архітектурні особливості цих систем і підкреслено їхній потенціал створення реалістичних синтетичних голосів навіть на основі обмежених даних. Це посилює ризики атак на голосові біометричні системи та вимагає нових стратегій захисту.
5. Визначено основні виклики та ризики, пов’язані з використанням синтетичного мовлення, серед яких ключовими є атаки типу spoofing, можливість шахрайського використання клонованих голосів та зростаюча складність розпізнавання підробок. Наголошено на важливості впровадження антиспуфінгових технологій і комбінованих систем автентифікації для підвищення загальної безпеки.
6. Акцентовано увагу на етичних аспектах впровадження технологій клонування голосу, зокрема питаннях приватності, потенційних порушень цифрової ідентичності та загрозах використання синтетичних голосів для дискредитаційних або шахрайських дій. Підкреслено необхідність міжнародного регулювання та розроблення нормативних актів, що забезпечать баланс між розвитком технологій і захистом прав користувачів.

РОЗДІЛ 2. КОНЦЕПТУАЛЬНА МОДЕЛЬ ГОЛОСОВОЇ АВТЕНТИФІКАЦІЇ НА ОСНОВІ НЕЙРОМЕРЕЖЕВИХ ТРАНСФОРМЕРІВ

У контексті стрімкого розвитку інформаційних технологій та активного впровадження біометричних систем автентифікації особливого значення набуває питання вдосконалення методів і засобів обробки голосової біометрії. Існуючі підходи до аналізу біометричних даних потребують оновлення та підвищення ефективності в умовах зростаючих вимог до точності, стійкості до зовнішніх впливів та адаптивності до новітніх загроз. У цьому аспекті актуальним є застосування інформаційних нейромережевих технологій як інструменту інтелектуалізації процесів обробки та аналізу голосових характеристик. Запропонована концепція системи голосової біометричної автентифікації грунтується на системному підході до моделювання об'єктів дослідження, використання нейронних мереж для витягування релевантних ознак, побудови механізмів підтримки прийняття рішень та забезпечення відповідності міжнародним стандартам у галузі біометричної безпеки. Такий підхід створює наукову методологічну основу для підвищення ефективності технологій голосової біометрії в сучасному інформаційному середовищі.

2.1. Концептуальна модель реалзації вдосконалення методів і засобів голосової біометрії на основі інформаційних нейромережевих технологій

Запропонована концепція системи вдосконалення методів і засобів голосової біометрії орієнтована на забезпечення підвищення точності, надійності та стійкості обробки біометричних даних в умовах активного розвитку технологій штучного інтелекту. Ключовим завданням є створення такої інформаційної нейромережевої технології, яка дозволить суттєво поліпшити існуючі підходи до аналізу, класифікації та оптимізації голосових біометричних характеристик.

Концептуальна модель голосової біометричної автентифікації [56, 57] базується на кількох ключових структурно-функціональних блоках, кожен з яких виконує визначену роль у загальній архітектурі системи. Структурна схема

реалізації концепції вдосконалення голосової біометрії на основі нейромережевих технологій подана на рисунку 2.1.

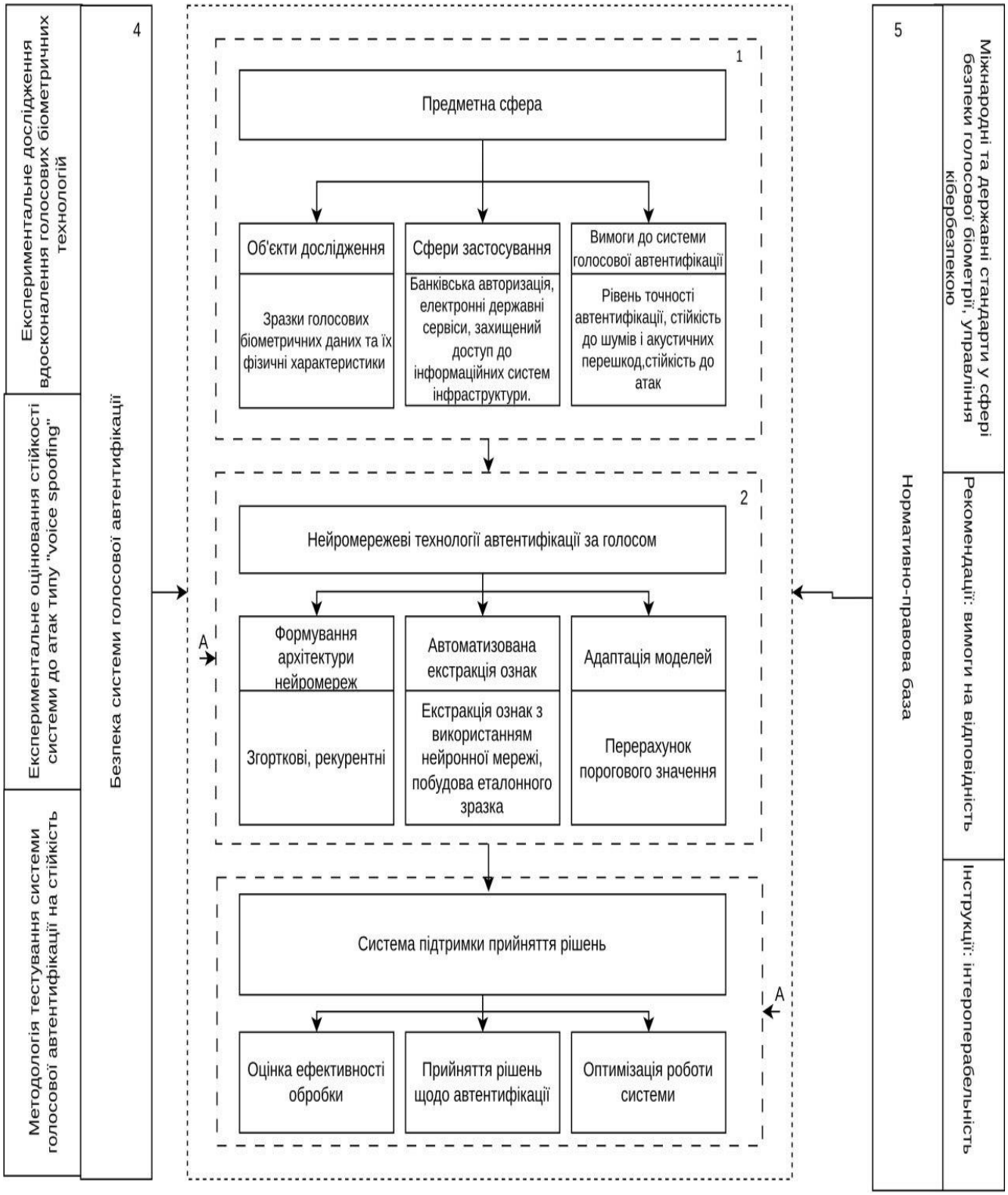

Рисунок 2.1. Концептуальна модель реалізації вдосконалення методів і засобів голосової біометрії на основі інформаційних нейромережевих технологій

Перший блок – це блок предметної сфери, що визначає об'єкти дослідження: зразки голосових біометричних даних, їх фізичні та цифрові характеристики, а також сфери практичного застосування, такі як банківська автентифікація, електронні державні сервіси, захищений доступ до інформаційних систем. В межах цього блоку встановлюються основні вимоги до системи – рівень точності автентифікації, стійкість до шумів і перешкод, здатність протистояти спробам штучної модифікації голосових даних.

Другою компонентою моделі концепції є блок нейромережевих технологій, який забезпечує інтелектуальне ядро системи голосової автентифікації. У межах цього блоку здійснюється вибір архітектури сучасних нейронних мереж, зокрема згорткових (CNN) для аналізу спектральних ознак аудіосигналу та рекурентних (LSTM, RNN) або трансформерних моделей для врахування часових залежностей мовлення. Основною метою є побудова системи, здатної автоматизовано виділяти інформативні ознаки голосу у вигляді векторних ембедінгів, що забезпечують сталу автентифікацію користувача.

На основі отриманих ознак формуються еталонні шаблони користувачів, які зберігаються у базі даних системи для подальшої автентифікації. Крім того, блок нейромережевих технологій забезпечує адаптацію системи до нових умов: зокрема, реалізується механізм перерахунку порогових значень для прийняття рішень про успішність або відхилення автентифікації. Завдяки цьому зберігається високий рівень точності навіть за змінних умов — зокрема, різного акустичного середовища, типів пристроїв або мовних характеристик користувача.

Важливим є блок підтримки прийняття рішень, який забезпечує системну інтеграцію результатів аналізу. На цьому рівні здійснюється оцінка ефективності методів обробки, порівняння показників точності, специфічності та чутливості, визначення оптимальних стратегій подальшої обробки даних. Система підтримки рішень допомагає визначати, чи потрібно здійснювати перенавчання моделі, змінювати алгоритми попередньої обробки або коригувати критерії класифікації.

Четвертим компонентом є блок забезпечення безпеки системи голосової автентифікації, який охоплює як теоретичні засади, так і експериментальні

процедури оцінювання стійкості до сучасних атак, зокрема до атак типу "voice spoofing". У межах цього блоку розробляється методологія тестування, що дає змогу кількісно оцінити здатність системи виявляти синтетичні або підроблені голосові сигнали за різних сценаріїв мовленнєвої взаємодії.

У дослідженні значний акцент зроблено на експериментальній оцінці ефективності сучасних біометричних технологій в умовах імітації атак з боку зловмисників.[58] Такий підхід дає змогу перевірити працездатність системи в умовах реальних загроз і виявити потенційні вразливості на рівні ембедінгів, моделей класифікації або порогових механізмів. Отримані результати дозволяють обґрунтувати шляхи вдосконалення архітектури системи та підвищення її стійкості до фальсифікацій і несанкціонованого доступу.

У концепції побудови системи голосової біометричної автентифікації передбачено модуль стандартизації та уніфікації, що забезпечує відповідність розробленого рішення чинним міжнародним стандартам у сфері біометричної безпеки, інформаційних технологій та кіберзахисту. Зокрема, інтеграція вимог стандартів ISO/IEC 30107 (Presentation Attack Detection) і C2PA (Coalition for Content Provenance and Authenticity) гарантує нормативну сумісність системи та її підвищену стійкість до маніпуляцій із голосовими даними. Це створює передумови для впровадження розробленої технології у критично важливих інформаційних інфраструктурах, де на перший план виходять вимоги до автентичності користувача, цілісності біометричних шаблонів і захищеності від атак, зокрема спуфінгу та підміни даних.

Таким чином, запропонована концепція передбачає цілісне, багаторівневе вдосконалення методів та засобів голосової біометричної автентифікації шляхом використання інформаційних нейромережевих технологій. Такий підхід дає змогу не лише реагувати на сучасні виклики безпеки, але й формувати резерв можливостей для майбутнього розвитку галузі біометричної автентифікації.

2.2. Процес розпізнавання особи за голосом: структура та етапи реалізації

На основі проведеного аналітичного огляду існуючих підходів до побудови біометричних системи автентифікації на основі голосового сигналу узагальнено структурну схему таких систем (рис. 2.2).

Рисунок 2.2. Структурна схема процесу обробки голосового сигналу в системах біометричної автентифікації

У ланці цифрової обробки сигналу отриманий голосовий сигнал оцифровується і піддається попередній обробці. Далі оброблений сигнал опрацьовується алгоритмом екстракції ознак, після чого векторизоване представлення зразку опрацьовується за допомогою одного з алгоритмів класифікації. Нижче подано детальний опис для кожного із елементів наведеної вище структурної схеми.

2.2.1. Відбір мовного сигналу

Першим і фундаментальним етапом функціонування будь-якої системи голосової автентифікації є відбір мовного сигналу, який забезпечує первинне отримання акустичної інформації від користувача. Цей етап має ключове значення, оскільки точність подальшого аналізу прямо залежить від достовірності та чіткості отриманого звукового сигналу. В реальних умовах запис мовлення може супроводжуватись різноманітними акустичними спотвореннями, такими як фоновий шум, реверберація, відлуння, нестабільна амплітуда, дефекти вимови або варіативність у тембрі голосу. Для мінімізації впливу таких факторів доцільно застосовувати мікрофони з високою роздільною здатністю та вузькою діаграмою спрямованості, а також забезпечувати контроль над умовами середовища запису. Додатково важливими є вимоги до стійкої гучності, стабільного положення джерела мовлення та фіксованої дистанції до мікрофона, особливо при реєстрації зразків для подальшого використання як еталонів автентифікації.

Аудіосигнал, за своєю фізичною природою, є аналоговим, тобто являє собою безперервну функцію часу. Оскільки цифрові обчислювальні системи не можуть безпосередньо обробляти аналогові сигнали, виникає необхідність у виконанні аналогово-цифрового перетворення (АЦП). Це перетворення забезпечує представлення аналогового сигналу в цифровій формі шляхом дискретизації у часі та квантування за рівнем амплітуди. АЦП включає три основні етапи: дискретизацію, квантування та кодування.

Під час дискретизації відбувається вибірка значень амплітуди звукової хвилі з фіксованим часовим інтервалом. Частота дискретизації є критичним параметром, що визначає точність подання сигналу в часовій ділянці. У системах обробки мовлення зазвичай використовується частота 16 кГц, яка згідно з теоремою Найквіста–Шеннона забезпечує коректне відтворення частотного спектру мовлення, що в основному міститься в межах до 8 кГц. Для високоточних задач або при роботі з широкосмуговими голосовими зразками можуть застосовуватися вищі частоти, зокрема 22.05 кГц або 44.1 кГц [59].

Квантування передбачає округлення неперервних значень амплітуди до найближчих рівнів у фіксованій шкалі. Ступінь деталізації цього процесу задається бітовою глибиною – найчастіше використовується 16-бітне квантування, що забезпечує представлення 65536 можливих рівнів сигналу. Вища глибина квантування знижує рівень квантувального шуму та збільшує динамічний діапазон, але потребує більших обсягів пам’яті для зберігання.

Кодування – це завершальна фаза АЦП, під час якої дискретизовані в часі й квантовані за рівнем амплітуди значення перетворюються в цифрову бітову послідовність, придатну для зберігання, обробки або передавання в інформаційних системах. Отриманий цифровий сигнал (наприклад, у форматі WAV або PCM) є об’єктом подальшої обробки, зокрема попередньої фільтрації, виділення ознак та порівняння з еталонними зразками. Для забезпечення сумісності з трансформерними архітектурами зазвичай використовується формат WAV із частотою 16 кГц.

Таким чином, етап відбору мовного сигналу разом із процедурою аналогово-цифрового перетворення формує вхідні дані для всієї системи голосової автентифікації. Від точності виконання цього етапу залежить успішність наступних фаз обробки та загальна надійність розпізнавання користувача за голосом.

Після оцифрування сигналу система виконує попередню обробку для покращення якості вхідних даних.

2.2.2. Попередня обробка мовного сигналу

Попередня обробка є критично важливим етапом у структурі системи голосової автентифікації, що слугує для підвищення якості сигналу перед його аналізом і використанням у процедурах розпізнавання та верифікації. Метою цього етапу є усунення або зменшення небажаних компонентів мовного сигналу, таких як шум, мовчазні інтервали, артефакти запису тощо, а також нормалізація параметрів сигналу для приведення їх до єдиного масштабу, що підвищує узагальнювальну здатність системи та стійкість до акустичної варіативності.

Попередня обробка безпосередньо впливає на якість ознак, що виділяються у наступних етапах, і відповідно – на загальну ефективність моделі.

Оскільки мовний сигнал є нестаціонарним, високовимірним та схильним до сильних спотворень в умовах реального середовища, до його попередньої обробки висуваються підвищені вимоги. Вона зазвичай реалізується за допомогою комбінації методів, таких як пригнічення шуму (noise reduction), виявлення мовної активності (voice activity detection, VAD) та нормалізація амплітуди (amplitude normalization).

Пригнічення шуму є одним із ключових етапів обробки, особливо в системах, призначених для роботи в неконтрольованих акустичних умовах. Під шумом зазвичай розуміють адитивні або мультиплікативні спотворення, які не несуть мовної інформації. У разі адитивного шуму, наприклад фонового гулу, ефективно застосовуються методи спектрального віднімання (spectral subtraction), коли спектр фону оцінюється на основі мовчазних ділянок, а потім віднімається від загального спектру сигналу. Більш складні методи включають алгоритми фільтрації Вінера (Wiener filtering), які мінімізують середньоквадратичну помилку між очищеним і бажаним сигналом, або байєсівські методи, що враховують апріорні ймовірності мовлення та шуму.

В останнє десятиліття отримали широке застосування глибокі нейронні мережі, навчені розділяти мовний та шумовий компоненти на основі великих корпусів пар "чистий-зашумлений" сигналів. Зокрема, ефективно використовуються моделі типу DNN, LSTM або U-Net для знешумлення спектрограм. Такі підходи дозволяють зменшити шум навіть у важких умовах із низьким відношенням сигнал/шум, покращуючи якість ознак, що будуть виділятись на наступному етапі.

Виявлення мовної активності є критичною процедурою, що відокремлює ділянки сигналу, які містять інформативне мовлення, від тиші або фонового шуму. Основна мета VAD – зменшити обсяг оброблюваної інформації та виключити фрагменти, які можуть мати негативний вплив на побудову ознак або модель користувача [60]. Прості енергетичні детектори базуються на порівнянні середньої

або миттєвої енергії сигналу із заданим порогом. Більш складні алгоритми використовують частотні ознаки (наприклад, співвідношення енергії у високих і низьких діапазонах), спектральну ентропію, кумулятивні характеристики або машинне навчання (GMM, SVM, DNN). Алгоритми VAD можуть працювати у реальному часі й бути адаптивними до змін умов середовища. Наприклад, моделі з LSTM або Transformer здатні враховувати контекст у часі та розрізняти мовлення навіть при коротких і неповних фрагментах. Якісний VAD забезпечує кращу узгодженість ситеми, зменшує кількість хибно прийнятих шумових фрагментів та підвищує ефективність наступних етапів.

Умови запису мовлення можуть суттєво відрізнятися: різна гучність голосу, чутливість мікрофона, відстань до джерела, тощо. Усі ці фактори впливають на амплітуду сигналу, яка без нормалізації може викликати розкид у статистичних характеристиках ознак та ускладнити навчання моделей.

Нормалізація амплітуди – це процес вирівнювання рівня сигналу до певного стандартного діапазону. Зазвичай використовується пікова нормалізація (до максимальної амплітуди), середньоквадратична нормалізація або z-нормалізація для забезпечення однакової масштабності між записами. У деяких випадках застосовуються адаптивні алгоритми, що аналізують статистику всього сигналу або окремих фреймів для більш точного вирівнювання.

Результатом попередньої обробки є очищений, сегментований і нормалізований мовний сигнал, стійкіший до варіативності вхідних умов. Такий підхід забезпечує високу узгодженість і достовірність під час подальшого виділення ознак, а також сприяє зниженню загального рівня помилок при верифікації та ідентифікації користувача. У сучасних реалізаціях попередня обробка часто виконується у вигляді автономного модуля або інтегрується в структуру глибокої нейронної мережі як окремий препроцесинговий шар, оптимізація якого здійснюється одночасно з навчанням основної моделі.

2.2.3. Виділення ознак для біометричної автентифікації

Після етапу попередньої обробки, в результаті якого отримується очищений, сегментований та нормалізований мовний сигнал, система голосової автентифікації переходить до етапу виділення ознак (feature extraction). Основне завдання цього етапу полягає в перетворенні вхідного одномірного аудіосигналу у компактне, але водночас максимально інформативне векторне представлення, яке відображає характерні особливості голосу конкретного користувача. З погляду інформаційної теорії, процес виділення ознак спрямований на зменшення розмірності простору ознак при збереженні максимальної кількості дискримінативної інформації, релевантної для розпізнавання особи.

У мовних біометричних системах широко застосовуються методи обробки сигналів у частотній та часово-частотній областях, які дозволяють ураховувати спектральні характеристики мовлення, тембр, інтонаційні особливості, резонансні властивості вокального тракту тощо. Виділені ознаки повинні відповідати кільком ключовим вимогам: бути стійкими до зовнішніх шумів, варіативності мовлення (швидкість, гучність, емоційність), зміни обладнання та каналу зв’язку, а також забезпечувати достатній ступінь розрізнення між голосами різних користувачів.

Таким чином, етап виділення ознак слугує своєрідним містком між сирим цифровим аудіо та високорівневими статистичними або нейронними моделями, забезпечуючи максимально щільне і стійке до перешкод описання голосу як біометричного маркера.

2.2.4. Побудова голосового відбитка

Побудова голосового відбитка (voiceprint) реалізується як окремий функціональний модуль у структурі системи голосової автентифікації і відіграє ключову роль у процесі формування біометричного шаблону користувача. Його основне завдання полягає в перетворенні послідовності ознак мовного сигналу, отриманої на попередньому етапі, у компактне векторне представлення фіксованої розмірності, яке відображає унікальні характеристики голосу конкретного користувача. Отриманий вектор часто називають ембедінгом голосу (speaker

embedding), або ж голосовим відбитком, за аналогією з іншим біометричним параметром – відбитком пальця.

На відміну від ознак, які є локальними в часовій ділянці (наприклад, MFCC для одного фрейму), голосовий відбиток агрегує інформацію на глобальному рівні: з усього мовного сегменту – незалежно від його довжини – формується єдиний вектор. Таким чином, побудова відбитка виконує функцію узагальнення динамічного, нестаціонарного процесу у форматі векторної репрезентації, придатної для порівняння, класифікації або зберігання у базі даних. У науковій літературі цей процес також класифікується як задача embedding learning або representation learning у контексті персональних акустичних моделей.

Залежно від покоління системи, архітектури та методу навчання, розрізняють кілька основних підходів до побудови голосового відбитка:

Gaussian Mixture Model – Universal Background Model (GMM-UBM) – один із перших статистичних підходів, що отримав широке застосування в класичних системах біометричної автентифікації за голосом. У рамках цієї моделі вектор ознак, зокрема коефіцієнти мел-частотного кепстрального перетворення (MFCC), апроксимується сумішшю багатовимірних нормальних (гаусівських) розподілів.

На першому етапі формується універсальна фоново-нормативна модель (Universal Background Model, UBM), яка відображає усереднені статистичні характеристики мовлення великої вибірки користувачів. Далі для кожного конкретного користувача здійснюється персоналізація моделі шляхом адаптації параметрів (середніх значень, коваріацій та вагових коефіцієнтів) на основі наданого зразка мовлення. Як правило, така адаптація виконується за допомогою алгоритму максимізації апріорної правдоподібності [61].

Незважаючи на інтерпретованість, відносну простоту реалізації та низьку обчислювальну складність, підхід GMM-UBM має низку обмежень. Зокрема, він демонструє нестійкість до змін акустичних умов запису та обмежену здатність до узагальнення в умовах дефіциту навчальних даних.

i-vector (Identity Vector). Розвитком статистичних методів стало впровадження так званих і-векторів, які реалізують ідею факторизації варіацій

мовного сигналу у низьковимірному просторі. Суть методу полягає у проєкції супер-вектора GMM-параметрів (отриманого після адаптації UBM) у лінійний підпростір, з урахуванням тотальної варіації (total variability space).

Таким чином, i-vector є компактним фіксованим представленням мовлення, що поєднує як інформацію про ідентичність мовця, так і про канальні особливості (шум, мікрофон, середовище тощо).

Хоча цей підхід показав високу ефективність у багатьох задачах, він вимагає подальшого застосування методів нормалізації (наприклад, LDA або PLDA), щоб відокремити канальні й персональні варіації. i-vector системи демонструють обмеження при коротких зразках або складних акустичних умовах.

x-vector (DNN-based embeddings). Сучасні системи побудови голосового відбитка значною мірою орієнтовані на використання глибоких нейронних мереж. Архітектура x-vector передбачає подачу послідовності ознак (наприклад, MFCC або мел-спектрограм) на вхід багатошарової нейромережі (зазвичай TDNN або CNN), яка виконує агрегацію інформації у часовій ділянці та формує вектор фіксованої довжини.

Мережа навчається у задачі класифікації за ідентифікатором користувача (supervised learning), після чого проміжний шар, що відповідає за векторне представлення, використовується як голосовий ембедінг. Такий підхід забезпечує високу стійкість до міжсесійних змін, адаптивність до різних довжин мовлення та хорошу здатність до узагальнення.

Удосконаленням цього підходу є моделі типу ECAPA-TDNN, які використовують механізми уваги (attention mechanisms), залишкові блоки, канальну нормалізацію та інші сучасні архітектурні прийоми для отримання більш інформативного відбитка.

Трансформерні ембедінги (Transformers / Self-Supervised models). Останнім часом у задачах голосової біометрії активно застосовуються попередньо навчені трансформерні моделі (Wav2Vec 2.0, HuBERT, Whisper), які здатні формувати глибокі багаторівневі представлення мовлення. Ці моделі тренуються у самонавчальному режимі (self-supervised learning) на великих корпусах і

демонструють здатність витягати семантичні та акустичні характеристики з високим ступенем дискримінативності. У задачі побудови голосового відбитка використовується або усереднений вектор ознак з одного з шарів, або спеціально навчений проєкційний блок поверх трансформера.

Таким чином, побудова голосового відбитка є процесом формування фіксованого представлення особистості мовця в багатовимірному ознаковому просторі. Якість цього представлення визначає ефективність усієї системи, оскільки подальше порівняння виконується саме між такими векторами. На цьому етапі відбувається стиснення, узагальнення та виділення найбільш релевантної персональної інформації, що забезпечує диференціацію між користувачами за їх акустичним профілем.

2.2.5. Верифікація особи за допомогою порівняння тестового та реєстраційного зразків мовлення

У системах голосової автентифікації порівняння голосових відбитків є завершальним і критично важливим етапом, від якого безпосередньо залежить точність розпізнавання особи. Його основне завдання – визначення ступеня відповідності між отриманим вектором ознак мовлення (тестовим зразком) та шаблоном, попередньо зареєстрованим у базі даних (референтним зразком). За результатами порівняння тестового голосового зразка з еталонними шаблонами система приймає рішення щодо результату біометричної перевірки. Залежно від режиму роботи розрізняють два підходи до встановлення відповідності:

Верифікація (1:1) – перевірка відповідності між тестовим зразком та еталоном, пов’язаним з конкретною особою.

Ідентифікація (1:N) – пошук найбільш схожого еталону серед усіх зареєстрованих користувачів.

У першому випадку порівнюється пара векторів $e_{тест}$ та $e_{еталон}$, у другому – тестовий ембедінг порівнюється з усім набором $\{e_i\}_{i=1}^{N}$ у базі.

