Київський національний університет імені Тараса Шевченка
Факультет інформаційних технологій
Кафедра програмних систем і технологій

УДК 004.056.55 *На правах рукопису*

ВИПУСКНА КВАЛІФІКАЦІЙНА МАГІСТЕРСЬКА РОБОТА

Тема: “ Ідентифікація голосу диктора ”

Спеціальність – 121 “Інженерія програмного забезпечення”

ПОЯСНЮВАЛЬНА ЗАПИСКА

БР.ІПЗ – 30.00.00.000

Студент

ІПЗм-21 _____/Станіслав СИРОЇД/

Науковий керівник

к.т.н., доц. _____/Тетяна КОВАЛЮК/

Консультант

з питань нормоконтролю

к.т.н., асистент _____/ Анастасія ВЕЧЕРКОВСЬКА /

Допускається до захисту

Завідувач кафедри

д.т.н., проф. _____/Олексій БИЧКОВ/

Київ – 2022

Рішенням Екзаменаційної комісії

випускна кваліфікаційна робота студента

_____

захищена з оцінкою

_____

Голова Екзаменаційної комісії

д.т.н., проф. Андрій БОНДАРЧУК

Київський національний університет імені Тараса Шевченка

Факультет інформаційних технологій

Кафедра програмних систем і технологій

Спеціальність 121 “Інженерія програмного забезпечення”

ЗАТВЕРДЖУЮ

Завідувач кафедри

програмних систем і технологій

_____

«_____»_____2022 р.

ЗАВДАННЯ

НА ВИПУСКНУ КВАЛІФІКАЦІЙНУ МАГІСТЕРСЬКУ РОБОТУ

СТУДЕНТУ

Сироїда Станіслава Олеговича

(прізвище, ім’я, по батькові)

1. Тема випускної кваліфікаційної магістерської роботи: “ Голосова ідентифікація диктора”

затверджена наказом вищого навчального закладу від "__" грудня 2021 року №

2. Строк здачі студентом закінченої роботи: з _____ до _____

3. Вихідні дані до роботи: з використанням відкритих бібліотек tensorflow, numpy розробити програмне забезпечення що ідентифікує диктора по голосу

4.Зміст пояснювальної записки (перелік питань, що їх належить розробити):

1) сучасні системи та технології розпізнавання голосу;

2) методи реалізації нейронних мереж для розпізнавання голосу;

3) програмне забезпечення голосової ідентифікації

5. Перелік графічного матеріалу (з точним забезпеченням обов'язкових креслень):

1) діаграма прецедентів використання системи;

2) Функціональна схема системи голосової ідентифікації/верифікації особи

3) Схема навчання та функціонування UBM-GMM;

4) Схема алгоритму роботи системи

6. Консультанти з роботи із зазначенням розділів роботи, що їх стосуються

Розділ	Консультант	Підпис, дата	
		Завдання видав	Завдання прийняв
I	Ковалюк Т.В.		
I I	Ковалюк Т.В.		
I I I	Ковалюк Т.В.		
I V	Ковалюк Т.В.		

7. Дата видачі завдання «___» _____2022 р.

Керівник кваліфікаційної роботи _____ к.т.н. доц. Тетяна Ковалюк

(підпис, дата)

Завдання прийняв до виконання _____Станіслав Сироїд

(підпис, дата)

КАЛЕНДАРНИЙ ПЛАН

/п	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1	Аналітична частина	16.12.2021-24.12.2021	Виконано
2	Огляд та порівняння форм біометричної ідентифікації	05.01.2022-15.01.2022	Виконано
3	Огляд методів та існуючих підходів для голосової ідентифікації	15.01.2022-30.01.2022	Виконано
4	Огляд існуючих сервісів та сфер застосування	31.01.2022-12.02.2022	Виконано
5	Теоретична частина	13.02.2022-30.02.2022	Виконано
6	Постановка задачі голосової ідентифікації	05.03.2022-16.03.2022	Виконано
7	Опис способу представлення голосів	18.03.2022-21.03.2022	Виконано
8	Опис способу порівняння голосів	22.03.2022-04.03.2022	Виконано
9	Опис використовуваних метрик якості	05.03.2022-15.03.2022	Виконано
10	Інженерна частина	25.03.2022-15.04.2022	Виконано
11	Вибір мови та допоміжних модулів для програмної реалізації системи	16.04.2022-20.04.2022	Виконано
12	Програмна реалізація моделей та алгоритмів	25.04.2022-30.04.2022	Виконано

13	Технологічна та практична частина	30.04.2022-15.05.2022	Виконано
14	Опис тестових даних, обсягу та особливостей	15.05.2022-16.05.2022	Виконано
15	Тестування роботи системи ідентифікації	16.05.2022-18.05.2022	Виконано
16	Оформлення пояснювальної записки (ПЗ) та ілюстративного матеріалу для доповіді.	24.05.2022	Виконано

РЕФЕРАТ

Пояснювальна записка до кваліфікаційної роботи “Ідентифікація голосу диктора” : 75 с., 20 рис., 19 літературних джерел, 1 додаток.

НЕЙРОННА МЕРЕЖА, ІДЕНТИФІКАЦІЯ ГОЛОСУ, МАШИННЕ НАВЧАННЯ.

Мета дослідження – покращити методи обробки звуку для реалізації програмного забезпечення розпізнавання голосу.

Об’єкт дослідження – голоси людей, що диктують текст.

Предмет дослідження – алгоритми та методи для розробки системи що ідентифікує голос.

Розпізнавання диктора — це завдання ідентифікувати людей за їхніми голосами. Останнім часом глибоке навчання кардинально зробило революцію розпізнавання мовця. У цій роботі розглянуто кілька основних підзадач розпізнавання диктора, включаючи перевірку мовця, ідентифікацію, діаризацію та надійне розпізнавання мовця, з акцентом на методах глибокого навчання.

У роботі розглядаються різні існуючі підходи до вирішення задачі ідентифікації за голосом, класифікація різних типів біометричної ідентифікації, методи машинного навчання для ідентифікації, що використовуються, існуючі готові системи та рішення на ринку, затребуваність системи голосової ідентифікації на ринку біометричних систем ідентифікації. У роботі наводиться опис розробленого методу ідентифікації мовця, етапи обробки аудіофайлу та способу представлення голосу у вигляді ознак вектора. Також у роботі дається опис розробки системи, вибір допоміжних бібліотек та модулів для розробки, опис

реалізованих функцій. Наводяться результати експериментальних досліджень роботи реалізованої системи ідентифікації диктора оцінюється вплив архітектури моделі на якість роботи. Після цього підбиваються підсумки проведеної роботи з коротким описом результатів по кожному розділу. Продуктивність системи показала середній рівень розпізнавання 77% для висловлювань із 5 слів і 43% для довжини кількість висловлювань було збільшено до 20-слівних висловлювань для випадків навчених мовленнєвих висловлювань. З невідомим висловлювань, рівень розпізнавання 18% досягнуто для висловлювань з 20 слів.

Програмне забезпечення реалізовано мовою Python та при використанні загальнодоступних фреймворків. У даній роботі використовується глибоке навчання, біболтеки Numpy, TensorFlow ,Keras.

Набір голосивих даних складається з 4040 wav файлів. Голосові записи було отримано з відкритих бібліотек : Audio MNIST, CHIME- Home, EMODB, MS SNSD, PDS.

ABSTRACT

Explanatory note to the qualification work "Identification of the speaker's voice": 75 pages, 20 figures, 19 references, 1 appendix.

NEURAL NETWORK, VOICE IDENTIFICATION, MACHINE LEARNING.

The purpose of the study - analysis and selection of sound processing methods for the implementation of voice recognition software.

The object of research is the voices of people who dictate the text.

The subject of research - algorithms and methods for developing a system that identifies the voice.

Speaker recognition is the task of identifying people by their voices. Recently, deep learning has revolutionized speaker recognition. However, there are no comprehensive reviews of exciting progress. This paper discusses several key sub-objectives of speaker recognition, including speaker verification, identification, diarrhea, and reliable speaker recognition, with an emphasis on in-depth learning methods. The ability of recognition systems to correctly recognize speakers based on the distribution of their speech signals largely depends on how the recognition system can teach the parameters of the model /.

The paper considers various existing approaches to solving the problem of voice identification, classification of different types of biometric identification, machine learning methods for identification used, existing ready-made systems and solutions on the market, demand for voice identification system in the market of biometric identification systems. The paper describes the developed method of speaker identification, stages of audio file processing and method of voice representation in the form of vector features. The paper also describes the development of the system, the choice of auxiliary libraries and modules for development, a description of the implemented functions. The results of experimental researches of work of the realized system of identification of the announcer are resulted, influence of architecture of model on quality of work is estimated. After that, the results of the work are summed up with a brief description of the results for each section.

The performance of the system showed an average level of recognition of 77% for 5-word utterances and 43% for length, the number of utterances was increased to 20-verbal utterances for cases of learned speech utterances. With

unknown utterances, a recognition rate of 18% was achieved for utterances of 20 words.

The software is implemented in Python and using public frameworks. This work uses deep learning, libraries Numpy, TensorFlow, Keras.
The set of voice data consists of 4040 wav files. Voice recordings were obtained from the public .

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СКОРОЧЕНЬ, ТЕРМІНІВ

WAV – формат аудіофайлу, розроблений компаніями Microsoft та IBM.

MFCC – Мелчастотні кепстральні коефіцієнти (англ. Mel-frequency cepstral coefficients (MFCCs)) — це коефіцієнти мел-частотного кепстру

GMM – Модель гаусової суміші

ЄБС – Єдиної біометричної системи

ЄСІА - Єдиної системи ідентифікації та аутентифікації

FFT - перетворення Фур'є

DCT - алгоритм стиснення даних.

UBM – Універсальна фонова модель (Universal Background Model)

ВСТУП

Голос мовця містить у собі його особисті риси, зумовлені унікальним будовою органів вимови, наприклад, унікальний голосової формою тракту, розміром гортані, особливостями дикції, акцентом і ритмом, індивідуальною манерою мови. Таким чином, можна автоматично ідентифікувати того, хто говорить по його голосу через комп'ютер.

Ідентифікація того, хто говорить - це фундаментальне завдання обробки мови і голосу, яка знаходить широке застосування в реальних задачах. Наприклад, вона використовується для підтвердження особи на гарячих лініях та при дзвінках до call-центрів, при голосовій аутентифікації в особистих інтелектуальних пристроях, таких як стільникові телефони, автомобілі та ноутбуки. Така ідентифікація не тільки гарантує безпеку транзакцій банківської торгівлі та віддалених платежів, але також автоматизує ряд завдань, веде до скорочення часу та зростання лояльності. Біометрична ідентифікація на основі голосу також може широко застосовуватися в судовій практиці для розслідування підозрюваного щодо визнання його винним, для спостереження та автоматичної ідентифікації особи на відстані (можливо, без відома самої особи). Вона також може бути частиною завдання автоматичного визначення кількості тих, хто говорить у записаному аудіофрагменті та сегментації фрагмента на частини, в яких звучить голос певного диктора.

Незважаючи на те, що було досягнуто багато технологічних успіхів у галузі біометричної ідентифікації, залишається ще багато дослідницьких проблем, які необхідно вирішити. Розробка системи для ідентифікації мовця дозволить автоматизувати та прискорити вирішення багатьох повсякденних завдань.

Метою даної є побудова системи, що виконує автоматичну ідентифікацію диктора з його промови в аудіосигналі. У межах роботи вирішуються такі завдання:

• Отримання навчальної вибірки.

• Вивчення та розробка методу ідентифікації диктора.

• Проектування та розробка оптимальної архітектури обраної моделі.

• Тестування роботи реалізованої системи.

У першому розділі наводиться опис та порівняльний аналіз різних існуючих підходів до ідентифікації диктора, опис можливих рішень, їх переваг та недоліків.

У другому розділі ставиться завдання ідентифікації диктора, виділяються підзадачі, наводиться опис методу ідентифікації та дослідження метрик якості розробленої моделі.

У третьому розділі наводиться список вибраних допоміжних модулів, а також вказується вибрана мова та програмне середовище. Описуються реалізовані функції та елементи системи.

У четвертому розділі наводяться результати експериментальних досліджень впливу параметрів архітектури обраної моделі на якість роботи, наводяться результати тестування системи ідентифікації, внаслідок чого визначається оптимальний метод вирішення поставленого завдання.

У висновку підбиваються підсумки проведеної роботи з коротким описом результатів по кожному розділу.

РОЗДІЛ 1. АНАЛІЗ ІСНУЮЧИХ ПІДХОДІВ ДО ГОЛОСОВОЇ ІДЕНТИФІКАЦІЇ

Наразі існує кілька платних закритих систем голосової біометричної ідентифікації, спроектованих на основі використання різних методів машинного навчання. Якість роботи систем не є ідеальною і продовжує вдосконалюватися в даний час. Інтерес до завдання зростає.

Незважаючи на існування готових рішень, всі системи залишаються закритими, їх методи роботи ніде не описані і зберігаються в таємниці. Описані в розділі існуючі методи розв'язання задачі ідентифікації диктора за його промовою є лише зразковим описом деяких існуючих рішень без деталей реалізації та опису недоліків обраних методів.

1.1. Огляд та порівняння форм біометричної ідентифікації

Біометрія – це наука, заснована на вимірі та описі фізичних характеристик живих істот. У рамках завдань автоматичної ідентифікації особистості біометричні дані – це унікальні біологічні та фізіологічні характеристики, які дозволяють встановити особу людини. Біометричні дані можна розділити на:

• фізіологічні (що стосуються фізичного тіла – відбитки пальців, долоня руки, райдужна оболонка, голос, ДНК тощо)

• поведінкові (що стосуються поведінки людини – хода, мова, почерк тощо).

Розглядаючи динаміку розвитку світового ринку біометричних технологій, можна побачити значне стабільне зростання річного доходу ринку від біометрії у всіх регіонах планети. Виходячи з цього можна дійти

невтішного висновку, що використання біометричних технологій стає дедалі популярним, які рішення залишаються затребуваними над ринком.

