Міністерство освіти і науки України

Харківський національний університет радіоелектроніки

Факультет комп'ютерної інженерії та управління
(повна назва)

Кафедра електронних обчислювальних машин
(повна назва)

# КВАЛІФІКАЦІЙНА РОБОТА

## Пояснювальна записка

Рівень вищої освіти другий (магістерський)

Підвищення точності голосової автентифікації користувачів за рахунок використання мел-кепстральних коефіцієнтів
(тема)

Виконав:

студент II курсу, групи СПм-22-3

Риндик І.В.
(прізвище, ініціали)

Спеціальність

123 «Комп'ютерна інженерія»
(код і повна назва спеціальності)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма

Системне програмування
(повна назва освітньої програми)

Керівник: доц. Барковська О.Ю.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри ЕОМ _____ Коваленко А.А.
(підпис) (прізвище, ініціали)

2024 р.

Харківський національний університет радіоелектроніки

Факультет комп'ютерної інженерії та управління

Кафедра електронних обчислювальних машин

Рівень вищої освіти другий (магістерський)

Спеціальність 123 «Комп'ютерна інженерія»
(код і повна назва)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Системне програмування
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
"____" _____ 20__ р.

# ЗАВДАННЯ

## НА КВАЛІФІКАЦІЙНУ РОБОТУ

студенту Риндику Ігорю Вікторовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Підвищення точності голосової автентифікації користувачів за рахунок використання мел-кепстральних коефіцієнтів

затверджена наказом по університету від "01" квітня 2024 р. № 257Ст

2. Термін подання студентом роботи до екзаменаційної комісії 15 червня 2024 р.

3. Вхідні дані до роботи
датасет із записами голосву мовця, обчислювач на базі ЦПУ, методи попередньої обробки аудіофайлу, нейромережеві методи аналізу аудіофайлу

4. Перелік питань, що потрібно опрацювати у роботі
Огляд характеристик систем голосової ідентифікації
Розбір методів препроцесінгу
Створення функціональної моделі системи
Розробка методології проведення досліджень
Проведення експериментів

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (слайдів) 14

6. Консультанти розділів роботи (заповнюється за наявності консультантів згідно з наказом, зазначеним у п.1 )

| Найменування розділу | Консультант (посада, прізвище, ім'я, по батькові) | Позначка консультанта про виконання розділу | |
|---|---|---|---|
| | | підпис | дата |
| | | | |
| | | | |

## КАЛЕНДАРНИЙ ПЛАН

| № | Назва етапів роботи | Термін виконання етапів роботи | Примітка |
|---|---|---|---|
| 1 | Огляд характеристик систем біометричної | 02.04.24-08.04.24 | |
| | ідентифікації особистості | | |
| 2 | Розбір послідовності методів препроцесінгу | 09.04.24-16.04.24 | |
| 3 | Створення функціональної моделі системи | 17.04.24-22.04.24 | |
| 4 | Розробка методології проведення досліджень | 23.04.24-06.05.24 | |
| 5 | Проведення експериментів | 07.05.24-23.05.24 | |
| 6 | Оформлення матеріалів кваліфікаційної роботи | 24.05.24-03.06.24 | |
| 7 | Подання кваліфікаційної роботи керівникові та | 04.06.24-07.06.24 | |
| | її попередній захист | | |
| 8 | Подання кваліфікаційної роботи на рецензу- | 08.06.24-12.06.24 | |
| | вання | | |
| | | | |

Дата видачі завдання 01 квітня 2024 р.

Студент _____ (підпис)

Керівник роботи _____ (підпис) доц. Барковська О.Ю. (посада, прізвище, ініціали)

# РЕФЕРАТ

Пояснювальна записка кваліфікаційної роботи: 76 с., 30 рис., 14 табл., 1 дод., 34 джерел.

РОЗПІЗНАВАННЯ, КОМП'ЮТЕРНА СИСТЕМА, ПОПЕРЕДНЯ ОБРОБКА, НЕЙРОННА МЕРЕЖА, ТОЧНІСТЬ, МЕТОД, МОДЕЛЬ.

Метою роботи є підвищення точності голосової автентифікації мовця завдяки застосуванню двофакторної аутентифікації - ідентифікація мовця на основі спектрограми голосу та верифікації голосового пароля на основі сумісного використання двох нейромережевих моделей із згортковою архітектурою.

На основі запропонованої системи проведено ряд досліджень, а саме – проаналізовано та експериментально підтверджена ефективність одного із двох методів обробки аудіосигналу для виділення важливих характеристик (спектрограми та мел-кепстральні коефіцієнти), досліджена архітектура нейромережевих аналізаторів із згортковою архітектурою, вплив розподілу вхідного датасету на навчальну, тестову та валідаційну вибірки, проаналізована тривалість навчання та вимоги до апаратного забезпечення для забезпечення стійкості роботи системи.

Загальна оцінка результатів показує, що комбінуючи два запропонованих фактори, система підвищує рівень безпеки, оскільки для успішної автентифікації потрібно відповідати обом критеріям.

# ABSTRACT

Master’s thesis: 76 pages, 30 figures, 14 tables, 1 appendices, 34 sources.

RECOGNITION, COMPUTER SYSTEM, PRE-PROCESSING, NEURAL NETWORK, ACCURACY, METHOD, MODEL.

The purpose of the qualification work is to improve the accuracy of speaker voice authentication through the application of two-factor authentication - identifying the speaker based on the voice spectrogram and verifying the voice password by jointly using two convolutional neural network models.

A series of studies have been conducted based on the proposed system, namely: the effectiveness of one of the two audio signal processing methods for extracting important characteristics (spectrograms and mel-frequency cepstral coefficients) has been analyzed and experimentally confirmed, the architecture of convolutional neural network analyzers has been investigated, the impact of the distribution of the input dataset on training, testing, and validation sets has been studied, and the training duration and hardware requirements for ensuring system stability have been analyzed.

Overall assessment of the results shows that by combining the two proposed factors, the system increases the security level, as both criteria need to be met for successful authentication.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ ..... 8
ВСТУП ..... 9
1 ОГЛЯД ПРОБЛЕМНОЇ ОБЛАСТІ. ПОСТАНОВКА ЗАДАЧІ ..... 11
1.1 Обгрунтування актуальності обраної теми ..... 12
1.2 Особливості та відмінні риси ідентифікації за голосом ..... 17
1.3 Еволюція системи розпізнавання голосу ..... 21
1.4 Мета за задачі дослідження ..... 23
2 АНАЛІЗ АПАРАТНО-ТЕХНОЛОГІЧНОЇ БАЗИ ДЛЯ ВИКОНАННЯ ДОСЛІДЖЕННЯ ..... 24
2.1 Огляд методів вилучення ознак в аудіоряді. Етапи вилучення ознак в MFCC ..... 24
2.1.1 Лінійні прогнозні коефіцієнти (LPC) ..... 25
2.1.2 Лінійні прогнозні кепстральні коефіцієнти (LPCC) ..... 26
2.1.3 Лінійні коефіцієнти сприйняття (PLP) ..... 27
2.1.4 Кепстральні коефіцієнти Mel Frequency (MFCC) ..... 28
2.1.5 Відносна спектральна фільтрація коефіцієнтів логарифмічної області (RASTA) ..... 32
2.2 Методологічна база для проведення аналізу виділених ознак в аудіоряді ..... 37
3 АНАЛІЗ АПАРАТНО-ТЕХНОЛОГІЧНОЇ БАЗИ ДЛЯ ВИКОНАННЯ ДОСЛІДЖЕННЯ ..... 42
3.1 Експеримент 1. Порівняння точності ідентифікації спікера за голосом на основі мел-кепстральних коефіцієнтів та спектрограм ..... 48
3.1.2 Дослідження часу навчання та точності класифікації нейромережевих моделей у залежності від типу визначення характерних ознак ..... 50

3.1.3 Аналіз співвідношення навчальної, тестової та валідаційної вибірок у робочому датасеті ..... 53
3.1.4 Дослідження впливу частоти дискретизації на точність розпізнавання спікера ..... 56
3.1.5 Визначення архітектури нейронної мережі для досягнення найвищої точності розпізнавання спікера за голосом ..... 57
3.2 Експеримент 2. Розпізнавання цифрового паролю, названого голосом мовця ..... 61
ВИСНОВКИ ..... 64
ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ ..... 66
ДОДАТОК А Графічний матеріал кваліфікаційної роботи ..... 70

# ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

ANN – штучна нейронна мережа (англ, Artificial Neural Network)

ASR – розпізнавання мови (англ, Automatic Speech Recognition)

CNN – згорткова нейронна мережа (англ, Convolutional Neural Network)

HMM – прихована марковська модель (англ, Hidden Markov Model)

MFA – багатофакторна автентифікація (англ, Multi-Factor Authentication)

MFCC – мел-кепстральні коефіцієнти (англ, Mel-frequency Cepstral Coefficients)

MLP – багатошаровий перцептрон (англ, Multilayer Perceptron)

RNN – рекурентна нейронна мережа (англ, Recurrent Neural Network)

SVM – машина опорних векторів (англ, Support Vector Machine)

# ВСТУП

Взаємодія, яка відбувається між людиною та комп'ютером, є звичайною формою комунікації в сучасному світі. Цей вид взаємодії може бути здійснений лише за допомогою апаратних пристроїв, таких як клавіатура, сенсорний екран, миша тощо, але люди віддають перевагу більш природній формі спілкування, аніж використанню апаратних пристроїв для комунікації. Мова є найпоширенішим і значущим засобом комунікації, яким користуються люди. Натхненням для створення систем розпізнавання мови є взаємодія між людьми, причому голос є найважливішою ознакою цієї взаємодії. Розпізнавання мовлення може надати можливість функції диктування, за допомогою якої можна керувати пристроями і створювати документи. Створення документів за допомогою мови заощаджує час і зусилля людини. Бізнес-організації отримали значну користь від цієї галузі. Існує голосова технологія, за допомогою якої абоненти можуть вводити свої дані та отримувати відповідь пізніше, не спілкуючись фактично з живим оператором. Це призвело до зменшення витрат та оптимізації в сфері бізнесу, усуваючи витрати, які б витрачалися на збереження оператора. Невербальна комунікація між людиною та комп'ютером вже існувала, але використання голосу для цього призвело до розвитку галузі розпізнавання мовлення. Голосовий зразок містить багато інформації, починаючи від статі та віку особи, іноді навіть відображаючи стан душі спікера. Метою голосового розпізнавання є використання частини цієї інформації і ідентифікація спікера за однією або кількома з цих параметрів. Існують багато методів, які можуть бути використані для виділення ознак з голосового зразка, які можуть відноситися до спектрального чи кепстрального доменів.

Окрім голосового розпізнавання, розповсюдженою задачею аналізу голосу є біометрична автентифікація користувачів.

Голосова автентифікація використовує біометричні дані, які унікальні для кожної людини. Порівняно з паролями або PIN-кодами, які можна вкрасти, підглянути або відгадати, голосові дані складніше підробити. Це робить голосову автентифікацію ефективним засобом захисту конфіденційної інформації та особистих даних.

Використання голосових команд для доступу до систем або ресурсів є зручним для користувачів. Немає необхідності запам'ятовувати складні паролі або носити з собою додаткові пристрої, такі як токени або картки доступу. Просто вимовте фразу або слово, і система автоматично вас ідентифікує.

Голосова автентифікація може бути використана як один з елементів мультифакторної аутентифікації разом з іншими методами, такими як відбитки пальців, обличчя або PIN-коди. Це збільшує рівень безпеки, оскільки злочинцю потрібно буде подолати декілька бар'єрів для незаконного доступу.

Окрім того, завдяки постійному розвитку технологій штучного інтелекту, системи голосової автентифікації стають все точнішими та надійнішими. Вони можуть розпізнавати різні мови, акценти, інтонації та інші характеристики голосу, що робить їх досить універсальними для застосування у різних країнах та серед різних користувачів.

Сфера застосування методів голосової автентифікації є широкою. Це і фінансові установи, медичні системи, системи клієнтського обслуговування, смартфони та інші пристрої. За останні роки популярність цієї технології значно зросла, оскільки компанії та організації визнали її ефективність та переваги.

Отже, голосова автентифікація є актуальною та затребуваною технологією завдяки своїм безпечним і зручним властивостям, а також завдяки постійному розвитку технологій, що поліпшують її точність та надійність.

# 1 ОГЛЯД ПРОБЛЕМНОЇ ОБЛАСТІ. ПОСТАНОВКА ЗАДАЧІ

За сучасних умов розвитку суспільства проблема безпеки постає у новому аспекті – значна кількість об’єктів, яким потрібно забезпечити безпеку, надана у вигляді інформації, яка зберігається в електронних комп’ютерних системах та передається через мережі зв’язку. Тобто, виник новий аспект безпеки – захист інформації. Причому в даному випадку необхідно забезпечити декілька рівнів захисту – обмежити фізичний доступ до електронних комп’ютерних систем (серверів), де зберігається інформація, забезпечити доступ до роботи з інформацією тільки акредитованим особам, забезпечити контроль фізичного доступу до приміщень, де знаходяться сервери і т. ін. Проблема безпеки тісно зв’язана із обробкою певних ідентифікаторів, серед яких є і біометричні данні. Така обробка вижлива в медицині, освітньому процесі тощо [1]. Існує цілий комплекс заходів із забезпечення захисту інформації, проте вони не є на сто відсотків ефективними [2-3].

Для підвищення ефективності систем захисту останнім часом пропонується використовувати так звані біометричні системи ідентифікації. Біометричні системи ідентифікації встановлюють особу за індивідуальними біометричними параметрами людини.

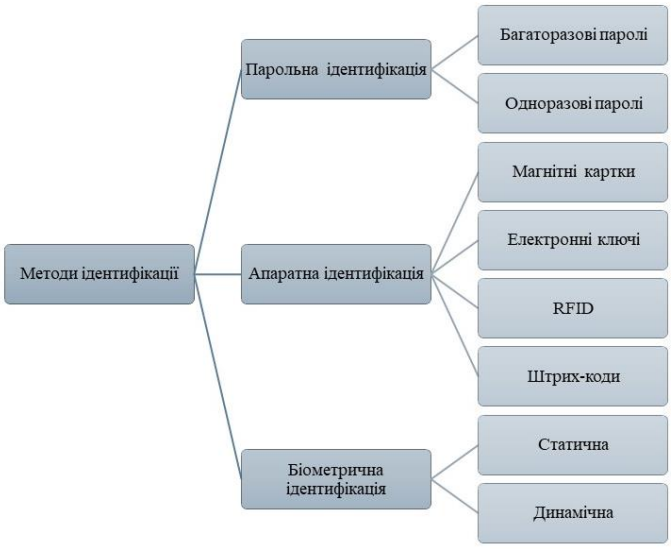



Рисунок 1.1 – Класифікація методів ідентифікації

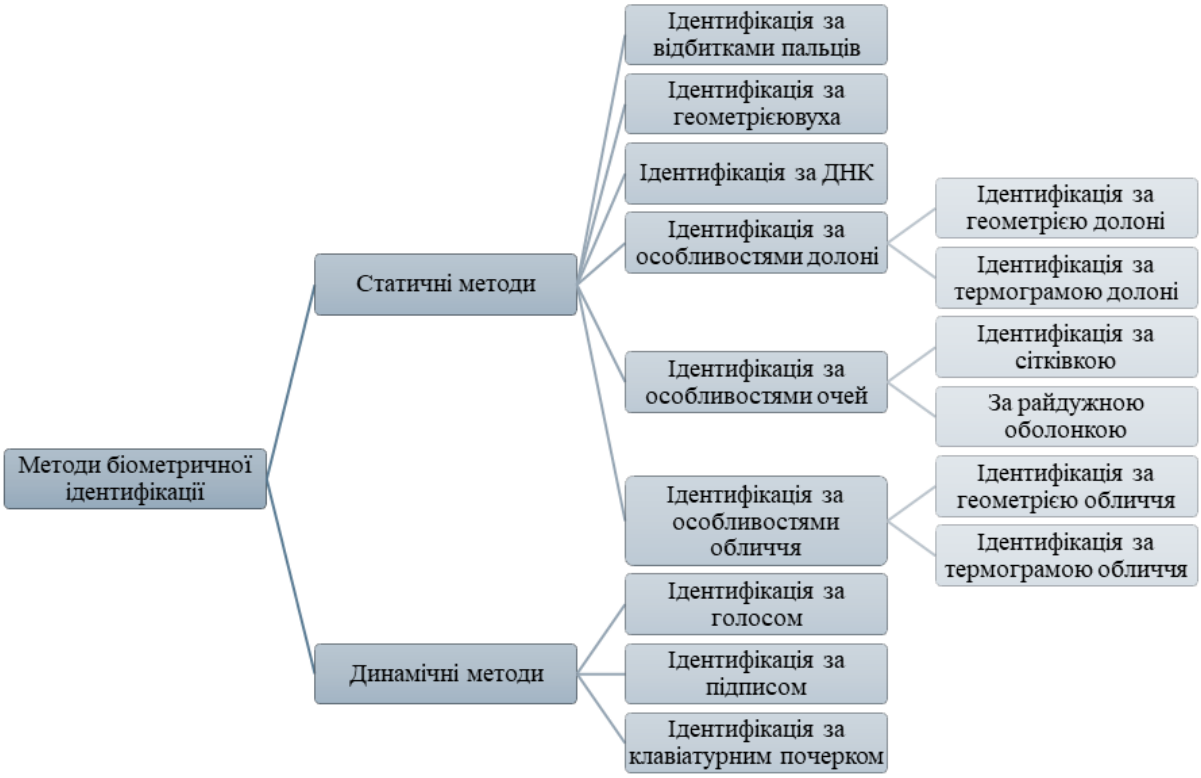



Рисунок 1.2 - Класифікація методів біометричної ідентифікації

Беззаперечною перевагою біометричної ідентифікації є те, що даний засіб неможливо забути чи загубити [4]. Біометричні ідентифікатори людини можна поділити на дві групи – статичні (також називаються фізіологічними) та динамічні (також називаються психологічними):

- фізіологічні (статичні) – засновані на фізіологічній (статичній) характеристиці людини, тобто унікальних властивостях, які властиві їй від народження і є невід'ємними від неї;
- психологічні (динамічні) – засновані на поведінковій (динамічній) характеристиці людини, особливостях, характерних для підсвідомих рухів у процесі відтворення будь-якої дії.