Верифікація може бути описана як задача перевірки статистичних гіпотез:

$$\begin{cases} H_0: e_{тест} \sim e_{еталон} \rightarrow \text{користувача авторизовано} \\ H_1: e_{тест} \nsim e_{еталон} \rightarrow \text{користувача не авторизовано’} \end{cases} \quad (2.1)$$

де $e_{еталон}$ – зареєстрований у системі зразок голосу користувача, $e_{тест}$ – тестовий зразок для верифікації [62].

Рішення приймається шляхом порівняння значення метрики подібності з порогом θ:

$$Accept \Leftrightarrow sim(\mathrm{e}_{\text{тест}} \sim \mathrm{e}_{\text{еталон}}) > \theta, \quad (2.2)$$

де sim – функція подібності або міра відстані, θ – порогове значення прийняття рішення про верифікацію.

Метричне порівняння відбитків. На цьому етапі застосовуються різні метрики подібності або відстані, що дозволяють оцінити, наскільки близькими є два вектори у заданому просторі ознак. Найбільш поширеними метриками є:

Косинусна подібність – визначає косинус кута між двома векторами, ігноруючи їхню абсолютну довжину. Широко використовується в системах на базі x-vector та інших нейронних ембедінгів, оскільки такі вектори зазвичай попередньо нормалізуються [63].

$$sim(e_1, e_2) = \frac{e_1 \cdot e_2}{||e_1|| \cdot ||e_2||}, \quad (2.3)$$

де e_1 – ембедінг першого зразка для порівняння, e_2 – ембедінг другого зразка для порівняння.

Евклідова відстань (L2-норма) – обчислює відстань між точками у просторі, враховуючи абсолютні значення різниць між усіма координатами. Менш стійка до масштабних деформацій, але використовується у простіших системах [64].

Манхеттенська відстань (L1-норма) – альтернативна метрика, яка використовується рідше, але іноді демонструє кращу стійкість до окремих шумових впливів [65].

Метод PLDA (Probabilistic Linear Discriminant Analysis) моделює розподіл ембедінгів у просторі ідентичності та фонових змін, що забезпечує кращу диференціацію міжособових відстаней навіть у присутності шумів або канальних варіацій. Він базується на гіпотезі, що кожен вектор ознак формується з двох джерел варіативності: (1) фактор, що залежить від особи, і (2) фактори, зумовлені умовами запису, шумами, каналом тощо. Ці компоненти моделюються як латентні

змінні у статистичній моделі, а сам PLDA виконує оцінювання апостеріорної ймовірності того, що два ембедінги належать одній особі. Такий підхід дає мзогу коректно порівнювати представлення користувачів навіть у складних умовах: при різному обладнанні, з різними фрагментами мовлення та навіть у присутності шуму. PLDA також легко комбінується з різними типами нейронних ознак, включаючи x-vector, ECAPA-TDNN, ResNet тощо [66].

Порогове прийняття рішення. Система виконує тестування на основі метрики схожості або ймовірнісної оцінки. Якщо обчислене значення схожості $s(x,y)$ перевищує заздалегідь встановлений або динамічно адаптований поріг τ, приймається позитивне рішення щодо автентичності користувача. У протилежному випадку зразок відхиляється як такий, що належить іншій особі або не проходить перевірку.

Адаптивне порогування та довірчі бали. У практичних реалізаціях широко застосовується індивідуалізоване порогування, при якому значення τ визначається для кожного користувача окремо, з урахуванням статистичних характеристик внутрішньокласових варіацій. Такий підхід є особливо ефективним у випадках високої міжкористувацької варіативності мовних шаблонів або при обмеженій кількості реєстраційних записів.

Крім того, сучасні системи часто доповнюють процес прийняття рішення розрахунком довірчої оцінки (*confidence score*) – числового індикатора ступеня відповідності вхідного зразка еталонному. Цей показник може слугувати основою для реалізації гнучких політик безпеки, наприклад, ініціювання додаткової перевірки в умовах невизначеності або автоматичного перенаправлення зразка до модуля виявлення атак типу спуфінг.

Таким чином, етап порівняння та прийняття рішення реалізує остаточну фазу роботи системи голосової автентифікації – прийняття класифікаційного рішення на основі математичного аналізу ембедінгів. Від коректності реалізації цього етапу залежить здатність системи ефективно розпізнавати користувачів, протистояти атакам, зберігати баланс між зручністю та безпекою, а також працювати в умовах акустичної нестабільності. Вибір метрики, методів порівняння та порогів повинен

узгоджуватися з архітектурою всієї системи та вимогами до її продуктивності в конкретній галузі застосування.

2.3. Класичні підходи до реалізації голосової автентифікації

Класичні підходи до реалізації систем голосової автентифікації становлять фундамент для розвитку сучасних біометричних технологій і залишаються актуальними як з теоретичної, так і з прикладної точки зору. Попри значний прогрес у сфері глибинного навчання, традиційні методи продовжують використовуватися як базові інструменти для попередньої обробки даних, побудови простих моделей або як еталонні методики для порівняльного аналізу нових рішень. Вони заклали основу для перших комерційних і академічних систем автентифікації, що підтверджується численними дослідженнями у сфері автоматичної обробки мовлення.

Одним із ключових методів є динамічне зіставлення часових рядів (Dynamic Time Warping, DTW), що дає змогу обчислювати мінімальну відстань між двома часовими рядами незалежно від їхньої довжини та локальних деформацій у часовій площині. Завдяки здатності адаптивно вирівнювати мовні сигнали, DTW є особливо ефективним у задачах голосової автентифікації, де природна варіативність темпу мовлення становить одну з основних проблем. Використання цього методу компенсує часові зсуви між тестовим і реєстраційним аудіозаписами, що сприяє підвищенню точності розпізнавання. [67].

Ще один важливий класичний підхід ґрунтується на Gaussian Mixture Models – статистичних моделях, що дозволяють описувати розподіл мовних ознак як суму кількох гаусових компонентів. Цей підхід продемонстрував високу ефективність у моделюванні складних акустичних залежностей та використовується у біометричних системах автентифікації завдяки своїй гнучкості та здатності працювати з невеликими вибірками. У сучасних системах GMM нерідко поєднується з нейронними мережами для побудови гібридних моделей, що сприяє значному підвищенню стійкості систем до змін мовного середовища та впливу шумів.

Не менш важливим аспектом класичних підходів є методи екстракції ознак, які забезпечують перетворення мовного сигналу у компактну та інформативну форму, придатну для подальшої класифікації. З-поміж них провідне місце посідають коефіцієнти мел-частотного кепстрального перетворення (Mel-Frequency Cepstral Coefficients, MFCC) та лінійне предиктивне кодування (Linear Predictive Coding, LPC). Метод MFCC, який імітує особливості людського сприйняття звуку через мел-шкалу, забезпечує формування коефіцієнтів, що компактно відображають спектральні властивості мовлення. LPC, у свою чергу, моделює мовний тракт як лінійну систему і ефективно апроксимує мовний сигнал через параметри лінійної предикції. Обидва методи стали стандартом де-факто в задачах мовної біометрії та досі використовуються як ключові блоки в численних системах.

У наступних підрозділах детально розглянуто математичні основи, особливості реалізації та порівняльні характеристики MFCC і LPC у контексті їхнього використання для побудови надійних систем голосової біометричної автентифікації. Особливу увагу буде приділено аналізу сильних і слабких сторін кожного методу та їхньому місцю в сучасних комплексних архітектурах.

2.3.1. Мел-частотні кепстральні коефіцієнти

Коефіцієнти мел-частотного кепстрального перетворення є одним із найпоширеніших методів вилучення ознак у системах автентифікації мовців. Їхня ефективність зумовлена здатністю компактно й інформативно відображати спектральні характеристики мовного сигналу, що істотно підвищує точність розпізнавання індивідуальних особливостей мовлення. Використання MFCC базується на психоакустичних особливостях людського слуху, зокрема на нелінійності сприйняття частоти та інтенсивності звукових сигналів. [68].

Процес обчислення MFCC складається з кількох послідовних етапів. На першому етапі аудіосигнал сегментується на короткі перекривні вікна тривалістю 20–40 мс, у межах яких сигнал можна вважати стаціонарним. До кожного вікна

застосовується віконна функція (зазвичай — Хеммінга) з метою зменшення спектральних витікань і підвищення точності подальшого спектрального аналізу.

Другий етап передбачає перетворення кожного вікна у частотну область за допомогою швидкого перетворення Фур'є (FFT), результатом чого є амплітудний спектр сигналу. Далі спектр інтерпретується у мел-шкалі, що моделює особливості сприйняття частот людським слухом. Перехід до мел-простору здійснюється за допомогою банку фільтрів, розташованих таким чином, щоб забезпечити високу роздільну здатність у низькочастотному діапазоні та нижчу — у високочастотному, що відповідає фізіологічній чутливості слухового апарату.

На третьому етапі здійснюється логарифмування енергетичних коефіцієнтів, отриманих після фільтрації, що відображає нелінійність сприйняття гучності та переводить мультиплікативні зміни спектра у зручну для обробки адитивну форму. Завершальним етапом є дискретне косинусне перетворення (DCT), яке зменшує кореляцію між коефіцієнтами та формує компактне векторне представлення. Як правило, використовують перші 12–13 коефіцієнтів MFCC, які найбільш інформативно відображають спектральну огинаючу мовного сигналу.

Процес побудови MFCC коефіцієнтів наведений на рисунку 2.3.

Рисунок 2.3. Процес побудови MFCC коефіцієнтів [69]

В автентифікації мовців MFCC використовуються для екстракції унікальних фізіологічних (структура голосових шляхів) та поведінкових (характер артикуляції, інтонація) ознак мовця. Ці коефіцієнти слугують вхідними ознаками для

різноманітних моделей класифікації, зокрема моделей на основі Gaussian Mixture Models, Hidden Markov Models, Support Vector Machines, а також сучасних глибоких нейронних мереж. Для підвищення стійкості моделей до динамічних змін у мовленні часто використовують похідні коефіцієнтів MFCC – дельта- (перший порядок) та дельта-дельта-коефіцієнти (другий порядок), які відображають зміни спектральних характеристик у часі.

Перевагами використання MFCC у системах автентифікації мовців є їхня обчислювальна ефективність, стійкість до змін інтенсивності сигналу та висока інформативність щодо спектральних особливостей мовлення. Разом із тим варто зазначити, що MFCC є чутливими до фонових шумів і реверберації, що вимагає застосування додаткових методів попередньої обробки сигналу або використання більш стійких до шуму варіантів ознак.

2.3.2. Лінійне предиктивне кодування

Лінійне передбачувальне кодування є одним із найстаріших і водночас ефективних методів вилучення ознак для обробки мовлення, який знайшов широке застосування в задачах ідентифікації та верифікації мовців. Основна ідея LPC полягає у моделюванні мовного сигналу як результату проходження білого шуму або збуджуючого сигналу через лінійний фільтр, параметри якого характеризують резонансні властивості голосового тракту мовця [70]. Це забезпечує ефективне вилучення інформації про формантну структуру мовлення, яка є унікальною для кожного індивіда.

Процес вилучення ознак за допомогою LPC починається з попередньої обробки аудіосигналу. Аналогічно до методів, що застосовуються у вилученні MFCC, мовний сигнал ділиться на короткі перекривні вікна тривалістю близько 20–30 мс. Це дає змогу враховувати стаціонарність мовного сигналу у межах кожного вікна. Для зменшення спектральних витікань до кожного сегмента застосовується віконна функція (зазвичай вікно Хеммінга або Хеннінга). Основним етапом є лінійне передбачення, у межах якого вважається, що поточне

значення мовного сигналу можна апроксимувати лінійною комбінацією кількох попередніх значень. Такий підхід можна описати рівнянням:

$$s(n) = -\sum_{k=1}^{p} a_k s(n-k) + e(n), \quad (2.4)$$

де $s(n)$ – поточне значення мовного сигналу, a_k - коефіцієнти лінійного передбачення (LPC-коефіцієнти), p - порядок моделі (кількість попередніх значень, що враховуються), $e(n)$ – похибка передбачення, або збуджувальний сигнал (excitation signal). Завдання полягає у знаходженні набору коефіцієнтів $\{a\ k\}$, які мінімізують середньоквадратичну похибку між фактичним значенням сигналу та його передбаченим значенням. Для цього найчастіше застосовується метод автокореляції, який приводить до системи нормальних рівнянь, що потім розв'язується з використанням алгоритму Левінсона-Дурбіна. Цей алгоритм забезпечує як високу обчислювальну ефективність, так і чисельну стабільність, що робить його стандартним вибором у практиці LPC-аналізу.

Отримані коефіцієнти LPC відображають форму мовного тракту, який діє як резонансний фільтр. Оскільки ця форма є фізіологічною характеристикою конкретної людини, LPC забезпечує ефективне захоплення індивідуальних ознак мовлення. Окрім самих коефіцієнтів, часто використовують кепстральні коефіцієнти на основі LPC (LPCC), які усувають кореляцію між параметрами і покращують розпізнавання мовця, особливо при використанні статистичних моделей класифікації.

У задачах ідентифікації та верифікації мовців LPC-ознаки зазвичай використовуються як вхідні параметри для різних класифікаторів, таких як Gaussian Mixture Models (GMM), Hidden Markov Models (HMM) або Support Vector Machines (SVM). Сучасні системи також інтегрують LPC-ознаки з іншими типами ознак (наприклад, MFCC чи періодичними ознаками) для покращення стійкості до шуму та змін у мовленні. Основними перевагами LPC є його обчислювальна ефективність, компактність представлення та здатність ефективно моделювати формантну структуру мовлення. Однак метод чутливий до фонових шумів і

реверберації, що може призводити до спотворення коефіцієнтів передбачення. Для подолання цього недоліку використовують попереднє шумозаглушення або комбінування LPC із більш стійкими методами вилучення ознак.

2.4. Методи екстракції голосових ознак на основі нейронних мереж

Упродовж останнього десятиліття методи екстракції ознак на основі нейронних мереж зазнали інтенсивного розвитку, істотно змінивши підходи до побудови систем голосової автентифікації. На відміну від класичних технік, що ґрунтуються на ручному проектуванні ознак (наприклад, MFCC або LPC), нейронні мережі забезпечують автоматизоване навчання ефективних репрезентацій мовного сигналу безпосередньо із сирих аудіоданих або їх спектральних перетворень. Використання технологій штучного інтелекту дозволило суттєво підвищити точність, безпеку та зручність процесу автентифікації особи за голосом. Сучасні підходи, побудовані на основі методів глибокого навчання та машинного аналізу даних, враховують не лише акустичні властивості мовлення, але й мовні, стилістичні та поведінкові особливості мовця, що сприяє підвищенню стійкості систем до спуфінгових атак і забезпечує їхню адаптивність до широкого спектра прикладних сценаріїв.

Одним із ключових досягнень є застосування архітектур трансформерів, таких як BERT та GPT, які здатні враховувати як акустичні, так і мовні ознаки голосу. Завдяки цьому вдається суттєво підвищити точність розпізнавання навіть у складних акустичних умовах [71]. Інтеграція мовного контексту в процес авторизації відкриває нові перспективи для побудови більш ефективних систем автентифікації користувачів.

Застосування моделей штучного інтелекту дає змогу враховувати не лише власне мовні характеристики голосу, а й контекст висловлювання, що сприяє підвищенню точності ідентифікації користувача та зменшенню кількості помилкових рішень.Сучасні мульти-модальні системи аналізують додаткові фактори, такі як мова тіла або відеодані з камер спостереження, щоб підтвердити

особистість користувача [72]. Це ускладнює підробку даних для зловмисників і підвищує загальний рівень безпеки системи.

Важливим напрямком розвитку є застосування генеративних моделей, зокрема GAN (Generative Adversarial Networks). Ці моделі дозволяють ефективно виявляти фальсифіковані аудіозаписи, створені за допомогою технологій маніпулювання звуком, і значно підвищують рівень захисту систем голосової авторизації [73].

Крім того, новітні методи враховують динамічні характеристики мовлення, такі як емоційний стан або рівень стресу користувача. Відомо, що емоційні зміни можуть суттєво впливати на тембр та інтонацію голосу. Тому врахування цих аспектів дає змогу створювати більш адаптивні системи, здатні зберігати точність навіть в умовах змін емоційного стану.

Актуальним також є використання метаданих голосу, таких як тембр, ритм, інтонація, акцент або рідна мова користувача. Врахування цих параметрів дає змогу суттєво підвищити точність автентифікації, особливо в умовах високої міжособистісної схожості голосів або в шумних середовищах. Таким чином, новітні підходи, що базуються на застосуванні штучного інтелекту, сприяють інтеграції різних типів біометричних, голосових та поведінкових даних у мульти-модальні системи ідентифікації та верифікації, що забезпечує високий рівень точності, надійності та безпеки [74].

Одним із найбільш перспективних напрямків розвитку систем голосової авторизації є використання генерування ембедінгів – компактних векторних представлень голосових характеристик користувача [75].

Генерування ембедінгів базується на використанні глибинних нейронних мереж, зокрема автоенкодерів та трансформерних моделей, для перетворення мовного сигналу у багатовимірний вектор ознак. Однією з головних переваг генерації ембедінгів є здатність систем до навчання та вдосконалення у процесі експлуатації. Застосування методів трансферного навчання дає змогу донавчати модель на нових даних без необхідності повного перенавчання системи, що істотно підвищує гнучкість і економічність розробки.

Індивідуальні векторні репрезентації також дозволяють враховувати поведінкові аспекти користувача, такі як типова манера мовлення або особливості інтонації, що суттєво покращує стійкість систем до атак на біометричні дані.

Крім того, ембедінги можуть інтегрувати різні типи даних: акустичні, лінгвістичні, паралінгвістичні та навіть соціальні сигнали. Це дозволяє створювати моделі, здатні ідентифікувати користувача в умовах фонових шумів, багатоголосся, змін емоційного стану тощо.

Таким чином, генерування ембедінгів є ключовим елементом для створення адаптивних, надійних та високоточних систем голосової авторизації, що можуть ефективно функціонувати в реальному середовищі.

У цьому контексті застосування технік на основі глибоких нейронних мереж, зокрема автоенкодерів і трансформерів, забезпечує можливість отримання високоякісних ембедінгів із голосових сигналів. Такі ембедінги являють собою багатовимірні вектори, що узагальнено відображають характеристики голосу — тембр, інтонацію, акцент, а також індивідуальні особливості мовного патерну користувача. Використання векторних представлень знижує вплив варіативності мовлення, зумовленої емоційним станом або зовнішніми чинниками, що є критичним для підтримання стабільності й точності авторизації.

Однією з ключових переваг ембедінгів є їх здатність до подальшого навчання в межах системи голосової авторизації. Залучення методів трансферного навчання забезпечує адаптацію моделі до нових мовних даних, змін у голосі користувача або варіацій умов комунікації. Це також сприяє підвищенню ефективності розпізнавання мовлення в реальному часі, з одночасним зменшенням потреби у великих обсягах анотованих даних для кожного індивідуального випадку [76].

Крім того, використання ембедінгів дає змогу ефективно вирішувати проблему схожих голосів або стилів мовлення шляхом побудови індивідуальних векторних представлень для кожного користувача. Завдяки високій чутливості до дрібних варіацій мовного сигналу ембедінги забезпечують точніше розрізнення навіть мінімальних відмінностей між голосами, що є важливою передумовою для підвищення рівня безпеки системи автентифікації. Процес генерації ембедінгів

також може бути розширений за рахунок інтеграції додаткових факторів, зокрема поведінкових характеристик, що уможливлює формування багатовимірних векторних репрезентацій, здатних враховувати не лише акустичні параметри голосу, а й контекстуальні особливості взаємодії користувача із системою [77].

Покращення системи голосової авторизації завдяки ембедінгам також включає інтеграцію різних типів даних, таких як акустичні, лінгвістичні та навіть соціальні сигнали. Вектори, що містять ці дані, дозволяють створювати моделі, які можуть працювати в складних умовах, наприклад, під час шумних розмов чи в ситуаціях, коли голос користувача зазнає значних змін через емоційний або фізіологічний стан.

Таким чином, генерування ембедінгів є важливим інструментом для покращення систем голосової авторизації, дозволяючи створювати моделі, що адаптуються до змін у голосі користувачів та працюють з різноманітними типами даних, підвищуючи точність і надійність процесу авторизації. В умовах постійного розвитку технологій штучного інтелекту, такі методи можуть стати основою для створення більш безпечних та адаптивних систем, здатних ефективно функціонувати в реальному середовищі.

2.4.1. Архітектура Titanet

У сучасних системах обробки мовної інформації однією з ключових задач є створення ефективних мовних представлень, які б зберігали індивідуальні характеристики мовців та забезпечували високу роздільну здатність для задач верифікації й ідентифікації. У цьому контексті архітектура Titanet демонструє значні переваги завдяки використанню спеціалізованих глибинних нейронних мереж з оптимізованими обчислювальними характеристиками [78]. На рисунку 2.4. представлено загальну архітектуру моделі Titanet.

Згідно з представленою структурою, модель складається з багаторівневого каскаду блоків, кожен з яких реалізує глибинну сепарабельну згортку, нормалізацію, нелінійну активацію та SE-модуль. Завдяки такій організації забезпечується баланс між здатністю моделі до локального та глобального

контекстного моделювання часової інформації та ефективністю її обчислювальної реалізації.

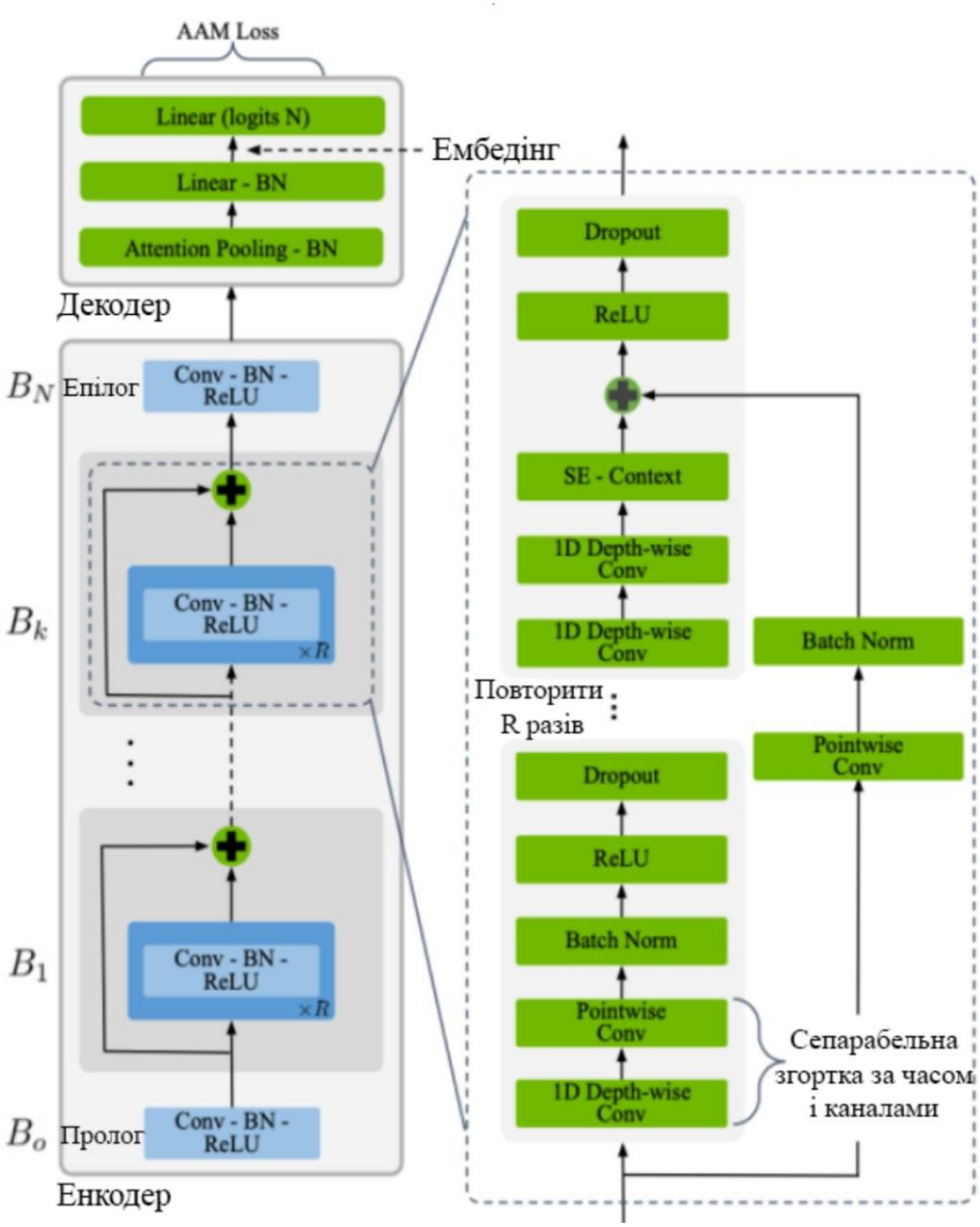

Рисунок 2.4 Архітектура моделі Titanet [77]

Titanet побудована на основі одновимірних глибинних сепарабельних згорток (1D depth-wise separable convolutions) у поєднанні з механізмами глобального контекстного моделювання Squeeze-and-Excitation (SE). Основна

концепція архітектури полягає у зменшенні кількості параметрів та обчислювальної складності без погіршення якості одержуваних мовних ознак, що має вирішальне значення для систем, орієнтованих на обмежені обчислювальні ресурси.

Використання глибинних сепарабельних згорток забезпечує суттєву економію обчислювальних ресурсів за рахунок декомпозиції класичної згортки на два послідовні етапи: перший етап (depth-wise convolution) передбачає обробку кожного каналу окремо, другий етап (point-wise convolution) здійснює лінеарне комбінування каналів за допомогою операції 1×1 згортки. Завдяки такому розділенню досягається ефективніше захоплення локальних часових залежностей мовного сигналу із суттєвим скороченням обсягу обчислень у порівнянні з традиційними згортковими шарами.

Інтеграція механізмів глобального контекстного моделювання через SE-модулі додатково підвищує репрезентативність отриманих ознак. SE-модуль реалізує двоетапний процес обробки інформації: на етапі «стиснення» (squeeze) виконується глобальне середнє згортання, яке агрегує часову інформацію у компактне векторне представлення; на етапі «збудження» (excitation) здійснюється моделювання вагових коефіцієнтів для кожного каналу з метою динамічного переналаштування їх важливості відповідно до глобального контексту вхідного сигналу. Такий механізм сприяє концентрації моделі на найбільш релевантних ознаках мовлення.