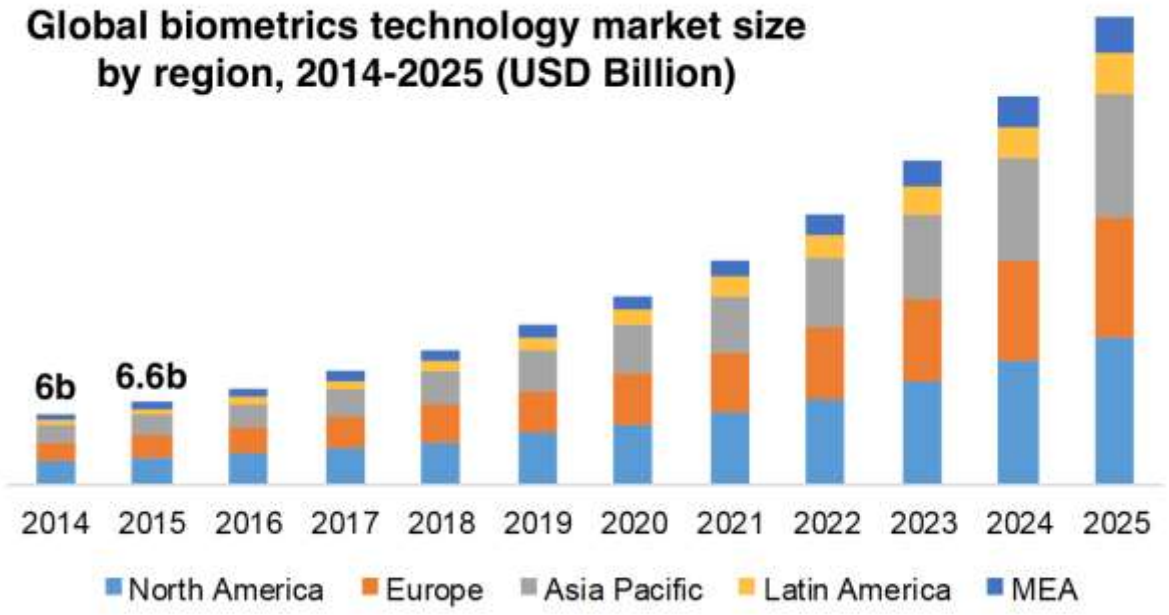


Рис.1 Річний прибуток від біометрії по регіонах. Прогноз. Світовий ринок: 2014-2025

Серед найпоширеніших типів фізіологічної біометрії – зображення обличчя людини, відбитки пальця, малюнок райдужної оболонки ока, голос, малюнки вен на долоні тощо.

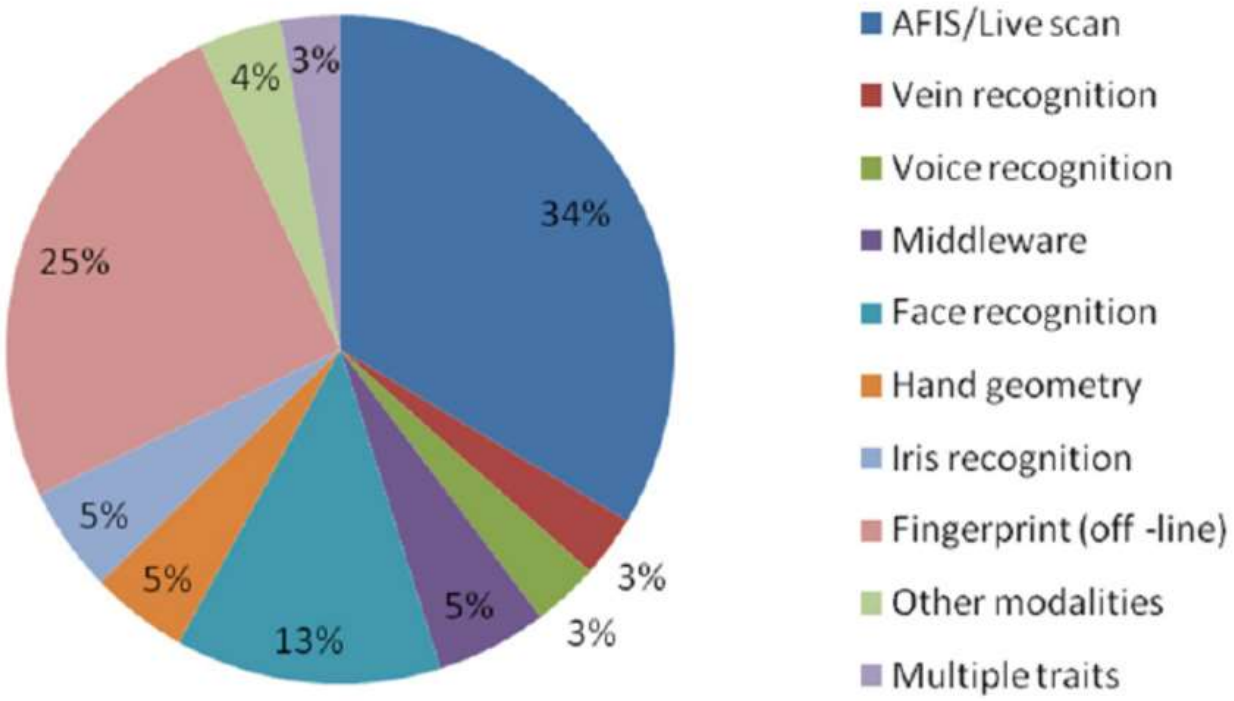


Рис.2 Світовий ринок біометричних технологій. Розподіл форм ідентифікації

Можна спостерігати переважання на ринку використання технології ідентифікації за відбитками пальця, що зрозуміло, оскільки за надійністю роботи саме вона займає лідируючу позицію. Однак такий метод ідентифікації є контактним, що тягне за собою деякі обмеження, в тому числі в умовах пандемії. Також можна побачити, частка ринку, використовує відбитки пальця як метод ідентифікації, поступово зменшується, поступаючись іншим методам.

В окремій доповіді Allied Market Research прогнозує, що ринок біометричних датчиків, який оцінювався в 1,16 мільярда доларів у 2020 році, зросте до 3,31 мільярда доларів до 2030 року при CAGR 11,8 відсотка.

Цей прогноз у поєднанні з загальним звітом про ринок передбачає, що вартість програмного забезпечення, інтеграції та послуг за вісім років загалом становить понад 123 мільярди доларів, що, можливо, дивно, враховуючи очікуване зростання сегменту апаратного забезпечення.

Фактори, які, за прогнозами, сприятимуть цьому росту, включають зростання загроз ідентичності та нові можливості використання біометричних даних, які є порушеннями конфіденційності, високий попит на безконтактні біометричні системи, а також біометричні носії.

З точки зору продуктивності за регіонами, Азіатсько-Тихоокеанський регіон збереже лідерство на ринку, за ним слідуватиме Північна Америка, причому перша, як очікується, зафіксує найшвидший CAGR у 13,2 відсотка.

У недавньому звіті Allied Market Research також прогнозується оцінка ринку автоматизованої системи ідентифікації відбитків пальців (AFIS) до 2030 року в 68 мільярдів доларів.

Таблиця 1. Значення помилок ідентифікації різних біометричних модальностей

Порівняльна характеристика біометричних систем	Відбиток пальця	Голос	Райдужна оболонка	Лице
Ймовірність «допуску чужого»	2,2%	1.0%	0.1%	0.1%
Ймовірність «відказу свому»	2,2%	до 4.0%	1.1-1.4%	до 4.0%

Вартість системи	Висока	Низька	Дуже висока	Висока

Проте темпи розвитку технологій голосової ідентифікації є досить великими. За прогнозами середньорічний темп зростання ринку голосових біометричних систем до 2022 року досягнуть більше 21% і займатимуть друге місце після технології, побудованої на використанні райдужної оболонки ока (рисунок нижче). Розвиток голосових технологій можна пояснити низькою вартістю системи проти іншими видами біометричних технологій.

Глобальний ринок біометричної аутентифікації та ідентифікації залишатиметься на піку своєї зрілості протягом найближчих років і продовжуватиме поширюватися завдяки новій сфері застосування, але такі сегменти, як відбитки пальців та розпізнавання обличчя, будуть свідками помітних змін у поведінці споживачів через COVID-19. . Інтеграція кількох технологій, таких як штучний інтелект (ШІ) та хмарні обчислення, у систему біометричної аутентифікації значно розширила бізнес-можливості ринку.

Останній тренд — біометрична аутентифікація, і ринок ідентифікації пропонує аутентифікацію як послугу, а не рішення «під ключ». Компанії вивчають можливості хмарних обчислень і хмарного сховища, щоб знизити загальні витрати на експлуатацію, пропонуючи споживачам бажані заходи безпеки. Зменшення надійності апаратного забезпечення та підвищення зрілості, оперативних можливостей та точності програмного забезпечення визначать сфери зростання ринку.

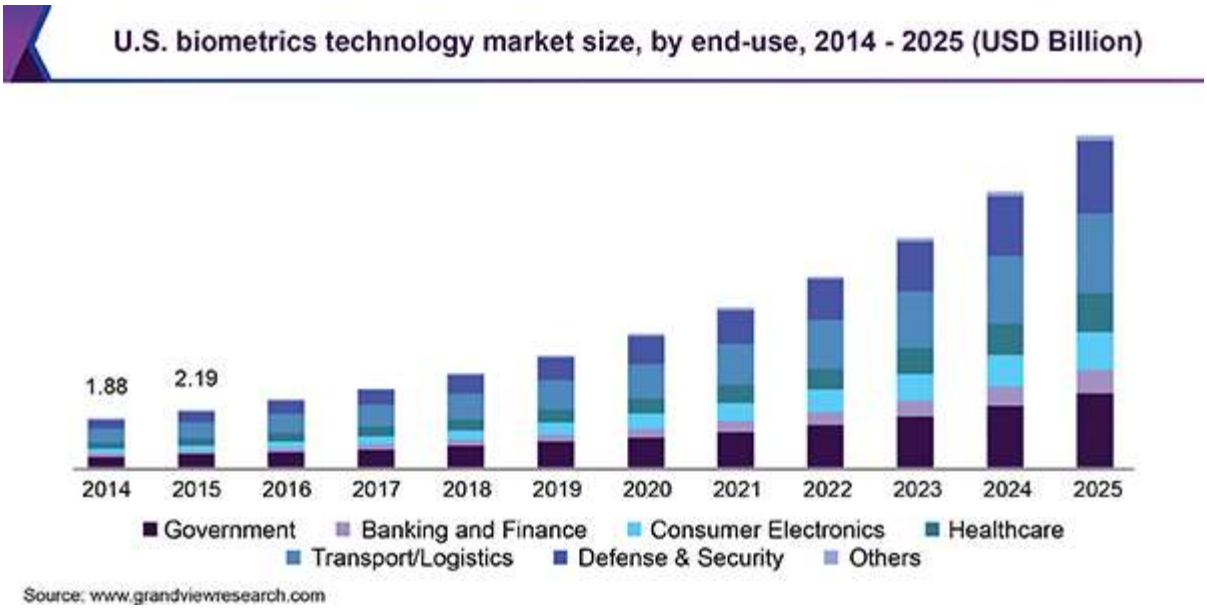


Рис.3 Прогноз середньорічного темпу зростання ринку біометричних систем у розрізі технологій до 2022 року.

Таким чином, технології на основі голосової біометрії зараз не є лідерами затребуваності на ринку, однак вони швидко розвиваються і мають низку переваг перед іншими біометричними технологіями, що робить їх потенційними лідерами у майбутньому [1].

1.2 *Існуючі системи та їх аналіз*

Нижче наведено список найпопулярнішого програмного забезпечення для голосового або мовного диктування, яке використовується користувачами по всьому світу, з повною інформацією.

Speech Recognition Software	Best For	Platform	Free Trial	Price
Dragon Professional	Overall dictation and voice recognition.	Windows OS	Yes	Dragon-Home is $150, Professional Individual is $300, Legal Individual is $500.
Dragon Anywhere	Professional speech recognition for your mobile.	Android & iOS devices	Yes	$15 per month or $150 per year.
Cortana	Windows devices.	Windows 10, iOS, Android, and Windows phone devices	-	Free
Amazon Lex	Creating Chatbot.	Used in the applications.	No	Based on the no. of speech requests processed.

Рис. 4 Існуючі системи голосої ідентифікації

Dragon Anywhere — це програма для диктування від Nuance для пристроїв iOS. Це хмарне рішення. Для диктування та редагування документів будь-якої довжини. Він надає вам хмарний інструмент розпізнавання мовлення. Це означає, що ви зможете отримати доступ до версій документів навіть з мобільного телефону. Ця програма дозволить вам зберегти ваш текст в Evernote. Також підтримуються такі формати документів, як .docx, .rtf, .rrtfd і текст.

Amazon Lex

Amazon Lex використовується в програмах для створення розмовного інтерфейсу. Розроблений бот можна використовувати на платформі Chat, пристроях IoT та мобільних клієнтах.

Cortana

Cortana — це віртуальний помічник, який постачається з системами Windows 10 і Windows phone. Він також доступний для пристроїв Android та iOS.

Як ми можемо бачити по рисунку 4, що відображений вище, більшість систем є платними та з високоми цінами для нашого регіону, більше того не всі системи пропонують конкретно вбудовану ідентифікацію диктора. Ці недоліки і є виправлені в системі що представляється в даній роботі.

1.3*Огляд методів та існуючих підходів для голосової ідентифікації*

Існує два типи систем ідентифікації того, хто говорить:

• Текстозалежна. Якщо вимовлений текст має бути однаковим для реєстрації диктора в основі та для ідентифікації, то цей підхід називається текстозалежним розпізнаванням. У такій системі ключове слово/фраза, що вимовляється, можуть бути або спільними для всіх дикторів, або унікальними (наприклад, унікальний пароль користувача).

• Текстонезалежна. Незалежні від тексту системи найчастіше використовуються для ідентифікації диктора, оскільки вони вимагають

меншої уваги з боку того, хто говорить. У цьому випадку текст при реєстрації та ідентифікації відрізняється, система вважається інваріантною до мови.

Розпізнавання мовця:

Ідентифікація голосу

Ідентифікація голосу => визначення, який зареєстрований мовець забезпечує дане висловлювання..

Перевірка голосом

Голосова перевірка => прийняття або відхилення заявки на ідентифікацію раніше ідентифікованого мовця.

Завдання ідентифікації людини за голосом, незалежно від обраного шляху її вирішення, можна розбити такі основні етапи:

1) Вилучення ознак з аудіо сигналу

2) Побудова моделі диктора, на основі отриманих на попередньому етапі ознак.

Процес визначення людини за голосом з представлених у системі зразків голосів, у всіх методах полягає у пошуку найбільш відповідного за моделлю диктора, виходячи з обраних критеріїв.

Мовні сигнали – це звукові коливання, які поширюються у повітряному середовищі. Вони характеризуються частотою (числом коливань на секунду), інтенсивністю (амплітудою коливань) та тривалістю. [2] Протягом всього мовного сигналу ці характеристики зазнають змін, і фіксуються за допомогою електронно-акустичних приладів, таких як осцилограф, спектрограф. Потім, за допомогою аналого-цифрового перетворювача, аналоговий мовний сигнал

переводиться в цифровий, який на ЕОМ представлений у вигляді WAV-файлу, з якого відбувається збір мовних ознак.

1.4 *Вибір системи ознак*

Характерні особливості голосу людини формується на основі розміру та форми мовного тракту, динамікою його зміни, пружністю тощо. Основними та найпопулярнішими характерними параметрами для опису голосу та мови людини, що враховують описані особливості, є MFCC. Практично у всіх дослідженнях робота ведеться з ними, а також іноді доповнюється набором з перших та других дельта-функцій – локальних оцінок похідних (у дискретному випадку похідна функції дорівнює різниці значень функції у послідовних точках).