## 1.1 Обгрунтування актуальності обраної теми

У випадку голосової ідентифікації використовуються частотні характеристики голосу, такі як тон, тембр – саме за ними створюється

цифрова модель. Голос формується з комбінації фізіологічних і поведінкових факторів. Привабливість даного методу – зручність у використанні. Основним занепокоєнням, пов'язаним з цим біометричним підходом, є точність ідентифікації. В даний час, ідентифікація за голосом використовується для управління доступом у приміщення середнього засобу безпеки, наприклад, лабораторії і комп'ютерні класи. Ідентифікація за голосом зручна, але, в той час, не така надійна як інші біометричні методи. Наприклад, людина з простудою або ларингітом може зазнавати труднощі при використанні даних систем. Статистика розповсюдження засобів біометричної ідентифікації на глобальному ринку видяляє лідируючі характеристики: відбиток пальця, характеристики обличчя, райдужна оболонка ока, голос (рисунок 1.3).

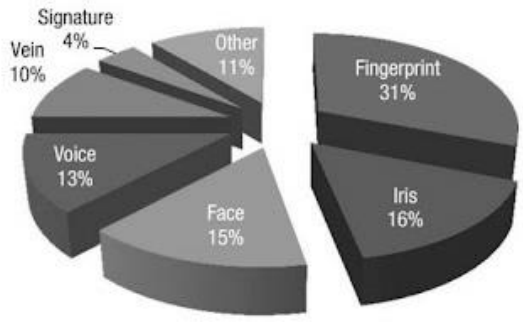



Рисунок 1.3 - Частка ринку біометричних технологій

Слід зазначити, що психологічні методи є менш надійними ніж фізіологічні.

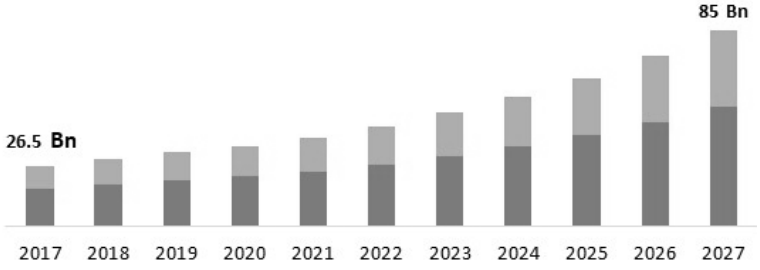



Рисунок 1.4 - Розмір ринку біометричних технологій за типом (2017-2027)

(джерело статистики – *www.kbvresearch.com*)

Основним рушієм ринку біометричних технологій (рисунок 1.4) є постійно зростаюче впровадження біометричних рішень у приватному та державному секторах. Широке впровадження вдосконалених систем безпеки в організаціях і використання електронних паспортів і електронних віз є іншими ключовими причинами, які спонукають до швидкого зростання ринку [5-7]. Швидко прогресуюча біометрична технологія, впровадження цих технологій на сайти електронної комерції та додатки, а також хмарні обчислення проклали шлях для основних гравців, які працюють на ринку біометричних технологій. Ряд досліджень стверджують, що до кінця 2024 року розмір глобального ринку біометричних технологій, ймовірно, досягне оцінки в 54,7 мільярда доларів США (Biometric Technology Market, published date: Jan. 2024). Біометрична ідентифікація увійшла в цсі сфери людського життя, наприклад (рисунок 1.5):

- банківський і фінансовий сегмент;
- урядовий сектор;
- споживча електроніка;
- охорона здоров’я;
- сегмент захисту та безпеки;
- інші кінцеві користувачі, включаючи такі сектори, як освіта, готельний бізнес і розваги.

У 2023 році банківський і фінансовий сегмент займав домінуючу позицію на ринку біометричних технологій, захоплюючи понад 25% частки. Лідируюча позиція цього сектору пояснюється дедалі більшим впровадженням біометричних систем для підвищення безпеки та досвіду клієнтів у банківських операціях. Біометричні технології, такі як відбитки пальців і розпізнавання обличчя, широко використовуються для автентифікації транзакцій і доступу до банківських рахунків, пропонуючи поєднання зручності та безпеки для клієнтів.

Рисунок 1.5 - Світовий ринок біометричних технологій у відношенні кінцевих споживачів у 2023 (%)

Урядовий сектор, інший значний кінцевий користувач, використовує біометричні технології для керування ідентифікацією, прикордонного контролю та реєстрації виборців, керуючись необхідністю ефективного та безпечного державного управління. Інвестиції цього сектора в біометричні системи є значними, спрямованими на посилення національної безпеки та надання державних послуг.

Споживча електроніка — сегмент, що швидко зростає, біометрія інтегрована в смартфони, ноутбуки та інші персональні пристрої для безпечного доступу та автентифікації користувачів. Ця тенденція викликана попитом на захист персональних даних і простотою використання таких технологій.

Охорона здоров’я — це новий сегмент застосування біометричних технологій, де вони використовуються для ідентифікації пацієнтів, контролю доступу до медичних записів і забезпечення конфіденційності інформації про здоров’я. Сектор транспорту/логістики також виграє від біометрії, особливо в покращенні заходів безпеки та ефективності роботи в аеропортах, морських портах і обробці вантажів.

У сегменті захисту та безпеки використовуються біометричні дані для контролю доступу, спостереження та перевірки особи, що є критично важливим для національної безпеки. Удосконалені біометричні системи все частіше застосовуються у військових і правоохоронних органах.

Інші кінцеві користувачі, включаючи такі сектори, як освіта, готельний бізнес і розваги, поступово використовують біометричні технології для перевірки особи, відстеження відвідуваності та персоналізованих послуг. Ці різноманітні додатки підкреслюють зростаюче проникнення та універсальність біометричних технологій у різних галузях промисловості, що зумовлено потребою у покращених рішеннях безпеки та автентифікації користувачів.

Останні дослідження в області біометричної ідентифікації проводяться наступними компаніями: NEC Corporation, HID Global Corporation, Thales Group, Gemalto (A Thales Company), Morpho (Safran Identity & Security), Fujitsu Limited, Crossmatch Technologies, Idemia (formerly OT-Morpho), Suprema Inc., Aware, Inc., Bio-key International, Inc., SecuGen Corporation.

У 2023 році корпорація NEC представила NeoFace Watch X9 із найшвидшим у світі алгоритмом розпізнавання обличчя. Дослідження зосереджено на мультимодальній біометрії, інтеграції відбитків пальців, райдужної оболонки ока та розпізнавання голосу для підвищення безпеки.

У 2023 році HID Global Corporation: запустила «bioConnect», модульну платформу для інтеграції різних біометричних модальностей і розробила потужні рішення для фізичного та логічного контролю доступу з розширеною біометрією.

У 2023 році Morpho (Safran Identity & Security) представив безконтактний зчитувач відбитків пальців для безперебійного контролю доступу та фокус на біометричних рішеннях для безпечних мобільних платежів і транзакцій електронної комерції.

Все вищезазаначене підкреслює актуальність та активний розвиток обраної для дослідження теми.

## 1.2 Особливості та відмінні риси ідентифікації за голосом

Біометрія залишається незамінною там, де необхідно забезпечити безпеку доступу до фізичних об'єктів та інформаційних ресурсів. Біометричні технології успішно застосовуються у правоохоронній діяльності, громадянській реєстрації, у сфері безпеки банківських звернень, інвестування, у питаннях охорони здоров'я та у багатьох сферах діяльності. Найчастіше використовуваними біометричними характеристиками є відбитки пальців, форма обличчя, райдужна оболонка ока, голос, підпис, геометрія руки. Ми не можемо сказати, що та чи інша характеристика є кращою за інших. Щоб вибрати відповідний біометричний метод ідентифікації, потрібно враховувати такі фактори, як область його застосування, необхідний рівень безпеки, цільову установку (верифікація або ідентифікація), очікуване число користувачів, практичність та інші. Біометрія - це розробка статистичних та математичних методів, що застосовуються до аналізу даних проблеми у біологічних науках. Впровадження цієї технології забезпечує нову безпеку підходів до комп'ютерних систем. Ідентифікація та перевірка - це два способи використання біометричних даних для ідентифікації особистості. Біометрія відноситься до використання фізичних чи фізіологічних, біологічних чи поведінкових характеристик для встановлення особистості людини. Ці характеристики є унікальними для кожної людини і залишаються частково незмінними протягом життя .

Будь-який фізіологічний або поведінковий атрибут може бути кваліфікований як біометрична ознака та має задовольняти критеріям, наведеним у Таблиці 1.1.

Мова – це універсальна форма спілкування. Розпізнавання за голосом – це процес ідентифікації того, хто говорить, за голосовими особливостями цієї мови [8]. Це відрізняється від розпізнавання мови, коли процес ідентифікації обмежений змістом, а не акустичними характеристиками мовця. Процес розпізнавання за голосом заснований на виявленні та виділення унікальних

характеристик мовлення. Характеристики голосу людини також називають голосовою біометрією.

Таблиця 1.1 – Критерії фізіологічних ознак

| | |
|---|---|
| Універсальність | Ознака властива всім людям. |
| Інваріантність | Обрана біометрична ознака має демонструвати незмінність у часі. |
| Доступність | Дані легко збирати, оцифровувати та виділяти з популяції. |
| Продуктивність | Відноситься до наявності ресурсів та обмежень у плані збору даних та гарантії досягнення високої точності. |
| Прийнятність | Готовність населення надати доступ до ознаки системі розпізнавання та обходу |
| Схильність до підробки | Може бути розглянуто як наслідок шахрайської атаки |

Таким чином, ідентифікація за голосом відома як процес ідентифікації того, хто говорить шляхом порівняння голосової біометрії даного мовця. Коли голос використовується для авторизації, то результатом має бути підтвердження або не підтвердження ідентифікованої особи та надання дозволу чи відмова у доступі до системи чи послуги. Після ідентифікації система перевіряє, чи має ця особа право на доступ до певного ресурсу чи функції. Голосова авторизація відбувається після ідентифікації та вимагає подальшого підтвердження користувача, наприклад, шляхом повторного проголошення певної фрази або виконання іншої дії.

Отже, голосова ідентифікація визначає особу за допомогою голосу, тоді як голосова авторизація підтверджує право цієї особи на доступ до системи чи послуги після успішної ідентифікації.

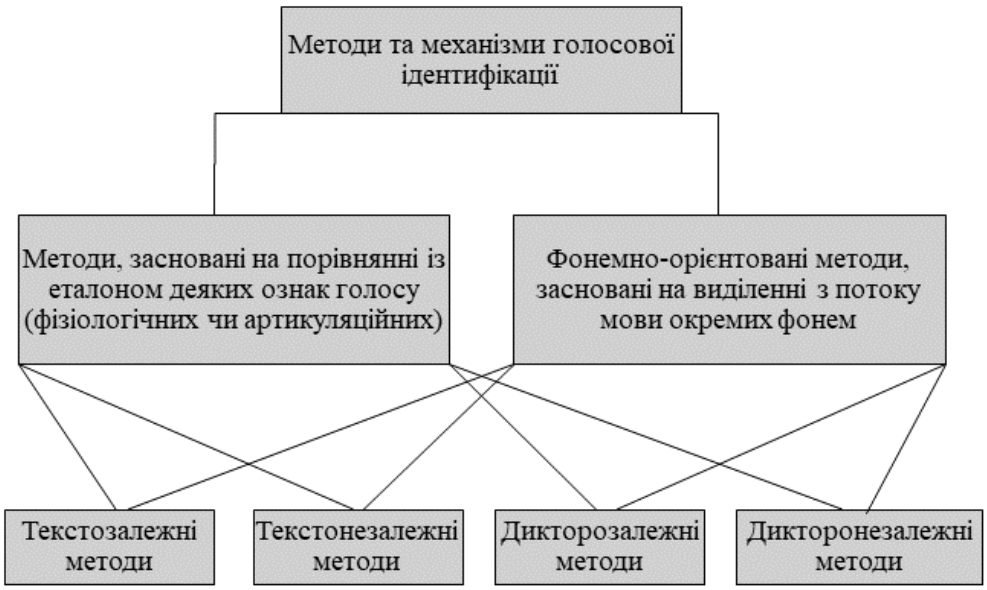



Рисунок 1.6 – Методи та механізми голосової ідентифікації

Сьогодні існує два головних методи створення систем голосової ідентифікації.

1 Еталонний метод засновано на порівнянні деяких ознак голосу (це можуть бути як фізіологічні, так і артикуляційні ознаки) з деяким еталоном. Як еталон використовують деяку групу окремих слів.

2 Фонемно-орієнтований метод засновано на виділенні з потоку мови окремих фонем.

Кожний з цих двох методів може бути застосовано для створення систем голосової ідентифікації, які можна поділити на наступні класи:

- текстозалежні;
- текстонезалежні;
- дикторозалежні;
- дикторонезалежні.

Дикторозалежність або дикторонезалежність системи. За визначенням, дикторозалежна система призначена для використання одним користувачем, тоді як дикторонезалежна система призначена для роботи з будь-яким диктором. Дикторонезалежність – труднодосяжна мета, оскільки під час навчання системи, вона налаштовується на параметри того диктора, з

прикладу якого навчається. Частота помилок розпізнавання таких систем зазвичай у 3-5 разів більша, ніж частота помилок дикторозалежних систем.

Текстозалежні системи – це системи голосової ідентифікації, які здійснюють ідентифікацію особи за допомогою певного ключового слова або ключової фрази, яку повинна вимовити особа, яка проходить ідентифікацію.

Текстонезалежні системи – це системи, які здійснюють ідентифікацію особи за допомогою голосу без прив’язки до будь-яких ключових слів. У даному випадку важливе значення мають артикуляційні ознаки голосу людини, саме вони використовуються як головні ознаки, а фізіологічні ознаки виступають як вторинні.

Ключовим етапом функціонування систем голосової ідентифікації є вилучення ознак, оскільки точність роботи системи напряму залежить від якості виділення унікальних характерних ознак голосового введення. Тому огляду літератури щодо різних підходів до вилучення ознак ідентифікації за голосом присвячено багато уваги в даній роботі (Розділ 2).

