КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 1 (25), 2024
ISSN 2663 - 4023
DOI 10.28925/2663-4023.2024.25.140160
УДК 004.8

Руда Христина Степанівна
аспірантка кафедри захисту інформації
Національний Університет «Львівська Політехніка», Львів, Україна
ORCID ID: 0000-0001-8644-411X
khrystyna.s.ruda@lpnu.ua

Сабодашко Дмитро Володимирович
доктор філософії, старший викладач кафедри захисту інформації
Національний Університет «Львівська Політехніка», Львів, Україна
ORCID ID: 0000-0003-1675-0976
dmytro.v.sabodashko@lpnu.ua

Микитин Галина Василівна
д.т.н., професор, професор кафедри захисту інформації
Національний Університет «Львівська Політехніка», Львів, Україна
ORCID ID: 0000-0003-4275-8285
halyna.v.mykytyn@lpnu.ua

Швед Марія Євгенівна
кандидат технічний наук, асистент кафедри захисту інформації
Національний Університет «Львівська Політехніка», Львів, Україна
ORCID ID: 0000-0003-0428-7777
mariia.y.shved@lpnu.ua

Бордуляк Святослав Михайлович
студент кафедри захисту інформації
Національний Університет «Львівська Політехніка», Львів, Україна
ORCID ID: 0009-0007-2076-9297
sviatoslav.borduliak.kb.2020@lpnu.ua

Коршун Наталія Володимирівна
д.т.н., професор, професор кафедри інформаційної та
кібернетичної безпеки імені професора Володимира Бурячка
Київський столичний університет імені Бориса Грінченка, Київ, Україна
ORCID ID: 0000-0003-2908-970X
n.korshun@kubg.edu.ua

ПОРІВНЯННЯ МЕТОДІВ ЦИФРОВОЇ ОБРОБКИ СИГНАЛІВ ТА МОДЕЛЕЙ ГЛИБИННОГО НАВЧАННЯ У ГОЛОСОВІЙ АУТЕНТИФІКАЦІЇ

Анотація. У цій статті розглядаються проблеми класичних методів аутентифікації, таких як використання паролів, які часто виявляються ненадійними через різноманітні уразливості. Основні недоліки цих методів включають втрату або крадіжку паролів, їх слабку стійкість до атак, а також складність управління паролями, особливо у великих системах. Біометричні методи аутентифікації, зокрема ті, що базуються на фізичних характеристиках, таких як голос, є перспективним рішенням, оскільки вони забезпечують високий рівень безпеки і зручності для користувачів. Біометричні системи аутентифікації мають переваги над традиційними методами, оскільки голос є унікальною характеристикою для кожної людини, що значно ускладнює можливість підробки або крадіжки. Проте, існують виклики щодо точності і надійності таких систем. Зокрема, голосові біометричні системи можуть стикатися з проблемами, пов'язаними зі змінами голосу через здоров'я, емоційний стан або навколишнє середовище. Метою статті є порівняння сучасних моделей глибинного навчання з традиційними методами цифрової обробки сигналів, які використовуються для розпізнавання особистості за голосом. Для даного дослідження були обрані текстозалежні методи (мел-частотні кепстральні коефіцієнти — MFCC, кодування з лінійним предиктором — LPC) та

текстонезалежні методи (ECAPA-TDNN, ResNet) з метою порівняння їхньої ефективності у задачах біометричної аутентифікації за голосом. Експеримент складався з реалізації систем біометричної аутентифікації, побудованих на основі кожного з описаних методів, та оцінки їхньої ефективності на спеціально зібраному наборі даних. Також в роботі детально розглянуто методи попередньої обробки аудіосигналів, які застосовуються в системах голосової автентифікації з метою забезпечення найкращої результативності в задачах розпізнавання мовця, зокрема такі як знешумлення методом спектрального віднімання, нормалізацію енергії, підсилювальну фільтрацію, фреймування та застосування віконного методу.

Ключові слова: біометричні технології; голосова автентифікація; цифрова обробка сигналів; мел-частотні кепстральні коефіцієнти; кодування з лінійним предиктором; глибинне навчання; нейронні мережі.

ВСТУП

На сьогоднішній день Інтернет є потужною платформою, що значно вплинула на спілкування та бізнес-процеси у сучасному світі. Зараз кількість його користувачів перевищує 2,4 мільярда, що сприяє значному росту популярності онлайн-торгівлі, обміну знаннями та соціальних мереж. Проте, з цим зростанням приходить і збільшення потреби в надійних заходах кібербезпеки та захисті конфіденційності.

Статистика на 2023 рік показує, що Facebook лишається найпопулярнішою соціальною мережею для зламів, які становлять понад 68 000 аккаунтів щомісяця. Часто це стає наслідком недостатньої уваги користувачів до кібербезпеки. Це підвищує значення трьох ключових аспектів безпеки: ідентифікації, автентифікації та авторизації. Ідентифікація — це процес визначення сутності, яка може бути людиною, машиною чи іншим активом, наприклад, програмним забезпеченням. У контексті безпеки, автентифікація та авторизація визначають, хто має доступ до інформаційних ресурсів у мережі. «Ідентифікація», «автентифікація» та «авторизація» є ключовими поняттями, що складають основу технологій безпеки інформаційних систем. Ідентифікація передбачає передавання ідентифікатора до системи. Після цього, перед автентифікацією заявник зазвичай надає інформаційній системі свій ідентифікатор (наприклад, логін або адресу електронної пошти), а монітор підтверджує ідентифікацію через процес автентифікації (наприклад, за допомогою пароля). Автентифікація є процесом доведення заявником своєї ідентичності, щоб монітор міг підтвердити, що він дійсно відповідає вказаній особі. Нарешті, авторизація визначає надані користувачеві привілеї.

Системи автентифікації відповідають на обидва ключові питання: «хто є користувачем?» і «чи є цей користувач дійсно тим, за кого себе видає?» Таким чином, автентифікація є одним із найперспективніших засобів підвищення довіри та безпеки для комерційних застосувань. Вона також гарантує забезпечення ідентичності згаданих суб’єктів. Одним з найкращих методів для забезпечення безпечної автентифікації користувача є використання біометричної автентифікації. В цьому контексті виникає необхідність дослідження оптимального методу для реалізації системи біометричної автентифікації.

Постановка проблеми. Класичні методи аутентифікації, такі як використання паролів, виявляються недостатньо надійними через велику кількість потенційних уразливостей, таких як втрати або крадіжки паролів, їх слабка стійкість і складність управління. Методи біометричної автентифікації, що базуються на фізичних характеристиках, таких як голос, є перспективним рішенням, оскільки здатні

забезпечувати високий рівень безпеки і зручності для користувачів. Однак, існують виклики щодо точності і надійності таких систем, що потребують додаткового дослідження та вдосконалення. В цьому напрямі актуальним і ефективним є дослідження сучасних моделей глибинного навчання та їх попередників, що базуються на цифровій обробці сигналів, які використовуються для біометричної автентифікації за голосом.

Аналіз останніх досліджень і публікацій. Згідно з [1], розпізнавання голосу (32%), відбитки пальців (27%), сканування обличчя (20%), геометрія руки (12%) і сканування райдужної оболонки ока (10%) були виявлені як п’ять найкращих біометричних вимірювань для споживача. Голос людини є складним фізіологічним явищем, яке включає взаємодію між ротовою порожниною та горловими органами з їх рухомими компонентами. Ці компоненти відображають як фізіологічні, так і поведінкові характеристики особистості. Голос можна розглядати як унікальну біометричну систему, засновану на звуковому складі мовлення, який включає різні частоти і амплітуди. Акустичні властивості голосу дозволяють ідентифікувати індивідуальні особливості голосового сигналу. Вони включають спектральний склад звуку, його часові характеристики та інші параметри, унікальні для кожної людини. Завдяки цим характеристикам голос може бути використаний для надійної ідентифікації та аутентифікації особи.

Голосова біометрична система є результатом інтердисциплінарного підходу, який поєднує інженерні розроблення з біологічними науками для створення ефективних і надійних методів ідентифікації особи за голосом [2].

Інноваційним рішенням у галузі біометричних технологій є впровадження систем, які поєднують одночасну верифікацію кількох біометричних ознак. Наприклад, підвищення рівня безпеки систем контролю доступу можна досягти шляхом заміни традиційних дверних ключів на надійні електромагнітні системи замків. Ці системи дозволяють доступ лише авторизованим особам за допомогою розпізнавання голосу або ідентифікації відбитків пальців [3].