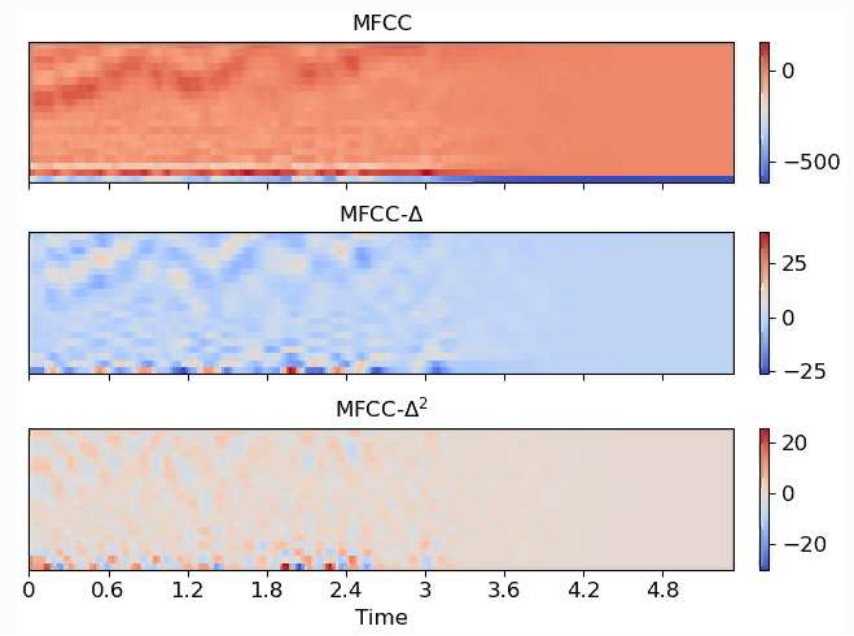


Рис.5 Візуалізація MFCC та дельта-функцій першого, другого порядку

1.5 *Вибір класифікатора ознак*

Сучасні системи голосової ідентифікації та верифікації працюють у двох режимах.

• Режим навчання. Виділяються характерні ознаки голоси людини, формується його голосова модель (голосовий відбиток) на основі цих ознак та виконується збереження моделі у базі даних.

• Робочий режим. Виділяються характерні ознаки голосового сигналу людини та виконується пошук у базі даних голосової моделі, що відповідає цим ознакам (ідентифікація особи).

1.6 *GMM*

В даний час найбільш результативним підходом до вирішення задач текстонезалежної ідентифікації особистості є побудова голосових моделей на основі моделей гаусових сумішей [3] (Gaussian Mixture Model, GMM). Самі моделі будуються з урахуванням деякого набору голосових ознак. Найбільш поширеним методом побудови голосових ознак є формування вектора крейдчастих кепстральних коефіцієнтів (Mel-Frequency Cepstral Coefficient, MFCC) з голосового запису.

Однак, незважаючи на досить хороші результати роботи в «лабораторних умовах», метод GMM-MFCC не може бути використаний для побудови реальних систем голосової ідентифікації, оскільки система має сильну залежність результатів від виду навчального матеріалу (на основі якого складається база голосових моделей та фонова модель), та умов запису голосового сигналу. Також недоліком є відносно велика потреба у великій кількості навчального матеріалу.

1.7 *Мережа Кохонена*

Як альтернативний спосіб порівняння голосових відбитків може використовуватися нейромережеве порівняння за допомогою

самоорганізованої мережі Кохонена. Мережа навчається без вчителя і має дві фази роботи – навчання та ідентифікацію. У процесі навчання навчальний алгоритм підлаштовує ваги мережі те щоб вийшли узгоджені вихідні вектори. Таким чином, процес навчання виділяє статистичні властивості навчальної множини і групує подібні вектори до класів. Пред'явлення на вхід вектора даного класу дає певний вихід. У ході ідентифікації відбуватиметься обчислення сум у вузлах мережі після подачі вектора на вхід, потім активація знайденого вузла та отримання класифікованого вектора.

1.8 *Глибокі нейронні мережі*

Окрім мережі Кохонена існуючі роботи засновані на використанні як класифікаторів глибокої нейронної мережі так й нейронної мережі з шаром згортки. Вона має кілька прихованих шарів, вхідний шар того ж розміру, що і вектор ознак, і вихідний шар з кількістю нейронів, що дорівнює кількості зареєстрованих дикторів у системі. [4] Незважаючи на різні способи оптимізації способу навчання мережі, такий підхід унеможливлює швидке додавання нових дикторів до системи. Для цього необхідно не лише наново навчити всю мережу, але й змінити її архітектуру (як мінімум – збільшити вихідний шар). Глибокі нейронні мережі останнім часом стали стандартним інструментом для вирішення різноманітних проблем комп’ютерного зору. У той час як навчання нейронної мережі виходить за рамки OpenVX, імпорт попередньо навченої мережі та виконання висновку в ній є важливою частиною функціональності OpenVX. Концепція Graph API вузлів, що представляють функції та посилань, що представляють дані, дуже зручна для реалізації глибоких нейронних мереж з OpenVX. Фактично, кожен блок нейронної мережі може бути представлений у вигляді вузла графа. OpenVX має спеціальний тип даних, що представляють тензори для забезпечення

обміну даними між цими вузлами, а самі вузли реалізовані в Розширенні нейронної мережі OpenVX. Іншим способом імпорту нейронної мережі в OpenVX є використання розширення імпорту ядра OpenVX. Розширення імпорту ядра може взяти попередньо навчену модель мережі та завантажити її в OpenVX як один вузол. Одним із форматів даних, які можна використовувати, є Neural Network Exchange Format (NNEF), стандарт, також розроблений Khronos Group. Дивіться Розділ 10, щоб дізнатися, як імпортувати попередньо навчену нейронну мережу в OpenVX.

1.9 *Огляд існуючих сервісів та областей застосування*

Більшість людей мають багато онлайн-рахунків, які потребують безпеки, і деякі з цих онлайн-рахунків, як-от програми онлайн-банкінгу, несуть високі ризики безпеки. Тепер, коли онлайн-банкінг настільки популярний, дуже важливо впровадити хороші системи ідентифікації, які забезпечують доступ до конфіденційної інформації лише власнику облікового запису. Однією з нових форм ідентифікації користувача є голосова ідентифікація. Подібно до помічників AI, які розпізнають ваш конкретний голос, фактор аутентифікації мовлення працює як унікальний «пароль», щоб розблокувати захищені облікові записи лише тоді, коли використовується ваш голос. Обліковий запис не буде доступним нікому, оскільки їхній голос звучить інакше.

Це забезпечує чудову безпеку, оскільки тепер можна використовувати багатофакторні системи, що поєднують багато різних функцій безпеки. Наприклад, скажімо, що для того, щоб отримати доступ до вашої програми для онлайн-банкінгу, вам потрібно відсканувати відбиток пальця та промовити

пароль власним унікальним голосом. Це набагато безпечніше, ніж використання традиційного пароля. Існують також системи, які використовують розпізнавання обличчя разом із розпізнаванням голосу.

Ідентифікація за голосом дедалі частіше використовується у системах безпеки банків, телекомів та інших організацій у вигляді компонента аутентифікації, яка є додатковим фактором парольного захисту. [5] Користувач повинен вимовити по телефону пароль, система визначає правильність парольної фрази і додатково перевіряє унікальний голосовий відбиток даного користувача.

Яскравим прикладом є «Chase Bank» (США), який впровадив у свою систему можливість запису голосу (збір голосів розпочався ще у 2018 році з метою створення Єдиної біометричної системи (ЄБС) та Єдиної системи ідентифікації та аутентифікації (ЕСІА)) та можливість голосового підтвердження операцій (технологія впроваджена з 21 січня 2021 року). Завдяки новій технології, організація сподівається захистити клієнтів від шахрайства. Також компанія почала пропонувати клієнтам на добровільній основі здати зразок голосу через мобільний додаток або записати голосовий зразок під час дзвінка до call-центру. Попередньо слід дати згоду на обробку персональних даних.

Створена ЕСІА зараз вбудована в систему авторизації користувачів на порталі Держпослуг та забезпечує доступ громадян за його бажанням та за його згодою до порталів Держпослуг. Після складання біометричних даних будь-який житель країни може користуватися послугами банків та державних установ без особистої присутності. Біометричну базу можна буде застосовувати для дистанційного складання іспитів, для проходження транспортного контролю в аеропортах, для підтвердження права

безкоштовного проїзду на громадському транспорті або оплати проїзду.

З найбільших рішень на світовому ринку голосової біометрії можна назвати наступних:

- Agnitio (Іспанія)
- Auraya Systems (Австралія)
- Authentify (ОАЕ)
- KeyLemon (Швейцарія)
- Nuance (США)
- ValidSoft (Велика Британія)
- Verint Systems (США)
- VoiceTrust (Німеччина)
- VoiceVault (Велика Британія)

Лідером ринку зараз зізнається Nuance. Її рішення встановлені iPhone, оскільки Siri побудована саме на технології Nuance.

Висновок : Завдання ідентифікації того, хто говорить по голосу, є в даний час дуже актуальним і продовжує набирати популярності. Ринок потребує нових якісніших рішень. Усі існуючі системи ідентифікації диктора закриті та не надають опису принципу своєї роботи.

Найцікавішим і практично цінним є завдання текстонезалежної ідентифікації. Також система повинна мати можливість швидко додавати нового диктора до своєї бази (без зміни архітектури всієї системи, що може спричинити погіршення якості або вимагати великого часу роботи) та не вимагати великого обсягу навчальних даних (не можна вимагати з нового диктора кілька годин його записаної мови). [6]

РОЗДІЛ 2. РОЗРОБКА ОПТИМАЛЬНОГО МЕТОДУ ГОЛОСОВОЇ ІДЕНТИФІКАЦІЇ

У розділі ставиться завдання ідентифікації за голосом та описується метод ідентифікації та метрики якості ідентифікації, що використовуються. Завдання ідентифікації розбивається на підзавдання, вирізняються етапи роботи системи ідентифікації. Наводяться методи вирішення поставлених підзадач.

2.1 *Постановка задачі голосової ідентифікації*

Поставимо завдання ідентифікації за голосом. Необхідно створити систему ідентифікації особи диктора за його голосом в отриманому аудіозаписі. Система повинна мати можливість зберігати голос диктора у своїй базі та виконувати ідентифікацію за запитом та отримання нового аудіозапису голосу. Допускається, що ідентифікація відбуватиметься по закритій множині збережених голосів – у цьому випадку відповідь завжди існуватиме і диктор завжди буде знайдено. Також необхідно, щоб на відповідь системи впливав лише голос диктора і ніяк не впливав вимовлений на аудіозаписі текст – система буде текстонезалежною.

Все завдання ідентифікації можна розбити на кілька етапів:

- Передобробка сигналу
- Виділення ознак
- Розпізнавання диктора

Після запису голосу засобами запису та отримання на вхід системи голос необхідно попередньо обробити. У цей етап входить чищення - видалення сторонніх шумів, мережевого гулу, виділення головного та

видалення інших голосів, видалення пауз, посилення високих частот та нормалізація сигналу.

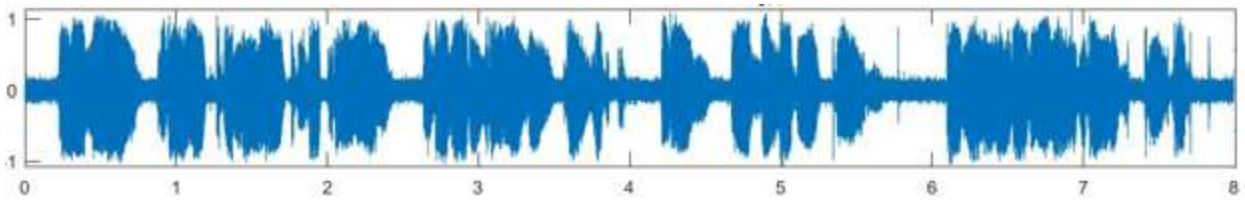


Рис.6 Початкова аудіодоріжка

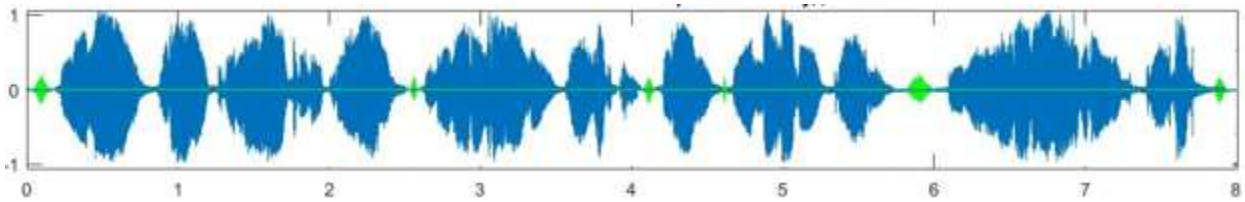


Рис. 7 Оброблений запис

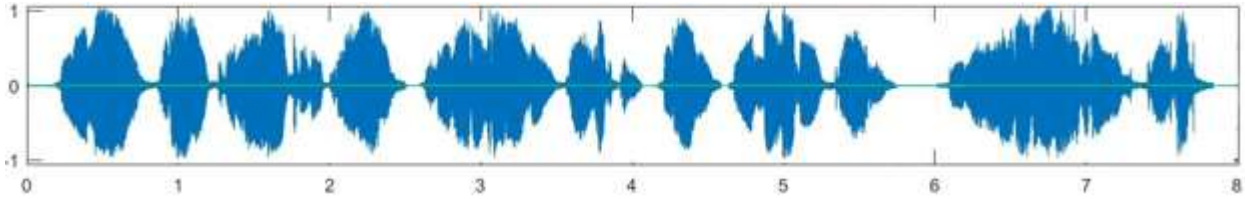


Рис.8 Аудіо без шумів

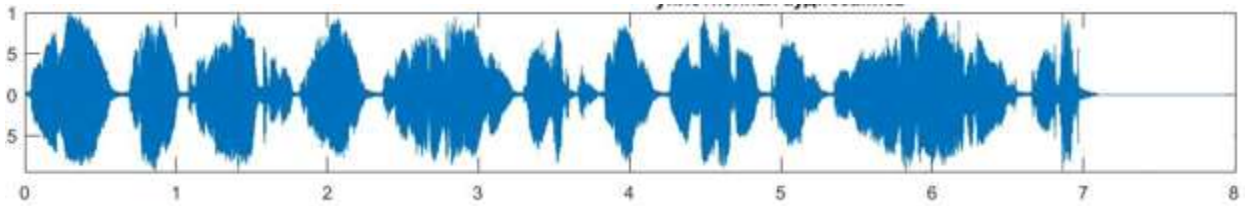


Рис.9 Ущільнене аудіо

Потім запис необхідно подати у вигляді багатовимірного вектора ознак, на основі якого буде виконуватися розпізнавання диктора.

У цій роботі основна увага приділятиметься останнім двом етапам, перший етап передобробки буде розглядатися побіжно, буде виконано лише обмежену кількість необхідних перетворень.