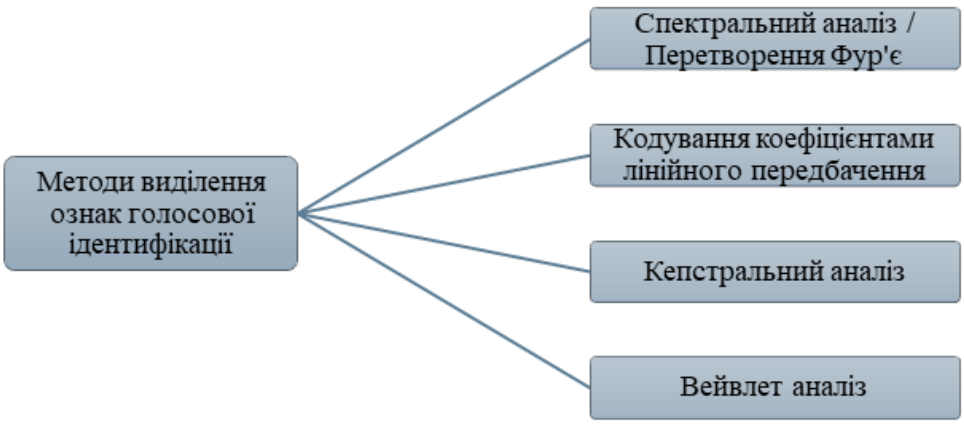



Рисунок 1.7 – Класифікація систем ідентифікації по голосу за методами виділення ознак

Існують методи виділення ознак, які допомагають вилучити деякі ознаки із зразка мовлення. Одним з таких методів є Мел-частотний кепстральний коефіцієнт (Mel-Frequency Cepstral Coefficient). Це найбільш

широко використовуваний метод для розпізнавання спікера через його простоту в реалізації та гнучкість [9].

## 1.3 Еволюція системи розпізнавання голосу

Початковою метою розпізнавання мовлення була розробка системи, яка імітує здатність людини до мовлення [10].

1773 р. – російський вчений Крістіан Краценштейн, професор фізіології в Копенгагені, створив голосний звук за допомогою резонансних трубок, які були приєднані до труб органу.

1791 – Вольфганг фон Кемпелен у Відні розробив акустико-механічний мовний апарат.

Середина 1800-х років – Чарльз Вітстон сконструював розмовну машину фон Кемпелена, використовуючи шкіряні резонатори, щоб можна було створювати різні звуки, схожі на мову, змінюючи конфігурації рукою.

1879 – Томас Едісон відкрив першу машину для диктування.

1930-ті роки – Гомер Дадлі розробив синтезатор мови під назвою Voice Operating Demonstrator (VODER), який був електричним прототипом роботи Вітстона.

1952 – Девіс, Біддулф і Балашек з лабораторій Белла розробили Одрі систему розпізнавання окремих цифр для конкретного розмовника.

1950-ті роки – Ослон і Белар з лабораторій RCA сконструювали розпізнавач десяти (10) складів для конкретного мовця.

1959 – Фрай і Денс з Університетського коледжу в Англії сконструювали розпізнавач фонем, який ідентифікує чотири (4) голосних і дев'ять (9) приголосних.

1960-ті роки – Атал та Ітакура винайшли базову теорію лінійного прогнозуючого кодування (LPD), яка скоротила наближення відповіді голосового тракту за формою мовного сигналу.

1962 – була розроблена коробка для взуття IBM, яка може розуміти 16

англійських слів.

1971 – Олександр Вайбель, розробив машину Гарпія в Університеті Карнегі-Меллона. Ця машина може розуміти 1011 слів з деякими фразами.

1970-ті роки – Том Мартінс заснував провідну компанію з розпізнавання мовлення під назвою Threshold Technology incorporation, створивши перший продукт автоматичного розпізнавання мовлення (ASR), відомий як система VIP-100. Крім того, Ленні Баум з Прінстонського університету виявив математичний розпізнавач мови під назвою «Прихована модель Маркова» (HMM). Крім того, було запроваджено використання технології розпізнавання образів для системи ідентифікації мовлення за допомогою LPC.

1980-ті – використання штучної нейронної мережі було запроваджено для мовлення; розпізнавання мови має тенденцію до передбачення.

1986 – IBM Targora була розроблена з використанням HMM для передбачення наступних фонем у мовленні. Ця система була найшвидшою в світі друкаркою на той час, оскільки вона могла розпізнавати до 20 000 англійських слів і ряд речень.

1990-ті роки – поява автоматичного розпізнавання мови (ASR).

1997 рік – провідний у світі розпізнавач безперервної мови було оголошено програмним забезпеченням Dragon для природної мови. Це програмне забезпечення могло розуміти 100 слів на хвилину без перерви між словами.

2006 – Агентство національної безпеки (NSA) почало використовувати розпізнавання мовлення для визначення ключових термінів у записаному мовленні.

2008 – Google представляє додаток голосового дослідження на мобільних пристроях.

2011 – Apple запускає Siri для голосового цифрового помічника.

## 1.4 Мета за задачі дослідження

Метою роботи є підвищення точності голосової автентифікації мовця завдяки застосуванню двофакторної аутентифікації - ідентифікація мовця на основі спектрограми голосу та верифікації голосового пароля на основі сумісного використання двох нейромережевих моделей із згортковою архітектурою.

Для досягнення поставленої мети мають бути вирішені наступні задачі:

- розробити модель системи двофакторної ідентифікації мовця на основі текстозалежних та текстонезалежних підходів;
- виконати аналіз ефективності методів обробки аудіосигналу для виділення ключових характеристик аудіосигналу (спектрограми та мел-кепстральні коефіцієнти);
- дослідити нейромережевих аналізаторів із згортковою архітектурою, дослідити вплив розподілу вхідного датасету на навчальну, тестову та валідаційну вибірки;
- проаналізувати вимоги до апаратного забезпечення під час навчання нейромережевих класифікаторів для забезпечення стійкості роботи системи.

Подальші дослідження будуть зосереждені на розширенні бази даних, включаючи мовців з різними вадами мови або незвичайними мовними патернами, інтеграції з іншими методами автентифікації (наприклад, розпізнавання обличчя або відбитки пальців, для створення більш надійної багатофакторної системи автентифікації), дослідженні методів адаптації до зміни динаміків, розробці прототипів систем голосової автентифікації для конкретних застосувань.

# 2 АНАЛІЗ АПАРАТНО-ТЕХНОЛОГІЧНОЇ БАЗИ ДЛЯ ВИКОНАННЯ ДОСЛІДЖЕННЯ

## 2.1 Огляд методів вилучення ознак в аудіоряді. Етапи вилучення ознак в MFCC

Системи розпізнавання голосу можуть бути залежні від тексту та незалежні від тексту. Є стандартний пайплайн, який дослідники переважно використовують у своїх дослідженнях (рисунок 2.1).

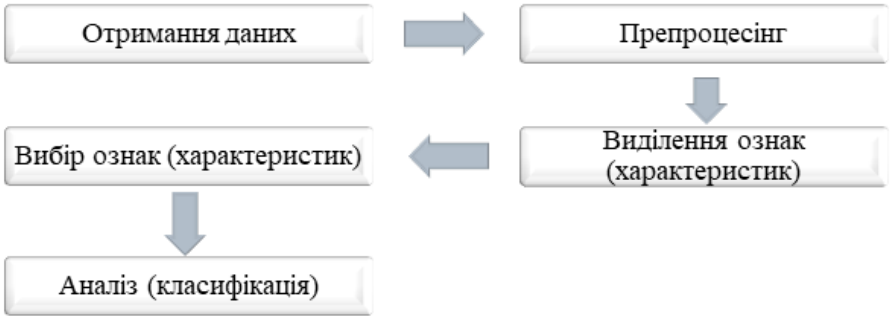



Рисунок 2.1 – Стандартний пайплайн для аналізу аудіоряду

Попередня обробка необхідна для видалення фонових шумів і обробки даних у формі, у якій їх можна використовувати для виділення ознак. Серед методів виділення ознак можна розглянути наступні: кодування з лінійним предиктором (LPC), перцептивне лінійне передбачення (PLP), відносна фільтрація спектрів (RASTA) і мел-частотний кепстральний коефіцієнт (MFCC). Переваги та недоліки цих методів дозволяють легко знайти компроміс для вибору методу вилучення ознак, який найкраще підходить для поточної ситуації та часу. Виділення ознак - це процес вилучення та роботи з прихованою інформацією в сигналі сирої даних [11]. Застосування виділення ознак дозволяє зробити дані керованими, оскільки вони вилучають всі неефективні ознаки з даних без втрати важливої або релевантної інформації.

Крім того, техніки виділення ознак допомагають розвивати систему з меншими зусиллями машини та збільшувати швидкість процесу навчання та узагальнення в машинному навчанні [12-13].

Наступним кроком є робота класифікатора. Різні дослідження показують, що класифікатори машинного навчання, такі як опорні векторні машини (SVM), K-найближчий сусід (K-NN), дерева рішень і приховані моделі Маркова (HMM), згорткові нейронні мережі (CNN), багатошаровий перцептрон (MLP) та рекурентні нейронні мережі (RNN) можуть бути використані в якості аналізатора (класифікатора).

### 2.1.1 Лінійні прогнозні коефіцієнти (LPC)

Linear Predictive Coefficients — це широко використовуваний кодер із низькою швидкістю потоку, який визначає спектр потужності цифрового мовного сигналу. Техніка LPC дуже залучена до формантного аналізу. У цій техніці, як альтернатива перетворенню всього цифрового мовного сигналу, різниця зразків генерується для прогнозування наступних голосових зразків і математично обробляється для виділення характеристики цифрового мовного сигналу. Цифровий мовний сигнал можна відновити шляхом об'єднання «n» кількості прогнозованих вибірок. Модель прогнозування проілюстрована наступним чином. Нехай x(n) і x(n-1) є поточним і попереднім зразками голосу, майбутній зразок можна передбачити за допомогою.

$$\mathrm{x(n)} = \sum_{i=1}^{n=\infty} pi\, x(n-i) \pm e(n)$$

,

де pi – коефіцієнт прогнозування. Помилка передбачення e(n) може бути обчислена за допомогою $\mathrm{e(n)} = x(n) - x(n)$

Помилки передбачення були очищені в послідовних вибірках, щоб

отримати точні прогнозовані вибірки, щоб уникнути шумових вибірок у функціях голосового друку. Процеси, пов’язані з виділенням ознак LPC, це попереднє виділення голосових кадрів, оцінка потужності.

### 2.1.2 Лінійні прогнозні кепстральні коефіцієнти (LPCC)

Техніка лінійних прогнозних кепстральних коефіцієнтів (LPCC) є версією LPC для подолання канальних ефектів шляхом виконання Cepstral Mean Subtraction (CMS). Cepstral — це послідовність спектральної щільності потужності, отримана з періодограми, яка використовується для відстеження висоти тону. Кепстр отримується шляхом виконання зворотного перетворення Фур’є для отриманого спектру потужності голосових сигналів і може класифікуватися як справжній кепстр, комплексний кепстр, фазовий кепстр і кепстр потужності, де кепстр потужності використовується в аналізі мовлення.

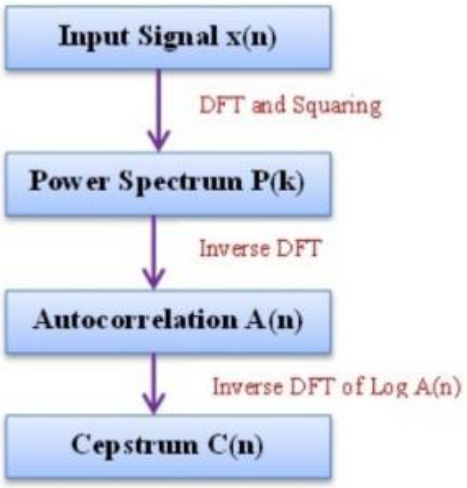



Рисунок 2.2 - Процес обчислення LPCC

На рисунку 2.2 нехай *x(n)* буде голосовим сигналом, що характеризується у часовій області, якщо його математично обробляти за допомогою дискретного перетворення Фур’є (DFT), а зведення в квадрат дає спектр потужності *P(k)* наступним чином.

$$\mathrm{P}(k) = |DFT(x(n))|^2$$

Обернене DFT *P(n)* не породило *x (n),* але генерує автокореляцію *A(n)* для мовного сигналу часової області *x(n).*

$$\mathrm{A}(n) = DFT^{-1}(P(k))$$

Кепстр *C(n)* отримується шляхом логарифмічного стиснення спектра потужності та обробки за допомогою інверсного DFT наступним чином.

$$C(n) = DFT^{-1}(\log P(k))$$

Обчислений кепстр є результатом логарифмічного спектра потужності *{log(P(k)}* замість отримання стандартного спектру потужності *{P(k)}* мовного сигналу *x(n).*

### 2.1.3 Лінійні коефіцієнти сприйняття (PLP)

Perceptual Linear Predictive (PLP) Coefficients — це імпровізована модель LPCC для зменшення шуму та зменшення невідповідностей тренувальних і тестових зразків голосу.

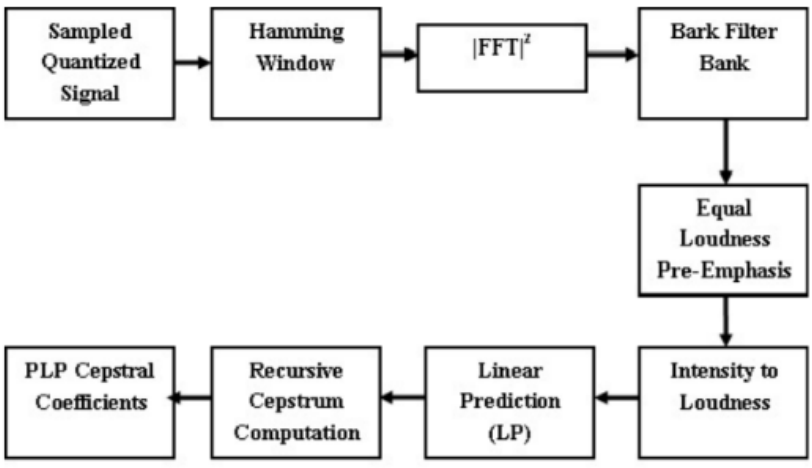



Рисунок 2.3 - Процес обчислення PLC Cepstral

Етапи, реалізовані в PLP, проілюстровані на рисунку 2.3, який включає (i) обчислення спектру потужності. (ii) Викривлення частоти. (iii) Згортка викривленого спектру зі спектром потужності для інтеграції критичного банку. (iv) Згладжування спектру. (v) Попереднє наголошення (vi) Лінійне передбачення та кепстральне вилучення. Інтеграція деформації частоти, згладжування та дискретизації дає банк частот кори. Квантований дискретизований сигнал зважується з використанням вікна Хеммінга з розміром вікна 20 мс, яке включає 256-точкове швидке перетворення Фур'є (ШПФ). ШПФ дискретизує 200 зразків мовного сигналу в 56 зразків із нульовим значенням.

$$\mathrm{P}(\omega) = Re[S(\omega)]^2 + Im[S(\omega)]^2.$$

Спектр потужності P(ω) спотворено для частоти гавкання **Ω** на

$$\Omega(\omega) = 61n\{\frac{\omega}{1200\pi} + \left[\left(\frac{\omega}{1200\pi}\right)\right]^2 + 1^{0.5}\}.$$

Виконується згортка викривленого спектра потужності та спектра потужності та отримана критична смуга. Коефіцієнти Cepstral для лінійного прогнозування були кращими на відміну від коефіцієнтів лінійного прогнозування, оскільки вони чутливі до числових помилок.

### 2.1.4 Кепстральні коефіцієнти Mel Frequency (MFCC)

Частотні кепстральні коефіцієнти Mel (MFCC), орієнтовані на лінійне косинусне перетворення логарифмічного спектра потужності. Частотні смуги в MFCC рівномірно розподілені на шкалі Мела, що було головною особливістю MFCC від методу LPCC.