Системна нерівність у біометричних системах, заснована на расових і гендерних відмінностях, останнім часом привернула багато уваги. Ці розбіжності були досліджені в існуючих біометричних системах, таких як біометрія обличчя (ідентифікація осіб на основі ознак обличчя). Ці расові чи гендерні відмінності можуть призвести до серйозних упереджень або інших соціальних проблем, коли голосові системи розгортаються у великих масштабах [4].

Додатковим викликом для систем голосової аутентифікації є технології клонування голосу. Ці системи можуть створити голосову модель мовця на основі лише кількох хвилин автентичного запису. Отримана модель здатна відтворювати будь-який текст обраним голосом, що становить значну загрозу для безпеки біометричних систем, заснованих на голосовій ідентифікації [5].

У контексті цих викликів майбутні дослідження в сфері голосової аутентифікації повинні зосередитися на розробці систем, здатних не лише точно ідентифікувати голос певного мовця, але й виявляти синтетичні аудіозаписи.

Мета статті — порівняння сучасних моделей глибинного навчання та їх попередників — методів, що базуються суто на цифровій обробці сигналів, що широко використовуються для розпізнавання людини за голосом; визначення ефективних підходів та інструментів для обробки аудіозаписів в процесі підготовлення їх для опрацювання системами голосової автентифікації з метою забезпечення найкращої результативності.

Основними завданнями статті є:

1. Огляд сучасних видів біометричної аутентифікації;
2. Дослідження методів обробки даних в голосовій аутентифікації;
3. Порівняння текстозалежних і текстонезалежних методів голосової аутентифікації.

РЕЗУЛЬТАТИ ДОСЛІДЖЕНЬ

Огляд сучасних видів біометричної автентифікації

Біометрична аутентифікація є методом ідентифікації особи на основі її унікальних фізичних, поведінкових або психологічних характеристик. Сьогодні цей метод знаходить широке застосування в різних сферах, включаючи смартфони, банківські системи та системи безпеки приміщень. Огляд видів біометричної аутентифікації дозволяє глибше зрозуміти їхні переваги та обмеження. Серед методів біометричної автентифікації найбільш популярними є:

1) розпізнавання обличчя — системи розпізнавання облич широко застосовуються, зокрема в проектах національної безпеки та у взаємодії людини з комп'ютером. Вони працюють за принципом попередньої обробки вхідного зображення, виділення облич з натовпу та подальшого аналізу для їх ідентифікації. Основні методи розпізнавання облич поділяються на два типи: методи, засновані на особливостях, та методи, засновані на зовнішньому вигляді. Методи, засновані на особливостях, використовують інформацію про структуру обличчя. Натомість методи, засновані на зовнішньому вигляді, розглядають задачу розпізнавання облич як проблему розпізнавання образів, базуючись на статистичному навчанні. В останні роки було запропоновано різні алгоритми для розпізнавання облич з високою точністю, кожен з яких має свої переваги та недоліки, що враховується при їх оцінці [6].

Одним із популярних підходів є розпізнавання облич на основі виділення та класифікації ознак, де спочатку виділяються риси обличчя, а потім класифікатор ідентифікує обличчя. Інший підхід передбачає використання глибокого навчання та нейронних мереж (DNN), що забезпечує високу точність розпізнавання, особливо при роботі з великими обсягами даних.

Алгоритми розпізнавання облич можуть автоматично визначати та ідентифікувати обличчя на великих масштабах, що робить їх корисними для безпеки, контролю доступу та відеоспостереження. Проте, ці алгоритми мають також певні обмеження, такі як вразливість до змін освітлення і ракурсу. Крім того, при їх використанні виникають питання щодо приватності та етики;

2) розпізнавання відбитків пальців — ідентифікація за відбитками пальців, яка базується на стійких особливостях папілярних ліній, здатна забезпечити надійний метод ідентифікації. Процес отримання зображень відбитків пальців включає виділення темних папілярних ліній та світлих проміжків, що може ускладнюватися факторами навколишнього середовища та поведінкою користувача, і часто потребує методів покращення зображень. Використання алгоритмів обробки зображень допомагає підвищити чіткість та контрастність відбитків, що забезпечує більш точне визначення унікальних характеристик. Крім того, сучасні системи ідентифікації за відбитками пальців використовують методи машинного навчання для підвищення точності розпізнавання та адаптації до різних умов зйомки.

На рис. 1 зображено інформативний характер відбитка капілярних ліній, що застосовуються дня ідентифікації за відбитком пальця.

Ідентифікаційна інформація за відбитками пальців поділяється на три рівні: макродеталі (рівень 1), детальні характеристики (рівень 2) та повний набір розмірних атрибутів (рівень 3). Рівень 1 включає глобальні шаблони відбитків пальців, які не є унікальними, тоді як рівень 2 має достатню дискримінаційну здатність завдяки дрібним деталям візерунка. Рівень 3, що включає структурні ознаки, є постійним і дійсно унікальним.

Рис. 1. Інформативний характер відбитка папілярних ліній.

Алгоритми ідентифікації відбитків пальців можна розділити на три групи: методи на основі зв’язку, методи на основі деталей та методи на основі нечітких ознак. Методи на основі зв’язку такі, як генеративні антагоністичні мережі (GAN), стекові автоенкодери (SAE) та мережі глибокого вірування (DBN), встановлюють зв’язки між невеликими точками відбитків пальців та їх навколишніми ознаками [7]. Методи на основі деталей використовують техніки для захоплення та порівняння малих деталей, наприклад, обмежену машину Больцмана (RBM), рекурентну нейронну мережу (RNN) та мережу радіальних базисних функцій (RBFN);

3) сканування сітківки й розпізнавання райдужної оболонки очей — це автоматизований біометричний метод ідентифікації, що використовує математичні алгоритми для аналізу відеозображень однієї або обох райдужних оболонок очей. Складні візерунки райдужної оболонки є унікальними, стабільними протягом життя і можуть бути розпізнані з певної відстані. Дискримінаційна здатність біометричних технологій визначається кількістю ентропії, яку вони можуть кодувати та використовувати для зіставлення [8]. Розпізнавання райдужної оболонки є особливо ефективним у цьому відношенні, зводячи до мінімуму ймовірність помилкових збігів навіть у великих популяціях. Основне обмеження цієї технології полягає в тому, що отримання якісного зображення з відстані понад один-два метри або без активної співпраці суб’єкта може бути складним. Проте, технологія розвивається, і розпізнавання райдужної оболонки вже можливе на відстані до 10 метрів або за допомогою камер в режимі реального часу.

4) розпізнавання мовця — це процес ідентифікації особи на основі її унікальних голосових характеристик. Унікальність голосу обумовлена індивідуальними особливостями голосового тракту, розмірами гортані та інших органів голосового апарату. Крім фізичних особливостей, кожна людина має власний стиль мовлення, вимовні шаблони, та індивідуальний вибір лексики, що дозволяє використовувати голос як біометричний параметр для автентифікації [9]. Завдання розпізнавання мовця поділяються на верифікацію та ідентифікацію. Верифікація диктора полягає у визначенні того, чи є особа, яка стверджує свою ідентичність, дійсно тією, за кого вона себе видає. Ідентифікація диктора — це процес визначення, хто саме говорить серед відомих дикторів. Більшість систем розпізнавання мовця використовують мел-частотні кепстральні коефіцієнти (MFCCs) та лінійні передбачувальні кепстральні коефіцієнти (LPC), які представляють характеристики голосового тракту.

Загалом, кожна з цих біометричних технологій має свої специфічні сильні сторони та області застосування. Комбінація кількох методів може забезпечити підвищену надійність і точність ідентифікації, що робить біометричні системи ключовим компонентом сучасних рішень для забезпечення безпеки та зручності.

Дослідження методів обробки даних в голосовій аутентифікації

Система розпізнавання голосу складається з двох основних модулів: отримання ознак та порівняння ознак. Отримання ознак — це процес екстрагування обмеженого обсягу даних з аудіосигналу, які подальше можна використовувати для представлення кожного користувача. Порівняння ознак включає етап ідентифікації невідомого користувача шляхом зіставлення вилучених ознак з вхідного голосового сигналу з ознаками, що належать відомим користувачам в системі.

Попередня обробка даних

У системах голосової біометрії перший етап — це фаза попередньої обробки (Preprocessing). Попередня обробка мови є надзвичайно важливою, особливо у випадках, коли тиша або навколишній шум є небажаними. Виявлення голосової активності (VAD) є добре відомою технікою, яка використовується протягом багатьох років для попередньої обробки мовного сигналу. Це включає шумозаглушення, попереднє виділення та зменшення розмірності мовлення, що робить систему більш ефективною з обчислювальної точки зору. Цей тип класифікації мовлення на дзвінкі або тихі/недзвінкі звуки [10] знаходить широке застосування, зокрема в оцінці фундаментальної частоти, виділенні формантів, маркуванні складів, ідентифікації кінцевих приголосних та визначенні кінцевої точки, а також у виявленні ізольованих висловлювань.