Таким чином, завдання голосової ідентифікації полягає в ідентифікації диктора за отриманим аудіофайлом з його промовою без урахування сказаного тексту (в пасивному режимі). При цьому завдання є закритим – диктор завжди вибиратиметься з дикторів, зареєстрованих у системі.

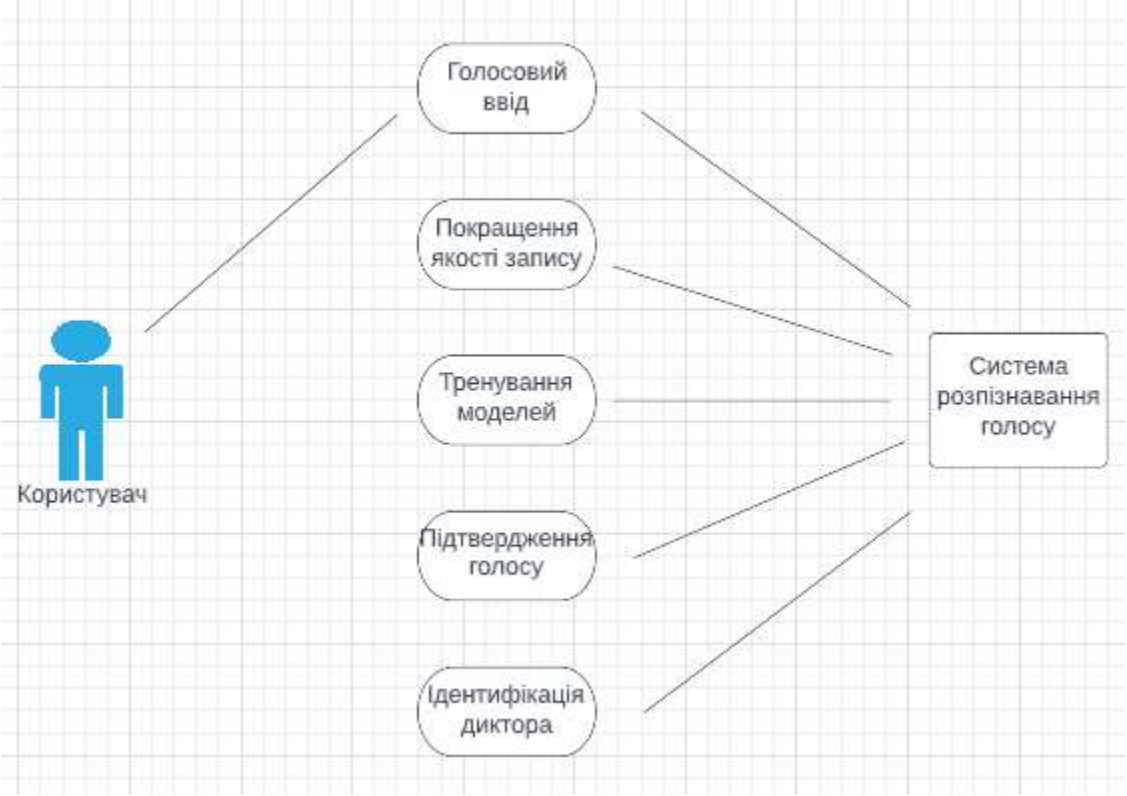


Рис. 10 Діаграма прецедентів

Опис способу представлення голосів

Звуковий сигнал являє собою впорядкований масив значень амплітуди звуку, до якого додано заголовок, що містить кількість каналів, частоту дискретизації та іншу інформацію. Однак аналізувати дані в такому вигляді неможливо - вони не містять ключових ознак, на основі яких метод зможе дати хороший результат.

MFCC (Mel Frequency Cepstrum Coefficients) був методом, який використовувався для виділення голосових характеристик

який широко використовувався в області мовленнєвих технологій, як для розпізнавання мовця, так і для мовлення визнання. Кепстральний коефіцієнт є функцією, яка зазвичай використовується в системах розпізнавання голосу.

Вихід MFCC це коефіцієнт ознаки, який містить значення, які можуть представляти сигнал мовлення . [7]

Одним із способів представлення аудіоданих та вилучення з них ознак є одержання крейдчастотних кепстральних коефіцієнтів (Mel Frequency Cepstral Coefficients).

Вихідний сигнал нарізають на фрейми (вікна) невеликої довжини (10-40 мс), що перекриваються. Потім до фреймів, що вийшли, застосовують вікно Хеммінга (щоб згладити значення на межах фреймів і зменшити витік спектру), роблять швидке перетворення Фур'є (FFT) [8]:

і отримують спектральну густину потужності (розподіл потужності сигналу в залежності від частоти, тобто потужність, що припадає на одиничний інтервал частоти).

Потім спеціальною «гребінкою» фільтрів, розташованих рівномірно за крейдою-шкалою (малюнок нижче) роблять крейд-спектрограму (кожен крейд-фільтр - це трикутна віконна функція, яка підсумовує кількість енергії на певному діапазоні частот і тим самим дає крейда-коефіцієнт), після логарифмують отримані результати (вважається, що таким чином знижується чутливість коефіцієнтів до шумів) і застосовують дискретне косинусне перетворення (DCT) - алгоритм стиснення даних. [9]

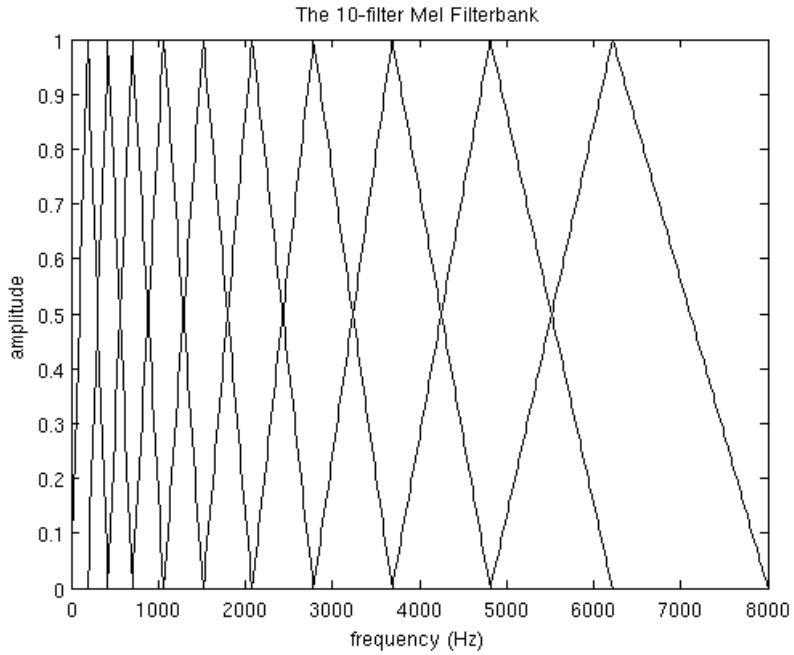


Рис.11 Десять крейд-фільтрів - трикутних віконних функцій.

Отримані таким чином коефіцієнти являють собою якусь стислу характеристику кадру. При цьому, оскільки фільтри, які ми застосовували, були розташовані в крейді, коефіцієнти враховують особливості сприйняття людського вуха, а значить несуть більше корисної інформації. Прийнято використовувати від 13 до 25 MFCC на кадрі.

Так як індивідуальність голосу часто залежить від швидкості мови, особливостей вимови та прискорення, також враховуватимемо похідні.

2.2 *Опис методу ідентифікації, що використовується.*

На даному етапі необхідно, маючи новий набір ознак невідомого диктора та безліч збережених у базі наборів відомих дикторів, віднести нові

дані до правильної категорії – диктора. Сформульоване завдання вирішуватиметься через побудову оптимальних для кожного диктора моделей гауссівської суміші розподілів.

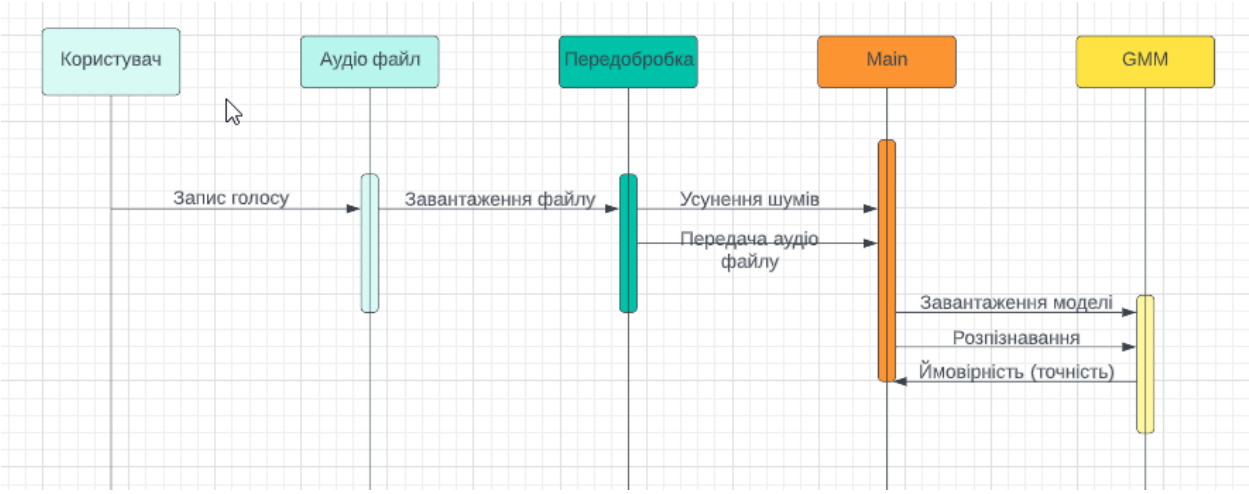


Рис. 12 Діаграма послідовності

Модель гаусової суміші розподілів [10] (Gaussian Mixture Model) - це функція, що складається з декількох гауссіанів, кожен з яких ідентифікується як k ∈ {1,…, K}, де K – кількість кластерів у наборі даних. Іншими словами, Модель Гаусової суміші (GMM) є сумішшю багатовимірних гаусових розподілів ймовірностей (багатомірних нормальних розподілів), які найкращим чином моделюють вхідний набір даних.

Хоча GMM часто класифікується як алгоритм кластеризації, у нашому випадку вона представлятиме оцінку щільності. Іншими словами, результатом припасування GMM до даних є технічно не кластерна модель, а ймовірнісна модель, що описує розподіл даних. На малюнку нижче дані представлені як точок, які розбиті на кластери. Крапки кожного кластера пофарбовані своїм кольором. Блакитні кола показують змодельований GMM розподіл вхідних даних [11]. Показано результати роботи для двох типів коваріації - covariance_type="spherical" (ліворуч) та covariance_type="full" (праворуч).

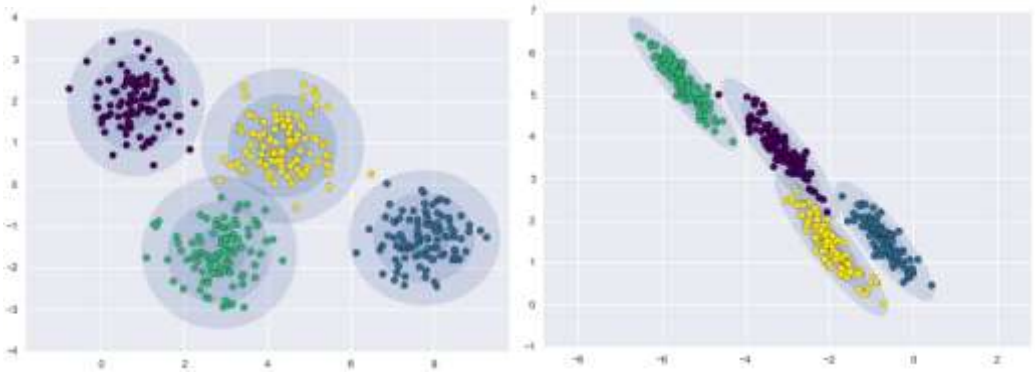
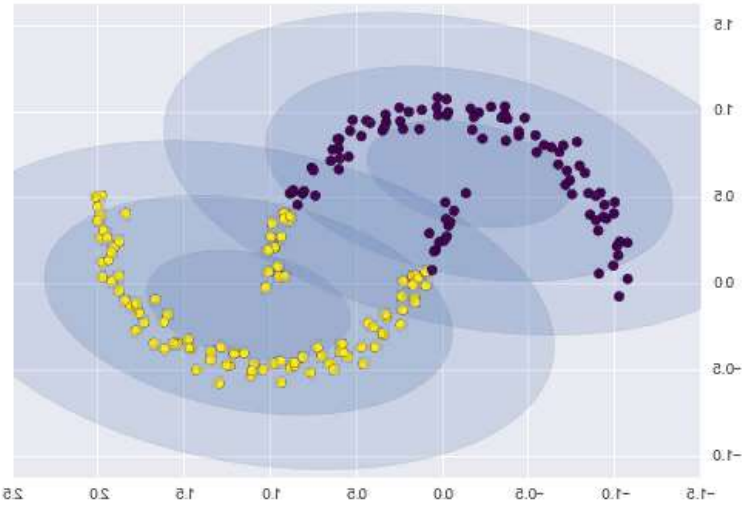


Рис.13 Візуальне представлення компонентів GMM залежно від даних та типу коваріації

Важливо вибрати достатню кількість компонентів GMM, інакше вона даватиме поганий результат (малюнок нижче).



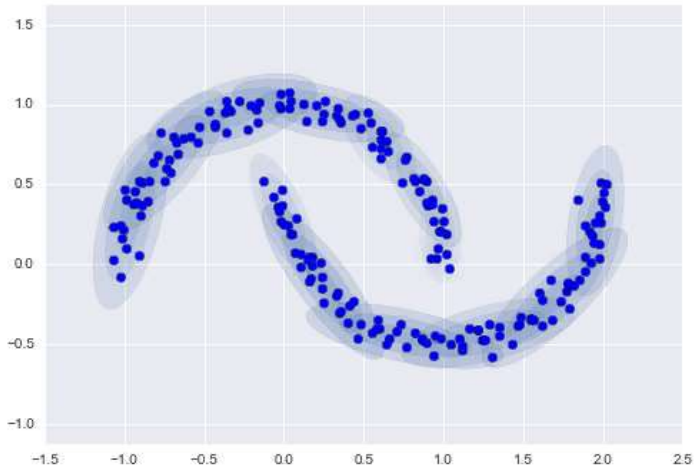


Рис.14 Поганий опис даних (ліворуч – 2 компоненти) та гарний (праворуч – 16)

2.3 *Опис способу порівняння голосів*

GMM зручна, оскільки є гнучким засобом моделювання довільного багатовимірного розподілу даних. Після цього моделювання будемо зберігати всі отримані моделі GMM кожного диктора. [12]

Отримуючи на вхід новий набір ознак, який необхідно віднести до вірної категорії - диктора, будемо для кожного збереженої моделі GMM диктора визначати ймовірні значення присутності серед даного розподілу записаного вектора ознак. Іншими словами, порівнюватимемо результат від кожної збереженої GMM. Вірною категорією буде визнано диктора з максимальним значенням ймовірності.