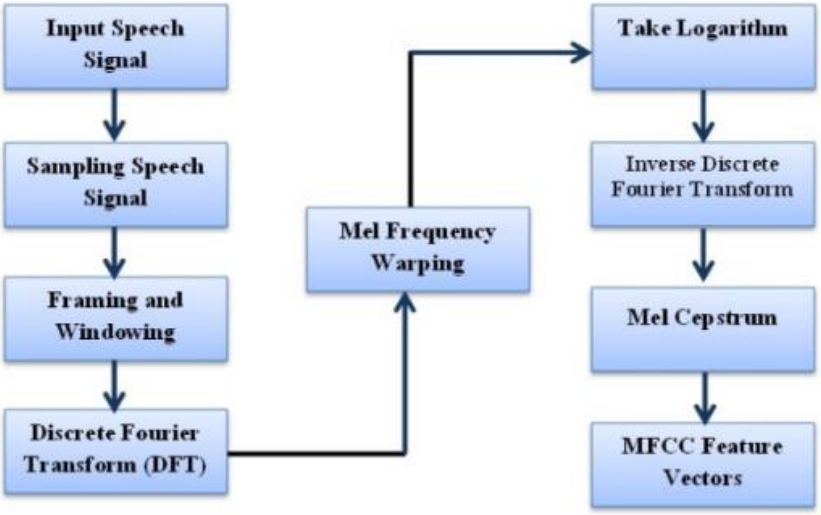



Рисунок 2.4 - Етапи вилучення MFCC

На рисунку 2.4 зображено кроки, пов’язані з виділенням MFCC із вхідного мовного сигналу. Мовний сигнал дискретизується та кадрується для обмеження 20-40 мс. Оцінка періодограми спектра потужності визначається для кожного кадру, а банк фільтрів Mel застосовується до спектру потужності. Логарифмічний результат усіх енергій банку фільтрів було оброблено за допомогою математичних інструментів дискретного косинусного перетворення (DCT).

Попереднє підкріплення є однією з загальновживаних практик передпроцесингу в області обробки сигналів, яка використовується для компенсації високочастотного сигналу, що був пригноблений під час виробництва сигналу. Попереднє підкріплення - це дуже перший крок під час адаптації MFCC і може бути реалізовано просто застосуванням фільтра високих частот із параметрами [1, −0,97]. Процес фільтрації змінює розподіл енергії по частотах, а також загальний рівень енергії [14].

Ідея розділення сигналів на відокремлені "кадри" полягає в тому, щоб розкласти сирі дані сигналу на кадри, де сигнал має тенденцію бути більш стаціонарним. Для стійких акустичних характеристик мови необхідно вивчати мову протягом достатньо короткого періоду. Процес кадрування сигналу - це процес розбиття сигналу на короткі кадри, причому кожен кадр,

як правило, триває від 20 до 40 мілісекунд. Щоб забезпечити неперервність між кадрами, часто вводиться перекриття. До кожного кадру потім застосовується функція вікна для зменшення період часу. Щодо мовного сигналу, період часу від 20 до 30 мс повідомляється як квазістаціонарний сегмент (QSS), оскільки час між двома замиканнями голосових складок приблизно 20 мс. Однак голосні голоси, як повідомляється, фіксуються від 40 до 80 мс [15]. Таким чином, вимірювання спектра короткострокового часу зазвичай проводяться протягом вікон тривалістю 20 мс, і кожен кадр перекривається на 10 мс з наступним. Перекриття кадрів на 10 мс дозволяє відстежувати часові характеристики мовного сигналу. З перекриттям кадрів звукове представлення мови буде приблизно центровано на якомусь кадрі.

На кожному кадрі застосовується вікно, щоб звузити сигнал до меж кадру. Загалом, вікна Хеннінга та Хеммінга [16] є одними з найвідоміших номінантів. Ці вікна можуть підсилити гармоніки, згладити краї та зменшити ефект краю під час виконання ДКП на сигналі. На рисунку 2.5 показані прямокутне вікно Хеммінга та вікно Хеннінга як у часовому, так і у частотному доменах.

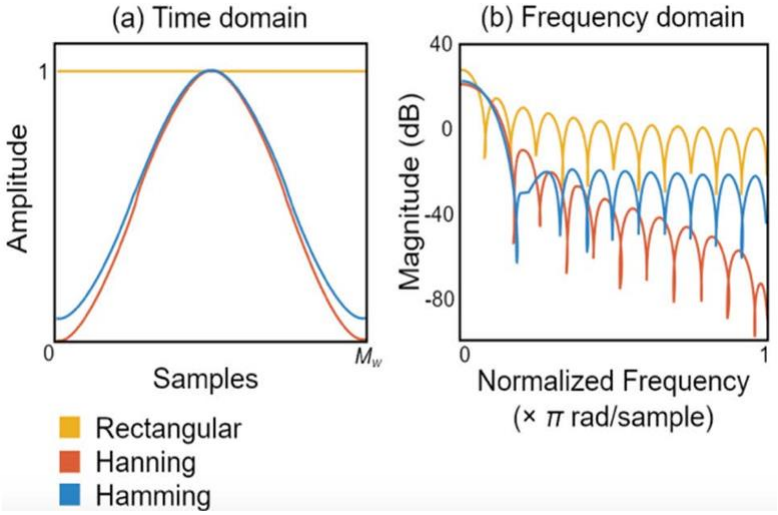



Рисунок 2.5 - Прямокутні вікна Ганніга та Хеммінга в часі та частотній області

Спектр потужності можна описати як розподіл потужності частотних

компонентів, що складають сигнал [17]. Традиційно для обчислення спектра потужності використовується Дискретне перетворення Фур'є (DFT). Спектр потужності для кожного з отриманих кадрів має бути визначений на основі рівняння (2.1):

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-\frac{2\pi j n K}{N}} \quad k = 1, 2, 3 \ldots N-1 \tag{2.1}$$

де x (n) – дискретний сигнал, а N – дожвина сигналу.

Фільтр-полоса Мела - це банк фільтрів, який конструюється на основі сприйняття висоти тона. Фільтр Мела був розроблений спочатку для аналізу мови і, подібно до слуху людини, спрямований на видобуток нелінійного представлення мовного сигналу. Звичайно банк фільтрів Мела складається з 40 трикутних фільтрів [18]. Функція передачі TF кожного фільтра m-тої лінії може бути обчислена за допомогою рівняння (2.2):

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \le k < f(m) \\ 1 & k = f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} & f(m) < k \le f(m+1) \\ 0 & k > f(m+1) \end{cases} \tag{2.2}$$

де $f$(m)- центральна частота трикутного фільтра, а $\sum_m^{M-1} H_m(k)=1$.

Шкала Мела до частоти відгуку і навпаки обчислюється за рівняннями (2.3) і (2.4).

$$m = 2595 log_{10}\left(1 + \frac{f}{700}\right)$$
$$f = 700\left(10^{m/2595} - 1\right) \qquad (2.3, 2.4)$$

Дискретне косинусне перетворення (DCT) виражає скінченну послідовність точок даних як суму косинусних функцій, що коливаються при різних частотах. DCT було представлено Насіром Ахмедом в 1972 році. У процесі обчислення MFCC DCT застосовується до фільтрів Мела для вибору найбільш прискорюючих коефіцієнтів або для відокремлення відносин в логарифмічних спектральних величинах від фільтр-банку [19]. DCT розраховується за допомогою рівняння (2.5):

$$X(k) = \sum_{n=0}^{N-1} x_n * COS\left(2\pi jnk/_N\right), \quad k = 1, 2, 3 \ldots N-1 \qquad (2.5)$$

де $x_n$ це дискретний сигнал і N це довжина сигналу.

MFCC — це багатоплатформна техніка, яка може працювати в MATLAB і python. Порівняння продуктивності дає кращі результати в MATLAB на відміну від програмування на Python.

### 2.1.5 Відносна спектральна фільтрація коефіцієнтів логарифмічної області (RASTA)

Метод RelAtive SpecTrA filtering of log domain coefficients (RASTA) усуває недолік придушення нульової частоти та повільно змінних компонентів у мовному сигналі, коли виділення ознак виконується методом MFCC. Фільтр RASTA випереджає MFCC, знижуючи вплив шуму в мовних сигналах і забезпечуючи високий рівень надійності.

Рисунок 2.6 - Процес фільтрації RASTA

На рисунку 2.6 показано етапи, які виконуються під час фільтрації мовного сигналу RASTA. Бібліотека аналізу надає потужність для компенсації результатів лінійного спотворення каналу в результаті пропускання кожного коефіцієнта під час процесу фільтрації. Оцінка чистої мови виконується шляхом вибору частоти «i» та індексу кадру «k».

$$S_i(k) = \sum_{j}^{M} w_i(j) Y_i(k-j)$$

Метод RASTA забезпечує кращий коефіцієнт продуктивності завдяки видаленню повільних і швидких варіацій. У той час як цей метод обґрунтовує незначні недоліки продуктивності для отримання чіткої інформації. Кроки, описані в RASTA, є:

- розрахунок алгоритму критичної зони;
- логарифмування критичної смуги;
- визначення часової похідної спектра логарифмічної критичної смуги;
- реінтеграція тимчасової похідної логарифмічної критичної смуги;
- додавання кривої рівної гучності;
- виконання оберненого логарифмічного процесу.

RASTA має однакові переваги з MFCC, тоді як попередній кращий для мовного сигналу, записаного в шумному середовищі. Фільтру MFCC не вдається зафіксувати хорошу продуктивність у фільтрації мовного сигналу, записаного в шумному середовищі.

Порівняння між LPC і MFCC наведено у таблиці 2.1

Таблиця 2.1 – Порівняння методів виділення ознак - LPC і MFCC

| | LPC | MFCC |
|---|---|---|
| | 1 | 2 |
| 1 | Коефіцієнти лінійного предиктора (LPC) імітують людський голосовий апарат і забезпечують надійну характеристику мовлення. | MFCC - це відтворення механізму людського слуху, розроблене для вибіркового застосування теорії роботи вуха, припускаючи, що людський голос є хорошим розпізнавачем спікера. |
| 2 | Він підтримує одного спікера та одну мову і є швидким у впровадженні. | Він підтримує кілька спікерів і кілька мов і трохи повільний у реалізації. |
| 3 | Невеликий обсяг словника робить його надійним. | Він надійно працює зі словниками від середнього до великого обсягу. |
| 4 | Метод лінійного предиктора використовується для досягнення параметрів фільтра, порівнянних з голосовим апаратом, шляхом зниження середньоквадратичної похибки між вхідним сигналом і спроектованим мовленням. | MFCC - це коефіцієнти, сформовані на основі слухового досвіду людини на основі викривленої частотної шкали. |
| 5 | Використовується в системах розпізнавання мовлення, де основною метою є отримання властивостей голосового апарату. | Зазвичай використовуються для розпізнавання спікерів і розуміння мовлення. |
| 6 | Це забезпечує дуже точні оцінки параметрів мовлення і відносно добре підходить для обчислень. | Певні проблеми розпізнавання образів, пов’язані з людським голосом, вважаються точними в кепстральних коефіцієнтах. |

Продовження Таблиці 2.1

| | 1 | 2 |
|---|---|---|
| 7 | Оцінки LPC є дуже чутливими до квантування шуму і можуть бути непридатними для узагальнення. | MFCC не є абсолютно точним, коли присутній фоновий шум, і може бути погано пристосований для узагальнення. |
| 8 | Підхід LPC широко використовується в музичних і технічних компаніях для створення мобільних роботів, у телефонних компаніях, для оцінки тональності скрипки та інших інструментальних струнних гаджетів. | З метою безпеки, MFCC використовується для класифікації бронювання авіаквитків, номерів, вимовлених по телефону, та програмного забезпечення для розпізнавання мовлення. |

Отже, в більшості випадків для отримання ознак голосу використовується кепстральний аналіз. Кепстральний аналіз було запропоновано математиком О. М. Колмогоровим. Він запропонував спеціальний функціональний ряд для обробки регулярних стаціонарних випадкових процесів. Цей ряд було названо в його честь – кепстр (Kolmogorov Equation Power Series Time Response). Кепстр визначає послідовність коефіцієнтів розкладання функції у ступінчатий ряд, який є енергетичним спектром цієї функції.

MFCC - одна з часто використовуваних ознак, яка застосовується в різних областях аудіоаналізу, особливо в обробці голосового сигналу, такій як розпізнавання говоріння, розпізнавання голосу та ідентифікація особи чи статі (рисунок 2.7). MFCC може бути розрахована за допомогою послідовних кроків:

- кадрування сигналу;
- обчислення спектру потужності;
- застосування фільтра Мела до отриманих спектрів потужності;

- розрахунок логарифмічних значень всіх фільтрів та, нарешті;
- застосування дискретного косинусного перетворення (ДКП).

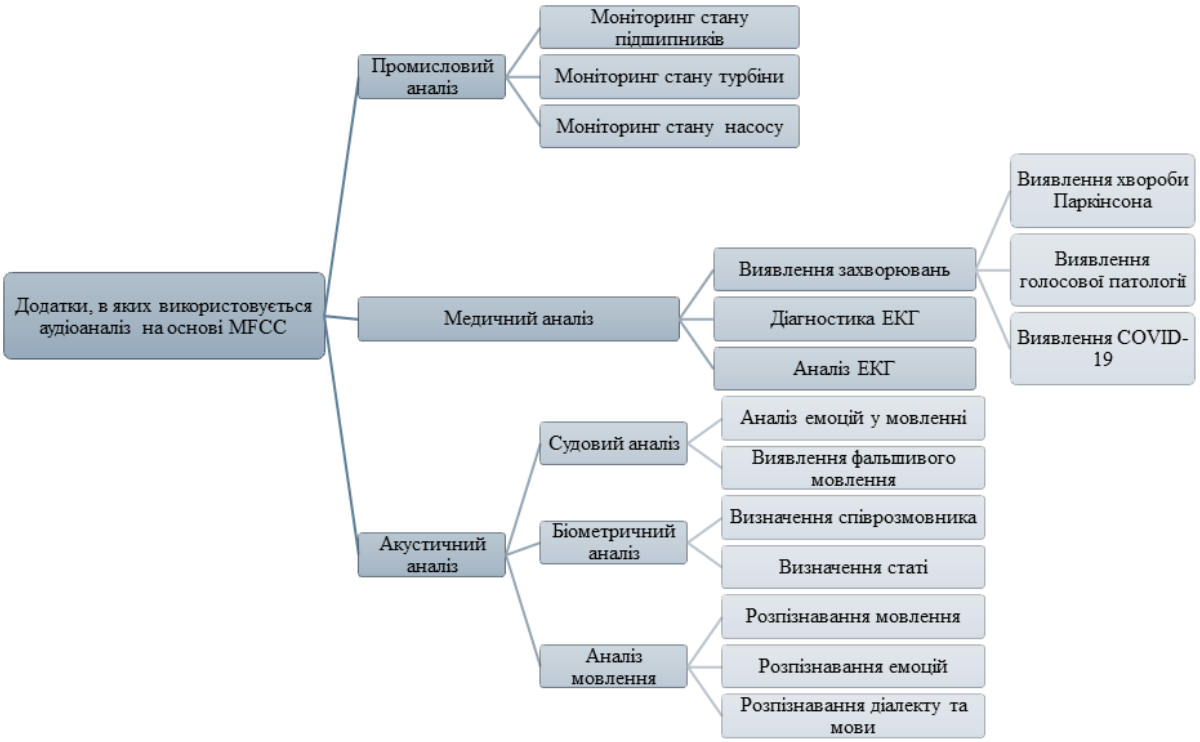



Рисунок 2.7 – Практичні задачі, реалізовані на основі MFCC

Застосування в біометрії - це застосування, яке дозволяє підтвердити або визначити ідентичність певної характеристики особи. Мета цих застосувань - підтверджувати, що надані послуги доступні лише законному користувачеві. Застосування в біометрії можна класифікувати за двома категоріями в залежності від їх завдання: верифікаційні та ідентифікаційні застосування. Верифікаційне застосування підтверджує ідентичність особи, порівнюючи зібрану біометричну інформацію з її власним біометричним шаблоном, який зберігається в базі даних. Відношення між об'єктом та базою даних називається "один до одного". У той час як ідентифікаційне застосування називається "один до багатьох", оскільки застосування намагається ідентифікувати особу, шукаючи шаблони всіх користувачів в базі даних. На підставі існуючої літератури можна визначити, що MFCC використовується в двох біометричних застосуваннях: розпізнавання диктора та визначення статі під час телефонного дзвінка.