Існує декілька методів класифікації (маркування) подій у мовленні. Загальноприйнятою є трирівнева репрезентація, яка включає наступні стани:

- *тиша*: відсутність мовлення;
- *беззвучний*: голосові зв’язки не вібрують, тому результуюча форма мовної хвилі є випадковою або неперіодичною;
- *дзвінкий*: голосові зв’язки вібрують періодично, коли повітря витікає з легень, тому результуюча форма хвилі є квазіперіодичною [10].

При розробленні системи біометричної автентифікації за голосом, попередня обробка є першою фазою інших етапів, що дозволяє розрізняти мовні та немовні сигнали та створювати вектори ознак. Попередня обробка коригує або модифікує мовний сигнал таким чином, щоб він був більш придатним для вилучення ознак аналізу. Серед етапів попередньої обробки голосового сигналу можна виділити наступні:

1) знешумлення методом спектрального віднімання — метод спектрального віднімання [11] підвищує якість вихідного зашумленого сигналу шляхом зменшення шуму в частотній ділянці. Він починається з перетворення зашумленого сигналу з часової області в частотну, розділяючи сигнал на окремі частотні компоненти. Шляхом оцінки спектру шуму (часто це робиться шляхом аналізу частин сигналу, де, як припускається, присутній шум, або за допомогою окремого запису шуму), метод визначає характеристики шуму. Цей розрахунковий спектр шуму потім віднімається із загального спектру шумового сигналу, що ефективно зменшує шум, одночасно прагнучи зберегти вихідний сигнал. Очищений частотний спектр нарешті перетворюється назад у часову область, що призводить до покращеного сигналу зі значно зниженим шумом. Загалом спектральне віднімання ізолює та видаляє шумові компоненти, покращуючи чіткість і якість сигналу;

2) нормалізація енергії — процес, який використовується для зміни амплітуди сигналу з метою балансування його енергетичних властивостей або силових параметрів. Цей процес спрямований на стандартизацію амплітуд сигналів для поліпшення їхнього вимірювання, аналізу та подальшої обробки [12]. Основна мета полягає у вирівнюванні енергії сигналу до заданого діапазону, що дозволяє зменшити вплив артефактів або шуму під час аналізу сигналу та забезпечити оптимальні умови для застосування алгоритмів обробки сигналів;

3) підсилювальна фільтрація — процедура, в якій розмовний аудіосигнал часто містить високочастотні компоненти, що можуть бути важливі для збереження при обробці. У деяких системах використовується передсигнальна фільтрація для компенсації високочастотних компонентів перед дискретизацією, щоб уникнути їхнього неправильного впливу на процес. Цей підхід спрямований на зменшення впливу шуму та артефактів на аналізований сигнал шляхом підкреслення високочастотних компонент і послаблення низькочастотних. Зашифрований мовний сигнал має високий динамічний діапазон і може бути супроводжений адитивним шумом. Для зменшення динамічного діапазону сигналу часто використовується передсигнальна фільтрація. Це включає застосування високочастотного фільтра, який зазвичай представляє собою фільтр скінченної імпульсної відповіді (FIR) з лінійними характеристиками та єдиним нулем близько до початку координат. Цей фільтр призначений для модифікації спектру мовного сигналу та підкреслення частот, на які найбільш чутлива людська слухова система [13];

4) *фреймування* — процес розбиття безперервного потоку мовних зразків на сегменти постійної довжини для полегшення блокової обробки сигналу. Мовний сигнал можна вважати квазістаціонарним, тобто стаціонарним лише протягом коротких періодів часу [14]. Як наслідок, мовний сигнал змінюється повільно з часом, і при дослідженні протягом короткого періоду (5–100 мс) він вважається стаціонарним. Тому, мовні сигнали часто аналізують у коротких часових інтервалах, що відомо як короткочасний спектральний аналіз у обробці мовлення. Це означає, що сигнал розбивається на кадри тривалістю зазвичай 20–30 мс. Сусідні кадри зазвичай перекривають один одного на 30–50% для запобігання втраті важливої інформації мовного сигналу, що може статися через застосування віконного методу;

5) віконний метод — процес множення сегмента мовного сигналу на вікно заданої форми, з метою підкреслення певних характеристик сигналу. На рис. 2 та 3 продемонстровано мовний сигнал до та після застосування вікна Хана відповідно.

ISSN 2663 - 4023

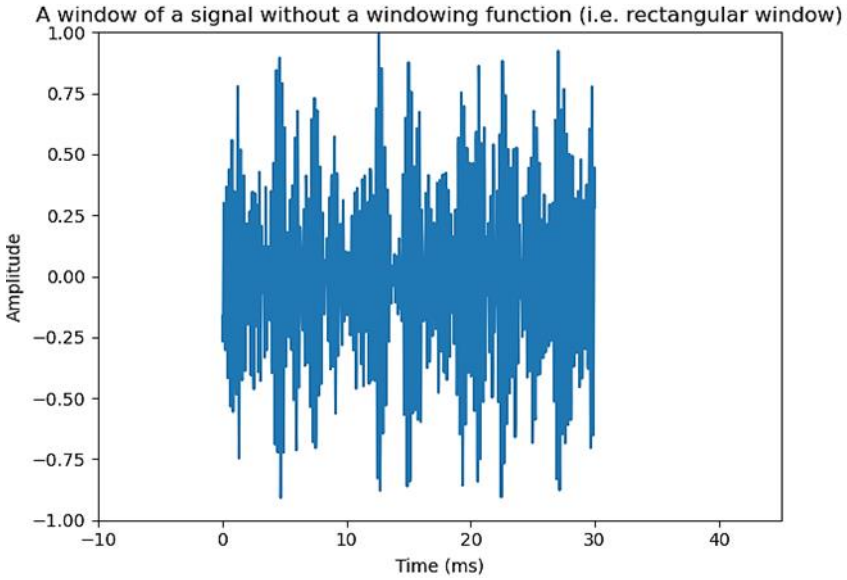

Рис. 2. Мовний сигнал без накладеного вікна

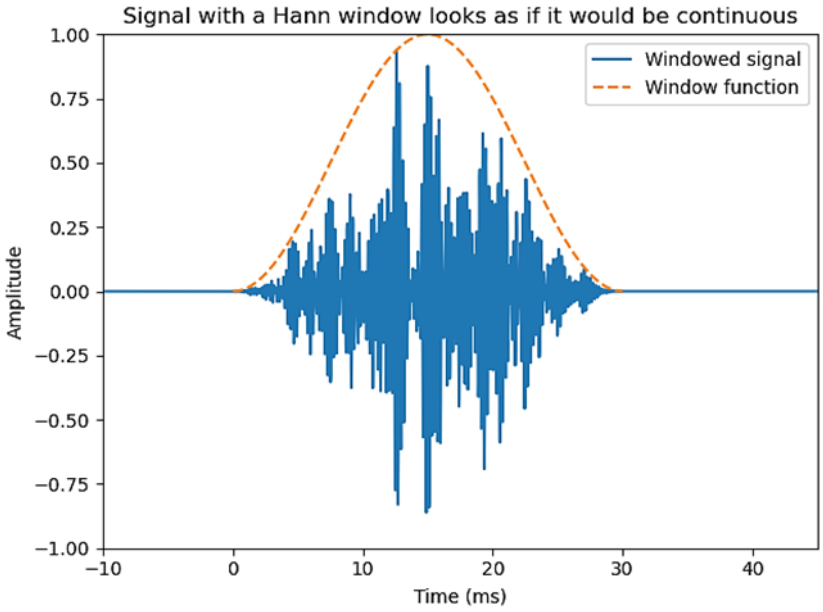

Рис. 3. Мовний сигнал після накладання вікна Хана

Для зменшення розривів мовного сигналу на початку і в кінці кожного кадру, сигнал має бути конічним або близьким до нуля. Це дозволяє мінімізувати неузгодженості. Даний ефект досягається шляхом «віконування» кожного кадру сигналу, що збільшує кореляцію Мел-частотних кепстральних коефіцієнтів (MFCC) та спектральних оцінок між послідовними кадрами [15].

Для отримання високої частотної роздільної здатності бажано використовувати довге вікно, однак, зважаючи на лінгвістичну важливість коротких перехідних процесів, коротке вікно є бажаним та ефективним. Оптимальним компромісом є довжина кадру приблизно 20–30 мс, з інтервалом між кадрами від 5 до 10 мс. Крім того, коротке вікно є достатнім для захоплення основних спектральних особливостей за умови, що інтервал

між кадрами також є достатньо коротким. Часто використовують вікно тривалістю 8 мс з інтервалом між кадрами 2 мс [16].