Фінальні шари TitaNet включають механізм вагового глобального пулінгу (weighted global pooling), який узагальнює часові ознаки у вектор фіксованої довжини, придатний для подальшого застосування в задачах класифікації, верифікації або пошуку за схожістю.

Для оптимізації процесу навчання застосовується функція втрат angular softmax margin loss, яка вводить додаткову кутову межу між представниками різних класів у просторі ознак. Це сприяє підвищенню міжкласової роздільності та зниженню внутрішньокласової дисперсії t-векторів, що є критичним для точності

систем верифікації мовців, особливо за умов роботи з короткими аудіофрагментами.

Однією з важливих особливостей архітектури Titanet є її здатність до ефективної генерації високоякісних мовних представлень при значно меншій кількості параметрів у порівнянні з іншими сучасними архітектурами, такими як ECAPA-TDNN або різні модифікації ResNet [50]. Це забезпечує широкі можливості для застосування Titanet у системах реального часу, включно із вбудованими пристроями, мобільними рішеннями, а також серверними системами з обробки великих обсягів мовних даних.

Загалом, методологія TitaNet поєднує ефективність у зменшенні обчислювальної складності завдяки глибинним сепарабельним згорткам і підвищену здатність до захоплення глобальних контекстних ознак через SE-модулі. Така архітектура забезпечує збалансованість між точністю та продуктивністю, що робить її придатною для використання як на серверних системах, так і на пристроях з обмеженими ресурсами.

2.4.2. Архітектура ECAPA

Архітектура ECAPA-TDNN (Emphasized Channel Attention, Propagation and Aggregation Time Delay Neural Network) була розроблена для покращення якості мовних представлень у задачах ідентифікації та верифікації мовців, базуючись на ідеях багато масштабного моделювання часових залежностей та вдосконаленого механізму агрегації ознак. Ключовим елементом архітектури є використання Res2Net-блоків, які дозволяють ефективно моделювати часові залежності на різних масштабах [79]. На рисунку 2.5. зображена архітектура мережі ECAPA-TDNN.

У кожному Res2Net-блоці вхідний канал розділяється на кілька підканалів, обробка яких відбувається паралельно за допомогою одновимірних згорткових операцій. Після локальної обробки результати з усіх підканалів агрегуються, що забезпечує багаторівневу репрезентацію часових ознак. Такий підхід дає змогу моделі ефективно захоплювати як короткострокові, так і довгострокові залежності у мовному сигналі, що є критичним для точності розпізнавання мовців.

Наявність залишкових (residual) з'єднань між шарами додатково покращує здатність мережі передавати градієнти у процесі навчання, підвищуючи стабільність оптимізації і сприяючи збереженню корисної інформації протягом глибинних шарів.

Рисунок 2.5. Архітектура мережі ECAPA-TDNN [78]

Завдяки цьому ECAPA-TDNN може бути навчена на більших обсягах даних без ризику деградації якості представлень.

Важливою особливістю архітектури є інтеграція механізмів Squeeze-and-Excitation (SE), які спрямовані на покращення взаємодії між каналами. SE-модуль виконує двоетапну операцію: спочатку за допомогою глобального середнього згортання здійснюється "стиснення" ознак для вилучення глобального контексту, а потім через "збудження" формується набір вагових коефіцієнтів, що масштабують кожен канал відповідно до його інформативності. Завдяки цьому модель отримує можливість динамічно адаптувати обробку сигналу, посилюючи важливі ознаки і пригнічуючи менш релевантні. Поєднання Res2Net-блоків та SE-модулів сприяє глибшому аналізу як часових, так і міжканальних залежностей, що безпосередньо впливає на якість отриманих мовних репрезентацій.

Ще одним важливим компонентом ECAPA-TDNN є вдосконалений модуль статистичного пулінгу з увагою, залежною від каналу. Традиційний статистичний пулінг, що використовує тільки середнє значення та дисперсію по часовій осі, не враховує важливість окремих тимчасових фреймів або каналів. Натомість вдосконалений пулінг у ECAPA-TDNN реалізує механізм, який призначає ваги різним фреймам і каналам залежно від їхнього внеску в загальне представлення. Це сприяє більш точному агрегуванню часових ознак у фіксований вектор, у якому зберігається найбільш інформативна частина сигналу для подальшої обробки.

На наступних етапах обробки цей вектор подається на вхід кільком шарам повнозв’язної нейронної мережі, яка відповідає за формування кінцевих мовних ембедінгів або виконання класифікації. Для навчання моделі використовується функція втрат Angular Additive Margin Softmax (AAM-Softmax) [51], яка вводить додаткову кутову межу між класами у векторному просторі. Це сприяє підвищенню міжкласової роздільності ембедінгів і одночасно зменшує внутрішньокласову варіацію. Такий підхід є надзвичайно важливим для задач верифікації мовців, оскільки забезпечує ефективне розпізнавання особи навіть за короткими аудіофрагментами.

Поєднання багатомасштабного моделювання часових залежностей через Res2Net-блоки, динамічного перерозподілу ваг через SE-механізми та вдосконаленого пулінгу ознак із каналозалежною увагою робить архітектуру ECAPA-TDNN однією з найбільш продуктивних у своєму класі. Вона демонструє переваги над традиційними моделями на основі Time Delay Neural Networks (TDNN), забезпечуючи при цьому прийнятну обчислювальну складність, що є критичним фактором для реального застосування в системах із обмеженими ресурсами.

Таким чином, архітектура ECAPA-TDNN поєднує в собі передові рішення щодо багатомасштабного аналізу часових ознак, ефективної міжканальної взаємодії та розширеної агрегації інформації, що забезпечує її високу ефективність у практичних задачах ідентифікації та верифікації мовців у різноманітних умовах експлуатації.

2.4.3. Мережа WavLM

У статті *"WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing"* представлено архітектуру WavLM – масштабовану нейронну модель для самонавчального попереднього тренування, орієнтовану на універсальне використання в широкому спектрі завдань обробки мовлення [80]. Модель була розроблена з метою покращення як контент-залежних завдань, таких як автоматизоване розпізнавання мовлення, так і контент-незалежних задач, включно з автентифікацією мовців та діаризацією мовлення. Основною ідеєю, що лежить в основі WavLM, є поєднання методів маскованого моделювання мовлення (Masked Speech Modeling) із денойзинговим навчанням, що забезпечує більш ефективне опрацювання мовних патернів моделлю навіть в умовах наявності шумів або при багатоголосих аудіофрагментах.

Архітектурно WavLM базується на модифікованій Transformer-моделі, яка була спеціально адаптована для задач обробки мовних сигналів. На відміну від класичних реалізацій Transformer-архітектур, WavLM використовує механізм відносної позиційної інформованості з вбудованими гейтами. Така модифікація

враховує взаємозв’язки між тимчасовими фреймами аудіосигналу без необхідності жорсткого абсолютного кодування позицій, що є особливо важливим для мовних сигналів із динамічною довжиною та варіативною швидкістю мовлення. Вбудовані гейти у позиційній моделі виконують роль фільтрації нерелевантної інформації, що сприяє підвищенню якості представлення мовних ознак, особливо при обробці довгих аудіофрагментів [81]. На рисунку 2.6. зображена архітектура моделі WavLM.

Рисунок 2.6. Архітектура моделі WavLM [80]

Ключовим компонентом самонавчального попереднього тренування у WavLM є масковане моделювання мовлення, що базується на принципі

приховування випадкових сегментів аудіосигналу під час навчання. Модель навчається відновлювати приховані ділянки, використовуючи контекстну інформацію з немаскованих частин сигналу. Цей підхід дозволяє формувати стійкі мовні представлення, здатні ефективно захоплювати як фонетичні характеристики, так і паралінгвістичні аспекти мовлення. Однак, на відміну від моделей попереднього покоління, таких як wav2vec 2.0, WavLM додатково інтегрує денойзингове навчання, що передбачає моделювання зашумлених аудіофрагментів із наступною реконструкцією "чистого" мовного сигналу. Така навчальна стратегія моделі підвищує стійкість до фонових шумів, реверберацій та інших акустичних завад, що часто трапляються у реальних умовах запису.

Для моделювання багатомовних та багатоголосих сценаріїв у WavLM реалізовано техніку *Mixing Augmentation*, за якої до вихідного аудіофрагменту додаються інші голоси або шумові компоненти. Це змушує модель навчатися ізолювати цільового мовця серед множини голосів або шумових сигналів, що особливо важливо для задач діаризації мовлення та верифікації мовців. У поєднанні з денойзинговим підходом це дає змогу моделі WavLM демонструвати високу якість розпізнавання мовних патернів у складних акустичних середовищах.

Важливою характеристикою моделі є масштаб навчання: WavLM була тренована на великому корпусі аудіоданих обсягом близько 94 тисяч годин, що охоплює широкий спектр мовних стилів, шумових умов та джерел запису. Такий обсяг даних дозволив моделі сформувати універсальні мовні представлення, що можуть бути безпосередньо використані для різних завдань без необхідності суттєвого додаткового донавчання. Більше того, використання оптимізованих архітектурних рішень та ефективних навчальних стратегій дозволило досягти конкурентоспроможних результатів при помірних обчислювальних витратах, що розширює можливості застосування WavLM у практичних системах обробки мовлення.

На етапі інференсу WavLM генерує мовні ембедінги, які можуть бути використані як вхідні ознаки для широкого спектру мовних систем. Завдяки високій дискримінаційній здатності ці ембедінги придатні як для контент-

орієнтованих завдань, таких як автоматизоване розпізнавання мовлення, так і для контент-незалежних задач, включно з верифікацією мовців, діаризацією та класифікацією емоційних станів. За результатами випробувань на загальноприйнятому бенчмарку SUPERB (Speech processing Universal PERformance Benchmark), WavLM перевершує або дорівнює рівню провідних на сьогодні моделей обробки мовлення, демонструючи універсальність і високу якість побудованих представлень.

Таким чином, архітектура WavLM об'єднує інноваційні підходи до моделювання мовних сигналів через масковане та денойзингове навчання, розширені трансформерні механізми обробки часових залежностей і багатомасштабну агрегацію аудіоінформації. Завдяки своїй універсальності, масштабованості та високій стійкості до реальних умов запису, WavLM є одним із найбільш перспективних рішень для універсальної обробки мовлення у широкому спектрі застосувань.

2.4.4. Система pyAnnote

У статті *"pyannote.audio 2.1 Speaker Diarization Pipeline: Principle, Benchmark, and Recipe"*[82] представлено оновлену архітектуру системи діаризації мовлення, реалізовану в інструментарії pyannote.audio версії 2.1. Основною метою запропонованого оновлення є інтеграція кращих практик кластеризаційних методів та нейронних мереж у єдину, ефективну та гнучку систему, здатну забезпечувати точне визначення тимчасових меж мовців і їхню автентифікацію у складних аудіосередовищах. Архітектура системи включає чотири послідовні етапи обробки: локальну нейронну сегментацію мовців, локальне отримання ембедінгів, глобальну агломеративну кластеризацію та фінальну агрегацію результатів.

На першому етапі здійснюється локальна нейронна сегментація мовців, яка базується на використанні енд-ту-енд нейронної моделі для обробки коротких аудіофрагментів тривалістю 5 секунд з кроком у 500 мілісекунд. Така розбивка із частковим перекриттям між вікнами сприяє більш точному виявленню активності мовців у часовому просторі й істотно знижувати складність завдання за рахунок

локалізації обробки. На цьому етапі впроваджується бінаризація ймовірностей активності мовця за заздалегідь визначеним порогом, що є першим серед основних гіперпараметрів системи. Сегментаційна модель забезпечує високу точність визначення моментів початку та закінчення мовлення, однак вона не здійснює глобального узгодження ідентифікаторів мовців, що залишає питання остаточного розпізнавання особистостей відкритим для наступних етапів.

Другий етап передбачає локальне отримання ембедінгів мовців із використанням тих фрагментів аудіо, де активний лише один мовець. Використання таких ізольованих сегментів підвищує якість ембедінгів, оскільки відсутність накладання голосів у навчальних даних зменшує ризик змішування ознак різних мовців. Формування ембедінгів здійснюється на основі фрагментів тривалістю до 5 секунд, що забезпечує стабільніші та більш репрезентативні векторні представлення. Водночас ефективність цього етапу істотно залежить від точності сегментації, виконаної на попередньому етапі: помилки у визначенні активності мовців можуть негативно впливати на якість отриманих ембедінгів.

Третій етап передбачає глобальну агломеративну кластеризацію ембедінгів з метою групування локальних представлень у глобальні кластери, кожен з яких відповідає окремому мовцю на протязі усього аудіофайлу. Для реалізації кластеризації використовується агломеративна ієрархічна кластеризація із центроїдним зв’язуванням, що було обрано через стабільність результатів та обмежену кількість гіперпараметрів у порівнянні з іншими методами, такими як спектральна кластеризація або варіаційні байєсівські підходи. Центроїдне зв’язування демонструє високу стійкість до варіацій у кількості доступних ембедінгів і зберігає конкурентоспроможність на різноманітних тестових наборах, що робить його оптимальним вибором для задач діаризації.

Фінальний етап системи полягає у проведенні фінальної агрегації результатів кластеризації з формуванням кінцевих розміток мовців у часовому просторі. На цьому етапі обчислюється миттєва кількість активних мовців у кожному часовому кадрі, після чого здійснюється узгодження локальних кластерів із часовими мітками. Важливою особливістю є реалізація механізму заповнення коротких пауз

між сегментами одного мовця. Згідно з рекомендаціями таких бенчмарків, як DIHARD та VoxConverse, паузи тривалістю менше 200–250 мілісекунд можуть бути об’єднані в один безперервний сегмент, що сприяє підвищенню природності та безперервності вихідних розміток.

Однією з ключових технічних особливостей системи pyannote.audio 2.1 є використання нейронної моделі ембедінгів на основі архітектури **ECAPA-TDNN**, імплементованої у бібліотеці SpeechBrain. У порівнянні з альтернативними рішеннями, такими як TitaNet або RawNet3, ECAPA-TDNN демонструє вищу якість моделювання як короткострокових, так і довгострокових часових залежностей, що забезпечує кращу дискримінаційну здатність ембедінгів. Завдяки високій роздільній здатності представлень, обрана модель сприяє підвищенню загальної точності діаризації, знижуючи кількість помилкових злиттів мовців або неправильних розділень.

Інструментарій pyannote.audio 2.1 також передбачає можливість використання як попередньо натренованих моделей "з коробки", так і їх адаптації під специфічні завдання користувача. Це досягається шляхом тонкого налаштування гіперпараметрів або донавчання моделі сегментації на спеціалізованих датасетах. Зокрема, оптимізація порогу сегментації, порогу кластеризації та максимально допустимої довжини паузи для об’єднання є критичними факторами для підвищення продуктивності системи при роботі з новими аудіоданими, особливо в умовах обмеженої кількості навчальних прикладів [83].

PyAnnote.audio 2.1 поєднує високу точність діаризації, обчислювальну ефективність та гнучкість конфігурації. Завдяки модульній структурі та широким можливостям адаптації, цей інструмент демонструє конкурентні результати на стандартних тестових наборах і придатний як для наукових досліджень, так і для застосування в системах реального часу. Гнучкість у налаштуванні компонентів дозволяє ефективно масштабувати систему відповідно до специфіки прикладних задач, що робить pyannote.audio 2.1 однією з провідних платформ для автоматизованої діаризації мовлення.

2.5. Порівняння традиційних і нейромережевих методів обробки та розпізнавання голосових зразків

У галузі голосової авторизації застосовуються різноманітні методи, які можна класифікувати на класичні методи, методи машинного навчання, методи глибокого навчання та підходи на основі генерації ембедінгів. Кожен з цих підходів має свої переваги та обмеження, що впливає на їхню ефективність у різних умовах та для різних типів завдань.

Класичні методи голосової авторизації, такі як використання MFCC LPC, є базовими методами для розпізнавання голосових характеристик. Вони використовують математичні перетворення для аналізу акустичних характеристик голосу, виділяючи важливі спектральні компоненти, що описують тембр і частотні ознаки. Ці методи відомі своєю простотою та швидкістю, що дає змогу їх ефективно використовувати у системах з обмеженими обчислювальними ресурсами.

Проте, MFCC та LPC мають суттєві обмеження. Вони погано справляються з шумами або варіаціями голосу в складних акустичних умовах, таких як реверберація, фонові шуми або багатоголосся. Крім того, ці методи не здатні ефективно обробляти зміни в голосі користувача, що можуть виникати через фізіологічні чи емоційні фактори, що обмежує їхню точність у реальних умовах.

Методи машинного навчання, зокрема GMM та HMM, забезпечують крок уперед у порівнянні з класичними підходами завдяки здатності моделювати ймовірнісні розподіли голосових ознак та часові залежності в мовленні. GMM дозволяють моделювати розподіли акустичних ознак для кожного мовця, що підвищує точність і дає змогу краще справлятися з варіаціями голосу. HMM, у свою чергу, забезпечують ефективну обробку часових залежностей, що є важливим для моделювання звукових послідовностей. Хоча ці методи показують кращі результати, ніж класичні підходи, вони все ж мають низку обмежень. Зокрема, вони є вразливими до атак, що використовують синтезовані або записані голоси. Крім того, для їхнього ефективного застосування необхідні великі обсяги даних для навчання моделей, що може бути складним у випадку обмежених ресурсів.

Методи глибокого навчання, зокрема CNN, RNN та трансформери, стали основою для новітніх систем голосової авторизації завдяки своїй здатності ефективно обробляти складні акустичні дані. CNN використовуються для виділення важливих просторових патернів у голосових сигналах, що сприяє зниженню впливу шумів і забезпечує високу точність. RNN, зокрема LSTM мережі, здатні зберігати інформацію про попередні елементи послідовності, що дає змогу моделювати часові залежності у мовленні. Трансформери, такі як Wav2Vec 2.0, дозволяють навчати моделі з використанням великих обсягів неаннотованих даних, що робить їх особливо корисними для розпізнавання мовлення та авторизації.

Однак, попри свою високу точність, методи глибокого навчання потребують значних обчислювальних потужностей і великих обсягів даних для тренування. Це створює проблеми для їх застосування в умовах обмежених ресурсів або в реальному часі.

Підхід на основі генерації ембедінгів є новітнім напрямком у голосовій авторизації. Цей підхід полягає у створенні багатовимірних векторних представлень голосових характеристик, які захоплюють такі ознаки, як тембр, інтонація, акцент та інші специфічні особливості голосу. Такі ембедінги дозволяють моделі знижувати вплив шуму, фізіологічних змін голосу та варіацій в умовах середовища.

Методи, що використовують автоенкодери або трансформери для генерації ембедінгів, дозволяють створювати моделі, які є стійкими до шуму та варіацій у голосі. Це забезпечує високу точність і гнучкість, оскільки такі моделі можуть адаптуватися до змін у голосі користувача без необхідності в значних обсягах аннотованих даних. Підхід на основі ембедінгів забезпечує високу точність і стабільність навіть у складних акустичних умовах, при цьому знижуючи вплив шумів та інших акустичних перешкод, що позитивно впливає на ефективність системи голосової авторизації. Однак цей підхід має й певні недоліки, зокрема потребу в значних обсягах даних для тренування моделей, що може бути проблемою при обмежених навчальних прикладах. Крім того, моделі, засновані на генеруванні ембедінгів, характеризуються високою обчислювальною складністю

під час тренування, що вимагає значних ресурсів і часу для обчислень. Огляд переваг, недоліків та типових сценаріїв використання різних методів обробки голосових сигналів наведено в таблиці 2.1.

Таблиця 2.1.

Порівняння основних методів обробки та розпізнавання голосу за критеріями ефективності, обчислювальних витрат та сфер застосування

Методи обробки та розпізнавання	**Переваги**	**Недоліки**	**Приклад застосування**
Класичні методи (MFCC, LPC)	Простота реалізації, швидкість обробки, низькі обчислювальні витрати	Погана стійкість до шумів, варіацій голосу, недостатня точність у складних акустичних умовах	Розпізнавання голосу в тихих умовах, базові системи авторизації
Методи машинного навчання (GMM, HMM)	Покращена точність, здатність моделювати часові залежності, краща стійкість до варіацій голосу	Вразливість до атак з використанням записів або синтезованого голосу, потреба у великому обсязі навчальних даних	Банківські системи, автентифікація в call-центрах
Методи глибокого навчання (CNN, RNN, Wav2Vec)	Висока точність, здатність працювати в складних акустичних умовах, стійкість до шумів	Великі обчислювальні витрати, потреба у великих обсягах даних для тренування	Сучасні системи розпізнавання мовлення, персональні асистенти

Методи на основі ембедінгів	Висока точність, стабільність у складних умовах, здатність адаптуватися до змін у голосі	Великі обчислювальні витрати, потреба у великих обсягах даних для тренування	Високоточні системи авторизації, багатомовні системи

Порівняння різних методів голосової авторизації показує, що кожен з має свої сильні та слабкі сторони в залежності від умов застосування. Класичні методи є простими і ефективними в випадках з мінімальними шумами та змінами в голосі, проте їх точність обмежена у складних умовах. Методи машинного навчання покращують точність і можуть працювати з більш складними даними, але вони все ще мають обмеження щодо стійкості до атак. Методи глибокого навчання забезпечують найвищу точність і стійкість до шумів, однак потребують значних обчислювальних ресурсів. Методи на основі генерування ембедінгів є найбільш перспективним завдяки своїй гнучкості, здатності адаптуватися до змін у голосі та високій точності, що робить його ідеальним для сучасних та майбутніх систем голосової авторизаціїПорівняння ключових характеристик традиційних та нейромережевих підходів до обробки голосових біометричних даних подане в таблиці 2.2.

Таблиця 2.2.

Порівняльна характеристика традиційних методів та методів на основі нейронних мереж у задачах голосової біометричної автентифікації

Критерій порівняння	**Традиційні методи**	**Методи на основі нейронних мереж**
Обчислювальна складність	Низька – підходять для пристроїв із обмеженими ресурсами.	Висока – потребують потужних обчислювальних ресурсів (GPU/TPU).
Потреба у навчальних	Мінімальна –	Значна – потрібні великі та

даних	ефективні навіть при малих обсягах даних.	різноманітні датасети.
Дискримінативна здатність	Помірна – обмежена чутливість до індивідуальних особливостей мовця.	Висока – добре виокремлюють індивідуальні характеристики голосу.
Стійкість до шуму	Низька – потребують додаткових методів шумозаглушення.	Висока – можуть бути навчені для роботи в зашумлених умовах.
Адаптивність	Обмежена – зміни в умовах потребують ручного налаштування.	Висока – мережі адаптуються до різних акустичних середовищ.
Складність впровадження	Низька – прості в реалізації та налаштуванні.	Висока – вимагають досвіду у розробленні та налаштуванні моделей.
Моделювання контексту	Відсутнє – аналізують сигнал без урахування послідовності.	Присутнє – RNN і трансформери враховують часові залежності.
Гнучкість використання	Обмежена – використовуються переважно для класичних задач.	Висока – підходять для широкого спектру застосувань.
Витрати на підготовку	Низькі – не потребують значних обчислювальних ресурсів.	Високі – навчання може займати значний час і ресурси.
Стійкість до міжмовних варіацій	Обмежена – можуть бути чутливі до зміни мови.	Висока – здатні навчитися універсальним ознакам.

2.6. Безпека системи голосової автентифікації

Блок концепції, представлений на схемі, узагальнює стратегічні напрями, що визначають науково-практичні засади побудови, перевірки й вдосконалення системи голосової біометричної автентифікації в умовах сучасного кіберсередовища. Основна ідея цього блоку полягає у забезпеченні високого рівня захищеності системи від зловмисних впливів, зокрема атак із застосуванням

синтетичних голосів, через поетапне поєднання теоретичних принципів, експериментальних перевірок і технологічного вдосконалення [84].

Початковим елементом блоку є загальна характеристика безпеки системи голосової автентифікації, де формулюються базові поняття, визначаються типові загрози для біометричних систем, а також окреслюється структура можливих векторів атак. Це створює підґрунтя для обґрунтованого вибору методів тестування й аналізу вразливостей.

У подальших підрозділах реалізується практичний аспект концепції. Зокрема, здійснюється експериментальне оцінювання стійкості системи до атак типу voice spoofing, що охоплює створення тестового середовища, генерацію синтетичних зразків голосу, проведення атаки та аналіз результатів. Важливим акцентом тут є використання реалістичних умов, максимально наближених до сценаріїв зловмисного використання технологій клонування голосу.

Наступною логічною ланкою виступає формування методології тестування системи на стійкість. У цьому підрозділі описуються підходи до побудови сценаріїв атак, критерії для оцінювання точності, повноти та надійності роботи системи, а також застосовувані метрики, серед яких FAR, FRR, EER та Accuracy. Запропонована методологія забезпечує відтворюваність експериментів і об'єктивність у порівнянні різних конфігурацій.

Завершальний підрозділ блоку присвячено вдосконаленню системи на основі результатів випробувань. Тут розглядаються можливості інтеграції додаткових засобів захисту, адаптації моделей до нових умов та автоматичного перенавчання у відповідь на зміну типу атак. Важливим є також врахування сучасних технологічних стандартів у сфері біометричної безпеки, що забезпечує відповідність розробленої системи міжнародним вимогам та її практичну придатність для інтеграції у реальні середовища.

Змістовна послідовність цього блоку забезпечує не лише виявлення слабких сторін системи, а й закладає підхід до її безперервного вдосконалення, що відповідає принципам гнучкої, динамічної та захищеної біометричної автентифікації нового покоління.

2.7. Нормативно-правова база біометричної автентифікації за голосом

У сучасних умовах активного розвитку біометричних технологій забезпечення правового регулювання їх впровадження набуває особливої актуальності. Використання голосових даних для автентифікації особи передбачає обробку чутливої персональної інформації, що обумовлює необхідність суворого дотримання міжнародних стандартів, вимог національного законодавства та етичних принципів.

Біометричні дані, зокрема голос, мають унікальну природу: на відміну від традиційних паролів або токенів, вони не можуть бути змінені у випадку компрометації. Це підвищує вимоги до безпеки їх обробки, зберігання та використання, а також до юридичного захисту користувачів. Наявність розвиненої нормативної бази є ключовим чинником довіри до систем біометричної автентифікації як з боку кінцевих користувачів, так і з боку органів регулювання.

На міжнародному рівні основні вимоги до обробки біометричних даних встановлюються стандартами серії ISO/IEC. Зокрема, стандарт ISO/IEC 19794-13:2021 визначає формат представлення біометричних даних голосу, забезпечуючи інтероперабельність між різними системами та платформами [85]. Це забезпечує уніфікацію процедур збору, обробки та передавання голосових відбитків у рамках інтегрованих рішень безпеки.