Опис використовуваних метрик якості

Допустимо, необхідно визначити, чи належить зразок мови Y диктор S (з набору збережених дикторів). Сформулюємо необхідні гіпотези.

H0 - мова Y належить диктору S

H1 - мова Y належить не диктор S

матрицю помилок класифікації представлено в таблиці нижче. У ньому Y – відповідь алгоритму на об'єкті, Y – справжня мітка об'єкта.

Таблиця 2. Матриця помилок класифікації

	$\mathcal{Y} = Y \in S$	$\mathcal{Y} = Y \notin S$
$\mathbb{Y} = Y \in S$	True Positive (TP)	False Positive (FP)
$\mathbb{Y} = Y \notin S$	False Negative (FN)	True Negative (TN)

Тоді помилка першого роду буде ймовірністю "відмови своєму" (False Negative), і помилка другого роду - ймовірність "допуску чужого" (False Positive). [13]

Введемо метрику точності. Вона визначається за такою формулою:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Для оцінки якості роботи алгоритму кожному з класів окремо введемо метрики precision (точність) і recall (повнота).

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

Precision можна інтерпретувати як частку об'єктів, названих класифікатором позитивними і при цьому дійсно позитивними, а recall показує, яку частку об'єктів позитивного класу з усіх об'єктів позитивного класу знайшов алгоритм. Введення precision не дозволяє нам записувати всі об'єкти в один клас, тому що в цьому випадку ми отримуємо зростання рівня

False Positive (FP). Recall демонструє здатність алгоритму виявляти цей клас взагалі, а precision – здатність відрізняти цей клас від інших класів.

2.6 Оптимізація методу

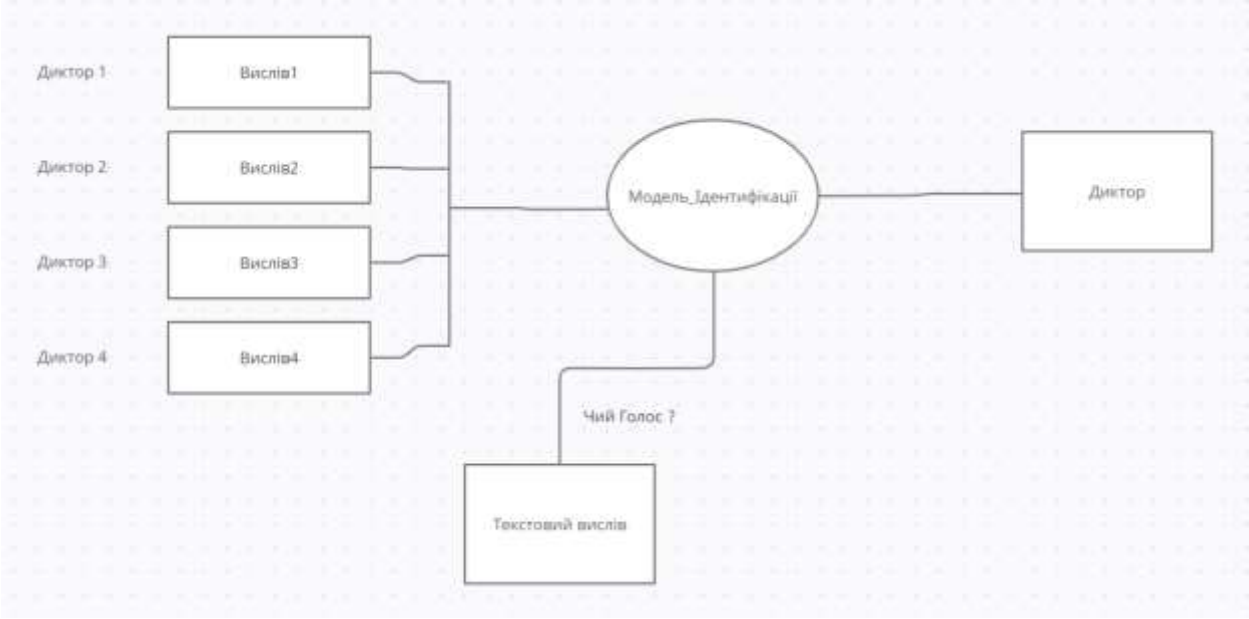


Рис.15 Функціональна схема ідентифікації диктора

Щоб зберегти зразок голосу нового диктора, необхідно створити та навчити GMM, а потім зберегти її в базі. Однак у процесі навчання виникає труднощі – голосових даних замало. Для хорошого навчання GMM бажано мати кілька годин промови диктора. У цьому завдання така вимога неможлива. Потрібно вміти будувати модель на невеликому обсязі даних – парі слів чи парі фраз. [14]

Вирішуватимемо цю проблему за допомогою Універсальної фонової моделі (Universal Background Model, UBM, УФМ).

Універсальна фонова модель (UBM) в задачі розпізнавання диктора є деякою моделлю, навченою на кінцевій великій кількості голосів певних дикторів і містить в собі апостеріорні знання про влаштування людського

голосу в цілому. UBM також є великою моделлю гаусової суміші (GMM), навченої для представлення диктонезалежного розподілу ознак. Модель навчається методом Expectation-maximization (EM) на навчальній множині. Вектор середніх, витягнутий із моделі після навчання, називається супервектором середніх. Коли необхідно отримати ознаки для вимови, що знову прийшла, параметри UBM [15] підлаштовуються методом оцінки апостеріорного максимуму (maximum a posteriori probability estimate, MAP) і отриманий вектор середніх моделі вже називається дикторським супервектором середніх. Підбудована модель буде кінцевою GMM для нового диктора, яка буде збережена.

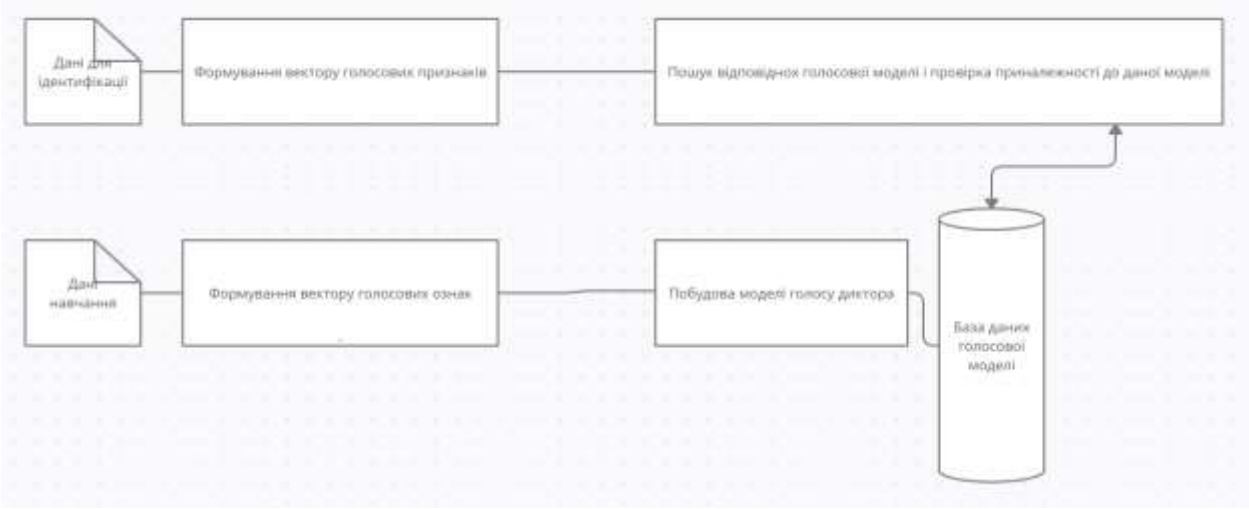


Рис.16 Функціональна схема системи голосової ідентифікації/верифікації особи

Таким чином, UBM дозволяє реалізовувати GMM на невеликій кількості вхідних даних нового диктора, а також дозволяє збільшити точність

розпізнавання диктора порівняно з неадаптованими моделями, які можна побудувати на невеликій навчальній вибірці.

Системи, створені за допомогою UBM, називаються системами верифікації диктора на основі моделі Гаусових сумішей та універсальної фонової моделі (GMM-UBM).

Принцип навчання та роботи UBM зображений на малюнку.

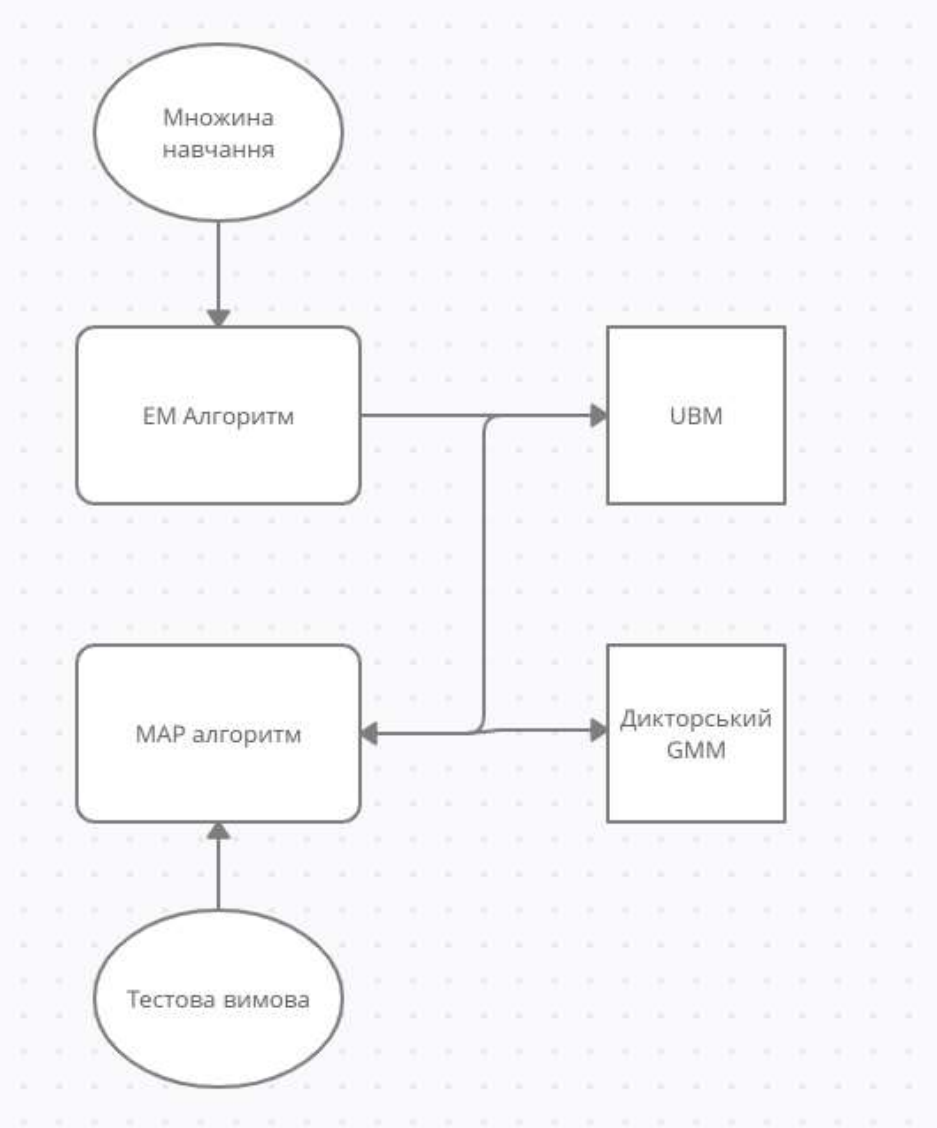


Рис.17 Схема навчання та функціонування UBM-GMM

При створенні UBM необхідно, щоб дані, що використовуються для навчання моделі, були збалансованими щодо подальшого застосування системи. Іншими словами, голоси дикторів повинні сильно відрізнятися, у вибірці повинні бути присутніми як високі голоси, так і низькі, з різною швидкістю мови та різними записуючими системами - збалансовані і на кшталт використовуваних при записі дикторів мікрофонів.

У показано, що системи голосової верифікації, що виробляють адаптацію моделі диктора з універсальної фонової моделі, мають набагато кращу точність, ніж системи, в яких модель диктора навчається окремо від УФМ. Це можна пояснити тим, що УФМ покриває більшість класів акустичних подій, що з'являються у промові дикторів. Відповідно, під час адаптації моделі диктора частина таких подій, що з'явилися в промові диктора, змінюють і підлаштовують компоненти суміші під конкретного диктора. Частина подій, що залишилася, не зустрічається в навчальній вибірці, копіюється з УФМ. Таким чином, це додає стійкості моделі до тих акустичних подій конкретного диктора, які були відсутні в навчальній вибірці. [16]

Для D-мірного вектора ознак x, поданого на вхід, щільність суміші розраховується як:

$$P(x|\lambda) = \sum_{k=1}^{M} w_k * g(x|\mu_k, \Sigma_k)$$

Тут:

x – є D-мірним вектором ознак

w_k,k=1,..,M – чи є суміш вагами, у сумі вони = 1

μ_k,k=1,..,M – маточування кожної Гауссіани

λ=(w_k,μ_k,Σ_k),k=1,..,M – параметри кожної GMM

Σ_k,k=1,..,M – коваріація кожної Гауссіани

g(x|μ_k,Σ_k) – густини Гауссіан, що визначаються як:

$$g(x|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_k|^{\frac{1}{2}}} exp^{-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)}$$

Зазвичай використовується діагональна коваріаційна матриця, а не повна матриця ковар, оскільки вона більш ефективна з обчислювальної точки зору і емпірично працює краще.

GMM навчається на наборі навчальних векторів. Параметри GMM обчислюються ітеративно з використанням алгоритму максимізації очікування (EM), і тому немає жодних гарантій, що він двічі перетворюється на те саме рішення в залежності від ініціалізації.

Алгоритм MAP є той самий алгоритм EM з початковими параметрами моделі GMM, рівними параметрами UBM. На кожній ітерації відбувається перерахунок та зміна маточікування моделі, зберігаючи значення коварації, оскільки зміна коварації не покращує якість роботи моделі.

Для перерахунку матожидання використовується максимальна апостеріорна адаптація (maximum a posteroiri adaptation) [17]:

$$\mu_k^{MAP} = \alpha_k\mu_k + (1-\alpha_k)\mu_k^{UBM}$$

Тут:

$\alpha_k = \frac{n_k}{n_k + r_k}$ – Коефіцієнт адаптації множення

n_k – кількість адаптаційних даних

r_k – фактор відповідності (relevance factor)

Висновок: Для реалізації системи ідентифікації використовуватимемо метод GMM на ознаках MFCC. Оптимізація у вигляді UBM дозволить позбутися недоліків та труднощів побудови GMM, [18] описаних у попередньому розділі. GMM, будучи засобом моделювання довільного багатовимірного розподілу даних, виконуватиме функцію класифікатора, видаючи щільність суміші на вхідному векторі ознак для кожного із збережених дикторів [19].