Правильний вибір вікна повинен враховувати баланс між різними факторами: форма вікна може зменшити відмінності, але водночас може змінити форму сигналу. Довжина вікна пропорційна частотній роздільній здатності й обернено пропорційна часовій роздільній здатності. Перекриття сигналу пропорційне частоті кадрів, але також пропорційне кореляції між наступними кадрами;

б) функція обчислення ознак — критично важливий етап в системах голосової автентифікації, який відіграє ключову роль у процесі ідентифікації особи за її голосом. Цей процес полягає у перетворенні вхідного аудіосигналу в набір числових ознак, що відображають його характеристики та специфічні особливості. Актуальність і значущість функції витягування ознак полягає в її здатності ефективно виділяти ключову інформацію з аудіосигналів, що дозволяє підвищити точність ідентифікації та забезпечити високий рівень безпеки у системах голосової автентифікації.

Передусім, функція витягування ознак забезпечує зменшення розмірності вхідних даних шляхом видалення шумів, що робить подальший аналіз більш ефективним та швидким. Відбір та перетворення лише найінформативніших ознак дозволяє використовувати менше ресурсів обчислення та зберігання, що є критично важливим для реалізації систем голосової автентифікації у реальному часі.

Крім того, функція витягування ознак допомагає відокремити суттєві акустичні характеристики голосового сигналу від шумів та інших непотрібних артефактів, підвищуючи стійкість та надійність системи автентифікації. Враховуючи лише ключові фізичні та акустичні параметри, такі як формантні частоти, енергія сигналу та інші, цей процес знижує вплив випадкових варіацій у голосових зразках та покращує точність ідентифікації.

Після проведення всіх етапів попередньої обробки даних та вилучення ознак, наступним кроком є процес класифікації. Класифікатор аналізує кожний зразок і визначає його належність до певного класу на основі попередньо визначених ознак. Рішення про класифікацію зразка приймається на основі порогового значення впевненості класифікатора. Якщо впевненість класифікатора в належності зразка до певного класу перевищує встановлене порогове значення, зразок вважається належним до цього класу. Такий підхід дозволяє зменшити ймовірність хибних класифікацій, забезпечуючи високу точність та надійність моделі.

Порівняння текстозалежних і текстонезалежних методів голосової аутентифікації

Голосова аутентифікація є важливою технологією, що забезпечує ідентифікацію користувачів на основі їхніх унікальних голосових характеристик. Ця технологія може бути реалізована двома основними методами: текстозалежними і текстонезалежними. Кожен метод має свої особливості, переваги та обмеження, які потрібно враховувати при їх впровадженні у системи безпеки.

Текстозалежні методи голосової аутентифікації

Текстозалежні методи передбачають, що користувач вимовляє заздалегідь визначену фразу або пароль, яку система використовує для порівняння з попередньо збереженим зразком голосу [17]. Цей підхід дозволяє системі фокусуватися на конкретних звукових патернах, що підвищує точність розпізнавання. Реалізація таких методів включає кілька важливих технічних етапів, кожен з яких потребує ретельної уваги для забезпечення точності та надійності аутентифікації.

На першому етапі, під час реєстрації користувача, він вимовляє визначену фразу кілька разів. Ці записи обробляються для створення еталонного зразка голосу, який зберігається у базі даних. Для забезпечення якісного запису необхідно використовувати мікрофони з високою роздільною здатністю, а також застосовувати методи попередньої обробки сигналу, такі як видалення шуму та нормалізація рівня гучності. Це дозволяє отримати чистий і чіткий запис голосу, який можна використовувати для подальшого порівняння.

Наступним важливим кроком є вилучення ознак з голосового сигналу. Цей процес є критично важливим для точності системи, оскільки він визначає, які характеристики голосу будуть використовуватися для порівняння. Одним з найбільш поширених методів вилучення ознак є використання коефіцієнтів мел-частотної кепстральної аналізи (MFCC), що дозволяє витягти характеристики голосу, стійкі до змін у тембрі та гучності. Інші методи, такі як лінійне предиктивне кодування (LPC), використовуються для моделювання вокального тракту і вилучення його характеристик. Також можна створювати унікальні відбитки голосу (voiceprint) на основі спектральних характеристик.

Після вилучення ознак система переходить до порівняння записаної фрази користувача з еталонним зразком. Для цього використовуються різні алгоритми, серед яких динамічне вирівнювання часу є одним з найбільш ефективних. Цей метод дозволяє вирівнювати часові відмінності між еталонним зразком та новим записом, забезпечуючи коректне порівняння. Крім того, використовуються методи машинного навчання, такі як підтримуючі вектори (SVM) або нейронні мережі, для класифікації голосових характеристик. Кореляційний аналіз також може бути застосований для вимірювання схожості між двома голосовими сигналами на основі кореляційних коефіцієнтів. Ці підходи забезпечують високу точність і надійність системи, що робить її більш стійкою до різноманітних викликів і змін у голосовому сигналі.

Переваги:

1) висока точність розпізнавання: оскільки система порівнює голос користувача з фіксованим еталоном, імовірність помилкового відторгнення або прийняття значно знижується;

2) простота реалізації: алгоритми текстозалежної аутентифікації зазвичай менш складні і вимагають менше ресурсів для обробки.

Недоліки:

1) вразливість до атак відтворення: зловмисники можуть записати голос користувача і використовувати цей запис для несанкціонованого доступу;

2) обмежена гнучкість: користувачі повинні вимовляти одну і ту ж фразу, що може бути незручно в певних сценаріях.

Текстонезалежні методи голосової аутентифікації

Текстонезалежні методи не вимагають від користувача вимовляти конкретну фразу. Натомість система аналізує унікальні характеристики голосу незалежно від змісту вимовленого тексту [18]. Цей підхід надає значну гнучкість і зручність у використанні, але вимагає врахування кількох складних технічних аспектів для забезпечення надійності та точності аутентифікації.

На етапі реєстрації користувач вимовляє кілька різних фраз. Ці записи обробляються для створення еталонного зразка голосу, який відображає унікальні характеристики користувача. Для забезпечення високої якості запису використовуються мікрофони з високою роздільною здатністю. Важливим є також попередня обробка

сигналу, яка включає видалення шуму, нормалізацію рівня гучності та видалення зайвих звукових артефактів. Ці кроки гарантують, що зразки є чистими і готовими для подальшого аналізу.

Наступним кроком здійснюється вилучення ознак, яке є критично важливим, оскільки саме на його основі здійснюється ідентифікація користувача. У текстонезалежних методах використовуються передові алгоритми для вилучення ознак, які залишаються стабільними незалежно від вимовленого тексту.

Один з найбільш популярних методів — це коефіцієнти мел-частотної кепстральної аналізи (MFCC), які дозволяють витягти основні характеристики голосу. Іншим важливим методом є спектральне злиття (spectral concatenation), яке забезпечує високу точність при аналізі голосових сигналів. Крім того, лінійне предиктивне кодування (LPC) використовується для моделювання вокального тракту і вилучення його характеристик. У деяких випадках застосовуються методи глибокого навчання, які автоматично вилучають і оптимізують ознаки з голосових даних.

Текстонезалежні методи потребують складних алгоритмів для порівняння зразків голосу. Одним з ключових підходів є використання моделей глибокого навчання, зокрема, рекурентних нейронних мереж (RNN) або згорткових нейронних мереж (CNN). Ці моделі навчаються на великій кількості даних і здатні розпізнавати складні патерни в голосових сигналах. Крім того, векторні квантизації (Vector Quantization) та алгоритми кластеризації, такі як k-means, використовуються для порівняння нових записів з еталонними зразками.

Важливим аспектом текстонезалежної аутентифікації є встановлення порогового значення, яке визначає рівень впевненості класифікатора у належності зразка до конкретного користувача. Це порогове значення налаштовується таким чином, щоб збалансувати між ймовірністю хибних відмов (false negatives) і помилкових спрацьовувань (false positives). Порогове значення можна адаптувати в залежності від конкретних вимог системи безпеки та рівня допустимих ризиків.

Переваги:

1) гнучкість у використанні: користувачі можуть говорити будь-який текст, що робить процес аутентифікації менш обтяжливим;

2) вищий рівень безпеки: текстонезалежні методи є більш стійкими до атак з використанням записів, оскільки базуються на унікальних властивостях голосу, які важко підробити.