Окрему увагу приділено питанням стійкості систем до атак. Стандарт ISO/IEC 30107-3:2017 встановлює методики тестування на вразливість біометричних систем до атак із використанням підроблених або синтетичних зразків. В умовах стрімкого розвитку технологій синтезу мовлення й клонування голосу ці вимоги набувають особливої важливості.

Паралельно із технічними аспектами, критичним є забезпечення конфіденційності біометричних даних. Стандарт ISO/IEC 24745:2011 визначає принципи захисту біометричної інформації, включно із шифруванням шаблонів, забезпеченням цілісності даних і захистом проти несанкціонованого доступу [86]. Впровадження цих заходів у практику систем автентифікації значно знижує ризики витоку даних та порушення приватності користувачів.

Окрім технічних стандартів, біометричні системи повинні відповідати вимогам законодавства щодо захисту персональної інформації. На рівні Європейського Союзу такі вимоги встановлені Загальним регламентом захисту даних (GDPR) [87]. Біометричні дані, відповідно до GDPR, віднесені до особливої категорії даних, обробка яких можлива лише за умови отримання явно вираженої згоди суб’єкта або за наявності іншої легальної підстави. Регламент вимагає застосування принципів мінімізації обробки даних, прозорості процесів і гарантування права користувача на доступ, виправлення та видалення його персональної інформації.

В Україні обробка біометричних даних регулюється положеннями Закону України "Про захист персональних даних" [88], який встановлює аналогічні принципи щодо обмеження обробки даних до мінімально необхідного обсягу та забезпечення їх захисту. Обробка біометричних даних можлива лише за наявності окремої письмової згоди суб’єкта даних, а у випадку обробки у сфері безпеки – в межах спеціальних законодавчих норм.

Додатково питання забезпечення інформаційної безпеки регулюються через впровадження стандартів управління безпекою інформації, зокрема ДСТУ ISO/IEC 27001:2015 [89]. Згідно з вимогами цього стандарту, організації мають впроваджувати комплексні заходи щодо контролю доступу, моніторингу обробки даних, захисту мережевої інфраструктури та управління ризиками інформаційної безпеки.

Таким чином, сучасна нормативно-правова база визначає багаторівневу систему вимог до систем біометричної автентифікації за голосом, яка охоплює технічні аспекти формування, захисту і тестування голосових відбитків, а також правові й етичні аспекти обробки персональних даних. Забезпечення відповідності цим вимогам є необхідною умовою для створення ефективних, безпечних і правово обґрунтованих систем біометричної автентифікації, що здатні функціонувати в умовах зростаючих кіберзагроз і підвищених вимог до приватності.

2.8. Висновки до розділу 2

Розділ 2 присвячено побудові концептуальної моделі системи голосової біометричної автентифікації, орієнтованої на використання сучасних нейромережевих трансформерів. Основні висновки можна сформулювати так:

1. Сформовано концептуальну модель голосової автентифікації, яка охоплює ключові компоненти: предметну область (сценарії застосування, вимоги до точності та безпеки), блок нейромережевих технологій (CNN, LSTM, трансформери), систему прийняття рішень, блоки стандартизації й безпеки. Така архітектура забезпечує цілісне й адаптивне функціонування біометричної системи в умовах реальних загроз.
2. Розкрито етапи обробки голосового сигналу – від відбору, попередньої обробки (VAD, знешумлення, нормалізація), до виділення ознак та побудови векторного представлення голосу. Підкреслено важливість стабілізації мовного сигналу та приведення його до уніфікованого формату для забезпечення сумісності з нейромережевими моделями.
3. Проведено порівняльний аналіз традиційних та сучасних методів екстракції ознак. Виявлено, що класичні підходи (MFCC, LPC, GMM-UBM, i-vector) мають обмежену стійкість до варіативності мовлення й акустичного середовища. Натомість нейромережеві підходи (x-vector, ECAPA-TDNN, TitaNet, WavLM) демонструють кращу узагальнювальну здатність, особливо в умовах коротких зразків та зашумлених даних.
4. Описано процес побудови голосового відбитка як ембедінгу, що агрегує інформацію з усього мовного сегмента у компактне векторне представлення, придатне для класифікації та зберігання. Наголошено, що трансформерні архітектури забезпечують вищу дискримінативність ознак і краще відображають як семантичні, так і акустичні характеристики мовлення.
5. Розглянуто процедуру верифікації особи на основі порівняння голосових ембедінгів за допомогою метричних методів (косинусна відстань, PLDA). Підкреслено, що вибір метрики істотно впливає на точність системи й може вимагати додаткового калібрування порогів прийняття рішень.

6. Проаналізовано безпекові аспекти функціонування системи, включно з потребою протидії атакам типу voice spoofing. Зазначено, що нейромережеві моделі, крім точності, повинні забезпечувати стійкість до синтетичних підробок, зокрема клонованих голосів. У цьому контексті рекомендовано інтеграцію модулів антиспуфінгу та механізмів виявлення аномалій.

7. Обґрунтовано необхідність уніфікації та стандартизації: запропонована модель узгоджена з міжнародними вимогами у сфері біометричної безпеки, зокрема стандартами ISO/IEC 30107 та C2PA, що забезпечує її нормативну сумісність і практичну придатність для критичних інфраструктур.

Таким чином, запропонована концепція створює міцне наукове підґрунтя для реалізації ефективних, масштабованих і стійких до атак голосових біометричних систем, здатних функціонувати в умовах підвищених вимог до безпеки, точності та довіри до автентифікації.

РОЗДІЛ 3. ІМПЛЕМЕНТАЦІЯ ВДОСКОНАЛЕННЯ СИСТЕМИ ГОЛОСОВОЇ АВТЕНТИФІКАЦІЇ

У цьому розділі представлено архітектурне вдосконалення системи голосової автентифікації на основі ембедінгів, з описом механізмів реєстрації, калібрування та порівняння голосових зразків.

3.1. Вдосконалення системи голосової автентифікації на основі ембедінгів

Система автентифікації за голосом є прикладом біометричної системи, що базується на аналізі індивідуальних акустичних характеристик мовлення. Архітектурно вона складається з двох основних компонентів: модуля реєстрації, у якому формується еталонне представлення голосу користувача, та модуля верифікації, що виконує порівняння нового зразка з еталонним і приймає управлінське рішення щодо допуску. Обидва етапи реалізуються на основі нейронних мереж та використовують багатовимірні векторні представлення голосу, відомі як ембедінги.

3.1.1. Формування бази зареєстрованих голосових профілів у системах верифікації

Реєстрація є першою та фундаментальною фазою функціонування системи автентифікації за голосом. Саме на цьому етапі закладається індивідуальний еталонний зразок, або ж референсне представлення голосу користувача, яке згодом використовуватиметься як основа для порівняння під час верифікації особи. Основна мета етапу реєстрації полягає в тому, щоби сформувати максимально стійкий і репрезентативний вектор ознак – так званий ембедінг, який точно відображає акустичні особливості мовлення конкретного користувача. [90] Структурна схема процесу реєстрації голосового профілю користувача подана на рисунку 3.1.

Рисунок 3.1. Процес реєстрації користувача в системі голосової біометричної автентифікації

Користувач, який проходить реєстрацію в системі, має надати серію з десяти голосових зразків. Ця кількість зумовлена прагненням досягти балансу між мінімально необхідною варіативністю даних для формування стійкого профілю голосу і прийнятним часом взаємодії з системою. Аудіозаписи, зазвичай, можуть містити або однакову, або різну текстову інформацію, однак критично важливим є те, щоби вони були промовлені у звичному для користувача стилі, з природною інтонацією та темпом, без надмірних емоційних або технічних спотворень.

Після отримання аудіозаписів вони надходять на попередню обробку, яка передбачає видалення зайвих фрагментів тиші, нормалізацію рівня гучності, приведення сигналу до єдиної частоти дискретизації, а також очищення від можливих шумів. Це дає змогу створити уніфікований вхід для подальших етапів, мінімізуючи вплив зовнішніх чинників на процес формування ембедінгів.

Далі оброблений сигнал перетворюється у спектральне представлення. Найчастіше для цього використовуються мел-кепстральні коефіцієнти, що дозволяють перевести мовлення з часової області у простір частотних ознак, які є наближеними до сприйняття людським слухом. Такий підхід дає змогу зосередитися саме на тих компонентах сигналу, що мають найбільшу інформативність у контексті розпізнавання мовця.

Після цього спектральні ознаки подаються на вхід глибокої нейронної мережі, зазвичай спеціалізованої архітектури, призначеної для розв’язання задачі ідентифікації або верифікації мовця. Мережа обробляє вхідні дані та екстрагує з кожного голосового зразка вектор ознак – ембедінг, що має фіксовану довжину (наприклад, 192 або 512 вимірів). Кожен такий ембедінг є стислим, але водночас достатньо інформативним відображенням індивідуального голосу: він охоплює особливості тембру, висоти тону, ритміки мовлення, інтонаційної структури, а також загального стилю вимови.

На наступному етапі всі отримані ембедінги усереднюються. Така процедура необхідна для зменшення впливу випадкових варіацій між окремими зразками, зумовлених такими факторами, як зміни емоційного стану, фонового шуму, швидкості мовлення або незначних фізичних коливань у вимові. Усереднений ембедінг виступає як узагальнене представлення голосу користувача – його цифровий голосовий «відбиток». Завдяки цьому система зберігає не один, потенційно випадковий зразок, а більш стійке векторне представлення, яке краще відображає сталі характеристики мовця.[91]

Завершальним кроком реєстраційної фази є збереження сформованого усередненого ембедінгу у базі даних системи. Цей вектор асоціюється з унікальним ідентифікатором користувача та надалі буде використовуватись як точка порівняння під час верифікації. Таким чином, процес реєстрації забезпечує основу для надійного й персоналізованого механізму автентифікації, дозволяючи системі в майбутньому зіставляти нові голосові зразки з еталонними профілями мовців з високою точністю та стійкістю до зовнішніх впливів.

3.1.2. Вибір і налаштування порогового значення для прийняття рішень у системах голосової автентифікації

Одним із ключових компонентів системи голосової біометричної автентифікації є визначення порогового значення, на основі якого здійснюється рішення про прийняття або відхилення запиту на автентифікацію. Надто низький поріг може призвести до високої ймовірності хибного прийняття сторонніх осіб

(False Acceptance Rate, FAR), тоді як надто високий – до надмірного відхилення справжніх користувачів (False Rejection Rate, FRR). У цьому дослідженні реалізовано адаптивний підхід до визначення індивідуального порогового значення для кожної моделі екстракції ознак, що враховує специфіку формованих ембедінгів та їх метричні властивості.

Для кожного користувача формувався еталонний голосовий шаблон на основі 10 аудіофайлів, які проходили попередню обробку та подальше перетворення в векторні представлення ознак (ембедінги). Залежно від досліджуваної моделі екстракції (наприклад, ECAPA-TDNN, WavLM, TitaNet, тощо), ембедінги мали різну структуру, розмірність і рівень дисперсії. З цієї причини усереднення 10 ембедінгів здійснювалося окремо для кожної моделі, з метою побудови стійкого векторного профілю користувача, який знижує вплив шумів, акцентів і варіацій мовлення.

Побудова пар для тестування. Для кожного користувача формувалося 90 пар зразків для порівняння:

45 пар "істинний–істинний" – кожен справжній тестовий зразок порівнювався з еталонним шаблоном того ж користувача. Це дозволяло оцінити здатність моделі розпізнавати автентичні зразки.

45 пар "істинний–чужий" – еталонний шаблон порівнювався з випадковими зразками інших користувачів, що забезпечувало перевірку на захищеність від атак на основі підміни голосу.

Кожна пара оцінювалася за допомогою косинусної відстані:

$$d_{cos}(x, y) = 1 - \frac{x \cdot y}{\|x\| \cdot \|y\|}, \tag{3.1}$$

де x – ембедінг еталонного шаблону, y – ембедінг тестового зразка.

На основі отриманих 90 значень косинусних відстаней для кожної моделі було сформовано два розподіли: позитивний (істинні збіги); негативний (збіги з іншими класами).

Для кожного з розподілів обчислювалися FAR та FRR при змінних значеннях порогу θ. Визначення оптимального порогу θ^* здійснювалося як значення, що мінімізує абсолютну різницю між FAR і FRR:

$$\theta^* = arg\min_{\theta}|FAR(\theta) - FRR(\theta)| \quad (3.2)$$

У результаті експериментального дослідження для кожної моделі екстракції ознак було визначено окреме оптимальне порогове значення, адаптоване до статистичних властивостей відповідного векторного простору ознак. Такий підхід забезпечує підвищення точності прийняття рішень за рахунок урахування індивідуальних метричних характеристик ембедінгів, що генеруються різними архітектурами моделей. Застосування індивідуалізованих порогів сприяє зниженню рівня помилок критичного типу, що є особливо важливим для високонадійних застосувань, таких як системи фінансової автентифікації або контроль доступу до конфіденційних інформаційних ресурсів. Крім того, така стратегія сприяє масштабованості системи, оскільки уможливлює інтеграцію нових моделей без необхідності приведення їх до єдиного порогового критерію. Таким чином, роздільне порогове налаштування для кожної моделі є обґрунтованим методологічним рішенням, що підвищує адаптивність і загальну ефективність системи в умовах міжмодельної варіативності.

3.1.3. Порівняння підходів до оцінювання подібності векторів ознак у голосових біометричних системах

На етапі верифікації система виконує порівняння поточного голосового зразка користувача з його еталонним представленням, сформованим під час попередньої реєстрації. Для здійснення такого порівняння застосовується метрика подібності, яка кількісно оцінює ступінь відмінності або наближеності між відповідними векторами ознак у багатовимірному просторі. Вибір та точність цієї метрики мають вирішальне значення для функціонування системи, оскільки саме вона визначає здатність моделі коректно розмежовувати автентичні та неавтентичні запити, впливаючи тим самим на загальний рівень її надійності та захищеності.

Найбільш поширеними метриками, що застосовуються у подібних задачах, є косинусна відстань, евклідова відстань (L2-норма), мангеттенська відстань (L1-

норма) та Махаланобісова відстань. У таблиці 3.1 наведено порівняльний аналіз їхніх властивостей у контексті голосової автентифікації.

Таблиця 3.1.

Порівняльний аналіз метрик обчислення відстані між голосовими ознаками

Метрика	**Переваги**	**Недоліки**
Косинусна відстань	Інваріантність до масштабу, стійкість до шуму та флуктуацій, ефективність у високовимірному просторі	Ігнорує абсолютну відстань, може бути недостатньо чутливою в задачах з низькою різноманітністю ембедінгів
Евклідова (L2-норма)	Простота реалізації, інтуїтивна інтерпретація як геометрична відстань	Чутливість до змін гучності, зниження роздільної здатності у високих розмірностях
Мангеттенська (L1-норма)	Стійкість до викидів, менш чутлива до поодиноких змін у координатах	Менш ефективна для високорозмірних ембедінгів, гірша здатність моделювати складні розподіли
Махаланобісова відстань	Врахування кореляцій між ознаками, адаптація до структури даних	Висока обчислювальна складність, потреба у великій навчальній вибірці, нестабільність при мультиколінеарності

Серед зазначених метрик, саме косинусна відстань виявилася найбільш придатною для задачі порівняння голосових ембедінгів. Її головна перевага полягає у фокусуванні на напрямку вектора, а не на його довжині. Це дає змогу системі зосередитися на структурних характеристиках мовлення користувача, ігноруючи

такі неінформативні зміни, як варіації гучності або дистанції до мікрофона. Таким чином, косинусна відстань демонструє високу стійкість до зовнішніх акустичних впливів і є менш чутливою до флуктуацій, що часто виникають у реальних умовах запису голосу.

Додатковою перевагою цієї метрики є її обчислювальна простота, що дозволяє застосовувати її в режимі реального часу без значного навантаження на систему. Це робить її придатною не лише для академічних досліджень, але й для практичного використання в мобільних додатках, голосових інтерфейсах, а також у системах контролю доступу, де критично важливі швидкість і стабільність роботи.

Отже, використання метрики косинусної відстані у системі автентифікації за голосом є теоретично обґрунтованим і емпірично підтвердженим вибором, що забезпечує оптимальний баланс між точністю, стійкістю та обчислювальною ефективністю.

3.1.4. Процес верифікації користувача в системах голосової автентифікації

Модуль верифікації користувачів виконує ключову функцію в системі біометричної автентифікації, забезпечуючи перевірку належності голосового зразка до зареєстрованого мовця. Цей процес є критичним для забезпечення як безпеки, так і зручності автентифікаційної процедури, адже дає змогу з високою точністю підтверджувати особу без використання паролів чи фізичних ідентифікаторів [92].

Архітектура модуля, представлена на рисунку 3.2, реалізує повний цикл обробки від прийому голосового сигналу до прийняття остаточного рішення на основі обчисленої метрики схожості.

Основні етапи функціонування модуля верифікації:

1. Отримання та попередня обробка тестового зразка. Перший етап полягає в отриманні аудіозапису від користувача, який бажає пройти автентифікацію. Зразок

мовлення може бути записаний у реальному часі за допомогою мікрофона або завантажений з локального носія. Для забезпечення якості вхідного сигналу та мінімізації впливу зовнішніх факторів, на цьому етапі проводиться низка процедур попередньої обробки. Ці операції дозволяють зменшити міжсесійні варіації та підвищити стабільність ознак, що будуть витягнуті на наступному етапі.

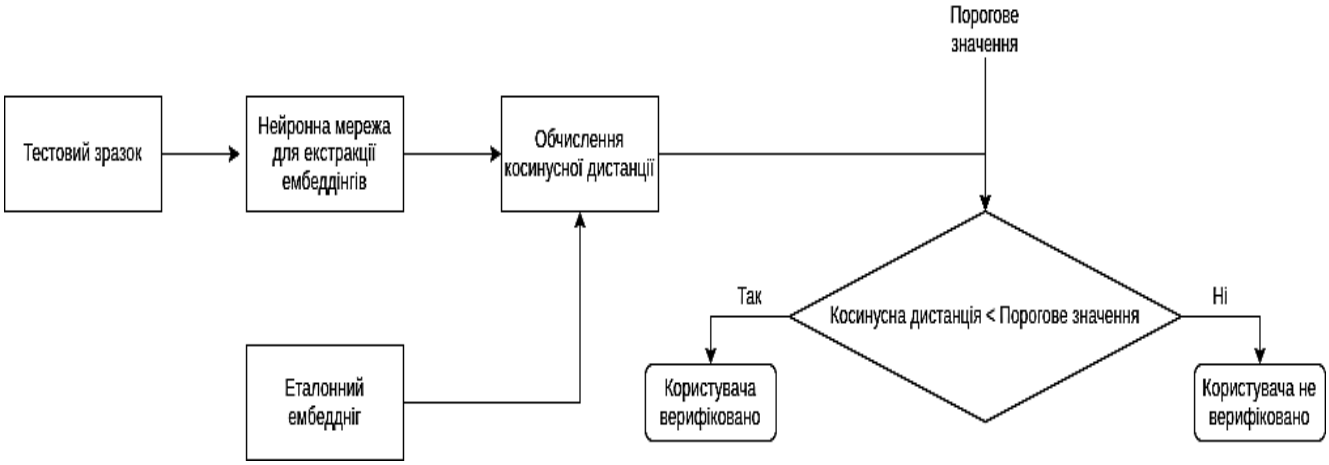

Рисунок 3.2. Схема модуля верифікації системи біометричної автентифікації [93]

2. Екстракція ембедінга. На другому етапі попередньо оброблений сигнал подається на вхід моделі екстракції ознак. Для цього зазвичай використовуються попередньо навчені нейронні мережі, зокрема архітектури типу ECAPA-TDNN, ResNet, WavLM або x-vector. Результатом є ембедінг – щільне векторне представлення, що відображає унікальні голосові характеристики мовця в багатовимірному ознаковому просторі.

3. Обчислення косинусної відстані. Отриманий ембедінг тестового зразка порівнюється із заздалегідь збереженим еталонним ембедінгом зареєстрованого користувача, що був сформований під час процесу реєстрації. Для оцінювання ступеня схожості між цими двома векторами використовується косинусна відстань:

4. Прийняття рішення. Останній етап – це прийняття бінарного рішення на основі значення косинусної відстані. Якщо обчислене значення менше або дорівнює заздалегідь встановленому пороговому значенню (threshold), вважається, що тестовий зразок відповідає еталонному, і користувач успішно верифікований.

Якщо відстань перевищує поріг, система відхиляє верифікацію, інтерпретуючи тестовий зразок як такий, що належить іншому мовцю.

3.2. Опис набору даних для побудови та експериментального оцінювання ефективності біометричної системи

У межах дослідження систем автоматичної автентифікації мовців було сформовано спеціалізований аудіодатасет, що складається виключно з автентичних мовних зразків, отриманих із відкритих джерел. Його створення мало на меті забезпечити контрольоване середовище для аналізу стійкості алгоритмів автентифікації до міжособистісних варіацій мовлення та зовнішніх акустичних факторів. Загальний обсяг датасету становить 600 аудіофрагментів, рівномірно розподілених між 20 унікальними мовцями, які розглядаються як окремі класи в задачі класифікації.

Дані були здебільшого отримані з публічного онлайн-архіву LibriVox, який містить аудіокниги, начитані волонтерами, та забезпечує доступ до широкого спектра англомовних голосів. Такий вибір джерела дозволив сформувати корпус мовлення з високим ступенем варіативності, який, з одного боку, є автентичним, а з іншого – достатньо якісним для автоматизованої обробки.

Кожного мовця представлено 30 аудіозаписами, що формують збалансовану вибірку з однаковою кількістю прикладів на кожен клас. З цієї кількості 10 записів увійшли до тренувального піддатасету, призначеного для побудови й налаштування системи, тоді як решта 20 – до тестового піддатасету, який використовувався для незалежної оцінки її якості. Усі файли мають формат .wav і середню тривалість приблизно 10 секунд, що забезпечує достатній обсяг мовного матеріалу для виявлення характерних акустичних і просодичних ознак кожного спікера.

Склад вибірки охоплює 10 чоловіків і 10 жінок, що забезпечує гендерний баланс. Спікери представляють різні вікові групи та демонструють широкий спектр акцентів і стилів англомовної вимови – від стандартної американської та

британської до регіональних діалектів. Це створює умови для оцінювання роботи системи в контексті лінгвістичної неоднорідності.

Крім того, аудіозаписи відрізняються за рівнем технічної якості: деякі мають студійні характеристики, інші містять фонові шуми, відлуння або інші акустичні артефакти. Така варіативність дає змогу моделювати різноманітні сценарії реального використання систем біометричної автентифікації, включно з умовами зниженої якості сигналу, що є типовими для практичних застосувань.

Узагальнені характеристики аудіодатасету наведено в таблиці 3.2.

Таблиця 3.2.

Характеристики аудіодатасету для тестування системи біометричної автентифікації за голосом

Назва характеристики	**Значення**
Загальна кількість аудіозаписів	600
Кількість унікальних спікерів (класів)	20
Стать спікерів	10 чоловіків, 10 жінок
Записів на кожного спікера	30
Формат аудіо	.wav
Середня тривалість одного запису	~10 секунд
Джерело аудіо	LibriVox (відкриті джерела)
Перший піддатасет (калібрування)	200 записів (10 на кожного спікера)
Другий піддатасет (тестування)	400 записів (20 на кожного спікера)

Особливу увагу було приділено підготовленню тренувального піддатасету. Записи, що до нього увійшли, були ретельно відібрані та попередньо оброблені –

зокрема, аудіо було розділено на окремі речення, щоб кожен фрагмент містив повноцінну мовну одиницю зі значущим обсягом голосової інформації. Така сегментація дозволила стандартизувати тривалість і зміст зразків, а також підвищити якість моделі за рахунок чітко структурованих даних. Крім того, ця процедура мала важливе експериментальне значення: речення, що увійшли до тренувального набору, були використані як текстова основа для створення синтетичного корпусу, озвученого за допомогою технологій клонування голосу. Це дало змогу створити контрольоване середовище для аналізу стійкості системи автентифікації до атак типу "voice spoofing", коли однаковий текст може бути промовлений як справжнім, так і згенерованим голосом.

Таким чином, сформований корпус є не лише збалансованим і репрезентативним, але й методологічно підготовленим для багаторівневого аналізу – від побудови моделей до перевірки їхньої стійкості в умовах наявності потенційних атак з боку систем генерування мовлення.

3.3. Оцінювання результативності голосової біометрії: методологічний підхід та метрики точності

Оцінювання точності біометричної автентифікації є фундаментальним аспектом під час аналізу ефективності систем, що базуються на голосових характеристиках користувача. У контексті практичного впровадження таких систем, особливо у сферах із підвищеними вимогами до безпеки, необхідним є не лише формальне вимірювання продуктивності алгоритмів, а й урахування вартості різних типів помилок, що можуть виникати під час автентифікації. Для цього використовують комплекс метрик, які дозволяють формалізувати поведінку системи при класифікації вхідних голосових відбитків як автентичних або підроблених [94]

Найбільш базовими показниками, що застосовуються для кількісної характеристики ефективності, є рівень хибного прийняття та рівень хибної відмови [95]. Обидві метрики є частковими похідними від загального співвідношення

коректних і некоректних рішень системи, однак кожна з них описує окремий тип помилок.

Показник FAR характеризує частку несанкціонованих спроб доступу, які були помилково класифіковані системою як автентичні. З математичної точки зору цей показник визначається як відношення кількості помилкових прийнять (false accepts) до загальної кількості спроб несанкціонованого доступу:

$$FAR = \frac{FP}{FP+TN}, \tag{3.3}$$

де FP – кількість хибно прийнятих неавторизованих користувачів, а FP+TN – загальна кількість неавтентичних спроб. Низьке значення FAR є критично важливим у системах, де насамперед необхідно забезпечити захист від зовнішніх атак, зокрема в банківській сфері або при дистанційній автентифікації в e-government сервісах.

На противагу цьому, показник FRR визначає ймовірність ситуації, коли справжній користувач системи не був успішно розпізнаний, тобто відбулося помилкове відхилення (false reject). Формально він розраховується як:

$$FRR = \frac{FN}{TP+FN}, \tag{3.4}$$

де FN – кількість невдалих спроб справжніх користувачів, а TP+FN – загальна кількість спроб справжніх користувачів. Високе значення FRR негативно впливає на зручність використання системи та може призводити до зниження довіри користувачів, особливо в мобільних додатках або побутових IoT-пристроях. У випадку голосової біометрії на FRR істотно впливають тимчасові зміни у голосі: хвороби, емоційний стан, фонова акустика або якість мікрофона.