РОЗДІЛ 3. РОЗРОБКА ПРОГРАМНОЇ СИСТЕМИ ГОЛОСОВОЇ ІДЕНТИФІКАЦІЇ

На даному етапі проводився вибір та вивчення мови та допоміжних бібліотек для реалізації системи, проведено та описано саму реалізацію та перевірено роботу методу ідентифікації.

3.1 *Вибір мови та допоміжних модулів для програмної реалізації системи*

Реалізація моделей, алгоритмів обробки аудіофайлу, функцій побудови коефіцієнтів, а також додаткових методів виконувалася мовою Python3.10. Як допоміжні бібліотеки використовувалися бібліотеки librosa, pydub і contextlib для завантаження та роботи з аудіофайлом, python_speech_features для побудови MFCC, webrtcvad для реалізації функції детектування наявності мови в аудіосигналі, joblib для запису даних та моделей, а також Numpy для реалізації функцій та алгоритму. Розробка проводилася у програмному середовищі Intellij Idea та PyCharm.

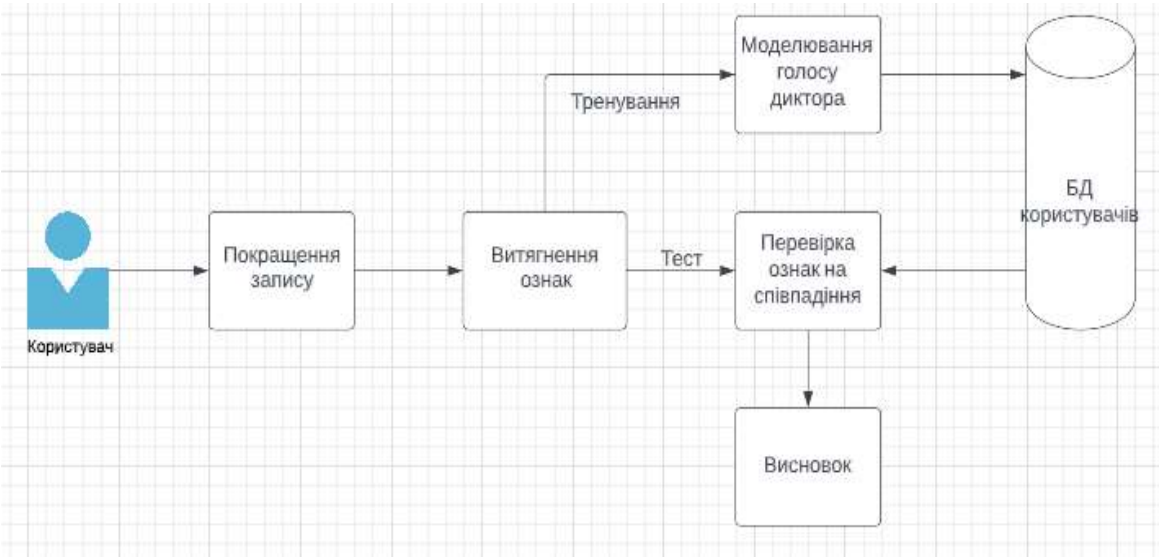


Рис.18 Процес роботи системи

3.2 *Функіональні вимоги*

Функціональні вимоги до системи розпізнавання голосу диктора наступні :

- Система повинна мати високий рівень правильних висновків про результати збігів в голосі
- Інформавання користувача про помилки
- Система повинна мати безпечна для даних користувача.
- Система повинна вираховувати до 500 зразків голосу без великих затрат часу
- Система повинна вміти обробляти формат wav

3.3 *Нефункціональні вимоги*

- Система повинна бути реалізована на мові python
- Процес розпізнавання не повинен перевищувати 1 хвилину.

3.4 *Програмна реалізація моделей та алгоритмів*

Опишемо реалізовані функції та класи. У процесі роботи було реалізовано клас Frame та функції:

def extract_features

Функція, яка обчислює ознаки з поданого на вхід аудіофрагменту. Функція обчислює MFCC і додає першу та другу похідну. Отримання ознак - даних, з якими працюватимуть моделі машинного навчання. В принципі, звуковий сигнал сам по собі - це вже дані, а саме впорядкований масив значень амплітуди звуку, до якого додається заголовок, що містить кількість каналів, частоту дискретизації та іншу інформацію. Але аналізувати ці дані

безпосередньо ми не зможемо, оскільки вони не містять таких речей, дивлячись на які наша модель може сказати — ага, ось ці шматки належать одній і тій самій людині.

def normalize

Функція нормалізації поданого на вхід ознак ознак. Багато методів машинного навчання можуть вести себе погано, якщо окремі ознаки не будуть більш-менш схожі на стандартні нормально розподілені дані: Гаусів з нульовим середнім та одиничною дисперсією. Фукція повертає нормалізований вектор. Центрує ознаки, видаляючи середнє значення кожної ознаки, а потім масштабує, ділячи ознаки стандартне відхилення.

def write_wave

Функція записує оброблений аудіофрагмент в .wav файл по вказаному шляху.

def execute

Функція проводить перевірку на наявність спільних ознак у тренованих моделях та провіряє на співпадіння з метою ідентифікації.

Також завантажує аудіофайл та обробляє його. Видаляє тишу та паузи, склеює та зберігає. Працює у двох режимах – зберігає або всі дані, які отримала після видалення пауз або якусь задану частину. Приймає на вхід ім'я файлу, який потрібно завантажити та шлях, яким потрібно зберегти оброблений файл.

def get_features

Функція витягує вектор ознак заданої розмірності з аудіофайлу або завантажує збережені ознаки файлу. Як параметри передається ім'я файлу, кількість MFCC. Повертає одержані ознаки

def get_gmm

Створює модель GMM з поданої на вхід UBM і навчає її на вхідних ознаках або завантажує збережену модель з файлу. Повертає модель GMM.

test_gmms

Функція призначена для тестування існуючих моделей. Вважає час роботи, а також метрики якості роботи кожної моделі, яка розташована у певній заданій директорії.

sklearn.mixture - це пакет, який використовує гауссівські змішані моделі для навчання без вчителя (підтримує чотири коваріаційні матриці, діагональні, сферичні, пов'язані та повні). Об'єкт GaussianMixture реалізує алгоритм очікуваного максимуму (EM) для припасування моделей гаусової суміші. Він також може малювати довірчі еліпсоїди для моделей з декількома змінними, а також розраховувати BIC (байесовський інформаційний критерій) для оцінки кількості кластерів даних. Подробиці дивіться на офіційному китайському сайті Sklearn 2.1. Модель гаусової суміші.

3.5 *Результат роботи системи*

Система показала себе надійною та з високою швидкодією. На настуному рисунку зображено процес тренування та створення моделей для

подальшої ідентифікації.

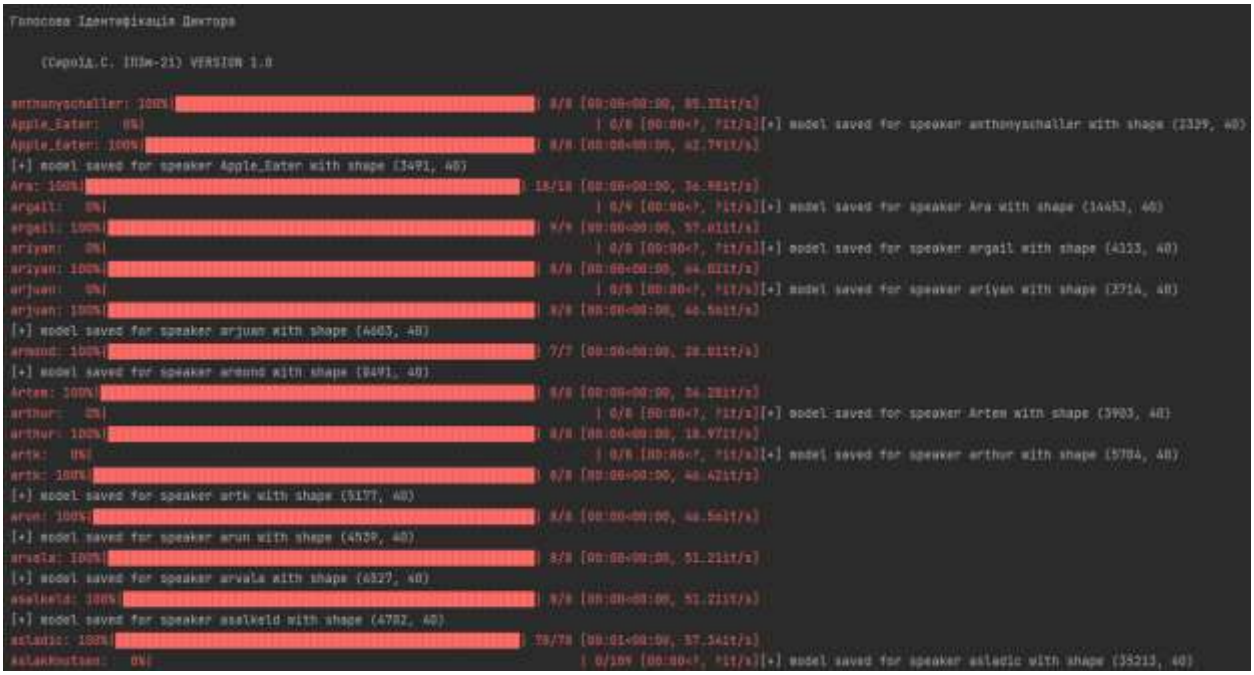


Рис. 19 Процес тренування моделей

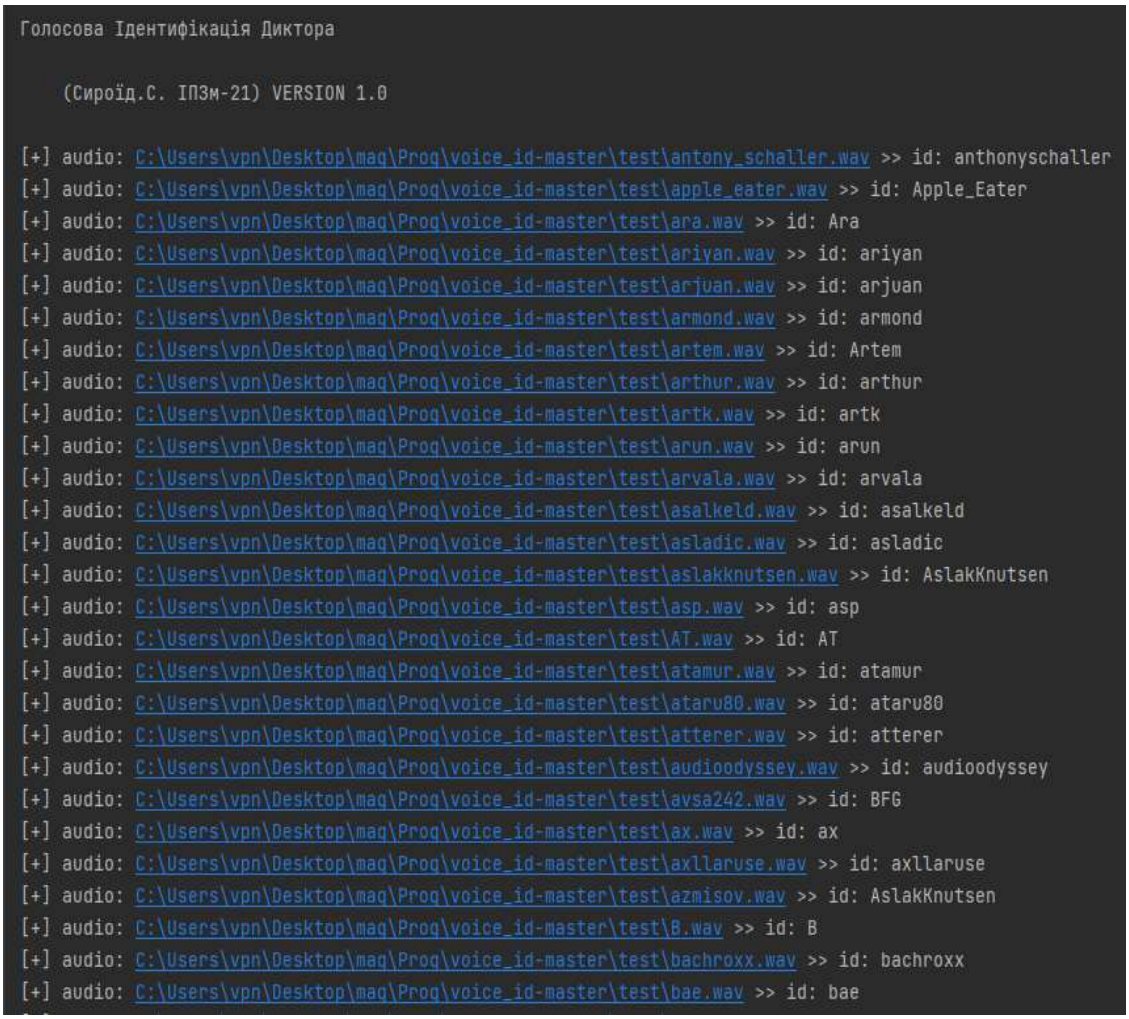


Рис. 20 Процес порівняння та ідентифікації голосів

Висновок : У цьому розділі вибрано програмну мову та середовище розробки, перераховано допоміжні бібліотеки та описано функції та класи, реалізовані в процесі створення системи ідентифікації.

РОЗДІЛ4. ЕКСПЕРЕМЕНТАЛЬНІ ДОСЛІДЖЕННЯ РОБОТИ СИСТЕМИ

У цьому розділі наводяться результати експериментальних досліджень методу ідентифікації диктора за допомогою конструювання моделі GMM- з різними архітектурними параметрами, а також вибір параметрів моделі для отримання системи з оптимальними показниками роботи, такими як точність, повнота. Була використана навчальна вибірка LibriSpeech об'ємом 4000 голосів різних дикторів, записаних за допомогою різних аудіопристроїв. На цій вибірці продемонстровано вплив різних параметрів архітектури на результат роботи. Проведені дослідження дали змогу визначити оптимальну архітектуру системи.

Після цього було проведено порівняння результатів роботи на вибірках різної тривалості задля встановлення мінімальної найкращої тривалості для хорошої роботи.

4.1 *Опис тестових даних, обсягу та особливостей*

Навчальна та валідаційна вибірки складаються з бази даних LibriSpeech. Вона включає кілька сотень годин англійської мови з різними діалектами, записаними на різні пристрої. Аудіофайли чисті, без шуму фону, сторонніх голосів і додаткових артефактів.

Попередня обробка полягала у видаленні пауз та аудіофрагментів тиші, склеюванні фрагментів та нарізці на фрагменти із заданою тривалістю для валідації та навчання GMM.

4.2 *Тестування системи ідентифікації*

Для визначення оптимальних параметрів моделі було збудовано 27 моделей різної архітектури, для кожної моделі проведено тестування роботи

при 40 доданих дикторах. Результати наведено у таблиці. Значення Accuracy, Precision, Recall обчислені як середнє арифметичне кожного класу (диктора). Також для порівняння було виміряно середній час прийняття рішення класифікатором залежно від тривалості аудіофрагменту промови, поданого на вхід для ідентифікації (стовпці «Час на 40 с» та «Час на 5 с»).