Недоліки:

1) зниження точності: варіативність мовлення та складність аналізу різних текстів можуть знижувати точність розпізнавання;

2) складність реалізації: алгоритми текстонезалежної аутентифікації є більш складними і вимагають значних обчислювальних ресурсів для обробки.

Певні етапи, такі як встановлення порогового значення впевненості класифікатора, є аналогічними для обох категорій методів голосої автентифікації. Це значення визначає, при якому рівні впевненості система вважає користувача автентичним. Порогове значення можна налаштувати відповідно до вимог системи безпеки, з метою зменшення кількості помилкових спрацьовувань (false positives) та хибних відмов (false negatives).

З часом голос користувача може змінюватися, тому важливо впровадити механізми адаптації системи. Періодичне оновлення зразків голосу дозволяє враховувати природні зміни голосових характеристик. Використання алгоритмів машинного навчання, що можуть адаптуватися до нових даних, забезпечує стабільну точність системи навіть за умови зміни голосу користувача.

Врахування всіх цих технічних аспектів є критично важливим для успішного впровадження і функціонування систем голосової аутентифікації. Забезпечення високоякісного запису, ефективного вилучення ознак, точного порівняння зразків, оптимальних порогових значень, безпеки даних та адаптації до змін голосу дозволяє створити надійну і точну систему аутентифікації, що відповідає сучасним вимогам безпеки.

Для даного дослідження було обрано текстозалежні методи (мел-частотні кепстральні коефіцієнти — MFCC, кодування з лінійним предиктором — LPC) та текстонезалежні методи (ECAPA-TDNN, ResNet) з метою порівняння їхньої ефективності у задачах біометричної автентифікації за голосом. Експеримент складався з реалізації систем біометричної автентифікації, побудованих на основі кожного з описаних методів, та оцінки їхньої ефективності на спеціально зібраному наборі даних. Порівняння проводилось за наступними критеріями: False Accept Rate (FAR), False Reject Rate (FRR), Equal Error Rate (EER), Detection Cost Function (DCF), та Accuracy.

Метод мел-частотних кепстральних коефіцієнтів

Використання мел-частотних кепстральних коефіцієнтів (MFCC) є одним зі стандартних методів вилучення ознак у системах голосової аутентифікації [19]. Зазвичай використовують близько 20 коефіцієнтів MFCC для обробки даних, хоча в багатьох випадках достатньо 10–12 коефіцієнтів. Основним недоліком методу MFCC є його чутливість до шуму, оскільки він залежить від спектральної форми сигналу. Для подолання цієї проблеми можуть застосовуватися методи, які враховують інформацію, що міститься у періодичності мовних сигналів, незважаючи на наявність у мовленні аперіодичних компонентів. Послідовність цих методів проілюстрована на рис. 4, де представлена блок-схема конвеєра MFCC.

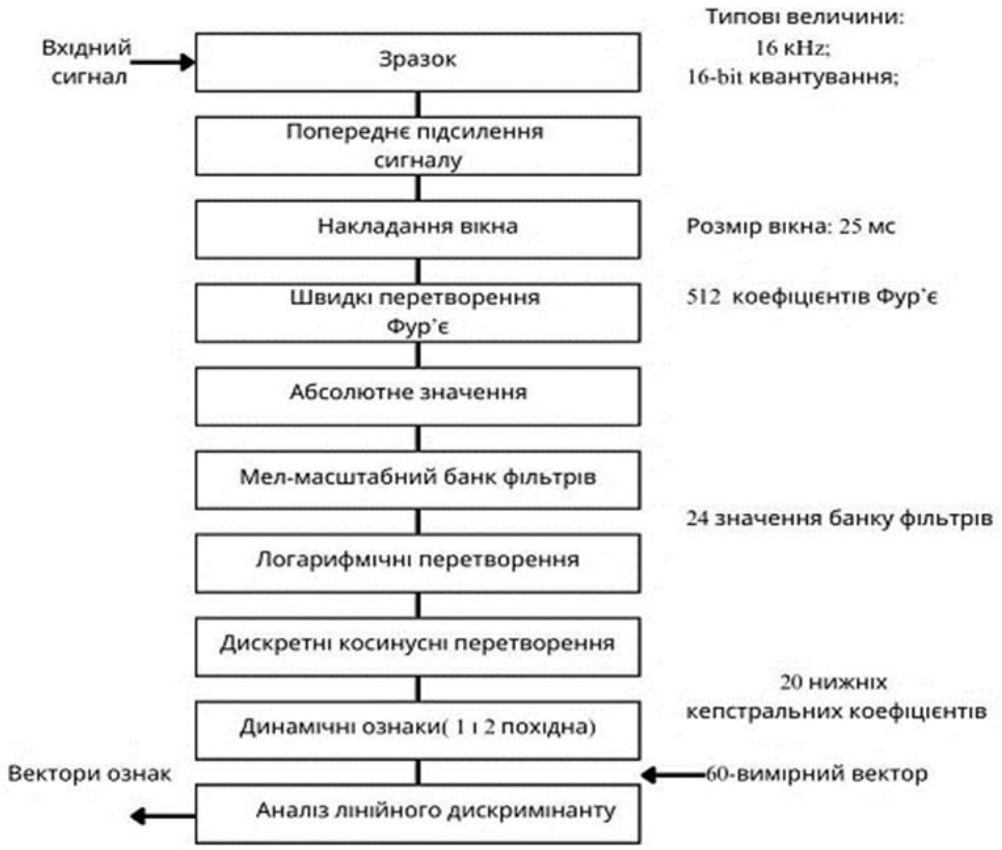

Рис. 4. Схема повного конвеєра MFCC

У нелінійній шкалі частот застосовується наближення до шкали мел-частот, яка є приблизно лінійною для частот нижче 1 кГц і логарифмічною для частот вище 1 кГц. Це пояснюється тим, що слухова система людини стає менш чутливою до частотних змін зі збільшенням частоти вище 1 кГц. Амплітудно-частотні характеристики (АЧХ) відповідають частотним характеристикам логарифмічно-нормального фільтра. Для їх обчислення спочатку обчислюється енергія логарифма з виходів фільтрів за формулою:

$$S_t[m] = ln(\sum |X_t[n]|^2 H_m[n]), 0 \le m \le M,$$

де $(X_t[n])$ — дискретне перетворення Фур'є (DFT) t-го вхідного мовного кадру, $(H_m[n])$ — частотна характеристика m-го фільтра, N — розмір вікна перетворення, а M — загальна кількість фільтрів. Потім дискретне косинусне перетворення (DCT) логарифмічної енергії обчислюється за формулою:

$$\overrightarrow{C_t}[m] = \sum S_t\,[n] cos\left(\pi m \left(\frac{n-0.5}{M}\right)\right), 0 \le m < M. \quad (1)$$

Оскільки слухова система людини чутлива до часової еволюції спектрального вмісту сигналу, часто включається аналіз змін цих характеристик. Для фіксації змін коефіцієнтів у часі обчислюються перший і другий різницеві коефіцієнти:

$$1.\ \left(\Delta\overrightarrow{C_t} = \overrightarrow{C_{t+2}} - \overrightarrow{C_{t-2}}\right);$$
$$2.\ \left(\Delta\Delta\overrightarrow{C_t} = \Delta C_{t+1} - \Delta C_{t-1}\right). \quad (2)$$

Ці динамічні коефіцієнти об'єднуються зі статичними коефіцієнтами для формування остаточного набору ознак, що представляють t-ий мовленнєвий фрейм.

Кодування з лінійним предиктором

Лінійне предикативне кодування (LPC) є одним із найпотужніших методів аналізу мовлення та корисним інструментом для кодування мовлення з низьким бітрейтом. Основна ідея лінійного предикативного аналізу полягає в тому, що поточний зразок мовлення можна апроксимувати як лінійну комбінацію попередніх зразків мовлення. LPC є моделлю, що базується на фізіології людського мовлення і використовує модель фільтра джерела, в якій функції передавання випромінювання горла, голосового тракту та губ інтегровані в один все полюсний фільтр, який імітує акустику голосового тракту [20].

Принцип роботи LPC полягає в мінімізації суми квадратів різниць між вихідним мовленнєвим сигналом і оціненим мовленнєвим сигналом протягом скінченного проміжку часу. Це дозволяє отримати унікальний набір предикативних коефіцієнтів, що обчислюються для кожного мовленнєвого кадру, який зазвичай має довжину 20 мс. Передавальна функція цифрового фільтра, що змінюється в часі, *H(z)*, описана формулою:

$$H(z) = \frac{G}{1-\sum a_k z^{-k}}, \quad (3)$$

де $k = 1,2,\ldots,p$, G — коефіцієнт підсилення, a_k — коефіцієнти LPC.