Оскільки FAR і FRR є взаємозалежними функціями від порогового значення прийняття рішення, важливою узагальненою метрикою, яка дає змогу здійснювати порівняльний аналіз різних моделей незалежно від конкретного вибраного порогу, є рівень рівноважної помилки (*Equal Error Rate*, EER). Цей показник відповідає такому значенню порогу, при якому кількість хибних прийнять дорівнює кількості хибних відмов, тобто FAR = FRR. Значення EER обчислюється чисельно на основі емпіричних даних тестування та широко використовується як еталон для

порівняння алгоритмів голосової автентифікації. Чим менше значення EER, тим вищою є загальна точність системи, з огляду на збалансованість між безпекою та зручністю.

Однак у реальних умовах використання систем біометричної автентифікації, особливо в комерційних і критичних інфраструктурах, самих лише показників EER, FAR та FRR недостатньо для повноцінної оцінки. Причиною цього є те, що різні типи помилок мають різну вартість у контексті практичного застосування. Наприклад, хибне прийняття неавторизованого користувача в системі онлайн-банкінгу може призвести до фінансових втрат, тоді як хибна відмова авторизованому користувачеві здебільшого спричиняє лише короткочасне незадоволення.

Для врахування цих факторів використовується функція вартості виявлення (*Detection Cost Function*, DCF), яка поєднує показники точності з апріорними ймовірностями класів і ваговими коефіцієнтами витрат [96]. DCF визначається як лінійна комбінація ймовірностей помилок, зважених на їхню вартість:

$$DCF = C_{miss} \cdot P_{miss} \cdot P_{target} + C_{FA} \cdot P_{FA} \cdot \left(1 - P_{target}\right), \qquad (3.5)$$

де C_{miss} і C_{FA} – відповідно вартість помилки пропуску (хибної відмови) та помилки хибного прийняття, P_{miss} і P_{FA} – ймовірності цих помилок, а P_{target} – апріорна ймовірність того, що запит автентичний. Завдяки введенню вагових коефіцієнтів ця метрика дає змогу адаптувати модель до конкретного застосування, наприклад, підвищити чутливість до безпеки у фінансових транзакціях або, навпаки, забезпечити мінімальний рівень відмов у споживчому додатку.

Окрім описаних вище показників, що характеризують поведінку біометричної системи з точки зору типових помилок, важливим є також загальний індикатор точності класифікації – метрика точності (accuracy). Вона визначається як відношення кількості правильно класифікованих спроб автентифікації до загальної кількості спроб, включаючи як автентичні, так і неавтентичні. Ця метрика відображає загальну ефективність системи, не залежачи від того, чи є рішенням

автентифікація користувача, чи відмова у доступі.Точність роботи системи голосової біометричної автентифікації можна визначити:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (3.6)$$

де TP — істинно позитивне спрацювання (коректна верифікація), FP — хибно позитивне, TN — істинно негативне (відмова у доступі сторонньому користувачу), FN — хибно негативне.

У системах голосової біометрії показник точності відіграє важливу роль при аналізі загальної надійності класифікації. Він демонструє здатність системи успішно розрізняти голосові шаблони автентичних і неавтентичних осіб у рамках заданого порогового значення. Застосування цієї метрики особливо доцільне на етапі порівняння продуктивності різних архітектур, моделей чи параметрів, оскільки вона забезпечує єдиний числовий індикатор узагальненої ефективності.

Узагальнюючи викладене, можна стверджувати, що кількісне оцінювання точності роботи системи біометричної автентифікації за голосовими ознаками повинне ґрунтуватися на комплексному аналізі низки ключових метрик. Показники хибного прийняття і хибної відмови відображають здатність системи приймати правильні рішення в умовах потенційних загроз або помилкових відмов. Рівень рівноважної помилки відображає збалансовану точність системи, характеризуючи точку, в якій частоти помилкових прийомів і відмов збігаються. Натомість функція вартості виявлення враховує контекстуальні чинники — зокрема вартість помилок і апріорну ймовірність автентичності — що має ключове значення для практичного застосування системи в умовах, де наслідки помилкових рішень є асиметричними.

Сукупне використання зазначених метрик є підґрунтям для методологічного підходу до об’єктивного оцінювання ефективності голосових біометричних систем. Це дає змогу не лише порівнювати результати різних підходів, а й адаптувати параметри системи до специфічних вимог певної сфери застосування, балансуючи між рівнем безпеки, точністю розпізнавання та користувацьким комфортом.

3.4. Оцінювання точності, продуктивності та масштабованості імплементованих систем

Оцінювання продуктивності систем біометричної автентифікації є критичним етапом у процесі їх розроблення, тестування та впровадження. Ефективність роботи таких систем визначається не лише їхньою здатністю правильно ідентифікувати або верифікувати особу, але й такими важливими характеристиками, як час обробки запитів та здатність масштабуватись без погіршення якості функціонування. Залежність загальної ефективності від множини технічних і алгоритмічних чинників вимагає багатовимірного аналізу ключових показників продуктивності в умовах реальної або змодельованої експлуатації.

Точність системи відображає її надійність і обумовлює рівень довіри користувачів до біометричної автентифікації. Високі показники точності мінімізують ризики хибних прийомів неавторизованих осіб або необґрунтованих відмов у доступі для легітимних користувачів, що є надзвичайно важливим у сферах, пов’язаних із забезпеченням конфіденційності, безпеки персональних даних та функціонування критичних об’єктів інфраструктури. Окрім цього, точність прямо впливає на прийнятність системи з правової, нормативної та етичної точок зору.

Час обробки запиту є визначальним фактором для забезпечення зручності та швидкості взаємодії користувача із системою. Надмірні затримки можуть суттєво знизити рівень користувацького досвіду, спровокувати відмову від використання біометричних рішень та негативно вплинути на ефективність процесів, що потребують оперативної аутентифікації у режимі реального часу. Особливої актуальності це набуває в мобільних або ресурсно обмежених середовищах, де кожна секунда впливає на стабільність роботи.

Масштабованість системи характеризує її здатність підтримувати прийнятний рівень точності та швидкості обробки при зростанні кількості користувачів або обсягу даних. Високий рівень масштабованості забезпечує адаптивність архітектури до умов динамічного навантаження і зберігає

стабільність функціонування у довгостроковій перспективі. З огляду на тенденції постійного розширення цифрових сервісів і збільшення кількості взаємодій у мережевих середовищах, забезпечення масштабованості є важливою передумовою для успішної інтеграції біометричних технологій у великі інформаційні системи, зокрема в державному, фінансовому та телекомунікаційному секторах.Таким чином, комплексне оцінювання точності, часу обробки та масштабованості є необхідним для об’єктивного аналізу продуктивності біометричних систем, виявлення їх сильних і слабких сторін та обґрунтування шляхів подальшого вдосконалення у відповідності до сучасних вимог кібербезпеки та інформаційно-технологічної інфраструктури

3.4.1. Результати оцінювання точності роботи систем біометричної автентифікації

З метою розроблення високонадійної системи голосової автентифікації було здійснено збір емпіричних даних від 20 респондентів, кожен із яких надав по 10 аудіозаписів із контрольованими мовними параметрами. Ці записи стали основою для побудови еталонних векторних репрезентацій (ембедінгів), що відображають індивідуальні вокально-акустичні характеристики кожного користувача. Основним завданням першого етапу експериментального дослідження було формування усереднених ембедінгів мовців та визначення порогових значень, які є критичними для реалізації процедури бінарної класифікації у процесах автентифікації.

Методологія передбачала багатоступеневу процедуру обробки даних. Початковий набір даних, що складався зі 200 аудіозаписів, був попередньо нормалізований та очищений від шумових артефактів для забезпечення високої якості виділення ознак. Подальша обробка включала екстракцію ембедінгів за допомогою обраних нейромережевих моделей, серед яких TitaNet, PyAnnote.audio, ECAPA-TDNN, WavLM-base-sv та WavLM-base-plus-sv. Векторні репрезентації формувалися у багатовимірному просторі ознак, що акумулювали ключові акустичні характеристики мовлення (тембр, інтонація, артикуляційна динаміка тощо).

Після побудови ембедінгів було сформовано файл формату *.pkl*, що містив по 90 пар ембедінгів для кожного респондента. Для забезпечення коректності експерименту ембедінг-пари були класифіковані на дві категорії: перша група (45 пар) включала порівняння ембедінгів одного користувача між собою, друга група (45 пар) – порівняння ембедінгів досліджуваного користувача з ембедінгами інших респондентів. Такий підхід дозволив згенерувати 1800 пар ембедінгів для повноцінного статистичного аналізу.

Оцінювання результатів здійснювалося за допомогою косинусної міри подібності між ембедінгами, що дозволило визначити ступінь наближеності векторних репрезентацій одного та різних користувачів. На основі отриманих значень було побудовано розподіли подібності для кожної категорії, що дозволило здійснити статистичне моделювання та визначити оптимальні порогові значення для процедури верифікації. Застосування різних моделей ембедінгів дало можливість порівняти їхню дискримінаційну здатність та виявити сильні й слабкі сторони кожної архітектури. Результати оцінювання точності роботи систем голосової біометричної автентифікації на етапі реєстрації зразків подані в таблиці 3.1.

Таблиця 3.1.

Результати оцінювання точності роботи систем голосової біометричної автентифікації на етапі реєстрації зразків

Модель	**FAR**	**FRR**	**EER**	**DCF**	**Порогове значення**	**Точність**
PyAnnote	4.44	4.44	4.44	0.78	0.30	0.96
WavLM-base-sv	7.00	6.89	6.94	0.75	0.93	0.93
WavLM-base-plus-sv	6.56	6.67	6.61	0.77	0.92	0.93
TitaNet	4.11	4.11	4.11	0.70	0.33	0.96
ECAPA	4.67	4.67	4.67	0.73	0.29	0.95

З використанням попередньо визначених порогових значень та усереднених векторних репрезентацій (ембедінгів) було реалізовано другий етап експериментального дослідження на новому наборі даних, що складався із 400 унікальних аудіозаписів із варіативним лінгвістичним наповненням. Основною метою цього етапу було завершальне тестування системи та виокремлення ключових емпіричних характеристик її функціонування. У процесі обробки зазначеного набору даних було сформовано файл у форматі .pkl, який містив 800 пар ембедінгів. Для кожного з досліджуваних спікерів формувалося по 20 пар, де усереднена векторна репрезентація порівнювалася з ембедінгами нових аудіозаписів цього ж мовця, а також ще 20 пар, у яких здійснювалося порівняння з ембедінгами аудіозаписів інших спікерів.

Узагальнені результати наведені в таблиці, яка демонструє незначні відмінності порівняно з першим експериментом, проте підтверджує високий рівень збіжності отриманих результатів. Це свідчить про успішність другого експерименту, ефективність застосування усереднення ембедінгів та налаштування обраних порогових значень.

Таблиця 3.2.

Результати оцінювання точності роботи систем біометричної верифікації на етапі верифікації

Модель	**FAR**	**FRR**	**EER**	**DCF**	**Точність**
PyAnnote	3.38	5.66	4.52	0.70	0.95
WavLM-base-sv	5.56	10.84	8.20	1.16	0.92
WavLM-base-plus-sv	5.31	8.87	7.10	1.09	0.93
TitaNet	2.42	5.91	4.17	0.52	0.96
ECAPA	3.62	6.65	5.14	0.75	0.95

Результати експериментального дослідження, представлені у таблицях 3.1. та 3.2., дозволяють здійснити комплексний аналіз ефективності моделей екстракції ембедінгів для завдань голосової верифікації.

Дані таблиці 3.2. демонструють, що моделі TitaNet та PyAnnote забезпечили найнижчі показники False Acceptance Rate і False Rejection Rate – по 4.11% та 4.44% відповідно, що корелює з високою точністю класифікації (96% для TitaNet і PyAnnote). Особливо варто відзначити модель TitaNet, яка показала найнижчий Equal Error Rate – 4.11%, а також найкращий показник Detection Cost Function – 0.70. Модель ECAPA також демонструє стабільну точність (EER = 4.67%, Accuracy = 95%), що підтверджує її придатність для використання в реальних сценаріях. Натомість моделі WavLM-base-sv та WavLM-base-plus-sv мали дещо вищі показники похибок (EER близько 6.6–6.9%) при аналогічному рівні точності (93%), що свідчить про їхню достатню, проте менш стабільну ефективність у порівнянні з лідерами.

Дані таблиці 3.2, що відображають результати на новому датасеті з варіативним текстовим наповненням, демонструють тенденцію: модель TitaNet знову утримує лідерство, забезпечуючи найнижчий рівень False Acceptance (2.42%) і Equal Error Rate (4.17%) та найменший показник DCF (0.52), що вказує на її високу стійкість навіть у більш складних умовах тестування. Модель PyAnnote також показала стабільно високі результати (EER = 4.52%, Accuracy = 95%), тоді як WavLM-архітектури продемонстрували значне зростання похибок (EER понад 7–8%), що свідчить про чутливість цих моделей до варіацій текстового контенту. ECAPA продовжує демонструвати хорошу ефективність із EER = 5.14% і точністю 95%, підтверджуючи свою надійність.

Загалом, результати обох експериментів підтверджують домінування моделі TitaNet у завданнях екстракції ембедінгів для голосової біометрії завдяки її найкращій узгодженості між низькими помилками та високою точністю. PyAnnote та ECAPA також демонструють конкурентоспроможну продуктивність і можуть розглядатися як альтернативні рішення в системах, де потрібна додаткова гнучкість

або модульність. Натомість моделі WavLM слід розглядати з обережністю в контексті сценаріїв із високим рівнем варіативності мовного контенту.

Отримані результати не лише підкреслюють важливість правильного вибору моделі ембедінгів, а й демонструють критичну роль емпіричного калібрування порогових значень для забезпечення максимальної ефективності системи автентифікації. Це створює міцне підґрунтя для впровадження вдосконалених голосових біометричних технологій із підвищеною стійкістю до різноманітних сценаріїв використання.

3.4.2. Оцінка продуктивності імплементованих систем

Продуктивність системи біометричної автентифікації є одним із ключових аспектів її ефективності у практичному застосуванні. У реальних умовах, особливо в масштабованих або критичних інфраструктурах, недостатня швидкодія системи може спричинити серйозні проблеми — від затримок в обслуговуванні користувачів до збоїв у роботі безпекових механізмів. Зокрема, для онлайн-сервісів, банківських додатків, аеропортових систем контролю доступу та інших сценаріїв, де автентифікація має відбуватися практично миттєво, надмірна затримка може стати фактором, що визначає непридатність рішення до впровадження. У таких випадках навіть різниця у сотих частках секунди може мати практичне значення як з погляду користувацького досвіду, так і з точки зору загальної надійності інформаційної системи.

Під час оцінки продуктивності системи розглядається загальний час обробки одного автентифікаційного запиту — від моменту надходження мовного зразка до прийняття рішення про підтвердження або відхилення доступу. Цей час включає кілька ключових етапів: попередню обробку аудіо (фільтрація шумів, нормалізація, виявлення голосового сегменту), екстракцію ознак (перетворення мовного сигналу у векторне представлення), порівняння з еталонним шаблоном користувача (векторною репрезентацією, збереженою під час реєстрації), а також обчислення метрики подібності та прийняття рішення згідно з визначеним порогом. Кожен із цих етапів має власне обчислювальне навантаження, тому навіть при однаковій

точності різні архітектури можуть суттєво відрізнятися за загальним часом відповіді.

У даному дослідженні для об’єктивного порівняння було протестовано декілька моделей екстракції ознак — серед них ECAPA-TDNN, WavLM та TitaNet. Кожна модель оцінювалася в однакових умовах із використанням єдиного набору тестових аудіозапитів, що забезпечило репрезентативність результатів і дозволило виключити вплив сторонніх факторів на виміри часу. Для кожної конфігурації було зафіксовано середній час обробки одного запиту, а також мінімальні та максимальні значення, що дозволило охарактеризувати не лише продуктивність, а й стабільність роботи моделей. Оцінювання продуктивності здійснювалося в ітераціях за секунду, що є загальноприйнятим критерієм у практиці впровадження розподілених сервісів реального часу. Порівняльні результати представлені у таблиці 3.3.

Таблиця 3.3.

Порівняння продуктивності моделей біометричної автентифікації за швидкістю обробки запитів (ітерацій за секунду)

Модель	**Налаштування порогового значення**	**Порівняння ембедінгів і прийняття рішення**
PyAnnote	11.07	15.06
WavLM-base-sv	9.5	4.55
WavLM-base-plus-sv	9.5	4.34
TitaNet	22.16	33.00
ECAPA	18.72	24.12

На основі експериментального тестування продуктивності різних моделей голосової біометричної автентифікації було отримано показники швидкості виконання двох ключових етапів: екстракції ембедінгів та порівняння ембедінгів із прийняттям рішення. Результати наведені в ітераціях за секунду (іт/сек), що відображає кількість запитів, які система здатна обробити за одну секунду на відповідному етапі.

Найвищі показники продуктивності продемонструвала модель TitaNet, яка забезпечила 22.16 іт/сек на етапі екстракції ембедінгів та 33.00 іт/сек при

порівнянні ембедінгів і прийнятті рішення. Це робить її найбільш ефективною в контексті швидкодії серед протестованих архітектур, що потенційно дозволяє її використання в системах із високим навантаженням або вимогами до швидкої обробки.

На другому місці за продуктивністю опинилася модель ECAPA, яка досягла 18.72 іт/сек на етапі екстракції та 24.12 іт/сек при порівнянні. Вона забезпечує добрий баланс між швидкістю та якістю ембедінгів, що робить її перспективною для більшості прикладних рішень.

Натомість моделі WavLM-base-sv та WavLM-base-plus-sv показали однакові результати на етапі екстракції (9.5 іт/сек), але дещо нижчі показники на етапі прийняття рішення (4.55 і 4.34 іт/сек відповідно), що вказує на потенційні обмеження в сценаріях з високими вимогами до часу відповіді.

PyAnnote, хоча й є досить популярною бібліотекою для верифікації мовців, демонструє порівняно помірні результати: 11.07 іт/сек для екстракції та 15.06 іт/сек для порівняння. Така продуктивність робить її придатною для використання в менш навантажених системах або як еталонний інструмент для порівняльного оцінювання. Загалом, результати показують, що вибір моделі має враховувати не лише її точність, але й швидкодію, особливо в реальному часі. Моделі TitaNet та ECAPA є найбільш придатними для практичного використання у сценаріях з високими вимогами до продуктивності, тоді як WavLM-архітектури можуть бути доцільнішими у задачах, де пріоритет надається якості розпізнавання.

3.4.3. Дослідження масштабованості систем голосової автентифікації на основі нейронних мереж

У процесі масштабування системи голосової біометричної автентифікації зростає кількість зареєстрованих користувачів, кожен з яких представлений власним голосовим профілем. Це призводить до збільшення загальної кількості класів, які система повинна точно розрізняти. У такому багатокласовому середовищі особливу складність становить той факт, що голоси різних людей можуть виявляти високий ступінь схожості за рядом акустичних параметрів –

зокрема, тембром, частотними характеристиками, інтонаційними моделями, манерою мовлення, швидкістю та акцентом.

Ці схожості зумовлюють зниження міжкласової відстані у векторному просторі ембедінгів, який формується в результаті проходження аудіозаписів через нейромережеву модель. Інакше кажучи, вектори, що репрезентують різних мовців, можуть частково перекриватися або розміщуватися на близькій відстані в багатовимірному ознаковому просторі. Така ситуація істотно підвищує ймовірність міжкласової плутанини, тобто хибної класифікації зразка як такого, що належить іншому користувачу.

З математичної точки зору, при зростанні кількості класів у фіксованому векторному просторі високої розмірності відбувається ущільнення розміщення класів, що знижує середню відстань між центроїдами цих класів. Це веде до того, що навіть малі флуктуації в тестовому ембедінгу, зумовлені фоновим шумом або варіацією мовлення, можуть спричинити перетин гіперплощини розмежування, що у свою чергу призведе до помилкового рішення.

У результаті – система стає менш стабільною до зростання кількості користувачів, і без відповідних компенсаторних механізмів (наприклад, адаптивного порогування, гібридних моделей чи агрегації кількох ембедінгів) точність автентифікації закономірно знижується.

З метою ґрунтовного дослідження масштабованості системи голосової біометричної автентифікації було реалізовано п’ять незалежних експериментальних конфігурацій, кожна з яких відповідала певному розміру користувацької бази – 10, 20, 50 та 70 зареєстрованих мовців. Такий підхід дозволив здійснити поетапну оцінку того, як зміни кількісного складу користувачів впливають на верифікаційну ефективність та стабільність роботи системи. У межах кожної конфігурації було проведено побудову еталонних голосових профілів шляхом формування середніх ембедінгів, отриманих з набору навчальних аудіозаписів тривалістю по 10 секунд у кількості десяти зразків на одного мовця. Аудіофрагменти оброблялися через одну з п’яти обраних нейронних архітектур, серед яких домінуючу роль відігравала модель TitaNet, однак також до тестування

було залучено чотири альтернативні мережі, що дозволяло комплексно порівняти ефективність різних стратегій екстракції голосових ознак [97].

Для забезпечення репрезентативності оцінювання до експериментів було включено мовців із різними характеристиками: перші 20 осіб були відібрані з відкритого корпусу LibriVox, що забезпечило контрольовані умови запису, однорідність мовного стилю та якісний звуковий супровід. Решта мовців становили різнорідну вибірку, сформовану з аудіозаписів, отриманих з різних публічних датасетів, зокрема записаних у різних акустичних середовищах, за допомогою різних типів мікрофонів, що дозволило змоделювати реалістичні сценарії застосування системи в умовах змінного рівня шуму, спотворень та технічної неоднорідності.

На основі обробки аудіозразків формувалися індивідуальні ембедінги, які потім усереднювалися для мінімізації впливу фонових шумів, інтонаційних варіацій та інших акустичних флуктуацій, властивих природному мовленню. Отримані усереднені вектори виступали як еталонні представлення мовців та використовувалися для виконання процедури верифікації. Наступним етапом стало калібрування кожної системи, що полягало у визначенні оптимального порогового значення косинусної відстані – критерію, який визначає, чи належить тестовий голосовий зразок до конкретного зареєстрованого користувача. У ході калібрування враховувалися обидва можливі сценарії – позитивна верифікація (зіставлення з еталоном того ж мовця) та негативна (порівняння із векторами інших осіб). Вибір порогу здійснювався на основі аналізу балансу між помилками першого роду (хибне прийняття неавторизованого користувача) та другого роду (відмова у верифікації легітимного користувача). Зі зростанням кількості зареєстрованих мовців спостерігалося поступове зниження оптимального порогу, що свідчить про необхідність більш суворої фільтрації у великих системах для уникнення хибних допусків.

Оцінювання результатів для кожної конфігурації здійснювалося за загальною точністю верифікації. Результати дослідження масштабованості наведені в таблиці 3.4.

Таблиця 3.4.

Точність роботи систем біометричної автентифікації зі зміною кількості користувачів

Модель	**10**	**20**	**50**	**70**
PyAnnote	98.38	95.49	84.56	80.14
WavLM-base-sv	92.58	91.83	86.42	82.65
WavLM-base-plus-sv	93.87	92.93	87.83	84.48
TitaNet	97.42	95.85	94.85	93.87
ECAPA	95.48	94.88	95.91	94.38

У таблиці представлено порівняльний аналіз п’яти сучасних моделей біометричної верифікації за голосом: PyAnnote, WavLM-base-sv, WavLM-base-plus-sv, TitaNet та ECAPA-TDNN. Дослідження проводилось в умовах змінної кількості зареєстрованих користувачів (10, 20, 50, 70), що дає змогу оцінити масштабованість системи та її стійкість до зростання кількості класів. Для кожної конфігурації обчислювалась загальна точність верифікації – частка правильно підтверджених автентичних користувачів та відхилених неавтентичних, виражена у відсотках.

Результати свідчать, що модель PyAnnote, яка демонструє найвищу точність (98.38%) при малій кількості користувачів (10), виявляє значне зниження ефективності зі збільшенням навантаження – до 80.14% при 70 користувачах. Такий спад продуктивності може бути пов’язаний із недостатньою здатністю моделі до узагальнення в умовах великого міжкласового різноманіття, а також з її меншою стійкістю до перенавантаження еталонного простору.

Моделі на базі WavLM (як WavLM-base-sv, так і WavLM-base-plus-sv) демонструють більш плавне падіння точності, зберігаючи прийнятну продуктивність навіть при 70 користувачах (82.65% і 84.48% відповідно). Водночас WavLM-base-plus-sv систематично перевищує свого "базового" варіанту, що вказує на позитивний ефект покращеної архітектури або попереднього навчання.

Модель TitaNet виявила найкращу масштабованість: її точність залишається на високому рівні незалежно від кількості користувачів – від 97.42% при 10 користувачах до 93.87% при 70, тобто зі спадом лише в ~3.5%. Це свідчить про

здатність моделі забезпечувати стабільне векторне представлення голосу навіть у складних умовах багатокласової сегментації.

Найвищу стійкість також продемонструвала модель ECAPA, яка, починаючи з 95.48% при 10 користувачах, досягає навіть кращої точності при 50 користувачах (95.91%) і залишається на рівні 94.38% при 70. Така поведінка може бути пов’язана з ефективними механізмами агрегації ознак, використанням каналів SE-Res2Net та вбудованих технік нормалізації, які дозволяють адаптувати модель до збільшення кількості класів без значної втрати дискримінативної здатності.

Результати проведеного тестування дають підстави для формулювання низки важливих висновків щодо придатності розглянутих моделей до використання в голосових біометричних системах, що функціонують у динамічних або масштабованих середовищах.

Найменшу деградацію точності при збільшенні кількості користувачів демонструють TitaNet та ECAPA, що є критично важливим для реальних застосунків, де система повинна підтримувати десятки або сотні одночасно зареєстрованих осіб.

TitaNet зберігає найвищу стабільність показників у широкому діапазоні класів, демонструючи майже незмінний рівень верифікації. Це дає змогу розглядати її як оптимальний варіант для великих розгортань, наприклад, у корпоративних або державних системах контролю доступу. ECAPA, своєю чергою, забезпечує не лише стабільність, а й аномально високий рівень точності в умовах зростаючої кількості користувачів, що може свідчити про добру адаптацію до різнорідних голосових характеристик і низьку міжкласову плутанину.

PyAnnote, хоча й демонструє найвищі результати у випадку з невеликою кількістю користувачів, швидко втрачає точність при масштабуванні системи, що ставить під сумнів її доцільність у задачах з великою кількістю класів. Моделі на основі WavLM залишаються помірним компромісом між точністю та обчислювальною складністю, проте поступаються ECAPA та TitaNet як за абсолютними показниками, так і за стабільністю.