Таблиця 3. Якість роботи побудованих моделей

Компоненти	Коваріація	MFCC	Accuracy	Precision	Recall	Час (на 40 с)	Час (на 5 с)
8	diag	13	0.9946	0.8125	0.8916	2.2191	1.6222
8	diag	20	0.9951	0.8125	0.9	2.2509	1.7025
8	diag	8	0.9942	0.8203	0.8833	2.1254	1.4941
8	full	13	0.9945	0.8375	0.8916	4.71203	3.1141
8	full	20	0.9945	0.8511	0.8916	5.2238	3.4823
8	full	8	0.9945	0.8125	0.898	3.3213	2.1422
8	tied	13	0.9945	0.8502	0.898	4.5222	3.0525
8	tied	20	0.9954	0.8875	0.9083	5.3009	3.5039
8	tied	8	0.9952	0.8527	0.878	3.2472	2.0821
16	diag	13	0.9937	0.8751	0.925	2.45507	1.67226
16	diag	20	0.9954	0.8875	0.9	2.5688	1.7152
16	diag	8	0.9937	0.8125	0.875	2.3292	1.6191
16	full	13	0.9951	0.8501	0.9	6.2387	4.1502
16	full	20	0.9951	0.8512	0.9	8.86304	5.6113
16	full	8	0.9951	0.8524	0.9	4.6579	2.96404
16	tied	13	0.9963	0.8503	0.925	6.2921	4.0401
16	tied	20	0.9963	0.8125	0.925	8.6095	5.5437
16	tied	8	0.9946	0.8208	0.875	4.6362	2.8073

32	diag	13	0.9954	0.8125	0.875	2.82457	2.05204
32	diag	20	0.9958	0.8875	0.925	2.8914	2.0058
32	diag	8	0.9951	0.8513	0.925	2.7116	1.9027
32	full	13	0.9954	0.8629	0.9083	12.5033	8.06306
32	full	20	0.9954	0.8629	0.9083	15.4039	9.91009
32	full	8	0.9954	0.8629	0.9083	7.4432	4.3401
32	tied	13	0.9951	0.8125	0.886	12.8066	8.0172
32	tied	20	0.9954	0.8875	0.9083	15.3268	9.8261
32	tied	8	0.9951	0.8208	0.965	7.3689	4.3601

Таким чином, описаний алгоритм навчання та побудови моделі GMM-UBM був випробуваний з різними параметрами:

1. Кількість MFCC: 8, 13, 20
2. Тип підступу в GMM: full, diag, tied
3. Кількість компонентів: 8, 16, 32
4. З різною тривалістю тестових даних: 5 сек, 20 сек, 40 сек

Для навчання GMM бралися аудіофрагменти промови тривалістю до 20 секунд.

4.3 *Висновки за результатами експериментальних досліджень*

Як видно з таблиці, всі збудовані моделі дають гарний результат роботи. Було помічено, що моделі з типом коваріації = full вимагають набагато більший об'єм пам'яті порівняно з моделями інших параметрів (наприклад, для 16 компонентів і MFCC = 20 об'єм пам'яті для моделей з типами diag, tied, full дорівнює відповідно 32, 93, 1359 КБ). Таким чином, за обсягом пам'яті найбільше вимагає модель full, найменше – diag. Причиною цього може бути більша кількість параметрів моделі, що настроюються. Також модель з типом

ковариації full найдовше навчається на вибірці (відповідно, модель з типом коваріації diag навчається найшвидше).

Час роботи процесу ідентифікації значно збільшується при збільшенні кількості компонентів суміші, збільшенні тривалості вхідного аудіофрагменту для ідентифікації, збільшенні кількості MFCC (розмірності вектора ознак) та виборі типу коваріації full.

Як результат, всі спроектовані моделі дають приблизно однаковий результат. Оптимальною моделлю може бути будь-яка модель з відносно невеликим часом роботи. Як приклад, оптимальною може бути прийнята модель з параметрами:

1. Кількість компонентів UBM: 16
2. Тип коваріації: diag
3. Кількість MFCC: 13
4. Тривалість навчання GMM: 20 с
5. Тривалість тестового фрагмента: 20 с

Важливо відмітити, що тривалість тестового фрагмента може відрізнятися від тривалості фрагмента, поданого на вхід моделі. У процесі попередньої обробки після видалення тиші і пауз фрагмент може значно скоротитися в розмірі, тому оптимальна тривалість обрана «із запасом».

Висновок

В результаті виконання даної навчально-дослідницької роботи буде досягнуто головної мети - створено систему ідентифікації особистості диктора за його голосом в отриманому аудіозаписі. У ході роботи було виконано такі завдання:

1. Отримано навчальну вибірку.
2. Вивчено та розроблено метод ідентифікації диктора.
3. Спроектовано та розроблено оптимальну архітектуру обраної моделі.
4. Пройдено тестування роботи реалізованої системи.

Досягнута ступінь успішного розпізнавання становить понад 77 відсотків на лідираційних даних при точності та повноті понад 87 та 90 відсотків відповідно. Надалі для поліпшення роботи необхідно збільшувати обсяг і різноманітність навчальної вибірки для побудови моделі UBM. Сферою застосування може бути будь-яке завдання ідентифікації, верифікації та аутентифікації користувача системи за його голосом, завдання діаризації та завдання визначення кількості тих, хто говорить у записаному аудіофайлі.

Література

1. Zhongxin Bai, Xiao-Lei Zhang, Jingdong Chen. Speaker Recognition Based on Deep Learning: An Overview. 2020. URL: https://arxiv.org/abs/2012.00931

3. Hajavi, A., Etemad A. 2019. A Deep Neural Network для ShortSegment Speaker Recognition. URL: https://arxiv.org/pdf/1907.10420.pdf

4. Joon Son Chung, Arsha Nagrani, Andrew Zisserman. 2018. VoxCeleb2: Deep Speaker Recognition. URL: https://arxiv.org/abs/1806.05622

5. Daniel Foley. Gaussian Mixture Modelling (GMM). URL: https://towardsdatascience.com/gaussian-mixture-modelling-gmm-833c88587c7f

6. T. Kinnunen, H. Li, на перевірці тексту-ідентичне співрозмовника: від особливих до supervectors, Speech communication. 2010. Стор. 12–40

7. Techportal Біометрія. Стандарти. Технології. Завдання. Рішення. URL: http://www.techportal.ru/security/biometrics/mirovoy-i-rossiyskiy-rynki-biometrii/mirovoy-rynok-marketsandmarkets

https://www.researchgate.net/publication/236142366_Tehnologii_biometricheskoj_identifikacii_licnosti_po_golosu_i_drugim_modalnostam_BIOMETRIC_TECHNOLOGIES_OF_PERSON_IDENTIFICATION_BY_VOICE_AND

8. Bimbot F. та ін. A tutorial on text-independent speaker verification 2004. Стор. 430-451. URL: https://studylib.net/doc/12607069/a-tutorial-on-text-independent-speaker-verification-pleas...

9. Reynolds D., Rose R. Robust text-independent speaker identification using Gaussian mixture speaker models. 1995. Стор. 72–83

10. Reynolds D. Experimental evaluation features for robust speaker identification. 1994. Стор. 639–643

11. Єдина біометрична система URL: https://bio.rt.ru/citizens/

12. Олексій Лукацький Голосова біометрія. Короткий огляд технології. 2015. https://www.securitylab.ru/blog/personal/Business_without_danger/147943.php

13. Oscar Contreras Carrasco Gaussian Mixture Models Explained. Від інтуїції до виконання. 2019. URL: https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95

14. Jake VanderPlas In Depth: Gaussian Mixture Models URL: https://jakevdp.github.io/PythonDataScienceHandbook/05.12-gaussian-mixtures.html

15. Stephen M Chu, Daniel Povey Universal background model заснований на speech recognition. 2008. URL: https://www.researchgate.net/publication/224762192_Universal_background_model_based_speech_recognition

16. Matsui T., Furui S. Likelihood normalization для speaker verification using phoneme-i speaker-independent model. 1995. Стор. 109-116.

17. Reynolds D. A. Comparison of background normalization methods for text-independent speaker verification. 1997 Стор. 963-966

18. Hermansky H., Malayath N. Speaker verification using speaker-specific mappings. 1998 стор.111-114.

19. Quatieri T. F. та ін. Speaker and language recognition using speech codec parameters. 1999 Стор. 787-790

Додаток

Додаток 1. Код скрипту тренування

```
import os
import pickle
import numpy as np
from scipy.io.wavfile import read
from sklearn.mixture import GaussianMixture as GMM
from sklearn import preprocessing
import python_speech_features as mfcc
from pprint import pprint
from tqdm import tqdm
#----------------------------------------------------------------------------------------
logo = """
Голосова Ідентифікація Диктора

  (Сироїд.С. ІПЗм-21) VERSION 1.0
"""
#----------------------------------------------------------------------------------------
def calculate_delta(array):
  """Розрахунок та повернення різниці даних ознак матриці"""
  rows,cols = array.shape
  deltas = np.zeros((rows,20))
```

```
    N = 2
    for i in range(rows):
        index = []
        j = 1
        while j <= N:
            if i-j < 0:
                first = 0
            else:
                first = i-j
            if i+j > rows -1:
                second = rows -1
            else:
                second = i+j
            index.append((second,first))
            j+=1
        deltas[i] = ( array[index[0][0]]-array[index[0][1]] + (2 *
(array[index[1][0]]-array[index[1][1]])) ) / 10
    return deltas

def extract_features(audio,rate):
    """Витягнення 20 mfcc ознак зі звуку, виконування CMS та
комбінація різниць
    щоб створити 40 dim ознаковий вектор"""
    mfcc_feat = mfcc.mfcc(audio,rate, 0.025, 0.01,20,appendEnergy = True)

    mfcc_feat = preprocessing.scale(mfcc_feat)
    delta = calculate_delta(mfcc_feat)
```

```
  combined = np.hstack((mfcc_feat,delta))
  return combined
#------------------------------------------------------------------------------------------
---------------------------


curr_dir = os.getcwd()
train_dataset_path = os.path.join(curr_dir,'train')
models_path = os.path.join(curr_dir,'models')

def execute():
  identities = {}
  for root, dirs, files in os.walk(train_dataset_path):
    for file in files:
      if file.endswith('.wav'):
        fpath = os.path.join(root,file)
        identity = fpath.split(os.sep)[-3].split('-')[0]
        if identity not in identities:identities[identity]=[]
        else:identities[identity].append(fpath)
  if identities:
    for identity in identities:
      features = np.asarray(())
      for sample in tqdm(identities[identity],desc=identity,ncols=100):
        sr,audio = read(sample)
        vector   = extract_features(audio,sr)
        if features.size == 0:features = vector
        else:features = np.vstack((features, vector))
```

```
        gmm = GMM(n_components = 16, max_iter = 200,
covariance_type='diag',n_init = 3)
        gmm.fit(features)
        model_file_path = os.path.join(models_path,'{}.gmm'.format(identity))
        pickle.dump(gmm,open(model_file_path,'wb'))
        print('[+] model saved for speaker {} with shape
{}'.format(identity,features.shape))
  return 1

if __name__ == '__main__':
  print(logo)
  execute()
```

Додаток 1. Код скрипту оцінки

```
import os
import pickle
import numpy as np
from scipy.io.wavfile import read
from sklearn.mixture import GaussianMixture as GMM
from sklearn import preprocessing
import python_speech_features as mfcc
from pprint import pprint
from tqdm import tqdm
#----------------------------------------------------------------------------------------
logo = """
Голосова Ідентифікація Диктора
```

```
  (Сироїд.С. ІПЗм-21) VERSION 1.0
"""
#-------------------------------------------------------------------------------
def calculate_delta(array):
  """Розрахунок та повернення різниці даних ознак матриці"""
  rows,cols = array.shape
  deltas = np.zeros((rows,20))
  N = 2
  for i in range(rows):
    index = []
    j = 1
    while j <= N:
      if i-j < 0:
        first = 0
      else:
        first = i-j
      if i+j > rows -1:
        second = rows -1
      else:
        second = i+j
      index.append((second,first))
      j+=1
    deltas[i] = ( array[index[0][0]]-array[index[0][1]] + (2 *
(array[index[1][0]]-array[index[1][1]])) ) / 10
  return deltas
```

```
def extract_features(audio,rate):
    """Витягнення 20 mfcc ознак зі звуку, виконування CMS та
комбінація різниць
    щоб створити 40 dim ознаковий вектор"""
    mfcc_feat = mfcc.mfcc(audio,rate, 0.025, 0.01,20,appendEnergy = True)

    mfcc_feat = preprocessing.scale(mfcc_feat)
    delta = calculate_delta(mfcc_feat)
    combined = np.hstack((mfcc_feat,delta))
    return combined
#-----------------------------------------------------------------------------------------
----------------------------


curr_dir = os.getcwd()
test_dataset_path = os.path.join(curr_dir,'test')
models_path = os.path.join(curr_dir,'models')
models = []
identities = {}
id_count=0
for root, dirs, files in os.walk(models_path):
    for file in files:
        if file.endswith('.gmm'):
            identity=file.replace('.gmm','')
            identities[id_count]=identity
            mpath=os.path.join(root,file)
            models.append(pickle.load(open(mpath,'rb')))
```

```
        id_count+=1

def execute():
    tests = []
    for root, dirs, files in os.walk(test_dataset_path):
        for file in files:
            if file.endswith('.wav'):
                fpath = os.path.join(root,file)
                tests.append(fpath)
    if tests:
        for test in tests:
            sr,audio=read(test)
            vector=extract_features(audio,sr)
            predict_probability = np.zeros(len(models))
            for i in range(len(models)):
                gmm=models[i]
                scores=np.array(gmm.score(vector))
                predict_probability[i] = scores.sum()
            prediction = np.argmax(predict_probability)
            subject_name = identities[prediction]
            print('[+] audio: {} >> id: {}'.format(test,subject_name))
    return 1

if __name__ == '__main__':
    print(logo)
    execute()
```

SOFTWARE ARCHITECURE DOCUMENT ABSTRACT

This paper presents the basic idea of speech recognition, proposed types of speech recognition, issues in speech recognition, different useful approaches for feature extraction of the speech signal with its advantage and disadvantage and various pattern matching approaches for recognizing the speech of the different speaker. Now day's research in speech recognition system is motivated for ASR system with a large vocabulary that supports speaker independent operations and continuous speech in different language.
General Terms Pattern Recognition, Automatic Speech Recognition (ASR), Acoustic Modeling, Language Modeling. Keywords Feature extraction, pattern matching, ANN, HMM, DTW, MFCC

1. INTRODUCTION

Automatic recognition of speech by machine has been a goal of research for more than four decades. In the world of science, computer has always understood human mimics. The idea which generated for making speech recognition system is because it is convenient for humans to interact with a computer, robot or any machine through speech or vocalization rather than difficult instructions. Human beings have long been inspired to create computer that can understand and talk like human. Since, 1960s computer scientists have been researching various ways and means to make computer record, interpret and understand human speech. The fundamental aspect of speech recognition is the translation of sound into text and commands. Speech recognition is the process by which computer maps an acoustic speech signal to some form of abstract meaning of the speech. This process is highly difficult since sound has to be matched with stored sound bites on which further analysis has to be done because sound bites do not match with pre-existing sound pieces. Various feature extraction methods and pattern matching techniques are used to make better quality speech recognition systems. Feature extraction technique and

pattern matching techniques plays important role in speech recognition system to maximize the rate of speech recognition of various persons.