LPC-аналіз кожного кадру також включає процес прийняття рішення про те, чи є кадр озвученим або неозвученим. Для цього використовується алгоритм визначення висоти тону, який коригує період і частоту тону. Параметри висоти тону, посилення та предикативних коефіцієнтів змінюються з часом від кадру до кадру, що відображає динаміку мовленнєвого процесу.

ECAPA-TDNN

Нейронні мережі, подібні до людського мозку, проте на початкових етапах їхня ефективність поступається традиційним моделям машинного навчання. У 1989 році Хінтон та його колеги запропонували метод Time Delay Neural Network (TDNN) для вирішення проблеми ефективного опрацювання динамічних характеристик

аудіосигналів та розпізнавання фонем із включенням контекстної інформації. TDNN має дві важливі властивості: здатність динамічно адаптуватися до часових змін ознак і мінімальну кількість параметрів. На відміну від традиційних глибоких нейронних мереж, у TDNN приховані шари спільно впливають на поточні та майбутні вхідні дані, ефективно використовуючи інформацію про тимчасовий контекст.

Модель Emphasized Channel Attention, Propagation, and Aggregation Time Delay Neural Network (ECAPA-TDNN) [21], представлена в 2020 році, об'єднує традиційну архітектуру TDNN з механізмами уваги, акцентуючи увагу на каналах, поширення та агрегацію в TDNN, що базується на верифікаторі мовців. ECAPA-TDNN покращує екстракцію та відображення ознак завдяки включенню розширених шарів контекстуальної агрегації.

Мел-акустична спектрограма представляє форму сигналу в часовій ділянці та є простою і легко доступною ознакою для розпізнавання типів аудіосигналів. Проте, інформація в часовій ділянці має тенденцію представляти неоптимальні результати через свою чутливість до впливів і нестабільність. Натомість частотна інформація забезпечує більшу точність у визначенні характеристик звуку і є менш схильною до впливу перешкод. Перетворення часової області аудіосигналу в частотну досягається за допомогою перетворення Фур'є або wavelet-перетворення, хоча ці підходи можуть призводити до втрати певних характеристик сигналу.

Акустична спектрограма, побудована на основі спектрального аналізу з включенням часового виміру, пропонує інтуїтивне зображення змін сигналу, включаючи як часову, так і частотну інформацію. Дослідження показують, що частотна роздільна здатність людського вуха є логарифмічною. Введення мел-частот враховує чутливість людського вуха до частот звукового сигналу. Логарифмічна залежність між лінійною частотою і частотою мел визначається формулою:

$$F_{mel} = 2595lg\left(\frac{1+f}{700}\right), \quad (4)$$

де F_{mel} — сприйнята частота в мелах, аf — частота в Гц.

Архітектура ECAPA-TDNN включає додатковий рівень вбудовування для вилучення мовленнєвих ознак з вхідного мовного сигналу та їх використання для розпізнавання мовця. Мережа вчиться відокремлювати специфічну для мовця інформацію від іншої акустичної інформації, використовуючи кілька згорткових шарів і шарів LSTM, що фіксують довгострокові залежності вхідного аудіосигналу. Шар вбудовування створює вектор із фіксованою розмірністю, що представляє ідентичність мовця.

Загалом, ECAPA-TDNN є ефективною моделлю з відносно невеликою кількістю параметрів, що дозволяє швидко працювати навіть на пристроях з обмеженими ресурсами.

ResNet

Архітектура ResNet [22], успішна у комп'ютерному зорі, також знайшла застосування в системах голосової автентифікації. Вона ефективно витягує та обробляє складні ознаки з аудіосигналів, підвищуючи точність і надійність розпізнавання голосу.

У системах голосової автентифікації аудіосигнали перетворюються на спектрограми або мел-частотні кепстральні коефіцієнти (MFCC), які стають вхідними даними для нейронної мережі. Спектрограма є двовимірним зображенням, що дозволяє ResNet обробляти їх аналогічно до зображень у комп'ютерному зорі.

Архітектура ResNet для голосової автентифікації складається з вхідного шару, де використовуються спектрограми або інші форми перетворених аудіосигналів. Початкові згорткові шари витягують базові ознаки зі спектрограм. Основною частиною ResNet є

залишкові блоки, які допомагають мережі вчитися глибоким патернам у даних, зберігаючи ефективність навчання. Кожен залишковий блок містить кілька згорткових шарів з прямими з'єднаннями, що обминають ці шари. Після кожного згорткового шару використовуються пакетні нормалізації та функції активації для стабілізації та прискорення навчання. Глобальний середній пулінг зменшує розмір просторових вимірів, зберігаючи важливу інформацію, а повнозв'язний шар і softmax завершують мережу для класифікації або інших завдань, таких як векторне представлення голосу (x-vector).

Основною перевагою використання ResNet у голосовій автентифікації є покращена здатність до узагальнення. Завдяки глибині та структурі залишкових блоків, ResNet може ефективно навчатися на великих наборах даних, витягуючи складні патерни, що покращує точність автентифікації. Залишкові блоки допомагають уникнути проблеми згасання градієнтів, що є особливо важливим для глибоких мереж, які обробляють аудіодані. Крім того, ResNet є гнучкою архітектурою, яка може бути адаптована для різних форматів вхідних даних, таких як спектрограми чи MFCC, і легко інтегрується в сучасні системи.

ResNet може використовуватися як частина більшої системи для витягування ознак з аудіосигналу, які потім подаються на вхід інших моделей, таких як опорні векторні машини (SVM). Вона також може використовуватися для кінцевої класифікації користувача на основі його голосу, замінюючи або доповнюючи традиційні методи. Використання ResNet для генерації векторних представлень голосу (x-vector) дозволяє створювати компактні та інформативні вектори, які можна використовувати для порівняння та автентифікації.

Таким чином, ResNet є важливим інструментом у системах голосової автентифікації, надаючи ефективний і надійний спосіб обробки складних аудіосигналів. Завдяки гнучкості, стабільності та здатності до витягування високорівневих ознак, ResNet допомагає покращити точність і надійність голосової автентифікації, роблячи її цінним компонентом у сучасних системах безпеки та ідентифікації.

Опис набору даних

У рамках цього дослідження було зібрано спеціальний набір даних обсягом 245 МБ, що складається із записів 10 англомовних знаменитостей, які начитували аудіокниги. Кожен мовець має 20 записів тривалістю до 16 секунд. Загалом набір даних містить 200 звукозаписів із загальною тривалістю приблизно 30 хвилин.

Кожен запис у наборі даних є моноканальним аудіо з частотою дискретизації 16 кГц у форматі .wav. Формат .wav використовує PCM-кодування (Pulse Code Modulation) без компресії, що забезпечує високу якість і точність звукових файлів.

Набір даних розроблено таким чином, щоб бути різноманітним, включаючи мовців різної статі та віку, для забезпечення комплексної оцінки моделей голосової автентифікації.

Метрики оцінювання

1) False Accept Rate (FAR) або частота помилкового допуску, визначає частку неавторизованих користувачів, які були помилково прийняті системою як авторизовані. Ця метрика важлива для оцінки надійності системи біометричної автентифікації в контексті запобігання несанкціонованому доступу:

$$FAR = \frac{FN}{N}, \tag{5}$$

де FN — кількість помилкових прийомів, N — загальна кількість спроб доступу неавторизованих користувачів.

2) False Reject Rate (FRR) або частота помилкового відхилення, визначає частку авторизованих користувачів, які були помилково відхилені системою. Ця метрика важлива для оцінки зручності використання системи для легітимних користувачів:

$$FRR = \frac{FP}{N}, \quad (6)$$

де FP — кількість помилкових відхилень, N — загальна кількість спроб доступу авторизованих користувачів;

3) Equal Error Rate (EER) або рівень рівних помилок, визначається як точка, в якій FAR і FRR є рівними. EER є узагальненим показником ефективності біометричної системи, де нижчі значення EER вказують на кращу загальну продуктивність системи;

4) Detection Cost Function (DCF) використовується для зваженої оцінки продуктивності біометричної системи, враховуючи різну вартість помилкових прийомів та відхилень. Ця метрика корисна для налаштування системи відповідно до специфічних вимог безпеки та користувачів:

$$DCF = C_{FA} * P_{FA} + C_{FR} * P_{FR}, \quad (7)$$

де C_{FA} — вартість помилкового прийому, P_{FA} — імовірність помилкового прийому,C_{FR} — вартість помилкового відхилення, P_{FR} — ймовірність помилкового відхилення;

5) Accuracy або точність класифікації, визначає загальну частку правильних рішень системи, включаючи як авторизованих користувачів, так і неавторизованих.

$$Accuracy = \frac{TP+TN}{N}, \quad (8)$$

де $TP + TN$ — кількість правильних рішень (авторизованих прийомів TP та неавторизованих відхилень TN), N — загальна кількість спроб доступу.