Таким чином, для задач біометричної верифікації з високими вимогами до масштабованості та стійкості оптимальними є моделі TitaNet та ECAPA, які демонструють найменшу чутливість до збільшення кількості користувачів без значної втрати точності.

3.5. Висновки до розділу 3

Розділ 3 присвячено реалізації технічних рішень для вдосконалення системи голосової біометричної автентифікації на основі нейромережевих ембедінгів. Основні висновки можна сформулювати так:

1. Розроблено архітектуру системи голосової автентифікації, яка включає модулі реєстрації, формування голосових профілів, верифікації користувачів та адаптивного порогового налаштування. Запропонована структура забезпечує високу точність, стабільність і масштабованість системи при зростанні кількості користувачів.
2. Показано ефективність підходу до формування узагальненого голосового відбитка шляхом усереднення ембедінгів, що компенсує акустичні варіації, шумові впливи та міжсесійні відмінності у мовленні користувачів.
3. Проаналізовано та порівняно популярні метрики оцінки подібності ознак у багатовимірному просторі – косинусну, евклідову, мангеттенську та Махаланобісову відстані. Встановлено, що саме косинусна відстань забезпечує оптимальний баланс між стійкістю до шуму, швидкістю обчислення та точністю верифікації.
4. Визначено індивідуальні порогові значення для різних моделей ембедінгів (ECAPA, TitaNet, WavLM), що враховують особливості кожної архітектури та підвищити ефективність прийняття рішень. Такий підхід демонструє перевагу над уніфікованим порогом у мультиархітектурному середовищі.
5. Запропоновано методику формування тестових пар для обчислення FAR та FRR, що дає змогу об’єктивно оцінити продуктивність системи у сценаріях як справжньої, так і несанкціонованої автентифікації.

6. Створено спеціалізований аудіодатасет для експериментального оцінювання системи, який базується на автентичних голосових зразках із відкритих джерел, що гарантує реалістичність тестового середовища.

7. Встановлено, що реалізація біометричної автентифікації на основі ембедінгів є перспективним напрямом для створення адаптивних, точних і безпечних голосових систем, здатних працювати в умовах реального середовища та зростаючих загроз.

РОЗДІЛ 4. ОЦІНЮВАННЯ СТІЙКОСТІ СИСТЕМИ ДО АТАК ТИПУ “VOICE SPOOFING” З ВИКОРИСТАННЯМ ТЕХНОЛОГІЙ КЛОНУВАННЯ ГОЛОСУ

У цьому розділі досліджується стійкість системи голосової автентифікації до атак, здійснених із використанням сучасних технологій клонування мовлення. З цією метою була сформована масштабна база синтетичних голосових зразків, згенерованих за допомогою чотирьох поширених моделей створення клонованого голосу: RVC, ElevenLabs, XTTS та Tortoise. Кожна з моделей використовувалася для синтезу мовлення, що імітує голоси 20 цільових спікерів, зареєстрованих у системі. Це дозволило побудувати експериментальне середовище, здатне відтворювати атаки з різним ступенем схожості та мовною варіативністю. Окрім базового аналізу стійкості до таких атак, у межах дослідження запропоновано підхід до інтеграції особистісних знань користувача як додаткового фактора автентифікації, а також реалізовано тестування моделей для визначення синтетичності мовлення. На основі отриманих результатів сформульовано низку практичних рекомендацій щодо підвищення захищеності біометричних систем у контексті загроз, пов’язаних із технологіями голосового клонування.

4.1. Експериментальний набір даних для перевірки стійкості системи автентифікації до атак на основі клонування мовлення

Оцінювання стійкості системи до атак із використанням технологій клонування голосу вимагало попередньої підготовки відповідних моделей та створення синтетичних аудіозаписів, що імітують голос цільового мовця. У рамках дослідження було застосовано низку сучасних моделей та сервісів голосового клонування, зокрема Retrieval-based Voice Conversion, ElevenLabs, Tortoise та XTTS. Для кожної моделі були використані специфічні підходи до підготовки даних, навчання моделей та синтезу клонованого мовлення.

У випадку моделі RVC процес навчання розпочався зі збирання аудіозаписів голосу цільової особи з відкритих онлайн-джерел. Отриманий датасет складався зі

240 зразків тривалістю по 10 секунд кожен, що сумарно забезпечило 40 хвилин різноманітного мовного матеріалу. Висока варіативність лінгвістичних та просодичних ознак у зібраних записах дала змогу моделі охопити широкий спектр вокальних характеристик мовця, включно з тембром, інтонацією, динамікою мовлення тощо. Навчання моделі здійснювалося протягом 400 епох, що дозволило уникнути перенавчання (overfitting) і забезпечити належну узагальненість. Після завершення тренування генерування мовлення відбувалося у два етапи: спочатку текстовий вхід опрацьовувався системою Coqui TTS, а згенерований аудіопотік передавався до RVC для перетворення на голос цільового користувача.

Для системи ElevenLabs було використано значно менший обсяг даних – лише 25 записів тривалістю по 10 секунд, що становило близько 4 хвилин мовного матеріалу. Незважаючи на обмеженість датасету, записи охоплювали широкий спектр лінгвістичних структур, що дозволило системі сформувати багатовимірний голосовий профіль мовця. Після завантаження аудіофайлів ElevenLabs автоматично аналізувала ключові акустичні характеристики, зокрема тембр, інтонаційні криві та швидкість мовлення, і на їх основі створювала індивідуальну віртуальну голосову модель. Процес генерування синтетичного мовлення відбувався у напівавтоматизованому режимі: користувач надавав текстовий вхід, а система генерувала аудіо, яке з високою точністю імітувало голос оригінального мовця.

Для моделей XTTS і Tortoise застосовувався альтернативний підхід до підготовки даних. Замість використання великої кількості коротких записів, як у випадку RVC чи ElevenLabs, було обрано один суцільний аудіофайл тривалістю три хвилини. У цьому записі були представлені різноманітні мовні конструкції, що дало змогу забезпечити належну фонетичну та просодичну насиченість матеріалу. Особливу увагу приділяли якості вхідного аудіо: вибиралися записи з мінімальним рівнем фонових шумів та високою чіткістю голосу. Це дало змогу покращити якість видобутих ознак голосу і, відповідно, підвищити точність синтезу. Обидві моделі, XTTS і Tortoise, реалізовані на основі сучасних глибоких нейронних архітектур і

здатні з високою достовірністю імітувати індивідуальні особливості мовлення на основі текстового вхідного сигналу.

У межах експериментального дослідження кожен із розглянутих методів клонування – RVC, ElevenLabs, XTTS та Tortoise – був використаний для створення окремих моделей, що імітують голоси 20 різних спікерів. Ці спікери були попередньо відібрані як референтні голоси для побудови та калібрування системи автентифікації. Таким чином, було згенеровано 80 клонованих голосових моделей (по 20 на кожну технологію), причому кожна навчалася виключно на даних одного конкретного мовця для досягнення максимальної індивідуалізації голосового профілю.

Після завершення навчання з кожної моделі було синтезовано по 40 аудіосемплів, що у сукупності сформувало корпус із 3200 зразків синтетичного мовлення. Половина з них (20 семплів) створювалася на основі текстів, що дослівно повторювали фрагменти оригінального тренувального набору відповідного спікера. Такий підхід дозволив оцінити здатність моделі точно відтворювати мовлення в умовах, наближених до навчального середовища. Решта 20 зразків були згенеровані на основі єдиного набору текстів, ідентичного для всіх 20 спікерів. Цей підхід унеможливлював вплив індивідуального мовного контексту і дозволяв здійснювати стандартизоване порівняння результатів між різними користувачами та моделями.

Таке комбіноване тестове середовище, що включало як знайомі, так і незнайомі текстові вхідні дані, забезпечило комплексну перевірку здатності клонованих голосів обходити механізми автентифікації. Воно також дозволило виявити потенційні закономірності у вразливості системи залежно від типу синтетичного аудіо, лінгвістичного контексту та особливостей застосованої технології клонування. Отримані результати стали основою для подальшого аналізу ефективності атак та формування рекомендацій щодо підвищення захищеності голосових біометричних систем. Структура сформованих датасетів представлена в Таблиці 4.1.

Таблиця 4.1.

Структура тестових датасетів для оцінювання стійкості до атак із використанням клонованого голосу

Опис датасету	**Походження аудіо**	**Кількість зразків для кожного класу**	**Тривалість зразка**
Датасет для створення моделі клонування голосу методом RVC	Оригінальне	240	10с
Датасет для створення моделі клонування голосу методом ElevenLabs	Оригінальне	25	10с
Датасет для створення моделі клонування голосу методом XTTS	Оригінальне	1	180с
Датасет для створення моделі клонування голосу методом Tortoise	Оригінальне	1	180с
Датасет для тестування системи голосової автентифікації до атак з застосуванням технологій клонування голосу з використанням фраз, які використовувались при реєстрації користувача	RVC	20	8-12с
	XTTS	20	8-12с
	ElevenLabs	20	8-12с
	Tortoise	20	8-12с
Датасет для тестування системи голосової автентифікації до атак з застосуванням технологій клонування голосу з використанням випадкових фраз	RVC	20	8-12с
	XTTS	20	8-12с
	ElevenLabs	20	8-12с

	Tortoise	20	8-12с

4.2. Методологія тестування системи голосової біометричної автентифікації на стійкість до атак із застосуванням технологій клонування голосу

Для системної та об’єктивної перевірки стійкості біометричної автентифікації за голосом до атак із використанням клонованого мовлення було реалізовано експериментальну методологію, яка моделює типову ситуацію реального функціонування системи в умовах, коли ймовірний зловмисник намагається пройти перевірку, використовуючи синтетично згенерований голос. У цьому сценарії досліджувалося, чи здатна система, що базується на заздалегідь сформованих голосових шаблонах зареєстрованих користувачів, відрізнити справжній голос від підробленого. Кожне вхідне аудіо подавалося як запит автентифікації від конкретного користувача, і система самостійно приймала рішення – приймати чи відхиляти запит.

Еталонні профілі користувачів формувалися на основі автентичних записів, які не використовувалися в тестуванні. Для кожного з 20 користувачів створювався індивідуальний шаблон – усереднений ембедінг, що акумулював характерні ознаки мовлення на основі кількох оригінальних аудіофрагментів. Надалі кожному користувачеві було призначено 40 синтетичних тестових зразків, що поділялися на дві групи. Перша група містила 20 зразків, згенерованих на основі його власного голосу (так звані "позитивні зразки"). Друга – 20 зразків, створених на основі голосів інших користувачів, але позначених як нібито належні цьому користувачеві ("імпостер-зразки").

Усі синтетичні голоси генерувалися чотирма різними технологіями – RVC, ElevenLabs, XTTS та Tortoise. Кожна модель синтезу створювала по 40 зразків на користувача, які додатково розділялися за лінгвістичним критерієм: половина з них містила тексти, які точно повторювали фрази, що використовувалися під час реєстрації (ідентичний текстовий сценарій), а решта – нові фрази, відсутні на

попередньому етапі (відмінний сценарій). Таким чином, загальний тестовий корпус включав 3200 синтетичних аудіозаписів для кожного з двох варіантів текстового наповнення, або 6400 зразків у сукупності на одну модель векторизації.

Кожен синтетичний запис векторизувався, тобто перетворювався у багатовимірне числове представлення – за допомогою однієї з п’яти моделей: Pyannote.audio, ECAPA-TDNN, TitaNet, WavLM-base-sv або WavLM-base-plus-sv. Отриманий ембедінг порівнювався з еталонним профілем користувача за допомогою косинусної міри подібності. Якщо подібність перевищувала заздалегідь встановлений поріг (визначений індивідуально для кожної моделі на етапі налаштування), система приймала рішення, що голос належить користувачеві.

На основі результатів зіставлення зразків і шаблонів здійснювалася класифікація тестів: прийняття свого голосу вважалося правильним позитивним результатом, відхилення чужого голосу – правильним негативним. Помилкове прийняття стороннього голосу або помилкове відхилення свого вважалися відповідно хибно-позитивними та хибно-негативними результатами.

Оцінка якості роботи системи проводилася за єдиною узагальненою метрикою – точністю класифікації (accuracy), яка визначалася як частка правильно класифікованих зразків серед усіх перевірок. На відміну від метрик FAR, FRR чи EER, які потребують побудови ROC-кривих, обрана метрика дозволяла зосередитися на загальній ефективності системи в умовах цілеспрямованої атаки.

Для кожної моделі ембедінгів здійснювалося по 3200 тестів у кожному сценарії текстового наповнення, що загалом становило 16 000 перевірок. З урахуванням двох типів текстів, загальний обсяг експерименту склав 32 000 автентифікацій. Це створює підґрунтя для переходу до аналізу емпіричних результатів, отриманих у ході тестування. Зібрані дані дозволили не лише визначити рівень стійкості кожної моделі до атак зі штучно згенерованими голосами, а й виявити залежність точності розпізнавання від типу клонувальної технології та характеру мовного контексту. Особливу увагу також приділено порівнянню поведінки системи у випадках, коли текст зразка збігається з реєстраційним сценарієм, і коли він відрізняється, що дає змогу оцінити чутливість

до контексту та потенціал генералізації. Отримані результати стали основою для подальшого вдосконалення архітектури системи та формування рекомендацій щодо підвищення її захищеності в умовах розвитку синтетичного мовлення.

4.3. Результати тестування стійкості системи голосової біометричної автентифікації до атак із застосуванням клонованого голосу

У цьому підрозділі наведено результати експериментального дослідження стійкості системи голосової біометричної автентифікації до атак із використанням синтетичного мовлення. Метою тестування було емпіричне оцінювання рівня вразливості системи до атак типу *voice spoofing*, що реалізуються шляхом генерування клонованого голосу із застосуванням сучасних генеративних моделей. Дослідження виконувалося відповідно до розробленої методики (підрозділ 4.2) та передбачало моделювання двох основних сценаріїв атак: із використанням ідентичного до реєстраційного лінгвістичного наповнення та із застосуванням нових фраз, відмінних від текстів навчальної вибірки.

Експериментальна база даних була сформована із залученням чотирьох поширених технологій синтезу клонованого мовлення: Retrieval-based Voice Conversion (RVC), ElevenLabs, XTTS та Tortoise. Для векторизації аудіоданих використовувалися п’ять моделей формування ембедінгів – PyAnnote.audio, ECAPA-TDNN, TitaNet, WavLM-base-sv та WavLM-base-plus-sv, що забезпечило можливість порівняльного аналізу їхньої стійкості до атак із різними рівнями акустичної складності. Оцінювання результатів здійснювалося за допомогою єдиної інтегральної метрики – точності класифікації, яка дозволила визначити частку правильно класифікованих запитів серед загальної кількості спроб автентифікації.

Представлені результати дозволяють об’єктивно оцінити вразливість голосових біометричних систем у контексті розвитку сучасних технологій генерування синтетичного мовлення. Також дослідження ідентифікує ключові чинники, що впливають на ефективність атак, зокрема тип генеративної моделі,

характеристики ембедінгів та особливості лінгвістичного контенту тестових зразків. Детальний аналіз отриманих даних наведено у підрозділах 4.3.1 та 4.3.2.

4.3.1. Тестування системи голосової автентифікації до атак з використанням ідентичних реєстраційним фраз

Метою першого етапу експериментального дослідження було визначення рівня стійкості системи біометричної автентифікації за голосом до атак із використанням клонованого мовлення у випадку, коли лінгвістичний зміст синтетичних аудіозаписів повністю відповідає мовному вмісту автентичних записів, що використовувалися під час навчання системи. Такий підхід максимально нівелює вплив змістових мовних характеристик на результат автентифікації й зосередитися виключно на акустичних та голосових ознаках, що відображають індивідуальні риси мовця.

У межах цього підходу було використано 20 моделей клонування голосу, згенерованих для 20 унікальних спікерів із навчальної вибірки. Для кожної моделі було синтезовано по 20 аудіофрагментів, текстовий зміст яких дослівно відповідав автентичним фразам із корпусу відповідного спікера. Загалом для кожної моделі ембедінгів було проведено по 400 експериментальних порівнянь, що забезпечило достатній рівень статистичної репрезентативності.

У Таблиці 4.2 наведено частку синтетичних записів, помилково прийнятих за автентичні, що відображає ймовірність успішної атаки.

Таблиця 4.2.

Частка синтетичних записів із ідентичним текстом, які успішно пройшли автентифікаційні випробування (%)

	PyAnnote	**wavlm-base-sv**	**wavlm-base-plus-sv**	**Titanet**	**Ecapa**
Tortoise	91.00	79.25	54.25	87.5	85.5
11labs	94.5	93.25	89	94.75	94.5

RVC	95	79.25	69.75	95	95
XTTS	95	73.25	66.75	95	95

Отримані результати демонструють високий ступінь уразливості системи до атак із застосуванням сучасних технологій клонування голосу, зокрема таких, як RVC та ElevenLabs, моделі яких досягли найвищих показників проходження автентифікації – до 95% успішності для деяких моделей ембедінгів.

Водночас, результати для моделей Tortoise та XTTS були дещо нижчими, що може свідчити про меншу точність відтворення вокальних характеристик або про наявність детектованих акустичних артефактів у синтезованих записах.

Моделі ембедінгів також виявили різну ефективність у протидії атакам. Зокрема, WavLM-base-plus-sv продемонструвала найкращу здатність до відхилення синтетичних голосів, фіксуючи найменший рівень помилкового прийняття (до 54.25% для Tortoise), тоді як TitaNet та ECAPA-TDNN показали високий рівень уразливості до всіх типів клонованих голосів.

4.3.2. Тестування системи голосової автентифікації до атак з використанням нових фраз

Другий етап експериментального дослідження був спрямований на оцінку стійкості біометричної системи голосової автентифікації до атак у випадках, коли синтетичні голосові зразки промовляють текст, відмінний від того, який використовувався для навчання системи. На відміну від попереднього сценарію, цей підхід унеможливлює опору системи на мовний контекст, що створює більш реалістичну модель потенційної атаки, де зловмисник має доступ до голосу, але не до тексту, який використовувався під час реєстрації користувача.

Такий сценарій моделює ситуацію, в якій клонований голос застосовується для імітації автентичного спікера в новому мовному контексті, наприклад, під час голосової автентифікації за довільним запитом. Він дає змогу оцінити, наскільки ефективно система може фокусуватися на незмінних акустичних параметрах мовлення, не покладаючись на подібність текстового вмісту.

Процедура клонування голосу, побудова тестових зразків і обчислення ембедингів повторювала підхід, реалізований на попередньому етапі. Для кожного з 20 клонованих голосів, створених із використанням моделей RVC, ElevenLabs, Tortoise та XTTS, було згенеровано по 20 синтетичних аудіозаписів, у яких використовувалися однакові, але нові текстові фрази, що не входили до мовного корпусу автентичного користувача.

У Таблиці 4.3 наведено результати автентифікації синтетичних голосів, згенерованих на основі відмінного лінгвістичного матеріалу. Показники відображають відсоток успішної автентифікації, тобто частку синтетичних записів, які система помилково класифікувала як справжні.

Таблиця 4.3.

Частка синтетичних записів із відмінним текстом, які успішно пройшли автентифікацію (%)

Метод клонування	**PyAnnote**	**wavlm-base-sv**	**wavlm-base-plus-sv**	**Titanet**	**Ecapa**
Tortoise	86	62.25	49	83.75	84.25
11labs	95	91.75	89	94.75	94.75
RVC	95	76.75	65.25	95	95
XTTS	95	74.25	68.25	95	94.75

4.3.3. Висновки за результатами тестування стійкості системи голосової автентифікації до атак із використанням клонованого голосу

Проведене експериментальне дослідження дозволило зробити низку ключових висновків щодо рівня вразливості біометричних систем голосової автентифікації до атак із застосуванням синтетичного мовлення:

1. Система виявила високий рівень вразливості до атак, реалізованих із застосуванням сучасних моделей клонування голосу. У випадках використання синтетичних аудіозаписів із лінгвістичним змістом, ідентичним до текстів реєстрації, ймовірність успішної атаки сягала до 95% для моделей RVC та

ElevenLabs. Це свідчить про суттєві недоліки в існуючих методах голосової автентифікації, які переважно спираються на акустичні характеристики голосу користувача.

2. Застосування синтетичних аудіозаписів із новим (відмінним від навчального) лінгвістичним контентом лише незначно зменшило ефективність атак, що свідчить про домінуючу роль акустичних ознак у процесі автентифікації. Це підтверджує недостатню здатність системи розпізнавати синтетичні голоси на основі суто акустичних характеристик і наголошує на необхідності врахування додаткових факторів.

3. Аналіз ефективності моделей ембедінгів виявив, що жодна з протестованих моделей не забезпечує повноцінного захисту від синтетичного мовлення. Модель WavLM-base-plus-sv продемонструвала порівняно кращу стійкість (найнижчий показник помилкових автентифікацій), але навіть її показники залишаються незадовільними з точки зору безпеки системи. Це свідчить про важливість продовження роботи над удосконаленням таких моделей і створення комплексних рішень для підвищення стійкості систем автентифікації.

4. Моделі TitaNet та ECAPA-TDNN виявили найвищу уразливість і не рекомендуються до застосування без додаткових механізмів захисту. Високий рівень помилкових автентифікацій за допомогою цих моделей робить їх використання ризикованим у реальних системах без відповідних заходів безпеки.

5. Отримані результати свідчать про значну вразливість голосових біометричних систем до атак, що базуються на використанні генеративних моделей синтезу голосу. Це створює високий ризик несанкціонованого доступу до захищених систем, що використовують автентифікацію за голосом. В умовах сучасного рівня розвитку генеративних технологій, що постійно вдосконалюються, необхідність підвищення рівня захисту голосових систем стає очевидною.

У контексті швидкого прогресу в галузі генеративних технологій необхідно зосередити подальші дослідження на розробці методів і технологій виявлення синтетичного мовлення. Зокрема, рекомендовано інтегрувати додаткові механізми детекції клонованих голосів, такі як багаторівнева автентифікація, аналіз

спектральних артефактів синтезу[98], використання мультимодальних біометричних рішень, а також розробляти спеціалізовані нейромережеві архітектури, адаптовані для виявлення штучних голосів.

Подальші дослідження повинні зосереджуватися на поєднанні різних біометричних ознак і контекстуальних факторів для зменшення ризиків, пов'язаних із використанням синтетичного голосу. Важливим напрямком є також створення спеціалізованих наборів даних, які дозволять проводити систематичні й комплексні дослідження стійкості голосових систем автентифікації до різноманітних видів атак.

4.4. Напрями вдосконалення системи голосової біометричної автентифікації в контексті протидії spoofing-атакам

Результати проведених експериментальних досліджень засвідчили високий рівень вразливості сучасних систем голосової біометричної автентифікації до атак типу *spoofing*, що здійснюються із використанням синтетичного мовлення, згенерованого за допомогою передових технологій клонування голосу. Виявлена тенденція підтверджує обмеженість традиційних методів автентифікації, що орієнтовані переважно на аналіз акустичних ознак мовця, та вказує на необхідність розроблення й впровадження нових підходів для підвищення стійкості таких систем.

У цьому підрозділі розглядаються ключові стратегічні напрями вдосконалення голосових біометричних технологій у контексті протидії spoofing-атакам. Зокрема, увага приділяється інтеграції приватних когнітивних факторів у процес автентифікації, що підсилює захищеність системи шляхом поєднання біометричних і семантично-когнітивних ознак користувача. Другий напрямок спрямований на дослідження можливостей удосконалення голосових біометричних технологій через впровадження спеціалізованих моделей для детекції синтетичного мовлення, які здатні ідентифікувати приховані ознаки штучного походження аудіосигналів [99]

Розроблення таких рішень є критично важливим кроком для забезпечення довготривалої безпеки біометричних систем автентифікації у світлі постійного вдосконалення методів генерування синтетичного мовлення та зростання ризиків несанкціонованого доступу. Представлені підходи спрямовані на формування багаторівневої моделі захисту, що поєднує фізіологічні, когнітивні та технологічні аспекти автентифікації, забезпечуючи комплексне підвищення стійкості систем до сучасних загроз.

4.4.1. Інтеграція приватних когнітивних факторів у голосову біометричну автентифікацію

Пропрієтарні знання, як чинник автентифікації, традиційно визначаються як такі, що доступні лише автентичному користувачеві та не підлягають об'єктивному спостереженню чи виведенню з відкритих джерел. На відміну від звичайного "секретного питання", пропрієтарне знання у системі голосової автентифікації має бути не лише семантично унікальним, але й реалізованим через мовлення, тобто через мовленнєву реакцію користувача на певний запит. Така реакція, окрім змістової (семантичної) складової, містить додаткову індивідуалізовану акустичну інформацію, яка пов’язана із голосовими особливостями суб’єкта: тембром, темпом, просодією, інтонацією, спектральною структурою, артикуляційною манерою тощо.

Однією з ефективних реалізацій такого підходу є використання персоналізованих динамічних голосових паролів. Замість використання фіксованої фрази, яку можна перехопити чи зімітувати, користувачу пропонується вимовити змінну фразу, яка є пов’язаною із його особистим досвідом або знанням. Наприклад, система може сформулювати питання типу: *"Назвіть улюблений дитячий фільм", "Опишіть свою першу подорож"*, або *"Як звали вашого першого вчителя з математики?"*. Відповідь на таке запитання матиме унікальну мовну структуру, що дозволить системі одночасно проаналізувати і зміст відповіді (семантичний аналіз), і мовленнєві характеристики (фонетико-акустичний аналіз),

і, у перспективі, навіть когнітивну динаміку користувача (емоційна забарвленість, реактивність, паузування тощо).

Іншим напрямом застосування пропрієтарних знань є моделювання індивідуального лінгвістичного профілю користувача на основі характерного слововживання, синтаксичних конструкцій, емоційної інтонації або навіть типових патернів стилістики мовлення. Якщо система вміє зіставляти сказане користувачем з його раніше створеним профілем, то вона здатна відрізнити навіть дуже схожі голоси або голоси, які належать до однієї соціолінгвістичної групи. Це сприяє підвищенню точності автентифікації за рахунок урахування множинних рівнів мовленнєвої інформації – від акустичного до семантичного.

Особливу увагу варто приділити захисту від атаки методом відтворення (replay attack). У таких атаках зловмисник намагається отримати доступ, відтворюючи раніше записане мовлення легітимного користувача. У випадку використання статичних фраз або шаблонних голосових команд, така атака має високі шанси на успіх. Впровадження пропрієтарних знань суттєво знижує ефективність подібних атак, адже кожна автентифікаційна сесія ґрунтується на новому, змінному, часто – непередбачуваному контексті. Крім того, пропрієтарне знання можна зробити часово-залежним, прив’язавши його до конкретного сеансу, транзакції або геолокації, що робить його ще більш унікальним і складним для фальсифікації.