2. CLASSIFICATION OF SPEECH RECOGNITION SYSTEM

2.1 Types of speech recognition system based on utterances

2.1.1 Isolated Words Isolated word recognition system which recognizes single utterances i.e. single word. Isolated word recognition is suitable for situations where the user is required to give only one word response or commands, but it is very unnatural for multiple word inputs. It is simple and easiest for implementation because word boundaries are obvious and the words tend to be clearly pronounced which is the major advantage of this type. The drawback of this type is choosing different boundaries affects the results.

2.1.2 Connected Words A connected words system is similar to isolated words, but it allows separate utterances to be "run-together" with a minimal pause between them. Utterance is the vocalization of a word or words that represent a single meaning to the computer.

2.1.3 Continuous Speech Continuous speech recognition system allows users to speak almost naturally, while the computer determines its content. Basically, it is computer dictation. In this closest words run together without pause or any other division between words. Continuous speech recognition system is difficult to develop.

2.1.4 Spontaneous Speech Spontaneous speech recognition system recognizes the natural speech. Spontaneous speech is natural that comes suddenly through mouth. An ASR system with spontaneous speech is able to handle a variety of natural speech features such as words being run together. Spontaneous speech may include mispronunciation, false-starts and non words

2.2 Types of speech recognition based on Vocabulary

The size of vocabulary of a speech recognition system can affect the complexity, processing and the rate of recognition of ASR system. So that ASR system are classified based on the vocabulary as following: • Small Vocabulary - 1 to 100 words or sentences • Medium Vocabulary - 101 to 1000 words or sentences • Large Vocabulary- 1001 to 10,000 words or sentences • Very-large vocabulary - More than 10,000 words or sentences

3.0 DIFFERENT FEATURE EXTRACTION TECHNIQUE USED IN SPEECH RECOGNITION

Table 1. Different Feature Extraction Methods Used In Speech Recognition System

Sr. No.	Method	Property	Advantage	Disadvantage
1	Principal component Analysis (PCA)	Nonlinear feature extraction method, Linear map; rapid; Eigen vector-based,	Good result for Gaussian data.	The directions maximizing variance do not always maximize information.
2	Linear Discriminate Analysis(LDA)	Supervised linear map, Depend on Eigen vector, Nonlinear feature extraction method.	Better than PCA for classification, Handles the case where the withinclass frequencies are unequal and their performance has been examined on randomly generated test data.	If the distribution is significantly non-Gaussian the LDA projection will not be able to preserve any complex structure of the data, which may be needed for classification.
3	Independent component Analysis(ICA)	Nonlinear feature extraction method, Linear map, iterative non-Gaussian.	Blind than PCA for classification	Extracted components are not ordered.

4	Linear Predictive coding	10 to 16 lower sequence coefficient, Static feature extraction method	Spectral analysis is done with a fixed resolution along a subjective frequency scale i.e. Mel frequency scale	Frequencies are weighted equally on a linear scale while the frequency sensitivity of the human ear is close to the logarithmic.
5	Filter Bank analysis	Filter tuned required frequencies	It provide a spectral analysis with any degree of frequency resolution (wide or narrow), even with nonlinear filter spacing and bandwidths.	always take more calculation and processing time than discrete Fourier analysis using the FFT
6	Mel-frequency Cepstrum Coefficients (MFCC)	Power spectrum is computed by implementing Fourier Analysis.	This method is used for find our features.	MFCC values are not very robust in the presence of additive noises it is common to normalize their values in speech recognition system to reduce the influence of noise.
7	Kernel based feature extraction method	Nonlinear transformations	Dimensionality reduction leads to better classification and it is used to remove noisy and redundant features and improvement in classification error.	Slow similarity calculation speed [14].
8	Wavelet	Better time resolution than Fourier Transform	It replaces the fixed bandwidth of Fourier transform with one proportional to frequency which allows better time resolution at high frequencies than Fourier transform	It requires longer compression time.
9	Cepstral Mean Subtraction	Robust Feature Extraction	It is same as MFCC but working on Mean statically parameter.	
10	RASTA Filtering	For Noisy speech	It find out feature in noisy data	It increases the dependence of the data on its previous context.

4.0 FUNCTIONING OF SPEECH RECOGNITION SYSTEM

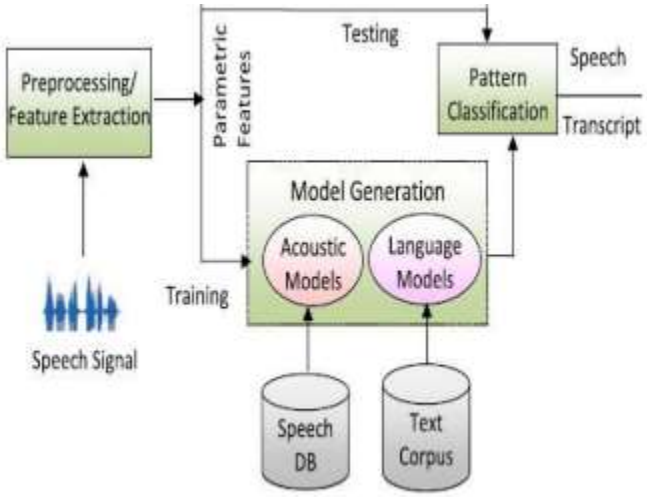


Fig 1: System Architecture of for Automatic Speech Recognition System

3.1 Feature Extraction

Feature extraction step finds the set of parameters of utterances that have acoustic correlation with speech signals and these parameters are computed through processing of the acoustic waveform. These parameters are known as features. The main focus of feature extractor is to keep the relevant information and discard irrelevant one. To act upon this operation, feature extractor divides the acoustic signal into 10- 25 ms. Data acquired in these frames is multiplied by window function. There are many types of window functions that can be used such as hamming Rectangular, Blackman, Welch or Gaussian etc. In this way features have been extracted from every frame. There are several methods for feature extraction such as Mel-Frequency Cepstral Coefficient (MFCC), Linear Predictive Cepstral Coefficient (LPCC), Perceptual Linear Prediction (PLP), wavelet and RASTA-PLP (Relative Spectral Transform)Processing etc.

Input: Digital speech signal (vector of sampled values)

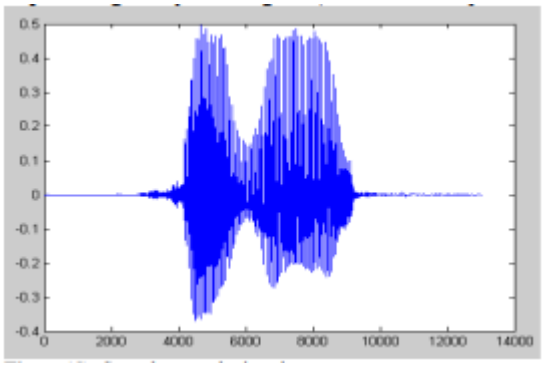


Fig 2. Sample speech signal.

3.2 Mel Frequency Cepstral Coefficients (MFCC's)

MFCC's are coefficients that represent audio, based on perception. It is derived from the Fourier Transform or the Discrete Cosine Transform of the audio clip. The basic difference between the FFT/DCT and the MFCC is that in the MFCC, the frequency bands are positioned logarithmically (on the mel scale).which approximates the human auditory system's response more closely than the linearly spaced frequency bands of FFT or DCT. This allows for better processing of data, for example, in audio compression. The main purpose of the MFCC] processor is to mimic the behaviour of the human ears.

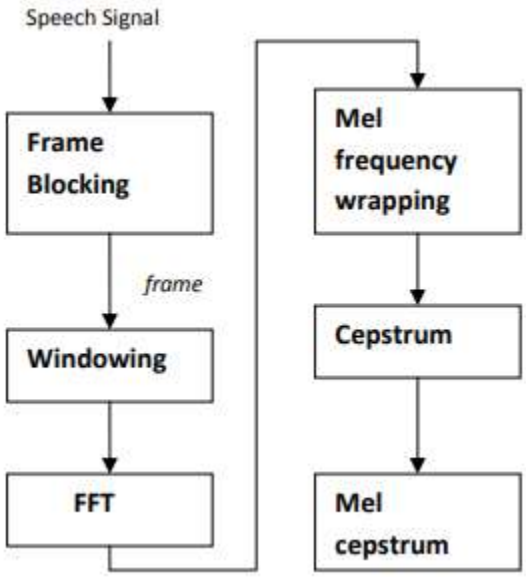


Fig 3. Figure(c): MFCC Block Diagram

4.0 TRAINING A NEURAL NETWORK WITH GMM LAYER Fig. 2 shows the structure of the embedded GMM within a Neural Network (NN) architecture during optimization with Cross Entropy (CE). There are three modules in this figure. The top module estimates priors online during CE training. This is done by inputting a constant value of 1 and the reference frame label into the softmax. The logarithm of the estimated prior distribution, log(p(s)), is then calculated. The GMM layer is used as the last layer of the network. If the GMM layer is used after the stack of hidden layers, the corresponding structure is referred to as Deep GMM (DGMM), otherwise it will be called Shallow GMM. The GMM layer outputs −log(p(x|s)) for each frame of the data. By subtracting this value from the logarithm of the prior, the logarithm of the joint distribution can be formed: log(p(x, s)) = log(p(s)) − (−log(p(x|s))). (4) The state posteriors for CE training can then be derived using p(s|x) = exp(log(p(x, s))) P s exp(log(p(x, s))). In this paper all experiments used CE training. However, the proposed design allows the use of any optimization criterion,

such as Maximum Likelihood (ML), CE, or Discriminative Sequence Training [18] [19]. (Use of ML to optimize the entire structure requires care, since the Jacobian of the bottleneck features has to be handled; this is not detailed here). The trainable sub-layers are initialized as follows. The μ-layer values are initialized to random numbers derived from a normal distribution, N (0, 1). All parameters in the Σ-layer are initialized with a constant value of 0 (corresponding to a transformed value of 1) and the parameters in the ω-layer are initialized with a uniform value of 1/g. We considered the effect on training of each parameter separately. Our observation was that training all parameters jointly achieves the best performance, but requires more overall training steps than first training just the means and weights while fixing the variances, before then training all parameters together

5.0 Testing the identification system

To determine the optimal parameters of the model, 27 models of different architecture were built, for each model, testing was performed with 40 added speakers. The results are shown in the table. The values of Accuracy, Precision, Recall are calculated as the arithmetic mean of each class (speaker). For comparison, the average time of decision-making by the classifier was measured depending on the duration of the audio fragment of the speech submitted to the input for identification (columns "Time for 40 s" and "Time for 5 s").

Компоненти	**Коваріація**	**MFCC**	**Accuracy**	**Precision**	**Recall**	**Час (на 40 с)**	**Час (на 5 с)**
8	**diag**	**13**	**0.9946**	**0.8125**	**0.8916**	**2.2191**	**1.6222**
8	**diag**	**20**	**0.951**	**0.8125**	**0.9**	**2.2509**	**1.7025**
8	**diag**	**8**	**0.942**	**0.8203**	**0.8833**	**2.1254**	**1.4941**

8	**full**	**13**	**0.945**	**0.8375**	**0.8916**	**4.71203**	**3.1141**
8	**full**	**20**	**0.945**	**0.8511**	**0.8916**	**5.2238**	**3.4823**
8	**full**	**8**	**0.945**	**0.8125**	**0.898**	**3.3213**	**2.1422**
8	**tied**	**13**	**0.945**	**0.8502**	**0.898**	**4.5222**	**3.0525**
8	**tied**	**20**	**0.954**	**0.8875**	**0.9083**	**5.3009**	**3.5039**
8	**tied**	**8**	**0.952**	**0.8527**	**0.878**	**3.2472**	**2.0821**
16	**diag**	**13**	**0.937**	**0.8751**	**0.925**	**2.45507**	**1.67226**
16	**diag**	**20**	**0.954**	**0.8875**	**0.9**	**2.5688**	**1.7152**
16	**diag**	**8**	**0.937**	**0.8125**	**0.875**	**2.3292**	**1.6191**
16	**full**	**13**	**0.951**	**0.8501**	**0.9**	**6.2387**	**4.1502**
16	**full**	**20**	**0.951**	**0.8512**	**0.9**	**8.86304**	**5.6113**
16	**full**	**8**	**0.951**	**0.8524**	**0.9**	**4.6579**	**2.96404**
16	**tied**	**13**	**0.963**	**0.8503**	**0.925**	**6.2921**	**4.0401**
16	**tied**	**20**	**0.963**	**0.8125**	**0.925**	**8.6095**	**5.5437**
16	**tied**	**8**	**0.946**	**0.8208**	**0.875**	**4.6362**	**2.8073**
32	**diag**	**13**	**0.954**	**0.8125**	**0.875**	**2.82457**	**2.05204**
32	**diag**	**20**	**0.958**	**0.8875**	**0.925**	**2.8914**	**2.0058**
32	**diag**	**8**	**0.951**	**0.8513**	**0.925**	**2.7116**	**1.9027**
32	**full**	**13**	**0.954**	**0.8629**	**0.9083**	**12.5033**	**8.06306**
32	**full**	**20**	**0.954**	**0.8629**	**0.9083**	**15.4039**	**9.91009**
32	**full**	**8**	**0.954**	**0.8629**	**0.9083**	**7.4432**	**4.3401**
32	**tied**	**13**	**0.951**	**0.8125**	**0.886**	**12.8066**	**8.0172**
32	**tied**	**20**	**0.954**	**0.8875**	**0.9083**	**15.3268**	**9.8261**

32	**tied**	**8**	**0.9951**	**0.8208**	**0.965**	**7.3689**	**4.3601**

Conclusion

Over the last decade, the GMM has become established as the standard classifier for text-independent speaker recognition . It operates on atomic levels of speech and can be effective with very small amounts of speaker specific training data. The primary focus of this work was on a task domain for a real application, such as voice mail labelling and retrieval. The Gaussian Mixture speaker model was specifically evaluated for identification tasks using short duration utterances from unconstrained conversational speech, possibly transmitted over noisy telephone channels. Gaussian mixture models were motivated for modelling speaker identity based on two interpretations. The component Gaussians were first shown to represent characteristic spectral shapes (vocal tract configurations) from the phonetic sounds which comprise a person's voice. By modelling the underlying acoustic classes, the speaker model is better able to model the short term variations of a person's voice, allowing high identification performance. In the end, created the system for voice identification purposes, which may be integrated to the future products.