Ці метрики разом забезпечують комплексну оцінку ефективності та надійності систем біометричної автентифікації.

Хід експерименту

Експеримент полягав у реєстрації класів у системі біометричної автентифікації та подачі на вхід системи зразків голосу як зареєстрованих користувачів, так і сторонніх осіб (зловмисників). Для кожного підходу була збудована окрема система автентифікації, яка аналізувала подані зразки голосу. Для оцінки продуктивності моделей верифікації користувачів для кожного мовця було сформовано спеціальні пари аудіозаписів в межах одного класу та пари з екземплярами різних класів.

Для кожного класу було проведено 20 експериментів: 10 експериментів із використанням зразків голосу валідних користувачів і 10 експериментів із використанням голосу сторонніх осіб. На основі отриманих результатів було побудовано табл. 1. Це дозволило оцінити ефективність кожної з розроблених систем у задачі біометричної автентифікації та порівняти їх між собою.

Таблиця 1

Результати експерименту

Назва методу/моделі	FAR(%)	FRR(%)	EER(%)	DCF	Точність (%)
MFCC	5	2.5	3.75	0.1	90
LPC	5	5	5	0.15	95
ECAPA-TDNN	0.2	0.2	0.2	3.6	97.6
ResNet	0.016	0.016	0.016	0.2	98.3

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

Вивчення біометричної автентифікації є важливим і актуальним через зростаючу потребу в надійних та зручних методах ідентифікації особистості в сучасному цифровому світі. Голосова автентифікація пропонує безконтактний, природний спосіб підтвердження особистості, що є особливо корисним у контексті мобільних пристроїв, смарт-динаміків та систем Інтернету речей (IoT). З розвитком технологій штучного інтелекту та машинного навчання підвищується точність і стійкість голосових біометричних систем, що дозволяє ефективно відокремлювати реальні голоси від підроблених або синтезованих. Крім того, голосова автентифікація має великий потенціал у забезпеченні додаткового рівня безпеки для фінансових транзакцій, доступу до конфіденційних даних та інших критичних додатків, де традиційні методи ідентифікації, такі як паролі або PIN-коди, можуть бути недостатньо надійними або зручними.

У системах голосової біометрії перший етап — це фаза попередньої обробки (Preprocessing). Попередня обробка коригує або модифікує мовний сигнал таким чином, щоб він був більш придатним для вилучення ознак аналізу, що дозволяє розрізняти мовні та немовні сигнали та створювати вектори ознак. Серед етапів попередньої обробки голосового сигналу можна виділити знешумлення методом спектрального віднімання, нормалізацію енергії, підсилювальну фільтрацію, фреймування та застосування віконного методу. Після проведення всіх етапів попередньої обробки даних та вилучення ознак, наступним кроком є процес класифікації, де класифікатор аналізує кожний зразок і визначає його належність до певного класу на основі попередньо визначених ознак.

Результати експерименту підтвердили ефективність різних підходів до біометричної автентифікації на основі голосу. Було збудовано окремі системи автентифікації для кожного підходу та проведено детальну оцінку їх продуктивності. У підсумку, найкращі результати продемонструвала система на основі мережі ResNet, досягнувши точності 98,3%. Це свідчить про високу надійність і потенціал цього підходу для застосування у реальних умовах.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Samuel, F. A., Titilayo, A. O., Abiodun, A. O., Modupe, A. O., Oyeladun, M. B., Mayowa, I. R., & Samuel, A. M. (2021). Voice recognition system for door access control using mobile phone. *International Journal of Science and Engineering Applications, 10(9),* 132–139. https://doi.org/10.7753/ijsea1009.1004
2. Amjad Hassan Khan, M. K., & Aithal, P. S. (2022). Voice Biometric Systems for User Identification and Authentication – A Literature Review. *International Journal of AppliedEngineering and Management Letters (IJAEML), 6(1),* 198–209. https://doi.org/10.5281/zenodo.6471040
3. Abe, B. C., Araromi, H. O., Shokenu, E. S., Idowu, P. O., Babatunde, J. D., Adeagbo, M. A., & Oluwole, I. H. (2022). Biometric Access Control Using Voice and Fingerprint. *Engineering And Technology Journal, 7(7),* 1376–1382. https://doi.org/10.47191/etj/v7i7.08
4. Chen, X., Li, Z., Setlur, S., & Xu, W. (2022). Exploring racial and gender disparities in voice biometrics. *Scientific Reports, 12(1).* https://doi.org/10.1038/s41598-022-06673-y
5. Inamdar, F. M., Ambesange, S., Mane, R., Hussain, H., Wagh, S., & Lakhe, P. (2023). Voice Cloning Using Artificial Intelligence and Machine Learning: A review. *Journal of Advanced Zoology, 44(S7),* 419–427. https://doi.org/10.17762/jaz.v44is7.2721
6. Dalvi, J., et al. (2022). *A survey on face recognition systems*. arXiv preprint.
7. Win, K., Li, K., Chen, J., Viger, P. (2020). Fingerprint classification and identification algorithms for criminal investigation: A survey. *Future Generation Computer Systems, 110,* 758–771. https://doi.org/10.1016/j.future.2019.10.019

8. Daugman, J. (2002). How iris recognition works. *Proceedings International Conference on Image Processing*. https://doi.org/10.1109/ICIP.2002.1037952
9. Poddar, A., Sahidullah, Md., Saha, G. (2017). Speaker verification with short utterances: a review of challenges, trends and opportunities. *IET Biometrics. 7(2),* 91–101. https://doi.org/10.1049/iet-bmt.2017.0065. ISSN 2047-4938
10. Childers, D. G., Hand, M., Larar-Silent, M. J. (1989). Voiced/Unvoiced/Mixed Excitation (Four Way), Classification of Speech. *IEEE Trans. On ASSP, 37(11).*
11. Upadhyay, N., & Karmakar, A. (2015). Speech Enhancement using Spectral Subtraction-type Algorithms: A Comparison and Simulation Study. *Procedia Computer Science, 54,* 574–584. https://doi.org/10.1016/j.procs.2015.06.066
12. Jakovljević, N., Janev, M., Pekar, D., & Mišković, D. (2008). Energy Normalization in Automatic Speech Recognition. In *Lecture Notes in Computer Science*, 341–347. https://doi.org/10.1007/978-3-540-87391-4_44
13. Hviyuzova, D., & Belitskiy, A. (2021). Development of a filter amplifier of the signal pre-processing device for the passive listening mode of the hydroacoustic complex (HAC). *E3S Web of Conferences, 266,* 04013. https://doi.org/10.1051/e3sconf/202126604013
14. Introduction to Speech Processing. (n. d.). https://speechprocessingbook.aalto.fi/Representations/Windowing.html
15. Junqua, J.-C., Mak, B., Reaves, B. (1994). A robust algorithm for word boundary detection in presence of noise. *IEEE Trans. on Speech and Audio Processing, 2,* 406– 412.
16. Junqua, J.-C., Mak, B., Reaves, B. (1994). A robust algorithm for word boundary detection in presence of noise. *IEEE Trans. on Speech and Audio Processing, 2*, 406–412.
17. Liu, Y., Qian, Y., Chen, N., Fu, T., Zhang, Y., & Yu, K. (2015). Deep feature for text-dependent speaker verification. *Speech Communication, 73,* 1–13. https://doi.org/10.1016/j.specom.2015.07.003
18. Heigold, G., Moreno, I., Bengio, S., & Shazeer, N. (2016). *End-to-end text-dependent speaker verification.* https://doi.org/10.1109/icassp.2016.7472652
19. Xu, M., Duan, L. Y., Cai, J., Chia, L. T., Xu, C., & Tian, Q. (2004). HMM-Based Audio Keyword Generation. In *Lecture Notes in Computer Science*, 566–574. https://doi.org/10.1007/978-3-540-30543-9_71
20. Wijoyo, S. (2011). *Speech Recognition Using Linear Predictive Coding and Artificial Neural Network for Controlling Movement of Mobile Robot.* http://fportfolio.petra.ac.id/user_files/97-031/E091%20full%20paper-Thiang%20-%20ICIEE%202011.pdf
21. Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). *ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification*. https://doi.org/10.21437/interspeech.2020-2650
22. Jakubec, M., Lieskovska, E., & Jarina, R. (2021). Speaker Recognition with ResNet and VGG Networks, *31st International Conference Radioelektronika (RADIOELEKTRONIKA),* 1–5. https://doi.org/10.1109/RADIOELEKTRONIKA52220.2021.9420202