Незважаючи на те, що багато останніх технологій були розроблені для розпізнавання голосу та верифікації мовника, все ще існують проблеми і відкриті питання, такі як універсальне фальшиве мовлення, яке потрібно виявляти. Основна мета виявлення фальшивого мовлення полягає в розрізненні між фальшивим мовленням та природнім мовленням [20]. Існує безліч характеристик, які використовуються для виявлення фальшивого мовлення, таких як MFCC та LPC. У [21] запропонувано модель на основі статистичного класифікатора для синтетичного мовлення, де MFCC використовувався як авторизована базова лінія. У [22] розробили модель для виявлення фальшивого стереоаудіо, де класифікатором був SVM, а характеристикою – MFCC. Згідно з їхніми результатами, MFCC може ефективно виявляти аудіо із фальшивим стерео. Автори у [23] розробили модель глибокого навчання, яка була навчена на основі MFCC, 1MFCC та 11MFCC для вирізнення синтезованого мовлення від реального мовлення людини. Вони виявили, що як 1MFCC, так і 11MFCC значно підвищують точність вирізнення порівняно з результатами використання лише MFCC.

## 2.2 Методологічна база для проведення аналізу виділених ознак в аудіоряді

Після виділення та вибору ознак, аудіо зразки класифікуються за різними категоріями. Існує безліч підходів до класифікації. На рисунку 2.8 зображено основні класифікатори, які далі будуть розглянуті у більших деталях.

Штучні нейронні мережі (ШНМ): ШНМ — це обчислювальні інструменти, які використовуються для аналізу багатьох складних проблем реального світу за допомогою біологічних нейронних мереж, заснованих на структурі та функціях нейронів. Він складається з шарів обчислювальних нейронів, які з’єднані між собою кількома зваженими лініями, які здатні виконувати дуже великі паралельні обчислення для обробки даних. ШНМ

використовується в різних сферах, починаючи від розпізнавання мови до прогнозування, а також у медицині.

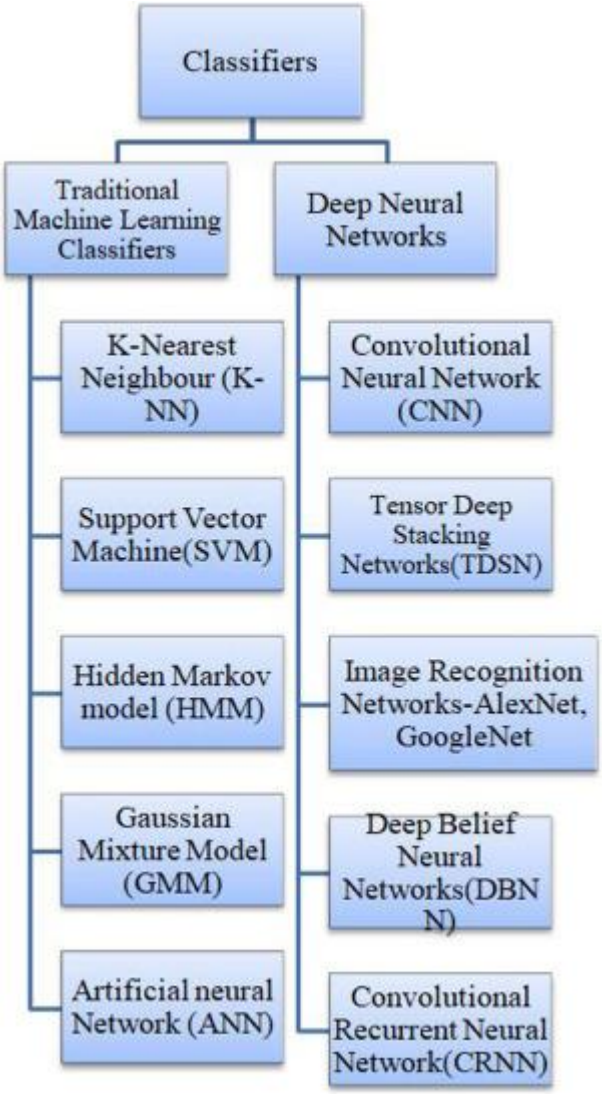



Рисунок 2.8 – Алгоритми класифікації, що використовуються в аудіоаналізі

Модель суміші Гауса (GMM): GMM — це функція щільності ймовірності на основі параметрів, яка представлена як зважена сума щільностей гаусових елементів. Він в основному використовується для вимірювання функцій безперервного розподілу в біометрії, а також пов’язаних з голосовим трактом спектральних характеристик у системі розпізнавання мовців. Його також можна використовувати в системі автоматичного розпізнавання сміху, а також визначення геометрії руки.

Машина опорних векторів (SVM): SVM — це алгоритм для вирішення

проблеми бінарної класифікації та регресійного аналізу. Цей алгоритм максимізує маржу шляхом визначення оптимальної гіперплощини. Використовується для нелінійно роздільних задач. Він також відображає дані в області вищих компонентів, де легко класифікує лінійні поверхні прийняття рішень. Його використання варіюється від розпізнавання рукописних цифр до розпізнавання обличчя та медицини.

Прихована модель Маркова (HMM): HMM — це статистичний інструмент, який використовується для демонстрації генеративних послідовностей, описаних набором спостережуваних послідовностей. HMM складається з двох стохастичних процесів, які включають невидимий процес прихованих станів і видимий процес спостережуваних символів. Прихований стан становить ланцюг Маркова, а розподіл ймовірностей спостережуваного символу залежить від основного стану. Він використовується для моделювання змінних у часі спектральних векторних послідовностей.

Глибинна нейронна мережа (DNN): глибоке навчання — це алгоритми машинного навчання, які впроваджуються в систему розпізнавання мовлення з великим словниковим запасом. МАють багато прихованих рівнів у порівнянні зі штучною нейронною мережею (ШНМ) і містить новий метод навчання своїх даних. Обговорення плюсів і мінусів цих алгоритмів представлено в таблиці 2.2.

Таблиця 2.2 - Вибрані алгоритми класифікації аудіоданих

| Алгоритм | Переваги | Недоліки |
|---|---|---|
| 1 | 2 | 3 |
| ANN | Використовувався для навчання дуже складних моделей.<br>Легко концептуалізувати.<br>Має багато бібліотек для легкої реалізації.<br>Можна широко використовувати для науково-дослідної роботи. | Для навчання потрібно більше даних.<br>Вибір правильних параметрів може бути дуже складним.<br>Остаточна модель може використовувати багато пам’яті.<br>Багатошарові нейронні мережі зазвичай важче навчати і вимагають налаштування багатьох параметрів. |

Продовження Таблиці 2.2

| 1 | 2 | 3 |
|---|---|---|
| GMM | Набагато гнучкіший щодо кластерної коваріації.<br>GMM дозволяє змішане членство точок у кластері. | Навчання може бути трохи складним.<br>Мало бібліотек для реалізації. |
| SVM | Зменшує надмірне навантаження під час тренувань.<br>Керує проблемами великого розміру.<br>Використовується для лінійних і нелінійно розділених даних.<br>Оптимальність гарантується завдяки опуклій оптимізації.<br>Може бути реалізований різними мовами програмування (Python, Matlab), а його бібліотеки легкодоступні. | Навчання може зайняти багато часу.<br>Це може бути недостатньо пояснювальним, оскільки не повертає ймовірнісне значення довіри.<br>Не підтримує структуровані представлення тексту, що забезпечує високу продуктивність у обробці природної мови (NLP). |
| HMM | Дуже гнучкий.<br>Це підходить для кращої підгонки даних.<br>Має дуже статистичний характер.<br>Він може навчатися з вихідних даних. | Має величезну кількість неструктурованих параметрів.<br>Може представляти лише невелику частину всіх розподілів.<br>Залежності між шарами можуть не бути представлені явно. |
| DNN | Постійно нарощує дані для покращення свого навчання.<br>Створює свої нові риси. | Для навчання потрібна дуже велика кількість немаркованих даних.<br>Для його реалізації потрібне дуже високопродуктивне обладнання.<br>Треба більше часу на навчання. |

Вибрати відповідний класифікатор складно, бо треба обирати компроміс між продуктивністю та обчислювальними витратами. Немає жодного дослідження, яке б порівнювало ефективність усіх класифікаторів. Для вибору класифікатора можна враховувати такі параметри:

- обчислювальна складність (хороший класифікатор повинен мати меншу обчислювальну складність. Обчислювальна складність означає час і

потужність, які споживає класифікатор. Час, який потрібний класифікатору для отримання результату, має бути меншим);

- точність розпізнавання (класифікатор повинен точно класифікувати вектори ознак);
- стійкість до шуму (хороший класифікатор повинен ігнорувати варіації, викликані масштабуванням амплітуди або смуги пропускання в аудіосигналі).

# 3 АНАЛІЗ АПАРАТНО-ТЕХНОЛОГІЧНОЇ БАЗИ ДЛЯ ВИКОНАННЯ ДОСЛІДЖЕННЯ

Зростаюча потреба у вдосконаленні комунікації між людьми та технологіями вимагає розуміння та інтерпретації людської мови, що є процесом автоматичної ідентифікації людини за допомогою інформації, яка міститься у мовних хвилях. Немає двох індивідів, голос яких звучав би однаково, через форму голосового тракту, розміри гортані та інші надсегментні особливості. На додаток до цього, кожна людина має свої власні фізичні відмінності, свій стиль мовлення, модель вимови, тощо.

За наявністю цих факторів голос людини можна використовувати як біометрику. Найбільш часто використованими короткочасними спектральними вимірюваннями є кепстральні коефіцієнти та їх коефіцієнти регресії. Коефіцієнти регрессії, як правило, беруться в кожному періоді кадру для представлення спектральної динаміки.

Актуальність застосування методів розпізнавання мовця обумовлена зростаючою потребу в безпеці, зручністю, персоналізацією та ефективністю та автоматизація бізнес-процесів в різних аспектах сучасного життя. Окремо можна відмітити найсучасніший фактор, який ще більше підвищує актуальність дослідження методів розпізнавання мовця – захист від фальсифікації та транскрибування голосових повідомлень у соціальних та комунікаційних платформах. Сфери застосування методів на практиці є вкрай поширеними (таблиця 3.1).

Ці приклади показують, наскільки широко та різноманітно можуть застосовуватися технології розпізнавання мовця для покращення безпеки, зручності користувачів та ефективності в різних сферах життя, що підтверджує актуальність дослідження методів розпізнавання мовця незалежно від сфери застосування.

Таблиця 3.1 – Сфери застосування методів розпізнавання мовця

| Сфера застосування | Приклади використання |
|---|---|
| 1 | 2 |
| Безпека та аутентифікація [24]:<br>- банківські та фінансові послуги;<br>- контроль доступу. | - для ідентифікації клієнтів при дзвінках у кол-центри або при доступі до банківських рахунків онлайн;<br>- для перевірки особи на основі голосу. |
| Правоохоронні органи та судова експертиза [25]:<br>- криміналістика;<br>- моніторинг підозрюваних. | - для ідентифікації підозрюваних або свідків на основі записів їх голосу;<br>- для відстеження та моніторингу підозрюваних через прослуховування телефонних розмов. |
| Користувацькі інтерфейси та персональні асистенти [26-27]:<br>- розумні пристрої та домашні помічники;<br>- мобільні телефони та додатки. | - у таких пристроях, як Amazon Alexa, Google Home, для персоналізації взаємодії з користувачами;<br>- аутентифікація користувача для доступу до пристроїв або певних додатків на основі голосу. |
| Телемедицина та охорона здоров’я [28]:<br>- дистанційний моніторинг пацієнтів;<br>- психологічні оцінки. | - для ідентифікації пацієнтів і надання їм персоналізованих медичних консультацій на основі їх голосу;<br>- для аналізу емоційного стану пацієнтів. |
| Комунікації та соціальні мережі:<br>- розпізнавання голосових повідомлень;<br>- захист від фальсифікації голосу. | - у таких додатках, як WhatsApp або Facebook Messenger, для автоматичного розпізнавання та транскрибування голосових повідомлень;<br>- для запобігання шахрайству на основі підробки голосу в соціальних мережах. |

Продовження Таблиці 3.1

| 1 | 2 |
|---|---|
| Бізнес та аналітика:<br>- аналітика кол-центрів;<br>- зворотній зв'язок від клієнтів. | - для автоматичного розпізнавання клієнтів та аналізу їхніх запитів для покращення обслуговування;<br>- для збору та аналізу голосових відгуків клієнтів. |
| Освіта та навчання [29-30]:<br>- ідентифікація студентів;<br>- персоналізоване навчання. | - для аутентифікації студентів при здачі онлайн-іспитів;<br>- у навчальних платформах для налаштування навчального матеріалу відповідно до голосових команд та запитів студентів. |

Актуальність проблеми ідентифікації та верифікації обумовлює різноманіття методів, що використовується для рішення задачі – від характеристи, обумовлених фізичною природою звуку до нейромережевих моделей глибокого навчання [30, 31]. Розглядаючи проблемну область та предмет дослідження, вдалося запропонувати розгорнуту систематизацію підходів ідентифікації та верифікації спікерів (рисунок 3.1).

Рисунок 3.1 – Запропонована систематизація підходів ідентифікації та верифікації спікерів

Завдання з розпізнавання мовця можна класифікувати на:

- ідентифікацію – це процес визначення того, хто говорить із групи вже відомих мовців. При ідентифікації кількість альтернативних рішень дорівнює розміру популяції;

- верифікацію – перевірка мовця – це процес прийняття або відхилення особи, заявленої мовцем. При верифікації є лише два варіанти, прийняти або відхилити, незалежно від розміру популяції.

Таким чином, продуктивність ідентифікації мовців зменшується зі збільшенням розміру популяції, тоді як ефективність верифікації мовців наближається до константи незалежно від розміру популяції, якщо тільки розподіл фізичних характеристик мовців не є надзвичайно упередженим.

Окрім того, системи розпізнавання мови можна класифікувати як:

- текстово-залежні системи розпізнавання мови (передбачено аналіз тексту для остаточного прийняття рішення);

- текстово-незалежні системи розпізнавання мови (не передбачено аналіз тексту для остаточного прийняття рішення).

У режимі текстової залежності обидва етапи, як навчання, так і тестування, ґрунтуються на одному й тому ж тексті, тоді як у випадку текстової незалежності подібних обмежень не існує.

У даній роботі розглядається гібридний підхід, що поєднує обидва підходи ідентифікації та верифікації людей за типом розпізнавання – незалежної від тексту та залежної від тектсу ідентифікації мовця, стійкої до шуму, за допомогою нейромережевої інтелектуальної моделі (рисунок 3.2). Проведені дослідження присвячені визначенню оптимальної архітектури нейронної мережі, яка забезпечить досягнення високої точності за мінімальну кількість епох навчання, а також дослідженню методів прискорення етапу препроцесінгу вхідного аудіоряду.

Запропоновану систему розпізнавання за голосом на основі спектрограм та верифікації за голосовим паролем можна назвати багатофакторною (MFA), оскільки запропонована система передбачає

використання більше одного методу для підтвердження особи користувача. У даному випадку є два фактори:

- спектрограма голосу – фізична характеристика, унікальна для кожної людини, яка використовується для розпізнавання голосу. Ідентифікація мовця на основі його голосу, використовуючи виділені характеристики (спектрограми або мел-кепстральні коефіцієнти) та нейромережеву модель із згортковою архітектурою, є першим фактором аутентифікації мовця;

- голосовий пароль – знання (фактор, що користувач знає), яке підтверджує автентичність. Верифікація голосового пароля за допомогою другої нейромережевої моделі із згортковою архітектурою, є другим фактором аутентифікації мовця.