З практичної точки зору, для успішної реалізації автентифікації з використанням пропрієтарних знань необхідно враховувати низку технічних та етичних факторів. По-перше, важливо забезпечити належний рівень захисту даних, які формують основу таких знань – зокрема, персональних історій, відповідей на запити, лінгвістичних шаблонів тощо. По-друге, система має бути здатною адаптуватися до варіативності мовлення користувача, яка може змінюватися в залежності від психоемоційного стану, фізичних обставин, оточення або навіть стану здоров’я. По-третє, доцільним є впровадження механізмів адаптивного навчання, які дозволяють системі з часом уточнювати модель користувача, враховуючи нові варіанти пропрієтарного мовлення.

Не менш важливою є й етична сторона питання. Оскільки пропрієтарне знання, втілене у мовленнєвій формі, часто відображає глибоко особисту інформацію, система автентифікації повинна гарантувати конфіденційність, прозорість обробки даних та відповідність нормам захисту приватності. Користувач має бути поінформований про те, які саме елементи його мовлення використовуються, як вони зберігаються і з якою метою застосовуються.

У підсумку, застосування пропрієтарних знань у системах голосової автентифікації відкриває нові горизонти у напрямі побудови високонадійних, адаптивних та когнітивно-стійких моделей перевірки ідентичності. Поєднання біометричних параметрів з когнітивно-семантичними ознаками формують багаторівневу систему захисту, яка не лише ускладнює можливість зовнішнього втручання, але й підвищує зручність та гнучкість у користуванні. У майбутньому, з розвитком технологій глибинного аналізу мовлення, системи, що базуються на пропрієтарних знаннях, можуть стати стандартом безпечної автентифікації у критично важливих галузях – від електронного банкінгу до управління інфраструктурними об’єктами.

4.4.2. Дослідження можливостей удосконалення голосових біометричних технологій через впровадження детекції синтетичного мовлення

У межах даної роботи було проведено експериментальне дослідження з метою оцінки потенціалу удосконалення систем голосової біометричної автентифікації шляхом інтеграції модулів детекції синтетичного мовлення. Зокрема, аналізувалась здатність таких модулів розпізнавати голосові зразки, згенеровані за допомогою сучасних технологій голосового клонування. Для цього було реалізовано дві окремі моделі класифікації, кожна з яких була протестована на однаковому наборі даних, що складався з 400 аудіофайлів, згенерованих кожною з чотирьох різних систем клонування, а також 400 оригінальних зразків.

У першому експерименті використовувалась модель, запропонована компанією ElevenLabs як інструмент виявлення синтетичного мовлення [100]. Згідно з результатами (див. таблицю 4.4.), дана модель продемонструвала майже

ідеальну здатність до виявлення аудіо, створених саме з використанням технології ElevenLabs, однак виявилась неспроможною ідентифікувати інші типи клонованих голосів: як файли, згенеровані альтернативними системами, так і справжні зразки, вона класифікувала як автентичні.

Таблиця 4.4.

Результати дослідження детектора 11labs

Дані	**Точність класифікації**
Оригінальні голоси	100%
11labs	98%
RVC	0%
XTTS	0%
Tortoise	0%

Це дозволяє припустити, що модель детекції ElevenLabs використовує приховані шаблони або артефакти, які характерні лише для її власного алгоритму синтезу мовлення. Відтак, її висока точність у межах вузького домену супроводжується повною відсутністю генералізації, поза межами цієї технології, що підтверджує важливість розроблення детекторів, орієнтованих на виявлення загальних ознак синтетичності, а не специфічних сигнатур окремих систем.

У другому експерименті було реалізовано власну модель детекції синтетичного мовлення, побудовану на основі мел-кепстральних коефіцієнтів (MFCC) як ознак, та класифікатора на основі опорних векторів (SVM). Для навчання моделі використовували збалансований набір: 400 оригінальних аудіофайлів та по 80 синтетичних зразків, згенерованих кожною з чотирьох розглянутих систем клонування. Така структура навчального набору забезпечувала рівномірне представлення кожного класу та дозволяла моделі формувати універсальні ознаки, релевантні для детекції різних типів синтетичного мовлення. Оцінка ефективності проводилась на незалежному тестовому наборі, що включав інші 400 оригінальних голосових зразків, а також по 400 згенерованих аудіофайлів

від кожного з вендорів. В таблиці 4.5. подані результати тестування власної моделі детекції на базі SVM.

Таблиця 4.5.

Результати дослідження детектора на основі SVM

Дані	**Точність класифікації**
Оригінальні голоси	25.96
11labs	87.75
RVC	65.50
XTTS	89.75
Tortoise	73.75

Як видно з отриманих значень, побудована модель виявила значно кращу здатність до класифікації синтетичних зразків у порівнянні з першою системою. Зокрема, точність виявлення аудіо, згенерованого моделями XTTS та ElevenLabs, становила 89.75% та 87.75% відповідно, тоді як для Tortoise та RVC вона склала 73.75% та 65.50%. Незважаючи на досить високу ефективність у виявленні синтетичних сигналів, модель також помилково класифікувала значну частину справжніх голосів як синтетичні – точність для оригінальних зразків склала лише 25.96%. Така невідповідність може бути пов’язана з перенавчанням на специфічні артефакти у навчальних даних або зі складністю коректного балансу між узагальненням і дискримінативністю ознак. У цілому, запропонована MFCC+SVM модель продемонструвала вищу узагальнювальну здатність, ніж детектор ElevenLabs, проте потребує подальшого вдосконалення для зниження частоти хибнопозитивних спрацювань на автентичних зразках.

Порівняльний аналіз двох експериментів засвідчив фундаментальні відмінності між підходами до детекції синтетичного мовлення. Модель ElevenLabs демонструє високу точність лише у межах власної технології генерації, що свідчить про орієнтацію на виявлення специфічних артефактів, притаманних її власним алгоритмам. Такий підхід є ефективним у вузькоспеціалізованих сценаріях, проте втрачає практичну цінність у гетерогенному середовищі, де можуть

використовуватись різні системи клонування. Натомість розроблена модель на основі MFCC і SVM виявила здатність до часткової генералізації, успішно розпізнаючи зразки, згенеровані декількома методами. Незважаючи на знижені показники точності для автентичного мовлення, ця модель відкриває перспективи для побудови універсальних детекторів синтетичних сигналів на основі акустичних ознак. Такий підхід є більш придатним для використання у реальних системах біометричної автентифікації, де ключову роль відіграє здатність виявляти фальсифікації незалежно від методу їх створення.

4.5. Висновки до розділу 4

Розділ 4 присвячено дослідженню стійкості системи голосової біометричної автентифікації до атак із застосуванням клонованого мовлення та формуванню підходів до підвищення її захищеності. Основні висновки можна сформулювати так:

1. Сформовано експериментальний корпус синтетичних голосових зразків на основі генеративних моделей RVC, ElevenLabs, XTTS та Tortoise, що охоплює як сценарії з ідентичним до реєстраційного, так і з відмінним лінгвістичним наповненням. Такий підхід дозволив забезпечити багатогранну перевірку вразливості системи до spoofing-атак різної складності.

2. Розроблено методику експериментального оцінювання, що моделює реальний процес автентифікації в умовах потенційного використання синтетичного мовлення. Застосування індивідуальних шаблонів ембедінгів на основі автентичних голосових зразків дало змогу здійснити об'єктивну верифікацію кожного синтетичного запиту та оцінити якість класифікації за узагальненою метрикою точності.

3. Проведено комплексне тестування системи з використанням п’яти різних моделей формування голосових ознак (ECAPA, TitaNet, PyAnnote, WavLM-base, WavLM-base-plus) та встановлено, що рівень помилкового прийняття синтетичних зразків сягає 95% у сценаріях з ідентичним лінгвістичним вмістом. Це

свідчить про суттєву уразливість системи до атак за умов повного відтворення мовних та акустичних характеристик.

4. Проведено порівняльний аналіз ефективності моделей ембедінгів у випадках, коли мовний контент синтетичних зразків відрізняється від реєстраційного. Показано, що зміна текстового наповнення лише незначно знижує ефективність атак, що вказує на домінуючу роль акустичних ознак у прийнятті рішень про автентичність.

5. Встановлено, що жодна з протестованих моделей ембедінгів не забезпечує задовільної стійкості до атак із застосуванням клонованого мовлення. Найменш уразливою виявилася модель WavLM-base-plus, тоді як моделі TitaNet та ECAPA-TDNN продемонстрували найвищу частку помилкових прийняттів, що вказує на їхню непридатність для використання в умовах високого рівня загроз.

6. Реалізовано два підходи до детекції синтетичного мовлення: з використанням фірмової моделі ElevenLabs та власної моделі на основі MFCC-ознак і SVM-класифікатора. Показано, що модель ElevenLabs демонструє надвисоку точність лише у вузькоспеціалізованому домені, тоді як власна модель має вищу узагальнювальну здатність, але потребує вдосконалення у частині зниження рівня хибнопозитивних рішень.

7. Доведено необхідність впровадження додаткових механізмів захисту у системи голосової автентифікації, зокрема – детекторів синтетичного мовлення загального призначення та багаторівневих схем автентифікації із залученням когнітивних ознак. Такий підхід пізвищує стійкість системи до сучасних загроз, пов’язаних з активним розвитком технологій голосового клонування.

ВИСНОВКИ

У дисертаційній роботі вирішено актуальне науково-практичне завдання вдосконалення методів і засобів біометричної автентифікації за голосом шляхом впровадження технологій машинного навчання, спрямованих на підвищення точності верифікації особи та забезпечення стійкості до атак із застосуванням синтетично згенерованого (клонованого) мовлення. У межах дослідження розроблено та експериментально обґрунтовано підходи, що забезпечують підвищення ефективності функціонування сучасних систем голосової автентифікації в умовах кіберзагроз. У результаті виконання цієї роботи отримано такі наукові результати:

1. Здійснено комплексний аналіз сучасних методів голосової автентифікації, здійснено їх класифікацію за принципами роботи (класичні, статистичні, нейромережеві) та проведено порівняльний аналіз текстозалежних і текстонезалежних підходів; виявлено їх переваги, недоліки та сфери ефективного застосування, що дозволило обґрунтувати доцільність використання трансформерних архітектур як основи для побудови високоточних і стійких систем автентифікації.

2. Досліджено актуальні загрози для голосової біометрії, пов’язані з розвитком технологій клонування мовлення; здійснено огляд сучасних генеративних моделей (RVC, Tacotron, VITS, XTTS, Tortoise) та проаналізовано вектори атак, наслідком чого стало формування вимог до підсилення архітектур безпеки та впровадження стійких до синтетичних впливів механізмів захисту.

3. Розроблено концептуальну модель системи голосової автентифікації, що інтегрує структурно-функціональні блоки на основі нейромережевих технологій (CNN, LSTM, Transformers); визначено ключові етапи обробки мовного сигналу та забезпечено відповідність моделі міжнародним стандартам біометричної безпеки, що є основою для вдосконалення методів і засобів голосової біометрії;

4. Виконано порівняльний аналіз класичних та нейромережевих методів виділення й обробки голосових ознак (MFCC, LPC, Titanet, ECAPA-TDNN,

WavLM); показано переваги трансформерних архітектур з метою забезпечення високої точності та стійкості автентифікації;

5. Реалізовано архітектурну модель системи голосової автентифікації на основі ембедінгів із впровадженням модулів реєстрації, налаштування та порівняння голосових зразків; обґрунтовано вибір метрики подібності у векторному просторі ознак, що створює передумови для підвищення точності і стабільності автентифікаційних рішень у багатокористувацьких динамічних середовищах.

6. Проведено експериментальне оцінювання точності, продуктивності та масштабованості системи голосової автентифікації на основі нейромережевих моделей. Досягнуто точності автентифікації до 96% при значеннях EER не нижче 4,17%. При масштабуванні до 70 зареєстрованих користувачів найстабільніші моделі демонстрували зниження точності не більше ніж на 3,5%. Час прийняття рішення в оптимальних конфігураціях не перевищував 4,5 мс. Отримані результати дозволили сформулювати рекомендації щодо адаптивного налаштування порогових значень і вибору архітектур, що забезпечують баланс між точністю, швидкодією та стійкістю до масштабування.

7. Запропоновано експериментальну методику оцінювання стійкості системи до атак із використанням сучасних технологій клонування голосу; сформовано тестовий аудіокорпус та проведено порівняльний аналіз моделей ембедінгів у сценаріях з ідентичним і новим лінгвістичним контентом, що дозволило встановити рівень вразливості системи до різних типів spoofing-атак.

8. Імплементовано підходи до детекції синтетичного мовлення на основі фірмового рішення ElevenLabs та авторської моделі MFCC+SVM. У результаті порівняльного тестування встановлено, що модель MFCC+SVM досягає точності виявлення синтетичного мовлення на рівні 89,75%, демонструючи кращу здатність до генералізації на сторонні голосові синтезатори. Натомість фірмове рішення ElevenLabs досягає до 98% точності на зразках, згенерованих власною технологією, однак не здатне виявляти зразки, згенеровані іншими методами клонування. На

підставі аналізу сформульовано рекомендації щодо інтеграції універсальних антиспуфінгових механізмів у архітектуру системи голосової автентифікації.

Загалом, результати дослідження підтверджують доцільність використання нейромережевих технологій для побудови високоточних і стійких до атак систем голосової автентифікації, що мають перспективи практичного застосування у сферах підвищеної безпеки.

СПИСОК ЛІТЕРАТУРИ

1. LeewayHertz. Data Security in AI Systems: An Overview [Електронний ресурс]. – Режим доступу: https://www.leewayhertz.com/data-security-in-ai-systems/
2. ENISA. ENISA Threat Landscape 2021: From April 2020 to mid-July 2021 [Електронний ресурс] / European Union Agency for Cybersecurity. – Luxembourg: Publications Office of the European Union, 2021. – 124 с. – Режим доступу: https://www.enisa.europa.eu/publications/enisa-threat-landscape-2021
3. Phonexia. Voice Biometrics: The Essential Guide [Електронний ресурс]. – Режим доступу: https://www.phonexia.com/knowledge-base/voice-biometrics-essential-guide/
4. MarketsandMarkets. Глобальний ринок голосової біометрії: прогноз до 2026 року [Електронний ресурс] / MarketsandMarkets. – 2021. – Режим доступу: https://www.marketsandmarkets.com/Market-Reports/voice-biometrics-market-104503105.html
5. Arik S., Chen J., Peng K., Ping W., Zhou Y. Neural Voice Cloning with a Few Samples [Електронний ресурс] // Advances in Neural Information Processing Systems 31 (NeurIPS 2018): Proceedings of the Annual Conference on Neural Information Processing Systems. – 2018.
6. Shen K., Ju Z., Tan X., та ін. NaturalSpeech 2: Latent Diffusion Models are Natural and Zero-Shot Speech and Singing Synthesizers [Електронний ресурс] // arXiv.org. – 2023. – Режим доступу: https://arxiv.org/abs/2304.09116
7. Qin Z., Zhao W., Yu X., Sun X. OpenVoice: Versatile Instant Voice Cloning [Електронний ресурс] / Zengyi Qin, Wenliang Zhao, Xumin Yu, Xin Sun // arXiv.org. – 2023. – Режим доступу: https://arxiv.org/abs/2312.01479
8. FIDO Alliance. Biometric Component Certification Program [Електронний ресурс]. – FIDO Alliance, 2024. – Режим доступу: https://fidoalliance.org/certification/biometric-component-introduction/ Ali M. et al. Synthetic speech detection for AI-generated audio. IEEE Access, 2023

9. FIDO Alliance. Biometric Component Certification Policy v1.4 [Електронний ресурс]. – FIDO Alliance, 2021. – Режим доступу: https://fidoalliance.org/specs/biometric/certificationpolicy/Biometric-Component-Certification-Policy-v1.4-fd-20211206.html
10. Evans N., Yamagishi J., Kinnunen T., Lee K. A., Delgado H., Nautsch A., Patino J., Sahidullah M., Todisco M., Wang X., Liu X. ASVspoof Challenges: Past, Present and Future // Interspeech 2021. – 2021. – С. 47–54. – DOI: 10.21437/ASVSPOOF.2021-8.
11. Європейська Комісія. PRIMA – Партнерство з досліджень та інновацій у Середземноморському регіоні [Електронний ресурс] // Дослідження та інновації. – 2023. – Режим доступу: https://research-and-innovation.ec.europa.eu/research-area/environment/prima_en
12. ISO/IEC 30107-3:2023. Інформаційні технології. Біометрична детекція атак представлення. Частина 3: Тестування та звітність [Електронний ресурс]. – Женева: ISO, 2023. – Режим доступу: https://www.iso.org/standard/79520.html
13. ISO/IEC 30136:2018. Performance Testing of Biometric Template Protection Schemes
14. Stan A. Residual Information in Deep Speaker Embedding Architectures [Електронний ресурс] // arXiv. – 2023. – Режим доступу: https://arxiv.org/abs/2302.02742
15. Potter R. K., Kopp G. A., Green H. C. Visible Speech. – New York: D. Van Nostrand Company, 1947. – xvi, 441 p.
16. Rabiner L. R., Schafer R. W. Digital Processing of Speech Signals. – Englewood Cliffs, N.J.: Prentice-Hall, 1978. – xvi, 512 p.
17. Furui S. Cepstral Analysis Technique for Automatic Speaker Verification // IEEE Transactions on Acoustics, Speech, and Signal Processing. – 1981. – Vol. 29, No. 2. – P. 254–272.
18. Furui S. Fifty Years of Progress in Speech and Speaker Recognition // The Journal of the Acoustical Society of America. – 2004. – Vol. 116, No. 4. – P. 2497–2498. – DOI: 10.1121/1.4784967.

19. Jain A. K., Ross A., Prabhakar S. An Introduction to Biometric Recognition // IEEE Transactions on Circuits and Systems for Video Technology. – 2004. – Vol. 14, No. 1. – P. 4–20. – DOI: 10.1109/TCSVT.2003.818349.
20. Hinton G. E., Deng L., Yu D., Dahl G. E., Mohamed A.-r., Jaitly N., Senior A., Vanhoucke V., Nguyen P., Sainath T. N., Kingsbury B. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups // IEEE Signal Processing Magazine. – 2012. – Vol. 29, No. 6. – P. 82–97. – DOI: 10.1109/MSP.2012.2205597
21. Statista. Speech Recognition – Worldwide [Електронний ресурс]. – Режим доступу: https://www.statista.com/outlook/tmo/artificial-intelligence/computer-vision/speech-recognition/worldwide
22. BBC News. MWC 2017: Google brings Assistant to more Android phones [Електронний ресурс]. – 2017. – Режим доступу: https://www.bbc.com/news/technology-39096355
23. US20120245941A1. Device Access Using Voice Authentication [Електронний ресурс] / Inventors: Adam J. Cheyer // Google Patents. – 2012. – Режим доступу: https://patents.google.com/patent/US20120245941A1/en
24. US9262612B2. Voice authentication for a computing device [Електронний ресурс] / Inventors: Adam J. Cheyer // Google Patents. – 2016. – Режим доступу: https://patents.google.com/patent/US9262612B2/en
25. Prasad R. “Ambient intelligence” will accelerate advances in general AI [Електронний ресурс] // Amazon Science. – 2021. – Режим доступу: https://www.amazon.science/blog/ambient-intelligence-will-accelerate-advancements-in-general-ai
26. Meta Reality Labs. Inside Facebook Reality Labs: The next era of human-computer interaction [Електронний ресурс] // Tech at Meta. – 2021. – Режим доступу: https://tech.facebook.com/reality-labs/2021/3/inside-facebook-reality-labs-the-next-era-of-human-computer-interaction/
27. Руда Х. С., Сабодашко Д. В., Микитин Г. В., Швед М. Є., Бордуляк С. М., Коршун Н. Порівняння методів цифрової обробки сигналів та моделей

глибинного навчання у голосовій аутентифікації // Кібербезпека: освіта, наука, техніка. – 2024. – № 5(25). – С. 140–160

28. Tu Y., Lin W., Mak M.-W. A Survey on Text-Dependent and Text-Independent Speaker Verification // IEEE Access. – 2022. – Vol. 10. – P. 99038–99049. – DOI: 10.1109/ACCESS.2022.3206541.
29. Xia W., Zhang C., Weng C., Yu M., Yu D. Self-supervised text-independent speaker verification using prototypical momentum contrastive learning // Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2021. – P. 6723–6727. – IEEE. – DOI: 10.1109/ICASSP39728.2021.9414722.
30. Tan X., Qin T., Soong F., Liu T.-Y. A Survey on Neural Speech Synthesis [Електронний ресурс] // arXiv preprint arXiv:2106.15561. – 2021. – 63 с. – Режим доступу: https://arxiv.org/abs/2106.15561
31. Muda L., Begam M., Elamvazuthi I. Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques [Електронний ресурс] // arXiv preprint. – 2010. – arXiv:1003.4083. – Режим доступу: https://arxiv.org/abs/1003.4083
32. Rabiner L. R. Linear Predictive Coding (LPC) [Електронний ресурс] // Лекція 13 курсу "Digital Speech Processing", University of California, Santa Barbara. – 2012. – Режим доступу: https://web.ece.ucsb.edu/Faculty/Rabiner/ece259/digital%20speech%20processing%20course/lectures_new/Lecture%2013_winter_2012_6tp.pdf
33. Ravindran S., Anderson D. V., Slaney M. Improving the Noise-Robustness of Mel-Frequency Cepstral Coefficients for Speech Discrimination [Електронний ресурс] // SAPA 2006 Workshop. – 2006. – Режим доступу: https://labrosa.ee.columbia.edu/SAPA2006/papers/131.pdf
34. Deka M.K., Nath C.K., Sarma S.K., Talukdar P.H. An Approach to Noise Robust Speech Recognition using LPC-Cepstral Coefficient and MLP based Artificial Neural Network with respect to Assamese and Bodo Language [Електронний ресурс] // International Symposium on Devices MEMS, Intelligent Systems &

Communication (ISDMISC) 2011. – Режим доступу: https://ijcaonline.org/isdmisc/number4/isdm088.pdf

35. Reynolds D. A. Automatic speaker recognition using Gaussian mixture speaker models // Lincoln Laboratory Journal. – 1995. – Vol. 8, No. 2. – P. 173–192. – [Електронний ресурс]. – Режим доступу: https://www.ll.mit.edu/sites/default/files/publication/doc/automatic-speaker-recognition-using-gaussian-reynolds-ja-7369.pdf
36. Rabiner L. R. A tutorial on Hidden Markov Models and selected applications in speech recognition // Proceedings of the IEEE. – 1989. – Vol. 77, No. 2. – P. 257–286. – [Електронний ресурс]. – Режим доступу: https://www.cs.ubc.ca/~murphyk/Bayes/rabiner.pdf
37. Paul D., Sahidullah M., Saha G. Generalization of Spoofing Countermeasures: a Case Study with ASVspoof 2015 and BTAS 2016 Corpora [Електронний ресурс] // arXiv preprint. – 2019. – arXiv:1901.08025. – Режим доступу: https://arxiv.org/abs/1901.08025
38. Dehghani A., Seyyedsalehi S. A. Time-Frequency Localization Using Deep Convolutional Maxout Neural Network in Persian Speech Recognition [Електронний ресурс] // arXiv preprint. – 2021. – arXiv:2108.03818. – Режим доступу: https://arxiv.org/abs/2108.03818
39. Sak H., Senior A., Beaufays F. Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition [Електронний ресурс] // arXiv preprint. – 2014. – arXiv:1402.1128. – Режим доступу: https://arxiv.org/abs/1402.1128
40. Radford A., Kim J. W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust Speech Recognition via Large-Scale Weak Supervision [Електронний ресурс] // arXiv preprint. – 2022. – arXiv:2212.04356. – Режим доступу: https://arxiv.org/abs/2212.04356
41. Frey N. C., Li B., McDonald J., Zhao D., Jones M., Bestor D., Tiwari D., Gadepally V., Samsi S. Benchmarking Resource Usage for Efficient Distributed Deep

Learning [Електронний ресурс] // arXiv preprint. – 2022. – arXiv:2201.12423. – Режим доступу: https://arxiv.org/abs/2201.12423

42. Brydinskyi V., Khoma Y., Sabodashko D., Podpora M., Khoma V., Konovalov A., Kostiak M. Comparison of Modern Deep Learning Models for Speaker Verification // Applied Sciences. – 2024. – Vol. 14, No. 4. – Article 1329. – https://doi.org/10.3390/app14041329
43. Wang Q., Lee K. A., Liu T. Scoring of Large-Margin Embeddings for Speaker Verification: Cosine or PLDA? [Електронний ресурс] // arXiv preprint. – 2022. – arXiv:2204.03965. – Режим доступу: https://arxiv.org/abs/2204.03965
44. Hong Q., Zhang J., Li L., Wan L., Tong F. A Transfer Learning Method for PLDA-Based Speaker Verification [Електронний ресурс] // arXiv preprint. – 2016. – arXiv:1604.02887. – Режим доступу: https://arxiv.org/abs/1604.02887
45. Микитин Г. В., Руда Х. С., Хома Ю. В. Categorization of deepfake methods: a comparison of single-shot and multi-shot transfer approaches // Захист інформації і безпека інформаційних систем : матеріали IX Міжнародної науково-технічної конференції (Львів, 25–26 травня 2023 р.). – 2023. – С. 109–110.
46. Mohammadi S. H., Kain A. An overview of voice conversion systems // Speech Communication. – 2017. – Vol. 88. – P. 65–82. – [Електронний ресурс]. – Режим доступу: https://doi.org/10.1016/j.specom.2017.01.008
47. Qian K., Zhang Y., Chang S., Yang X., Hasegawa-Johnson M. AutoVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss // Proceedings of the 36th International Conference on Machine Learning (ICML 2019). – In: Proceedings of Machine Learning Research. – Vol. 97. – P. 5210–5219. – [Електронний ресурс]. – Режим доступу: https://proceedings.mlr.press/v97/qian19c.html [Source Research Gate - https://www.researchgate.net/publication/312483751_An_Overview_of_Voice_Conversion_Systems]
48. Wang Y., Skerry-Ryan R. J., та ін. Tacotron: Towards End-to-End Speech Synthesis // Proceedings of Interspeech 2017. – 2017. – P. 4006–4010. –

[Електронний ресурс]. – Режим доступу: https://doi.org/10.21437/Interspeech.2017-1452

49. Jia Y., Zhang Y., Weiss R. J., та ін. Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018. – P. 5180–5184.
50. Qian K., Zhang Y., Chang S., Yang X., Hasegawa-Johnson M. AutoVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss // Proceedings of the 36th International Conference on Machine Learning (ICML 2019). – In: Proceedings of Machine Learning Research. – Vol. 97. – P. 5210–5219
51. Kong J., Kim J., Bae J. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis // Advances in Neural Information Processing Systems 33 (NeurIPS 2020). – 2020. – P. 17022–17033
52. Arik S. O., Chen J., Peng K., Ping W., Zhou Y. Neural Voice Cloning with a Few Samples // Advances in Neural Information Processing Systems 31 (NeurIPS 2018). – 2018. – P. 10019–10029.
53. Kim J., Kong J., Son J. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech // Proceedings of the 38th International Conference on Machine Learning (ICML 2021). – PMLR, 2021. – Vol. 139. – P. 5530–5540
54. Coqui.ai. Coqui TTS: A Deep Learning Toolkit for Text-to-Speech [Електронний ресурс]. – Режим доступу: https://github.com/coqui-ai/TTS
55. Kim S., Shih K., Badlani R., Santos J. F., Bakhturina E., Desta M., Valle R., Yoon S., Catanzaro B. P-Flow: A Fast and Data-Efficient Zero-Shot TTS through Speech Prompting // Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Main Conference Track. – 2023.
56. V. Dudykevych, H. Mykytyn and K. Ruda, "The concept of a deepfake detection system of biometric image modifications based on neural networks," 2022 IEEE 3rd KhPI Week on Advanced Technology (KhPIWeek), Kharkiv, Ukraine, 2022, pp. 1-4, doi: 10.1109/KhPIWeek57572.2022.9916378.