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 1 (25), 2024
ISSN 2663 - 4023

Khrystyna Ruda
Postgraduate student at Department of Information Security
Lviv Polytechnic National University, Lviv, Ukraine
ORCID ID: 0000-0001-8644-411X
khrystyna.s.ruda@lpnu.ua

Dmytro Sabodashko
PhD, Senior Lecturer at Department of Information Security
Lviv Polytechnic National University, Lviv, Ukraine
ORCID ID: 0000-0003-1675-0976
dmytro.v.sabodashko@lpnu.ua

Halyna Mykytyn
Dc.S., Professor at Department of Information Security
Lviv Polytechnic National University, Lviv, Ukraine
ORCID ID: 0000-0003-4275-8285
halyna.v.mykytyn@lpnu.ua

Mariia Shved
Candidate of Technical Sciences (PhD), Lecturer at Department of Information Security
Lviv Polytechnic National University, Lviv, Ukraine
ORCID ID: 0000-0003-0428-7777
mariia.y.shved@lpnu.ua

Sviatoslav Borduliak
Student at Department of Information Security
Lviv Polytechnic National University, Lviv, Ukraine
ORCID ID: 0009-0007-2076-9297
sviatoslav.borduliak.kb.2020@lpnu.ua

Nataliia Korshun
Doctor of Science, Professor, Professor of Head of the Department of
Information and Cyber Security named after Professor Volodymyr Buryachok
Borys Grinchenko Kyiv Metropolitan University, Kyiv, Ukraine
ORCID ID: 0000-0003-2908-970X
n.korshun@kubg.edu.ua

COMPARISON OF DIGITAL SIGNAL PROCESSING METHODS AND DEEP LEARNING MODELS IN VOICE AUTHENTICATION

Abstract. This paper addresses the issues of traditional authentication methods, such as the use of passwords, which often prove to be unreliable due to various vulnerabilities. The main drawbacks of these methods include the loss or theft of passwords, their weak resistance to various types of attacks, and the complexity of password management, especially in large systems. Biometric authentication methods, particularly those based on physical characteristics such as voice, present a promising alternative as they offer a higher level of security and user convenience. Biometric authentication systems have advantages over traditional methods because the voice is a unique characteristic for each person, making it substantially more challenging to forge or steal. However, there are challenges regarding the accuracy and reliability of such systems. Specifically, voice biometric systems can encounter issues related to changes in voice due to health, emotional state, or the surrounding environment. The primary objective of this paper is to compare contemporary deep learning models with traditional digital signal processing methods used for speaker recognition. For this study, text-dependent methods (Mel-Frequency Cepstral Coefficients — MFCC, Linear Predictive Coding — LPC) and text-independent methods (ECAPA-TDNN — Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network, ResNet - Residual Neural Network) were selected to compare their effectiveness in voice biometric authentication tasks. The experiment involved implementing biometric authentication systems based on each of the described methods and evaluating their performance on a specially collected dataset. Additionally, the paper

provides a detailed examination of audio signal preprocessing methods used in voice authentication systems to ensure optimal performance in speaker recognition tasks, including noise reduction using spectral subtraction, energy normalization, enhancement filtering, framing, and windowing.

Keywords: biometric technologies; voice authentication; digital signal processing; mel-frequency cepstral coefficients; linear predictive coding; deep learning; neural networks**.**

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Samuel, F. A., Titilayo, A. O., Abiodun, A. O., Modupe, A. O., Oyeladun, M. B., Mayowa, I. R., & Samuel, A. M. (2021). Voice recognition system for door access control using mobile phone. *International Journal of Science and Engineering Applications, 10(9),* 132–139. https://doi.org/10.7753/ijsea1009.1004
2. Amjad Hassan Khan, M. K., & Aithal, P. S. (2022). Voice Biometric Systems for User Identification and Authentication – A Literature Review. *International Journal of AppliedEngineering and Management Letters (IJAEML), 6(1),* 198–209. https://doi.org/10.5281/zenodo.6471040
3. Abe, B. C., Araromi, H. O., Shokenu, E. S., Idowu, P. O., Babatunde, J. D., Adeagbo, M. A., & Oluwole, I. H. (2022). Biometric Access Control Using Voice and Fingerprint. *Engineering And Technology Journal, 7(7),* 1376–1382. https://doi.org/10.47191/etj/v7i7.08
4. Chen, X., Li, Z., Setlur, S., & Xu, W. (2022). Exploring racial and gender disparities in voice biometrics. *Scientific Reports, 12(1).* https://doi.org/10.1038/s41598-022-06673-y
5. Inamdar, F. M., Ambesange, S., Mane, R., Hussain, H., Wagh, S., & Lakhe, P. (2023). Voice Cloning Using Artificial Intelligence and Machine Learning: A review. *Journal of Advanced Zoology, 44(S7),* 419–427. https://doi.org/10.17762/jaz.v44is7.2721
6. Dalvi, J., et al. (2022). *A survey on face recognition systems*. arXiv preprint.
7. Win, K., Li, K., Chen, J., Viger, P. (2020). Fingerprint classification and identification algorithms for criminal investigation: A survey. *Future Generation Computer Systems, 110,* 758–771. https://doi.org/10.1016/j.future.2019.10.019
8. Daugman, J. (2002). How iris recognition works. *Proceedings International Conference on Image Processing.* https://doi.org/10.1109/ICIP.2002.1037952
9. Poddar, A., Sahidullah, Md., Saha, G. (2017). Speaker verification with short utterances: a review of challenges, trends and opportunities. *IET Biometrics. 7(2),* 91–101. https://doi.org/10.1049/iet-bmt.2017.0065. ISSN 2047-4938
10. Childers, D. G., Hand, M., Larar-Silent, M. J. (1989). Voiced/Unvoiced/Mixed Excitation (Four Way), Classification of Speech. *IEEE Trans. On ASSP, 37(11).*
11. Upadhyay, N., & Karmakar, A. (2015). Speech Enhancement using Spectral Subtraction-type Algorithms: A Comparison and Simulation Study. *Procedia Computer Science, 54,* 574–584. https://doi.org/10.1016/j.procs.2015.06.066
12. Jakovljević, N., Janev, M., Pekar, D., & Mišković, D. (2008). Energy Normalization in Automatic Speech Recognition. In *Lecture Notes in Computer Science,* 341–347. https://doi.org/10.1007/978-3-540-87391-4_44
13. Hviyuzova, D., & Belitskiy, A. (2021). Development of a filter amplifier of the signal pre-processing device for the passive listening mode of the hydroacoustic complex (HAC). *E3S Web of Conferences, 266,* 04013. https://doi.org/10.1051/e3sconf/202126604013
14. Introduction to Speech Processing. (n. d.). https://speechprocessingbook.aalto.fi/Representations/Windowing.html
15. Junqua, J.-C., Mak, B., Reaves, B. (1994). A robust algorithm for word boundary detection in presence of noise. *IEEE Trans. on Speech and Audio Processing, 2,* 406– 412.
16. Junqua, J.-C., Mak, B., Reaves, B. (1994). A robust algorithm for word boundary detection in presence of noise. *IEEE Trans. on Speech and Audio Processing, 2,* 406–412.
17. Liu, Y., Qian, Y., Chen, N., Fu, T., Zhang, Y., & Yu, K. (2015). Deep feature for text-dependent speaker verification. *Speech Communication, 73,* 1–13. https://doi.org/10.1016/j.specom.2015.07.003
18. Heigold, G., Moreno, I., Bengio, S., & Shazeer, N. (2016). *End-to-end text-dependent speaker verification.* https://doi.org/10.1109/icassp.2016.7472652
19. Xu, M., Duan, L. Y., Cai, J., Chia, L. T., Xu, C., & Tian, Q. (2004). HMM-Based Audio Keyword Generation. In *Lecture Notes in Computer Science,* 566–574. https://doi.org/10.1007/978-3-540-30543-9_71

20. Wijoyo, S. (2011). *Speech Recognition Using Linear Predictive Coding and Artificial Neural Network for Controlling Movement of Mobile Robot.* http://fportfolio.petra.ac.id/user_files/97-031/E091%20full%20paper-Thiang%20-%20ICIEE%202011.pdf
21. Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). *ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification*. https://doi.org/10.21437/interspeech.2020-2650
22. Jakubec, M., Lieskovska, E., & Jarina, R. (2021). Speaker Recognition with ResNet and VGG Networks, *31st International Conference Radioelektronika (RADIOELEKTRONIKA)*, 1–5. https://doi.org/10.1109/RADIOELEKTRONIKA52220.2021.9420202