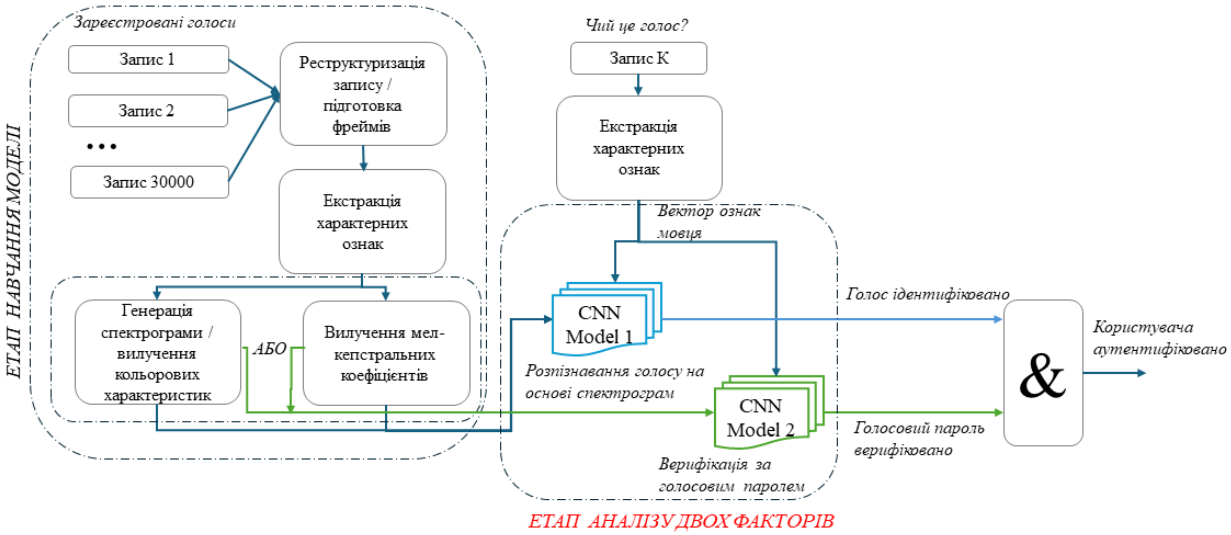



Рисунок 3.2 - Функціональна модель запропонованої системи двофакторної аутентифікації мовця

Комбінуючи ці два фактори, система підвищує рівень безпеки, оскільки для успішної автентифікації потрібно відповідати обом критеріям. Таким чином, така система є багатофакторною.

На основі запропонованої системи проведено ряд досліджень, а саме – проаналізовано та експериментально підтверджена ефективність одного із двох методів обробки аудіосигналу для виділення важливих характеристик

(спектрограми та мел-кепстральні коефіцієнти), досліджена архітектура нейромережевих аналізаторів із згортковою архітектурою, вплив розподілу вхідного датасету на навчальну, тестову та валідаційну вибірки, проаналізована тривалість навчання та вимоги до апаратного забезпечення для забезпечення стійкості роботи системи.

Таблиця 3.2 – Характеристики датасету для навчання нейромережевих аналізаторів

| Speaker_ID | Digit | File_ID | Volume | Pronunciation | Duration |
|---|---|---|---|---|---|
| Speaker_1 | 0 | File_1 | low | clear | 500ms |
| Speaker_1 | 1 | File_2 | medium | muffled | 501ms |
| Speaker_1 | 2 | File_3 | high | fast | 502ms |
| Speaker_1 | 3 | File_4 | low | slow | 503ms |
| Speaker_1 | 4 | File_5 | medium | clear | 504ms |
| ... | ... | ... | ... | ... | ... |
| Speaker_60 | 9 | File_30000 | high | slow | 2499ms |

Навчання нейромережевих моделей із згортковою архітектурою виконувалося на датасеті, який було організовано ноступним чином (таблиця 3.2):

- Speaker_ID – ідентифікатор спікера, в даному випадку у нас 60 спікерів (від Speaker_1 до Speaker_60);
- Digit – цифра, яку промовляє спікер (від 0 до 9);
- File_ID – унікальний ідентифікатор кожного аудіофайлу;
- Volume – гучність запису (low, medium, high);
- Pronunciation – вимова (clear, muffled, fast, slow);
- Duration – тривалість запису у мілісекундах (від 500ms до 2499ms).

## 3.1 Експеримент 1. Порівняння точності ідентифікації спікера за голосом на основі мел-кепстральних коефіцієнтів та спектрограм

Послідовність дій для реалізації системи з використанням мел-кепстральних коефіцієнтів (рисунок 3.3):

- початковий набір аудіофайлів, який складається із звукозаписів вимови кожної цифри англійською мовою 50 разів для 60 різних спікерів у форматі WAV, завантажується у пам’ять для подальшої обробки;
- кожен отриманий аудіоряд проходить процедуру повторної вибірки (resampling), для вирівнювання загальної кількості вибірок за секунду в рамках всього датасету;
- вилучення мел-кепстральних коефіцієнтів для кожного аудіоряду, та збереження тринадцяти перших у масив даних;
- формування набору даних за допомогою збережених коефіцієнтів та відповідного ідентифікатора спікера, розбиття початкового набору даних на три: навчальний, валідаційний та тестовий;
- навчання нейронної мережі конволюційного типу (CNN) з використанням сформованого набору навчальних та валідаційних даних;
- тестування навченої моделі з використанням тестової вибірки даних для отримання результатів точності ідентифікації спікерів.

Рисунок 3.3 – Послідовність дій для реалізації системи з використанням мел-кепстральних коефіцієнтів

Послідовність дій для системи з використанням спектрограм

(рисунок 3.4) має різні початкові кроки в попередній обробці навчальних даних, замість вилучення мел-кепстральних коефіцієнтів, потрібно виконати наступні кроки:

- за допомогою швидкого перетворення Фур'є, кожен аудіоряд перетворити на спектрограму, у вигляді зображення розміром 305*184 та зберегти її для подальшого використання;
- завантажити отримане зображення спектрограми та вилучити її кольорові характеристики у трьох каналах (рисунок 3.4);
- використання отриманих кольорових характеристик спектрограм для формування навчального датасету.

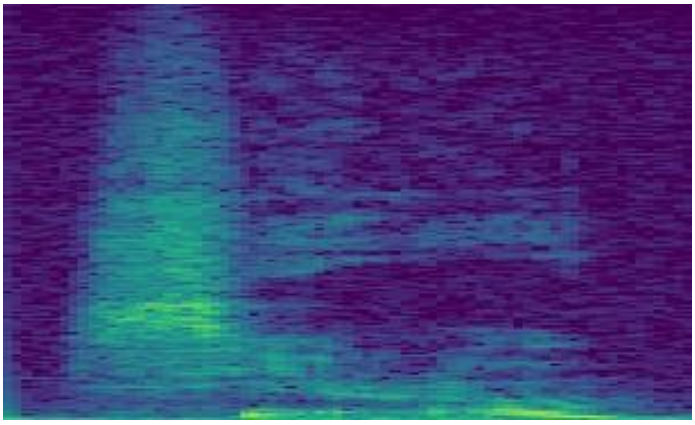


Рисунок 3.4 – Приклад спектрограми

Для порівняння точності систем з використанням мел-кепстральних коефіцієнтів та спектрограм було створено та навчено дві нейроні мережі з однаковими характеристиками, архітектурою та параметрами навчання. Дві системи відрізняються лише типом вхідних даних для нейронної мережі – мел-кепстральні коефіцієнти в першому випадку, та спектрограму у іншому.

Для навчання було використано однаковий початковий датасет, який складається з 30 000 аудіофайлів у форматі WAV, кожен файл представлений записом однієї з десяти цифр, для кожного з 60 спікерів, кожна з цифр має 50 різних записів кожного зі спікерів, з різною гучністю, вимовою та тривалістю. Навчання нейронної мережі проводилось з однаковою кількістю

епох (рисунок 3.5). Для отримання показників точності була проведена перевірка за допомогою тестового датасету, що не приймав участь у процесі навчання.

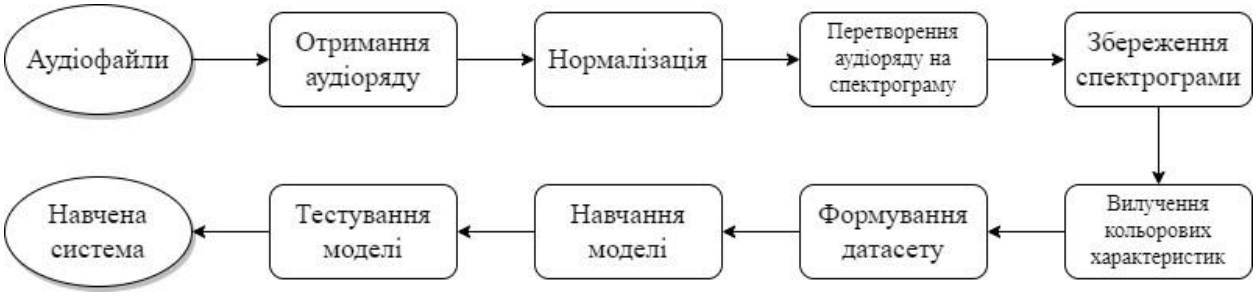



Рисунок 3.5 – Послідовність дій для реалізації системи з використанням спектрограм

3.1.2 Дослідження часу навчання та точності класифікації нейромережевих моделей у залежності від типу визначення характерних ознак

Порівняльна характеристика двох систем за точністю моделі при використанні тестового набору даних, а також за часом який потрібен для навчання та необхідний об’єм оперативної пам’яті наведена в таблиці 3.3.

Таблиця 3.3 – Результати навчання систем з використанням мел-кепстральних коефіцієнтів та спектрограм

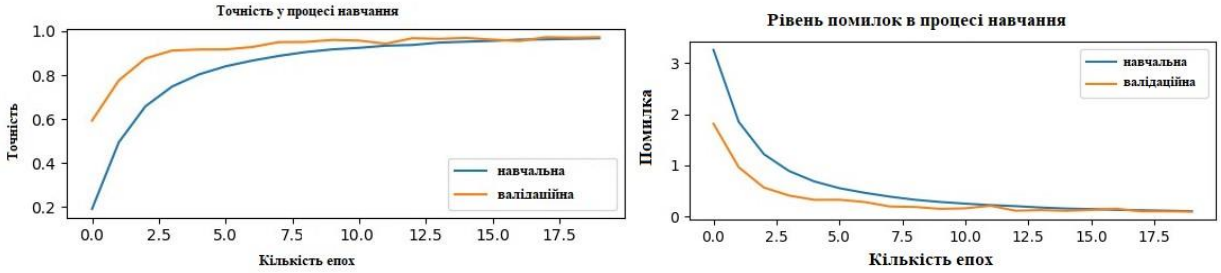
| | Спектрограма | | | Мел-кепстральні коефіцієнти | | |
|---|---|---|---|---|---|---|
| Кількість епох | Тривалість навчання | Об’єм пам'яті | Точність | Тривалість навчання | Об’єм пам'яті | Точність |
| 20 | 7 год. 48 хв. | 15 Гб | 73.13% | 6 хв. 43 сек. | 2 Гб | 96.84% |
| 60 | 23 год. 57 хв. | 15 Гб | 80.9% | 14 хв. 58 сек. | 2 Гб | 97.64% |

Модель, навчена з використанням мел-кепстральних коефіцієнтів, має вищу точність порівняно з моделлю, навченою зі спектрограмами на 17%. Окрім того, час навчання при використанні мел-кепстральних коефіцієнтів

відбувається на 23 години 42 хвилини швидше.

Результати навчання нейронної мережі з використанням мел-кепстральних коефіцієнтів з кількістю епох – 20:

- зміна точності під час навчання моделі, для навчальної та валідаційної вибірки даних (рисунок 3.6а);
- зміна похибки моделі для навчальної та валідаційної вибірки (Рисунок 3.6б).




а) б)

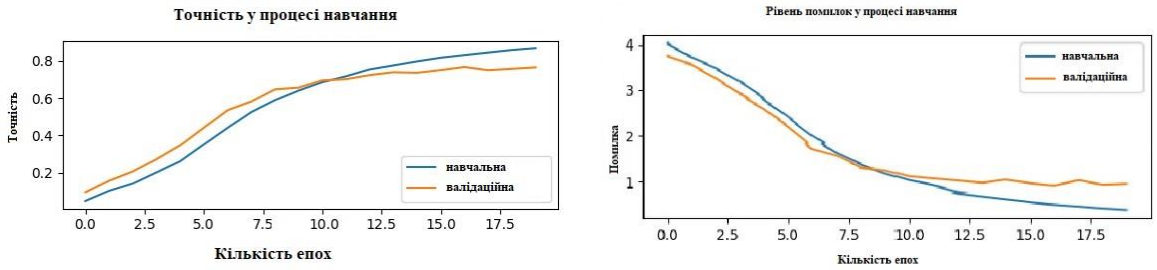
Рисунок 3.6 – Графік зміни точності моделі та похибки у процесі навчання на 20 епохах із використанням мел-кепстральних коефіцієнтів

Аналіз графіків на рисунку 3.6 показує, що модель демонструє хорошу здатність до навчання із використанням мел-кепстральних коефіцієнтів, про що свідчать зростаюча точність та зменшення рівня помилок.

Обидва показники навчання і валідації свідчать про ефективність моделі та її хорошу здатність до узагальнення з мінімальним перенавчанням до кінця періоду навчання.

Результати навчання нейронної мережі з використанням спектрограм з кількістю епох – 20:

- зміна точності під час навчання моделі, для навчальної та валідаційної вибірки даних (рисунок 3.7а);
- зміна похибки моделі для навчальної та валідаційної вибірки (рисунок 3.7б).

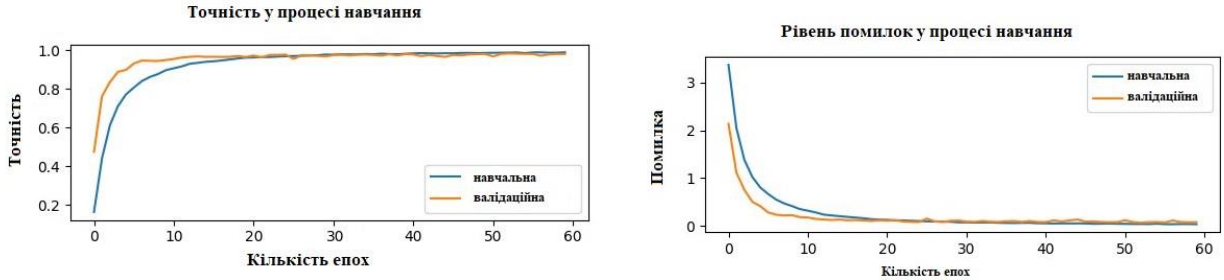



а) б)

Рисунок 3.7 – Графік зміни точності та похибки моделі у процесі навчання на 20 епохах із використанням спектрограм

Результати навчання нейронної мережі з використанням мел-кепстральних коефіцієнтів з кількістю епох – 60:

- зміна точності під час навчання моделі, для навчальної та валідаційної вибірки даних (рисунок 3.8а);
- зміна похибки моделі для навчальної та валідаційної вибірки (рисунок 3.8б).

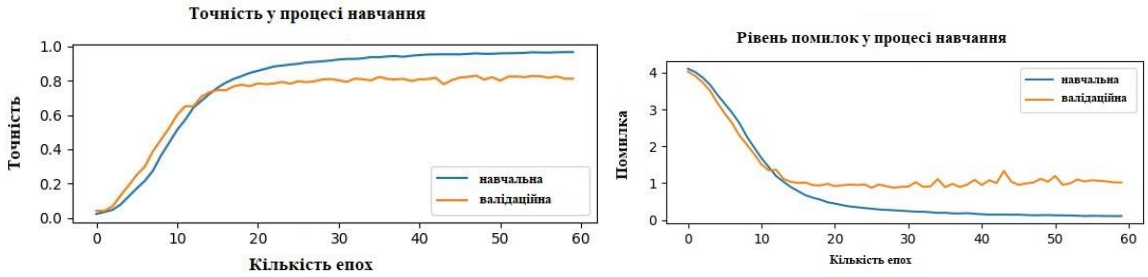



а) б)

Рисунок 3.8 – Графік зміни точності моделі та похибки у процесі навчання на 60 епохах із використанням мел-кепстральних коефіцієнтів

Результати навчання нейронної мережі з використанням спектрограм з кількістю епох – 60:

- зміна точності під час навчання моделі, для навчальної та валідаційної вибірки даних (рисунок 3.9а);
- зміна похибки моделі для навчальної та валідаційної вибірки (рисунок 3.9б).