57. Микитин Г. В., Руда Х. С. Концептуальний підхід до виявлення deepfake-модифікацій біометричного зображення засобами нейронних мереж // Комп'ютерні системи та мережі. – 2024. – Вип. 6, № 1. – С. 124–132.

58. Микитин Г. В., Руда Х. С., Яремчук Ю. Є. Методологія безпеки нейромережевих інформаційних технологій виявлення deepfake модифікацій біометричного зображення // Вісник Вінницького політехнічного інституту. – 2024. – № 1(172). – С. 74–80.

59. BAS Speech Processing Cookbook. Sampling Rate [Електронний ресурс] // Ludwig Maximilian University of Munich. – 2004. – Режим доступу: https://www.bas.uni-muenchen.de/forschung/BITS/TP1/Cookbook/node61.html

60. Тан З. Г., Саркар А. К., Дехак Н. rVAD : An Unsupervised Segment-Based Robust Voice Activity Detection Method : стаття / Computer Speech & Language. – 2020. – Vol. 59. – С. 1-21. – DOI: 10.1016/j.csl.2019.06.005.

61. Reynolds D. A., Rose R. C. Voice conversion using Gaussian Mixture Models : [стаття] / D. A. Reynolds, R. C. Rose // *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. – 1995. – Vol. 1. – P. 701–704. – DOI: 10.1109/ICASSP.1995.479300.

62. Schatten M., Baca M., Cubrilo M. Towards a general definition of biometric systems : [препринт] / M. Schatten, M. Baca, M. Cubrilo // *arXiv*. – 2009. – arXiv:0909.2365. – Режим доступу: https://arxiv.org/abs/0909.2365

63. George K. K., Kumar C. S., Ramachandran K. I., Panda A. Cosine Distance Features for Robust Speaker Verification : [тези конференції] / K. K. George, C. S. Kumar, K. I. Ramachandran, A. Panda // *INTERSPEECH 2015*. – 2015. – С. 234–237. – DOI: 10.21437/Interspeech.2015-91

64. Khosravy M., Gupta N., Patel N., Duque C. A. Recovery in compressive sensing: A review / M. Khosravy, N. Gupta, N. Patel, C. A. Duque // *Compressive Sensing in Healthcare* / [за ред. H. E. Egilmez, B. Sankur]. – Cambridge (MA, USA) : Academic Press, 2020. – С. 25–42.

65. Гасті Т., Тібшірані Р., Фрідман Дж. Основи статистичного навчання : [текст] / Т. Гасті, Р. Тібшірані, Дж. Фрідман. – 2-ге вид. – Нью-Йорк : Springer, 2009. – 745 с. – ISBN 978-0-387-84857-0.

66. Ramoji S., Krishnan P., Ganapathy S. NPLDA: A deep neural PLDA model for speaker verification : [препринт] / S. Ramoji, P. Krishnan, S. Ganapathy // *arXiv*. – 2020. – arXiv:2002.03562. – Режим доступу: https://arxiv.org/abs/2002.03562

67. Sakoe H., Chiba S**.** Dynamic programming algorithm optimization for spoken word recognition : [стаття] / H. Sakoe, S. Chiba // *IEEE Transactions on Acoustics, Speech, and Signal Processing*. – 1978. – Vol. 26, No. 1. – P. 43–49. – DOI: 10.1109/TASSP.1978.1163055.

68. Ali Sh., Tanweer S., Khalid S. S., Rao N. Mel Frequency Cepstral Coefficient: A Review / Sh. Ali, S. Tanweer, S. S. Khalid, N. Rao // Proceedings of the 2nd International Conference on ICT for Digital, Smart, and Sustainable Development. – New Delhi (India), 27–28 Feb. 2020. – EAI, 2021. – DOI: 10.4108/eai.27-2-2020.2303173.

69. Magre S. B., Deshmukh R. R., Shrishrimal P. P. A Comparative Study on Feature Extraction Techniques in Speech Recognition / S. B. Magre, R. R. Deshmukh, P. P. Shrishrimal // International Conference on Recent Advances in Statistics and Their Applications. – Department of Statistics, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (Maharashtra, India), Dec. 2013.

70. Fadhilah D. N., Magdalena R., Sa'idah S. Individual Identification System Design Through Voice Using Linear Predictive Coding Method and K-Nearest Neighbor / D. N. Fadhilah, R. Magdalena, S. Sa'idah // *Jurnal Teknik Informatika (JUTIF)*. – Telkom University (Indonesia), 28 Mar. 2021. – Vol. 2, No. 2. – P. 95–100. – DOI: 10.20884/1.jutif.2021.2.2.71.

71. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I. Attention is all you need / A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin // Advances in Neural Information Processing Systems. – 2017. – Vol. 30.

72. Hild T., Moore P., Powell M. Multimodal Authentication / T. Hild, P. Moore, M. Powell // Proceedings of the Annual General Donald R. Keith Memorial Conference. – West Point, New York, USA, May 2, 2024. – Institute for Industrial and Systems Engineering, 2024. – P. 150–157. – ISBN 978-1-938496-24-0.

73. Dahea W. A., Fadewar H. S. Multimodal biometric system: A review / W. A. Dahea, H. S. Fadewar // International Journal of Research in Advanced Engineering and Technology. – Nanded (Maharashtra, India) i Dhamar (Yemen), Jan. 2018. – Vol. 4, No. 1. – P. 25–31. – ISSN 2455-0876.

74. Dahea W. A., Fadewar H. S. An Overview of Deep Learning Techniques for Biometric Systems / W. A. Dahea, H. S. Fadewar // *International Journal of Research in Advanced Engineering and Technology*. – Jan. 2021. – Vol. 5, No. 1. – P. 15–22

75. Tripathi D. R., Nishad D. K. Biometric Authentication Systems: A Survey / D. R. Tripathi, D. K. Nishad // *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*. – Vol. 11, No. 3 (2020). – P. 2878–2884. – DOI: 10.61841/turcomat.v11i3.14653.

76. Li C., Ma X., Jiang B., Li X., Zhang X., Liu X., Lane I., Zhu Z. Deep speaker: an end-to-end neural speaker embedding system / C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, I. Lane, Z. Zhu // arXiv preprint arXiv:1705.02304. – 2017. – [Електронний ресурс]. – URL: https://arxiv.org/abs/1705.02304

77. Lambamo W., Srinivasagan R., Jifara W. Speaker Identification Under Noisy Conditions Using Hybrid Deep Learning Model / W. Lambamo, R. Srinivasagan, W. Jifara // Pan-African Conference on Artificial Intelligence (PanAfriConAI 2023). – Communications in Computer and Information Science, Vol. 2068. – Springer Nature Switzerland, Apr. 7, 2024. – P. 154–175. – DOI: 10.1007/978-3-031-57624-9_9.

78. Koluguri N. R., Park T., Ginsburg B. Titanet: Neural Model for Speaker Representation with 1D Depth-Wise Separable Convolutions and Global Context / N. R. Koluguri, T. Park, B. Ginsburg // *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*. –

Singapore, 22–27 May 2022. – IEEE, 2022. – P. 8102–8106. – DOI: 10.1109/ICASSP43922.2022.9747012.

79. Cornell University. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN-Based Speaker Verification [Електронний ресурс]. – Режим доступу: https://arxiv.org/abs/2010.12653.
80. Chen, S., Cheng, M., He, J., Wu, Y., and Yu, D. (2022). WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing [Електронний ресурс]. – Режим доступу: https://arxiv.org/abs/2110.13900.
81. ResearchGate. Introducing ECAPA-TDNN and Wav2Vec 2.0 Embeddings to Stuttering Detection [Електронний ресурс]. – Режим доступу: https://www.researchgate.net/publication/359728980.
82. Hervé Bredin, et al. "pyannote.audio 2.1 Speaker Diarization Pipeline: Principle, Benchmark, and Recipe" [Електронний ресурс]. – Режим доступу: https://arxiv.org/abs/2210.15765
83. Hervé Bredin. "How I Won the 2022 Diarization Challenges" [Електронний ресурс]. – Режим доступу: https://herve.niderb.fr/posts/2022-12-02-how-I-won-2022-diarization-challenges.html.
84. Mykytyn H., Ruda K., Shved M. The paradigm of building secure information technologies for detecting deepfake modifications of biometric images based on neural networks // *Informatyka techniczna i sztuczna inteligencja: колективна монографія*. – Chapter in a collective monograph. – Bielsko-Biała: Wydawnictwo naukowe Uniwersytetu Bielsko-Bialskiego, 2024. – P. 43–50
85. ISO/IEC 19794-13:2021. Information technology – Biometric data interchange formats – Part 13: Voice data. – Geneva: International Organization for Standardization, 2021. – 26 p.
86. ISO/IEC 24745:2011. Information technology – Security techniques – Biometric information protection. – Geneva: International Organization for Standardization, 2011. – 36 p.
87. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of

personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). – Official Journal of the European Union, L 119, 4.5.2016, p. 1–88..

88. Закон України "Про захист персональних даних" № 2297-VI від 01 червня 2010 р. // Відомості Верховної Ради України. – 2010. – № 34. – Ст. 481.
89. ДСТУ ISO/IEC 27001:2015. Інформаційні технології. Методи захисту. Системи управління інформаційною безпекою (ISO/IEC 27001:2013, IDT). – Київ: Мінекономрозвитку України, 2016. – 34 с.
90. Zaiets I., Brydinskyi V., Sabodashko D., Khoma Y., Ruda K. Integrated system for speaker diarization and intruder detection using speaker embeddings // CEUR Workshop Proceedings. – 2024. – Vol. 3654 : Cybersecurity providing in information and telecommunication systems 2024. Proceedings of the workshop cybersecurity providing in information and telecommunication systems (CPITS 2024), Kyiv, Ukraine, February 28, 2024 (online). – P. 228–238.
91. Сабодашко Д. В., Руда Х. С., Швед М. Є., Бордуляк С. М. Технологія ембедінгів голосу та діаризація мовців // XX International Scientific and Practical Conference «Scientific Research: Modern Challenges and Prospects» : collection of abstracts, April 24–26, 2024, Prague, Czech Republic. – 2024. – P. 72–74.
92. Дудикевич В. Б., Микитин Г. В., **Руда Х. С.** Застосування глибокого навчання для виявлення deepfake модифікацій біометричного зображення // Сучасна спеціальна техніка. – 2022. – № 1(68). – С. 13–22.
93. Dudykevych V., Yevseiev S., Mykytyn G., Ruda K., Hulak H. Detecting deepfake modifications of biometric images using neural networks // CEUR Workshop Proceedings. – 2024. – Vol. 3654 : Cybersecurity providing in information and telecommunication systems 2024. Proceedings of the workshop cybersecurity providing in information and telecommunication systems (CPITS 2024), Kyiv, Ukraine, February 28, 2024 (online). – P. 391–397.
94. Заєць І. С., Бридінський В. А., Сабодашко Д. В., Руда Х. С., Хома Ю. В., Швед М. Є. Використання ембедінгів голосу в інтегрованих системах для діаризації

мовців та виявлення зловмисників // Комп’ютерні системи та мережі. – 2024. – Вип. 6, № 1. – С. 54–66.

95. ISO/IEC 25010:2023. Systems and software engineering – Systems and software quality requirements and evaluation (SQuaRE) – System and software quality models. – Geneva: International Organization for Standardization, 2023. – 34 p.
96. Renals S. Lecture 17: Speaker verification [Електронний ресурс] // Automatic Speech Recognition (ASR) course. – School of Informatics, University of Edinburgh, 2019. – Режим доступу: https://www.inf.ed.ac.uk/teaching/courses/asr/2018-19/asr17-speaker.pdf, вільний.
97. Руда Х. С. Дослідження масштабованості біометричних систем автентифікації на основі вбудовування голосу // Social Development and Security. – 2025. – Vol. 15, iss. 1. – P. 161–170.
98. Руда Х. С. Використання візуальних і неявних артефактів для виявлення deepfake-модифікацій у біометричних зображеннях // Сучасні методи, інформаційне, програмне та технічне забезпечення систем керування організаційно-технічними та технологічними комплексами : матеріали VIII Міжнародної науково-технічної Internet-конференції (Київ, 26 листопада 2021 року). – 2021. – С. 207–208.
99. Руда Х. С., Микитин Г. В. Методи детекції синтетичних аудіозаписів, створених системами клонування голосу // Scientific Research in the Age of Virtual Reality: Exploring New Frontiers : collection of abstracts of LII International Scientific and Practical Conference (December 18–20, 2024), Montreal, Canada. – 2024. – P. 132–136.
100. ElevenLabs. AI Speech Classifier [Електронний ресурс]. – Режим доступу: https://elevenlabs.io/ai-speech-classifier

ДОДАТОК А. АКТИ ВПРОВАДЖЕННЯ

ЗАТВЕРДЖУЮ

Проректор з науково-педагогічної роботи

Національного університету

«Львівська політехніка»

доц. _____ Олег ДАВИДЧАК

"22" 05 2025 р.

АКТ

про впровадження результатів дисертаційної роботи в навчальний процес

Рудої Христини Степанівни

«Удосконалення методів та засобів біометричної автентифікації за голосом з застосуванням технологій машинного навчання» представленої на здобуття наукового ступеня доктора філософії за спеціальністю 125 – *Кібербезпека*

Комісія НУ «Львівська політехніка» у складі:

Голова комісії – голова науково-методичної ради інституту комп'ютерних технологій та метрології, д.т.н., проф. Яцук В.О.

Члени комісії:

Завідувач кафедри "Захист інформації", д.т.н., проф. Опірський.І.Р, професор кафедри "Захист інформації", д.т.н. проф. Микитин Г.В. і доцент кафедри "Захист інформації", к.т.н., Горпенюк А.Я. даним актом підтверджує, що проведені дисертантом наукові дослідження виконувалися на кафедрі «Захист інформації» Національного університету «Львівська політехніка». Основні положення та результати дисертаційної роботи впроваджені у навчальний процес кафедри «Захист інформації» Національного університету «Львівська політехніка» при вивченні дисципліни: «Гарантоздатність автоматизовпних систем» для студентів напрямку підготовки 125 "Кібербезпека та захист інформації", спеціалізації «Системи технічного захисту інформації, автоматизація її обробки», тема №4 «Методи захисту інформації в АСОІ.».

Голова комісії,
голова науково-методичної ради ІКТА
д.т.н., проф. Василь ЯЦУК

Члени комісії:
зав. каф. ЗІ, д.т.н., проф. Іван ОПІРСЬКИЙ
проф. каф. ЗІ, д.т.н. проф. Галина МИКИТИН
доц каф. ЗІ, к.т.н Андрій ГОРПЕНЮК

ЗАТВЕРДЖУЮ
Проректор з наукової роботи
Національного університету
«Львівська політехніка»
проф. _____ Іван ДЕМИДОВ
"13" 05 2025р.

АКТ

про використання результатів дисертаційної роботи

Рудої Христини Степанівни

«Удосконалення методів та засобів біометричної автентифікації за голосом з застосуванням технологій машинного навчання» представленої на здобуття наукового ступеня доктора філософії за спеціальністю 125 – *Кібербезпека*

Комісія у складі – голови начальника науково-дослідної частини, д.т.н., ст. досл. Небесного Р.В. та членів: завідувача кафедри захисту інформації, д.т.н, професора Опірського І.Р., завідувача відділу науково-організаційного супроводу наукових досліджень, к.т.н. Лазько Г.В. і в.о. заступника начальника планово-фінансового відділу Фаст І.І., цим актом підтверджують, що окремі результати дисертаційної роботи Рудої Х.С., виконано в межах держбюджетної науково-дослідної роботи: «Дослідження стійкості біометричних систем автентифікації до атак із застосуванням технології клонування голосу на основі глибинних нейронних мереж» (№ держреєстрації 0124U000407).

Рудої Х.С. проведено експериментальне оцінювання точності, продуктивності та масштабованості системи голосової автентифікації з використанням нейромережевих моделей; проаналізовано вплив кількості зареєстрованих користувачів на ефективність функціонування системи, на основі чого сформульовано практичні рекомендації щодо адаптивного налаштування порогових значень та підвищення стійкості до акустичних і міжсесійних варіацій.

Голова комісії,
начальник науково-дослідної
частини, д.т.н. ст. досл. _____ Роман НЕБЕСНИЙ

Члени комісії:
зав. каф. захисту інформації, д.т.н. проф. _____ Іван ОПІРСЬКИЙ

зав. відділу науково-організаційного
супроводу наукових досліджень, к.т.н. _____ Галина ЛАЗЬКО

в.о. заст. нач. планово-фінансового відділу _____ Ірина ФАСТ

UNIDATA LAB

ЗАТВЕРДЖУЮ
Директор
UNIDATALAB LTD
Company number: 14551292
_____Гліб ЖЕБРАКОВ
41 Devonshire Street, London,
W1G 7AJ, United Kingdom

ДОВІДКА
про впровадження результатів дисертації
Рудої Христини Степанівни
«Удосконалення методів та засобів біометричної автентифікації за голосом з застосуванням технологій машинного навчання» представленої на здобуття наукового ступеня доктора філософії за спеціальністю 125 «Кібербезпека»

Цією довідкою підтверджується використання результатів наукових досліджень, отриманих у дисертаційній роботі аспірантки Рудої Христини Степанівни на тему «Удосконалення методів та засобів біометричної автентифікації за голосом з застосуванням технологій машинного навчання».

Зокрема, застосовано окремі розроблені методи, програмні рішення та науково обґрунтовані підходи у рамках дослідницьких і прикладних проєктів, пов’язаних із побудовою, тестуванням та удосконаленням систем біометричної автентифікації. Отримані результати сприяли підвищенню ефективності технічних рішень, пов’язаних із обробкою голосових даних, підвищенням точності розпізнавання та забезпеченням стійкості до потенційних загроз.

Аспірантка також надавала консультативну допомогу щодо підходів та методів побудови, навчання та перевірки штучних нейронних мереж.

Директор
UNIDATALAB LTD _____ **Гліб ЖЕБРАКОВ**

ДОДАТОК Б. СПИСОК НАУКОВИХ ПРАЦЬ

Наукові праці, в яких опубліковано наукові результати дисертації:

1. Дудикевич В. Б., Микитин Г. В., Руда Х. С. Застосування глибокого навчання для виявлення deepfake модифікацій біометричного зображення // Сучасна спеціальна техніка. – 2022. – № 1(68). – С. 13–22. *Особистий внесок здобувача: розроблено архітектуру системи аналізу біометричних даних; проведено експериментальну оцінку ефективності моделі у контексті інформаційної безпеки.*

2. Dudykevych V., Yevseiev S., Mykytyn G., Ruda K., Hulak H. Detecting deepfake modifications of biometric images using neural networks // CEUR Workshop Proceedings. – 2024. – Vol. 3654 : Cybersecurity providing in information and telecommunication systems 2024. Proceedings of the workshop cybersecurity providing in information and telecommunication systems (CPITS 2024), Kyiv, Ukraine, February 28, 2024 (online). – P. 391–397. *Особистий внесок здобувача: виконано адаптацію архітектури системи для обробки біометричних даних; проаналізовано результати роботи моделі за різними конфігураціями параметрів.*

3. Zaiets I., Brydinskyi V., Sabodashko D., Khoma Y., Ruda K. Integrated system for speaker diarization and intruder detection using speaker embeddings // CEUR Workshop Proceedings. – 2024. – Vol. 3654 : Cybersecurity providing in information and telecommunication systems 2024. Proceedings of the workshop cybersecurity providing in information and telecommunication systems (CPITS 2024), Kyiv, Ukraine, February 28, 2024 (online). – P. 228–238. *Особистий внесок здобувача: реалізовано модуль виділення ембедінгів мовців; проведено тестування інтегрованої системи діаризації та виявлення зловмисників.*

4. Заєць І. С., Бридінський В. А., Сабодашко Д. В., Руда Х. С., Хома Ю. В., Швед М. Є. Використання ембедінгів голосу в інтегрованих системах для діаризації мовців та виявлення зловмисників // Комп’ютерні системи та мережі. – 2024. – Вип. 6, № 1. – С. 54–66. *Особистий внесок здобувача:*

досліджено ефективність використання ембедінгів мовлення у багатокомпонентних системах; здійснено оцінку точності та продуктивності методів діаризації.

5. Микитин Г. В., Руда Х. С. Концептуальний підхід до виявлення deepfake-модифікацій біометричного зображення засобами нейронних мереж // Комп'ютерні системи та мережі. – 2024. – Вип. 6, № 1. – С. 124–132. *Особистий внесок здобувача: сформульовано концептуальну модель системи обробки біометричних даних; реалізовано її прототип із використанням згорткових нейромереж.*
6. Микитин Г. В., Руда Х. С., Яремчук Ю. Є. Методологія безпеки нейромережевих інформаційних технологій виявлення deepfake модифікацій біометричного зображення // Вісник Вінницького політехнічного інституту. – 2024. – № 1(172). – С. 74–80. *Особистий внесок здобувача: досліджено загрози безпеці в біометричних системах; запропоновано методологічні засади захисту таких систем на основі нейромережевих рішень.*
7. Руда Х. С., Сабодашко Д. В., Микитин Г. В., Швед М. Є., Бордуляк С. М., Коршун Н. Порівняння методів цифрової обробки сигналів та моделей глибинного навчання у голосовій аутентифікації // Кібербезпека: освіта, наука, техніка. – 2024. – № 5(25). – С. 140–160. *Особистий внесок здобувача: проведено порівняльний аналіз традиційних та глибинних методів обробки мовного сигналу; запропоновано критерії вибору оптимальної архітектури для задачі голосової автентифікації.*
8. Руда Х. С. Дослідження масштабованості біометричних систем автентифікації на основі вбудовування голосу // Social Development and Security. – 2025. – Vol. 15, iss. 1. – P. 161–170. *Особистий внесок здобувача: виконано дослідження масштабованості систем на основі ембедінгів голосу; проведено серію експериментів з різною кількістю користувачів.*
9. Руда Х. С. Використання візуальних і неявних артефактів для виявлення deepfake-модифікацій у біометричних зображеннях // Сучасні методи,

інформаційне, програмне та технічне забезпечення систем керування організаційно-технічними та технологічними комплексами : матеріали VIII Міжнародної науково-технічної Internet-конференції (Київ, 26 листопада 2021 року). – 2021. – С. 207–208. *Особистий внесок здобувача: реалізовано попередню обробку вхідних даних для покращення якості екстракції ознак біометричного зображення.*

10. Dudykevych V., Mykytyn H., Ruda K. The concept of a deepfake detection system of biometric image modifications based on neural networks // 2022 IEEE 3rd KhPI Week on Advanced Technology : conference proceedings, Kharkiv, 3–7 October 2022. – 2022. – P. 585–588. *Особистий внесок здобувача: сформульовано загальні засади побудови систем аналізу біометричних даних із використанням нейромережевих моделей; здійснено підбір архітектури відповідно до вимог системи.*

11. Микитин Г. В., Руда Х. С., Хома Ю. В. Categorization of deepfake methods: a comparison of single-shot and multi-shot transfer approaches // Захист інформації і безпека інформаційних систем : матеріали IX Міжнародної науково-технічної конференції (Львів, 25–26 травня 2023 р.). – 2023. – С. 109–110. *Особистий внесок здобувача: класифіковано методи генерації синтетичних даних та проведено їх систематичний аналіз; визначено ефективність підходів до передачі знань у нейромережевих структурах.*

12. Сабодашко Д. В., Руда Х. С., Швед М. Є., Бордуляк С. М. Технологія ембедінгів голосу та діаризація мовців // XX International Scientific and Practical Conference «Scientific Research: Modern Challenges and Prospects» : collection of abstracts, April 24–26, 2024, Prague, Czech Republic. – 2024. – P. 72–74. *Особистий внесок здобувача: досліджено механізми формування голосових ембедінгів у задачах розділення мовців; виконано аналіз параметрів, що впливають на якість діаризації.*

13. Руда Х. С., Микитин Г. В. Методи детекції синтетичних аудіозаписів, створених системами клонування голосу // Scientific Research in the Age of Virtual Reality: Exploring New Frontiers : collection of abstracts of LII

International Scientific and Practical Conference (December 18–20, 2024), Montreal, Canada. – 2024. – P. 132–136. *Особистий внесок здобувача: виконано огляд сучасних методів обробки синтетичного мовлення; запропоновано підхід до оцінювання достовірності голосових зразків у системах автентифікації*

14. Mykytyn H., Ruda K., Shved M. The paradigm of building secure information technologies for detecting deepfake modifications of biometric images based on neural networks // *Informatyka techniczna i sztuczna inteligencja: колективна монографія*. – Chapter in a collective monograph. – Bielsko-Biała: Wydawnictwo naukowe Uniwersytetu Bielsko-Bialskiego, 2024. – P. 43–50. *Особистий внесок здобувача: проаналізовано вразливості біометричних систем до атак із застосуванням штучного інтелекту; обґрунтовано підходи до побудови систем захисту на основі глибинного навчання.*