а) б)

Рисунок 3.9 – Графік зміни точності та похибки моделі у процесі навчання на 60 епохах із використанням спектрограм

Порівнюючи точності навчання на 20 та 60 епохах видно, що графік на 60 епох демонструє більш тривалий період стабільності, що дозволяє краще оцінити довгострокову поведінку моделі. Для більш повного аналізу довгострокової стабільності моделі слід використовувати триваліші періоди навчання (як у графіку на 60 епох). Таблиця 3.3 показує, що використання мел-кепстральних коефіцієнтів дозволяє досягти найбільшої точності 97.64% розпізнавання мовця на 60 епохах навчання.

Подальші експерименти проводяться із використанням мел-кепстральних коефіцієнтів.

### 3.1.3 Аналіз співвідношення навчальної, тестової та валідаційної вибірок у робочому датасеті

Для встановлення оптимального відсоткового поділу початкового датасету на три окремі групи, було проведено набір тестів з почерговим збільшенням або зменшенням відсотка кожної з груп: навчальної, валідаційної та тестової.

Спершу було проведено набір експериментів з поступовим збільшенням відсотка валідаційної вибірки за рахунок зменшення навчальної, результат яких наведений у таблиці 3.4.

Таблиця 3.4 – Залежність точності моделі від відношення навчальної та валідаціної вибірок

| Навчальна вибірка у % | Валідаційна вибірка у % | Тестова вибірка у % | Час навчання | Точність у % |
|---|---|---|---|---|
| 80 | 10 | 10 | 17 хв. | 97.26 |
| 75 | 15 | 10 | 16 хв. | 97.3 |
| 70 | 20 | 10 | 15 хв. 10 сек. | 96.5 |
| 60 | 30 | 10 | 15 хв. | 96.2 |
| 50 | 40 | 10 | 14 хв. 20 сек. | 96.16 |
| 40 | 50 | 10 | 14 хв. | 95.59 |
| 30 | 60 | 10 | 13 хв. 25 сек. | 93.73 |
| 20 | 70 | 10 | 10 хв. 58 сек. | 93.66 |
| 10 | 80 | 10 | 8 хв. 43 сек. | 88.43 |
| 5 | 90 | 5 | 11 хв. 19 сек. | 77.1 |

Також було проведено ряд експериментів з поступовим збільшенням тестової вибірки за рахунок зменшення навчальної, результати представлені у таблиці 3.5.

Таблиця 3.5 – Залежність точності моделі від відношення тестової та навчальної вибірок

| Навчальна вибірка у % | Валідаційна вибірка у % | Тестова вибірка у % | Час навчання | Точність у % |
|---|---|---|---|---|
| 80 | 10 | 10 | 17 хв. | 97.26 |
| 70 | 10 | 20 | 14 хв. | 97.06 |
| 60 | 10 | 30 | 13 хв. | 96.42 |
| 50 | 10 | 40 | 11 хв. | 95.34 |
| 40 | 10 | 50 | 9 хв. 18 сек. | 94.62 |
| 30 | 10 | 60 | 7 хв. 33 сек. | 94.3 |
| 20 | 10 | 70 | 4 хв. 44 сек. | 92.85 |
| 10 | 10 | 80 | 3 хв. 52 сек. | 86.06 |
| 5 | 5 | 90 | 2 хв. 18 сек. | 77.49 |

Заключний етап експериментів передбачав встановлення оптимального співвідношення тестового та валідаціного набору даних, за рахунок поступового зменшення навчальної вибірки, результат наведено в таблиці 3.6.

Таблиця 3.6 – Залежність точності моделі від відношення навчальної та валідаційної вибірок

| Навчальна вибірка у % | Валідаційна вибірка у % | Тестова вибірка у % | Час навчання | Точність у % |
|---|---|---|---|---|
| 70 | 15 | 15 | 15 хв. | 97.4 |
| 60 | 20 | 20 | 13 хв. | 96.5 |
| 40 | 30 | 30 | 12 хв. 30 сек. | 94.91 |
| 20 | 40 | 40 | 8 хв. 30 сек. | 93.51 |
| 10 | 45 | 45 | 8 хв. 27 сек. | 92 |

Результати експериментів показали (рисунок 3.10), що час навчання зменшується зі зменшенням відсотка навчальної вибірки. Це очікувано, оскільки менша кількість даних для навчання означає менше часу, необхідного для опрацювання моделі.

Проте, і точність моделі також зменшується зі зменшенням відсотка навчальної вибірки. Це вказує на те, що більша навчальна вибірка покращує здатність моделі робити точні прогнози. Найвища точність (97.4% точності) досягнута на співвідношенні при 70% даних для навчання: 15% даних для валідації : 15% даних для тестування.

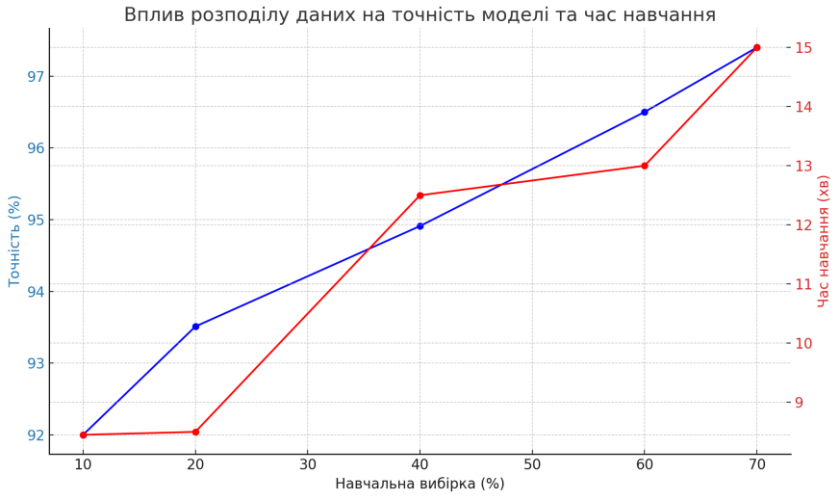



Рисунок 3.10 - Вплив розподілу даних на точність моделі та час навчання

Необхідно враховувати, що час навчання та ресурси можуть бути обмеженими. Тому, вибір розподілу даних повинен базуватися на компромісі між бажаною точністю та допустимим часом навчання. Проведений експеримент дозволяє оцінити тенденцію зміни точності при зміні відношення між навчальною, валідаційною та тестовою вибірками.

### 3.1.4 Дослідження впливу частоти дискретизації на точність розпізнавання спікера

В ході виконання роботи для встановлення оптимальної частоти дискретизації, що забезпечує найвищий показник точності розпізнавання спікера, було проведено тестування впливу різних значень частоти дискретизації на показник точності моделі.

Початковий набір аудіофайлів складається з записів, що мають частоту дискретизації – 48 kHz. Для проведення тестування також було взято популярні частоти дискретизації: 8 kHz, 16 kHz, 20.05 kHz, 32 kHz. Результати проведених експериментів наведені в таблиці 3.7

Таблиця 3.7 – Залежність точності моделі від показнику частоти дискретизації

| Частота дискретизації, kHz | Час навчання | Точність у % |
|---|---|---|
| 8 | 14 хв. 57 сек. | 97.62 |
| 16 | 15 хв. 13 сек. | 98.02 |
| 20.05 | 16 хв. 59 сек. | 97.84 |
| 32 | 17 хв. 05 сек. | 98.06 |
| 48 | 17 хв. 52 сек. | 98.28 |
| 88.2 | 19 хв. 35 сек. | 97.77 |

За результатами проведених експериментів було встановлено, що найбільш оптимальною частотою дискретизації є 48 kHz, яка забезпечує найвищий показник точності та оптимальний час навчання та завантаження

аудіоряду. Вищі частоти дискретизації фіксують більше інформації про сигнал. Для завдань обробки аудіо вища ЧД може допомогти моделі вловлювати тонкі нюанси та деталі аудіосигналу, потенційно підвищуючи точність. Хоча використовувати такі високі частоти дискретизації для розпізнавання спікерів не так традиційно, кожен набір даних і завдання можуть мати унікальні характеристики та впливати на налаштування системи.

### 3.1.5 Визначення архітектури нейронної мережі для досягнення найвищої точності розпізнавання спікера за голосом

У ході роботи було використано згорткову нейронну мережу (CNN) з різною кількістю згорткових шарів, епох навчання та параметрами налаштування для уникнення перенавчання моделі.

Впродовж процесу підбору оптимальної архітектури нейронної мережі було проведено ряд експериментів зі зміною кількості згорткових шарів. Для початку тестування було вибрано мережу з трьома згортковими шарами, кожен з яких супроводжується MaxPooling та BatchNormalization шарами, для покращення процесу навчання і точності вилучення важливих ознак. Кожен наступний згортковий шар має вдвічі більшу розмірність ніж попередній, що забезпечує підвищення кількості гіперпараметрів та дозволяє запам’ятовувати більш складні шаблони. Впродовж тестування було виконано поступове збільшення кількості згорткових шарів для того, щоб знайти оптимальну архітектуру нейронної мережі та кількість епох для навчання. Модель спочатку було навчено за допомогою початкового датасету, що складається з 60 спікерів, використовуючи оптимальний розподіл даних: 70% начальна вибірка, 15% валідаційна та тестова.

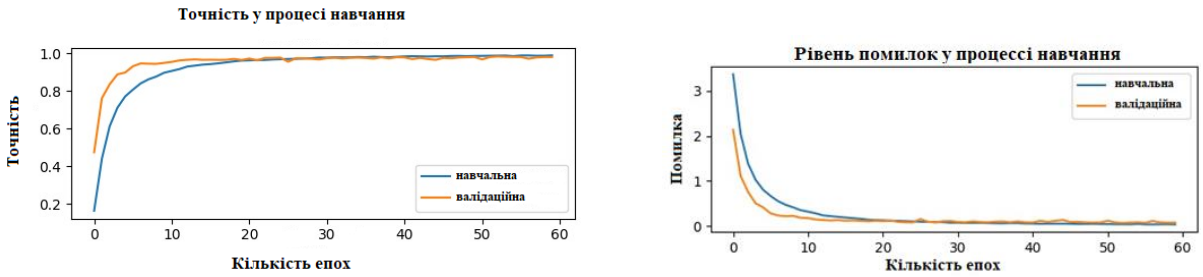
Результати проведених експериментів наведено у таблиці 3.8.

Таблиця 3.8 – Результати тестування різних архітектур нейронної мережі та кількість епох навчання

| Кількість шарів | Кількість епох | Час навчання | Точність у % |
|---|---|---|---|
| 3 | 20 | 4 хв. 45 сек. | 94.3 |
| 3 | 60 | 12 хв. | 96.1 |
| 3 | 150 | 30 хв. | 96.4 |
| 4 | 20 | 5 хв. 32 сек. | 95.66 |
| 4 | 60 | 14 хв. 58 сек. | 97.95 |
| 4 | 150 | 39 хв. | 97.64 |
| 5 | 20 | 6 хв. 24 сек. | 94.6 |
| 5 | 60 | 19 хв. 25 сек. | 96.18 |
| 5 | 150 | 45 хв. | 97.1 |
| 6 | 20 | 7 хв. 43 сек. | 93.4 |
| 6 | 60 | 22 хв. 57 сек. | 96.2 |
| 6 | 150 | 53 хв. | 97.05 |

За результатами попереднього тестування було визначено оптимальну архітектуру нейронної мережі – чотири згорткові шари та 60 епох навчання, забезпечують найкраще співвідношення точності моделі та часу навчання. Отримана точність – 97.95 % та час навчання – 14 хвилин 58 секунд.

Графіки точності (рисунок 3.11а) та рівня помилки (рисунок 3.11б) у процесі навчання для моделі з чотирма шарами згортки та 60 епох навчання.




а) б)

Рисунок 3.11 – Графік зміни точності моделі та похибки у процесі навчання на 60 епохах із використанням мел-кепстральних коефіцієнтів для 4 шарів нейронної мережі зі згортковою архітектурою

Детальний опис нейронної мережі наведено в таблиці 3.9.

Таблиця 3.9 –Детальний опис кожного шару моделі

| Тип шару | Вихідна розмірність | Кількість параметрів |
|---|---|---|
| Conv2D | (None, 87, 13, 32) | 320 |
| MaxPooling2D | (None, 44, 7, 32) | 0 |
| BatchNormalization | (None, 44, 7, 32) | 128 |
| Conv2D | (None, 44, 7, 64) | 18496 |
| MaxPooling2D | (None, 22, 4, 64) | 0 |
| BatchNormalization | (None, 22, 4, 64) | 256 |
| Conv2D | (None, 22, 4, 128) | 73856 |
| MaxPooling2D | (None, 11, 2, 128) | 0 |
| BatchNormalization | (None, 11, 2, 128) | 512 |
| Conv2D | (None, 11, 2, 256) | 295168 |
| MaxPooling2D | (None, 6, 1, 256) | 0 |
| BatchNormalization | (None, 6, 1, 256) | 1024 |
| Flatten | (None, 1536) | 0 |
| Dense | (None, 128) | 196736 |
| Dropout | (None, 128) | 0 |
| Dense | (None, 61) | 7869 |

Впродовж роботи над реалізацією системи, було виконано процес донавчання існуючої моделі, що була попередньо навчена на шістдесяти спікерах та має точність – 97.95%. Для донавчання моделі було використано інший датасет, що має три тисячі аудіозаписів шістьох нових спікерів, та таку саму структуру, як і початковий.

Процес донавчання складається з наступних кроків (рисунок 3.12):

- завантаження нового набору аудіофайлів для формування датасету;
- нормалізація набору аудіофайлів для вирівнювання загальної кількості вибірок за секунду в рамках всього датасету;
- вилучення мел-кепстральних коефіцієнтів та збереження тринадцяти перших;

- формування датасету: об’єднання нового набору даних з попереднім, що використовувався для навчання початкової моделі;
- завантаження початкової, вже натренованої моделі;
- зміна вихідного шару в завантаженій моделі – потрібно збільшити кількість нейронів на кількість нових спікерів, у даному випадку на шість.
- донавчання моделі з використання об’єднаного датасету;
- тестування моделі для отримання точності.

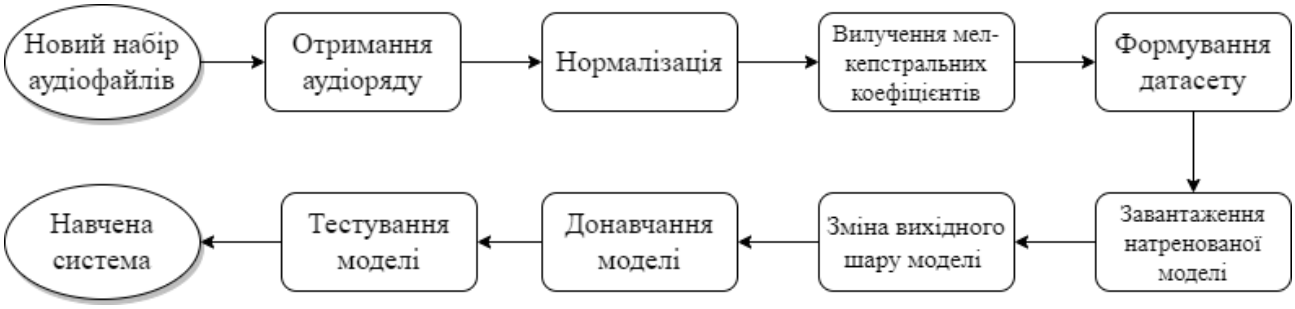


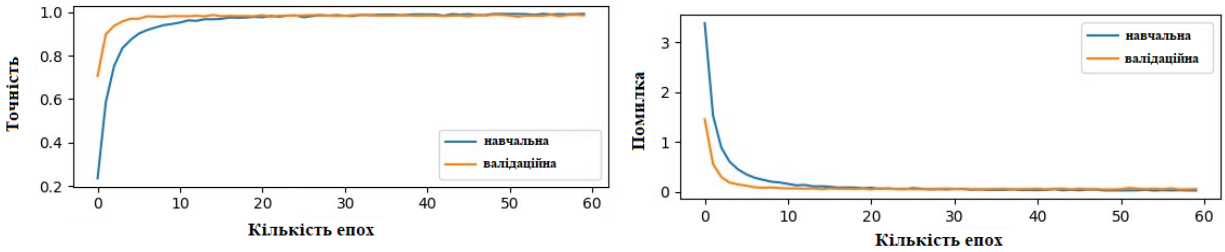

Рисунок 3.12 – Послідовність дій для процесу донавчання моделі

Результати порівняння точності моделі та часу навчання донавченої моделі та попередньо натренованої моделі наведені в таблиці 3.10.

Таблиця 3.10 – Порівняння результатів навчання моделей

| Модель | Кількість спікерів | Кількість епох | Час навчання | Точність у % |
|---|---|---|---|---|
| Існуюча | 60 | 60 | 14 хв. 58 сек. | 97.95 |
| Донавчена | 66 | 60 | 16 хв. 35 сек. | 99.5 |

Графіки точності (рисунок 3.12а) та рівень помилки (рисунок 3.12б) у процесі донавчання існуючої моделі з чотирма шарами згортки та 60 епох навчання.




а) б)

Рисунок 3.12 – Графік зміни точності моделі та похибки у процесі донавчання на 60 епохах із використанням мел-кепстральних коефіцієнтів для 4 шарів нейронної мережі зі згортковою архітектурою

Таким чином, в результаті додавання нових спікерів у систему, точність розпізнавання істотно збільшилась за рахунок більшого різноманіття даних для навчання. Було досягнуто рівень точності – 99.5%, час навчання залишився майже незмінним, що дозволяє додавати нових спікерів за відносно невеликий проміжок часу.

## 3.2 Експеримент 2. Розпізнавання цифрового паролю, названого голосом мовця

В ході виконання роботи було реалізовано систему автентифікації користувачів, що працює наступним чином: спікер називає цифровий пароль, який складається з чотирьох цифр, система перевіряє чи належить голос даному користувачеві, а також перевіряє правильність названого паролю, якщо і голос розпізнано правильно і пароль співпадає зі збереженим у базі даних, то користувач вважається автентифікованим, у іншому випадку доступ не надається.

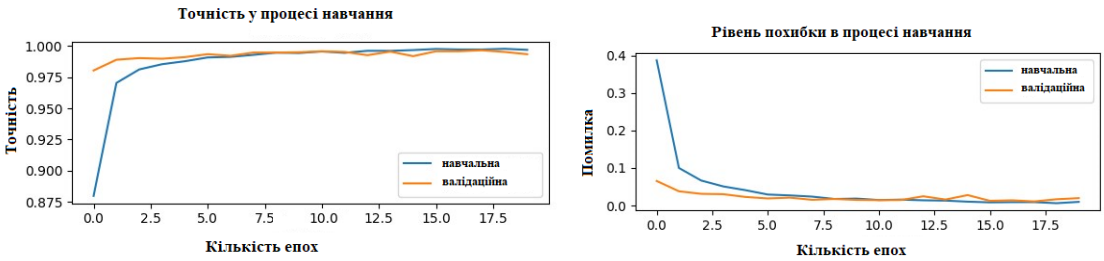
Для реалізації даної системи, було створено окрему нейронну модель, що дозволяє розпізнавати названу цифру. Нейронна мережа має схожу архітектуру моделі, що була застосована для розпізнавання спікера, яка складається з чотирьох згорткових шарів та 20-ти епох навчання. Єдиної відмінністю є останній(вихідний) шар, що у даному випадку має десять

нейронів, для кожної з десяти цифр, які потрібно розпізнати. Також, було використано спільний датасет, що складається з 30 300 аудіофайлів, 66 різних спікерів, але у даному випадку кожен аудіофайл має відповідність до однієї з десяти цифр. Отримані результати навчання наведені в таблиці 3.11.

Таблиця 3.11 – Результати навчання нейронної моделі для розпізнавання названої цифри

| Кількість шарів | Кількість епох | Час навчання | Точність у % |
|---|---|---|---|
| 4 | 20 | 6 хв. 22 сек. | 99.82 |
| 4 | 60 | 18 хв. 4 сек. | 99.4 |

В результаті кращу точність (Рисунок 3.13) має модель з меншою кількістю епох навчання, через те, що модель з шістдесятьма епохами навчання стикається з проблемою перенавчання на початковому наборі даних.




а) б)

Рисунок 3.13 – Графік зміни точності моделі та похибки у процесі донавчання на 60 епохах із використанням мел-кепстральних коефіцієнтів для 4 шарів нейронної мережі зі згортковою архітектурою

```
Execution time: 1194.94 ms
Password: 9876
Voice verified: [6, 6, 6, 6] True
b'$2b$12$SNQhSCrPW5NVI084b6gAteqM5ZN4hlL0K.pJW8fLva45OpgcNiH6K'
Password verified: True
```

Рисунок 3.14 – Результат роботи системи багатофакторної автентифікації

Результат роботи всієї системи багатофакторної автентифікації для одного з користувачів, якого було розпізнано за голосом та верифіковано правильність названого паролю (рисунок 3.14).

На основі наданих результатів консолі можна зробити узагальнений висновок про успішне виконання всіх етапів експерименту:

- процес зайняв 1194.94 мс, що свідчить про ефективність виконання;
- перевірка пароля "9876" виконана успішно;
- голосова верифікація голосу успішно завершена з відповідними значеннями [6, 6, 6, 6], що підтверджує автентичність голосу;
- пароль успішно захешовано за допомогою алгоритму bcrypt, що забезпечує високий рівень безпеки;
- пароль успішно підтверджено, що засвідчує його правильність та відповідність очікуваному значенню.

Таким чином, експеримент підтверджує, що система успішно виконує обробку і верифікацію як паролю, так і голосу, забезпечуючи належний рівень безпеки і надійності.

Говорячи про переваги запропонованої системи, варто відмітити, що максимальна точність верифікації голосового паролю була досягнута на нейромережевому класифікаторові зі згортковою архітектурою із чотирма шарами, навченому на 20 епохах навчання за 6 хвилин 22 секунди, використовуючи мел-кепстральні коефіцієнти для екстрації характерних ознак. Точність верифікації голосового паролю становить 99,82%.

Максимальна точність ідентифікації мовця за голосом становить 97,95%. Дана точність була досягнута на нейромережевому класифікаторові зі згортковою архітектурою із чотирма шарами, навченому на 60 епохах навчання за 14 хвилин 58 секунд.

Середня точність аутентифікації мовця на основі запропонованої системи двофакторної біометричної аутентифікації становить 98,75%

# ВИСНОВКИ

Метою роботи є підвищення точності голосової автентифікації мовця завдяки застосуванню двофакторної аутентифікації - ідентифікація мовця на основі спектрограми голосу та верифікації голосового пароля на основі сумісного використання двох нейромережевих моделей із згортковою архітектурою.

На основі запропонованої системи проведено ряд досліджень, а саме – проаналізовано та експериментально підтверджена ефективність одного із двох методів обробки аудіосигналу для виділення важливих характеристик (спектрограми та мел-кепстральні коефіцієнти), досліджена архітектура нейромережевих аналізаторів із згортковою архітектурою, вплив розподілу вхідного датасету на навчальну, тестову та валідаційну вибірки, проаналізована тривалість навчання та вимоги до апаратного забезпечення для забезпечення стійкості роботи системи.

Загальна оцінка результатів показує, що комбінуючи два запропонованих фактори, система підвищує рівень безпеки, оскільки для успішної автентифікації потрібно відповідати обом критеріям.

Модель, навчена з використанням мел-кепстральних коефіцієнтів, має вищу точність порівняно з моделлю, навченою зі спектрограмами на 17%. Окрім того, час навчання при використанні мел-кепстральних коефіцієнтів відбувається приблизно на 23 години 42 хвилини швидше.

Результати експериментів показали, що час навчання зменшується зі зменшенням відсотка навчальної вибірки. Це очікувано, оскільки менша кількість даних для навчання означає менше часу, необхідного для опрацювання моделі.

За результатами тестування було визначено оптимальну архітектуру нейронної мережі – чотири згорткові шари та 60 епох навчання, забезпечують найкраще співвідношення точності моделі та часу навчання.

Отримана точність – 97.95 % та час навчання – 14 хвилин 58 секунд досягнута на співвідношенні при 70% даних для навчання: 15% даних для валідації : 15% даних для тестування

На основі наданих результатів консолі можна зробити узагальнений висновок про успішне виконання всіх етапів експерименту:

- процес зайняв 1194.94 мс, що свідчить про ефективність виконання;
- перевірка пароля "9876" виконана успішно;
- голосова верифікація голосу успішно завершена з відповідними значеннями [6, 6, 6, 6], що підтверджує автентичність голосу;
- пароль успішно захешовано за допомогою алгоритму bcrypt, що забезпечує високий рівень безпеки;
- пароль успішно підтверджено, що засвідчує його правильність та відповідність очікуваному значенню.

Таким чином, експеримент підтверджує, що система успішно виконує обробку і верифікацію як паролю, так і голосу, забезпечуючи належний рівень безпеки і надійності.

Говорячи про переваги запропонованої системи, варто відмітити, що максимальна точність верифікації голосового паролю була досягнута на нейромережевому класифікаторові зі згортковою архітектурою із чотирма шарами, навченому на 20 епохах навчання за 6 хвилин 22 секунди, використовуючи мел-кепстральні коефіцієнти для екстрації характерних ознак. Точність верифікації голосового паролю становить 99,82%.

Максимальна точність ідентифікації мовця за голосом становить 97,95%. Дана точність була досягнута на нейромережевому класифікаторові зі згортковою архітектурою із чотирма шарами, навченому на 60 епохах навчання за 14 хвилин 58 секунд.

# ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Olesia Barkovska, Yaroslav Liapin, Tetiana Muzyka, Ihor Ryndyk, Pavlo Botnar, "GAZE DIRECTION MONITORING MODEL IN COMPUTER SYSTEM FOR ACADEMIC PERFORMANCE ASSESSMENT. CIVIL LAW ASPECT" , Information Technologies and Learning Tools, 2024, Vol 99, №1, 63-75pp DOI: 10.33407/itlt.v99i1.5503.

2. Shaheed K. et al. A systematic review on physiological-based biometric recognition systems: current and future trends //Archives of Computational Methods in Engineering. – 2021. – С. 1-44.

3. Alsaadi I. M. Study on most popular behavioral biometrics, advantages, disadvantages and recent applications: A review //Int. J. Sci. Technol. Res. – 2021. – Т. 10. – №. 1.

4. Palma D. et al. Biometric-based human recognition systems: an overview //Recent Advances in Biometrics. – 2022. – Т. 27. – С. 1-21.

5. Khan S. et al. Biometric systems utilising health data from wearable devices: applications and future challenges in computer security //ACM Computing Surveys (CSUR). – 2020. – Т. 53. – №. 4. – С. 1-29.

6. Marani M. et al. The role of biometric in banking: A review //EAI Endorsed Transactions on AI and Robotics. – 2023. – Т. 2. – №. 1.

7. Hernandez-de-Menendez M. et al. Biometric applications in education //International Journal on Interactive Design and Manufacturing (IJIDeM). – 2021. – Т. 15. – С. 365-380.

8. Jahangir R. et al. Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges //Expert Systems with Applications. – 2021. – Т. 171. – С. 114591.

9. Abdul Z. K., Al-Talabani A. K. Mel frequency cepstral coefficient and its applications: A review //IEEE Access. – 2022. – Т. 10. – С. 122136-122158.

10. Hanifa R. M., Isa K., Mohamad S. A review on speaker recognition:

Technology and challenges //Computers & Electrical Engineering. – 2021. – T. 90. – C. 107005.

11. G. Krishna, C. Tran, M. Carnahan, Y. Han, and A. H. Tew k, «Gen erating EEG features from acoustic features», in Proc. 28th Eur. Signal Process. Conf. (EUSIPCO), Jan. 2021, pp. 11001104, doi: 10.23919/Eusipco47968.2020.9287498.

12. A. Sharma, M. Amarnath, and P. Kankar, «Feature extraction and fault severity classi cation in ball bearings», J. Vib. Control, vol. 22, no. 1, pp. 176192, 2016, doi: 10.1177/1077546314528021.

13. C. Cremona, S. Yang, and L. Dieleman, «Overview of feature extraction techniques», in Proc. 3rd Int. Oper. Modal Anal. Conf., Sep. 2009, pp. 381390 doi: 10.2139/ssrn.3568122.

14. X. Zhao and D. Wang, Analyzing noise robustness of MFCC and GFCC features in speaker identi cation, in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., May 2013, pp. 72047208, doi: 10.1109/ICASSP.2013.6639061.

15. M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, «Automatic speech recognition and speech variability: A review», Speech Commun., vol. 49, nos. 1011, pp. 763786, Oct. 2007, doi: 10.1016/j.specom.2007.02.006.

16. K. S. Rao and K. E. Manjunath, Speech Recognition Using Articulatory and Excitation Source Features. Cham, Switzerland: Springer, 2017.

17. H. A. Feldman, N. Kaiser, and J. A. Peacock, «Power-spectrum analysis of three-dimensional redshift surveys», Astrophys. J., vol. 426, p. 23, May 1994, doi: 10.1086/174036.

18. H. Yin, V. Hohmann, and C. Nadeu, «Acoustic features for speech recognition based on gammatone lterbank and instantaneous fre quency», Speech Commun., vol. 53, no. 5, pp. 707715, May 2011, doi: 10.1016/j.specom.2010.04.008.

19. G. Strang, «The discrete cosine transform», SIAM Rev., vol. 41, no. 1,

pp. 135147, 1999, doi: 10.1137/S0036144598336745.

20. S. Camacho, D. M. Ballesteros, and D. Renza, „Fake speech recognition using deep learning”, in Proc. Workshop Eng. Appl., 2021, pp. 3848.

21. J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, and D. Erro, „A cross vocoder study of speaker independent synthetic speech detection using phase information”, in Proc. Annu. Conf. Int. Speech Commun. Assoc., 2014, pp. 16631667, doi: 10.21437/interspeech.2014-393.

22. T. Liu, D. Yan, R. Wang, N. Yan, and G. Chen, „Identification of fake stereo audio using SVM and CNN”, Information, vol. 12, no. 7, p. 263, Jun. 2021, doi: 10.3390/info12070263.

23. A.K.Singh,P.Singh, and K.Nathwani, „Using deep learning techniques and inferential speech statistics for AI synthesised speech recognition”, 2021, arXiv:2107.11412.

24. MK A. H. K., Aithal P. S. Voice Biometric Systems for User Identification and Authentication–A Literature Review //International Journal of Applied Engineering and Management Letters (IJAEML). – 2022. – Т. 6. – №. 1. – С. 198-209. **DOI:** https://doi.org/10.47992/IJAEML.2581.7000.0131

25. Faqir R. S. A. Digital Criminal Investigations in the Era of Artificial Intelligence: A Comprehensive Overview //International Journal of Cyber Criminology. – 2023. – Т. 17. – №. 2. – С. 77–94-77–94. DOI:10.5281/zenodo.4766706

26. Yang W. et al. Biometrics for internet-of-things security: A review //Sensors. – 2021. – Т. 21. – №. 18. – С. 6163. https://doi.org/10.3390/s21186163

27. Lien C. W., Vhaduri S. Challenges and opportunities of biometric user authentication in the age of iot: A survey //ACM Computing Surveys. – 2023. – Т. 56. – №. 1. – С. 1-37. https://doi.org/10.1145/3603705

28. P. Spachos, S. Gregori and M. J. Deen, "Voice Activated IoT Devices for Healthcare: Design Challenges and Emerging Applications," in *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 69, no. 7, pp. 3101-3107, July 2022, doi: 10.1109/TCSII.2022.3179680.

29. Havrashenko A., Barkovska O. Analysis of text augmentation algorithms in artificial language machine translation systems //Advanced Information Systems. – 2023. – Т. 7. – №. 1. – С. 47-53. https://doi.org/10.20998/2522-9052.2023.1.08

30. Tandel N. H., Prajapati H. B., Dabhi V. K. Voice recognition and voice comparison using machine learning techniques: A survey //2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). – IEEE, 2020. – С. 459-465. **DOI:** 10.1109/ICACCS48705.2020.9074184

31. N. Chandolikar, C. Joshi, P. Roy, A. Gawas and M. Vishwakarma, "Voice Recognition: A Comprehensive Survey," *2022 International Mobile and Embedded Technology Conference (MECON)*, Noida, India, 2022, pp. 45-51, doi: 10.1109/MECON53876.2022.9751903.