МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Київський національний університет імені Тараса Шевченка
Навчально-науковий інститут філології
Кафедра української мови та прикладної лінгвістики

ВПЛИВ АКУСТИЧНИХ ХАРАКТЕРИСТИК МОВЛЕННЄВИХ ДАНИХ НА ЯКІСТЬ АВТОМАТИЧНОГО РОЗПІЗНАВАННЯ МОВЦІВ

Кваліфікаційна робота
освітнього ступеня «бакалавр»
студентки IV курсу
ОПП «Прикладна (комп'ютерна) лінгвістика та англійська мова»,
спеціальності 035 «Філологія»,
спеціалізації 035.10 «Прикладна лінгвістика»,
галузі знань 03 «Гуманітарні науки»
Юліани КАРПЦОВОЇ

Науковий керівник:
асист. кафедри
української мови
та прикладної лінгвістики
Валентина РОБЕЙКО

«Допущено до захисту»
Протокол засідання кафедри
української мови та прикладної лінгвістики
від 5.06. № 16
Завідувач кафедри **Сергій РІЗНИК**

КИЇВ – 2026

АНОТАЦІЯ

Ця робота присвячена дослідженню впливу акустичних характеристик мовленнєвих даних на якість автоматичного розпізнавання мовців на матеріалі українськомовних корпусів.

Об’єктом дослідження є системи автоматичного розпізнавання мовців.

Предметом є вплив акустичних і кількісних характеристик мовленнєвих даних на точність роботи таких систем.

Методи, застосовані в цьому дослідженні для аналізу результатів, належать до традиційних лінгвістичних методів (перцептивний аналіз, аналіз формантної структури та паузації, показник зашумленості) та статистичних методів (оцінка ступеня впливу аналізованого явища на значення метрик, порівняння розбіжностей всередині груп та між групами).

У першому розділі міститься роз’яснення основних термінів та розгляд принципів роботи систем розпізнавання голосів. Також розділ містить огляд наукових джерел задля визначення акустичних та кількісних параметрів, які впливають на результати систем, і обґрунтування вибору моделей розпізнавання мовця для тестування.

Другий розділ містить опис створеної системи та її інтерфейсу, а також описи сформованих для тестування датасетів. У цьому розділі описано хід експерименту та здійснено попередній аналіз отриманих результатів.

У третьому розділі міститься лінгвістичний аналіз використаних для експерименту даних, порівняння отриманих значень метрик та перцептивний аналіз помилок, яких припустилася кожна модель при її використанні у створеній системі.

В кінці роботи міститься загальний висновок та список використаних наукових джерел, який містить 41 посилання. До роботи докладено Додаток, де розміщено посилання на github-репозиторій зі створеною системою.

Ключові слова: Автоматичне розпізнавання мовця, Ідентифікація мовця, Кластеризація ембедингів, Обробка мовлення, Акустичні параметри, Моделі автоматичного розпізнавання мовця, ReDimNet, ResNet, ECAPA-TDNN.

ABSTRACT

This paper investigates the impact of the acoustic characteristics of speech data on the accuracy of automatic speaker recognition using a corpus of Ukrainian-language recordings.

The object of the study is automatic speaker recognition systems.

The subject is the influence of acoustic and quantitative characteristics of data on the accuracy of such systems.

The methods used in this study to analyze the results include traditional linguistic methods (perceptual analysis, analysis of formant structure and pauses, signal-to-noise ratio) and statistical methods (evaluating the influence of the analyzed phenomenon on metric values, comparing discrepancies within and between groups).

The first chapter provides an explanation of key terms and discusses the principles of voice recognition systems. The section also includes a review of scientific sources to identify the acoustic and quantitative parameters that influence system performance, as well as a justification for the selection of speaker recognition models for testing.

The second section describes the created system and its interface, as well as the datasets formed for testing. This section describes the course of the experiment and provides a preliminary analysis of the results obtained.

The third section contains a linguistic analysis of the data used for the experiment, a comparison of the obtained metric values, and a perceptual analysis of the errors made by each model when used in the system.

The conclusion and a list of references, containing 41 citations, are provided at the end. An Appendix is attached to the work, containing a link to the GitHub repository with the developed system.

Keywords: Automatic speaker recognition, Speaker identification, Embedding clustering, Speech processing, Acoustic parameters, Automatic speaker recognition models, ReDimNet, ResNet, ECAPA-TDNN.

Зміст

Вступ

Темою цього дослідження є вплив акустичних характеристик мовленнєвих даних на якість автоматичного розпізнавання мовців.

Актуальність теми полягає в тому, що з розвитком машинного навчання зростає потреба в голосових системах розпізнавання особи, що базуються на цих технологіях (Fortune Business Insights, 2024, p. 1). Розпізнавання мовців у великих масивах аудіозаписів є важливим завданням для аналізу корпусів телефонних розмов, записів зустрічей, публічних виступів, застосування в системах моніторингу та інших ситуаціях, коли записи не мають позначок про те, хто говорить (Ghaemmaghami et al, 2016, p. 1). Точність роботи моделей автоматичного розпізнавання мовців значною мірою залежить від характеристик вхідних даних, що використовуються для їх навчання та тестування – наприклад, модель, навчена на дуже якісних записах переважно чоловічих голосів, може демонструвати гірший результат на телефонних записах чи аудіофайлах із жіночими голосами. Тому постає проблема дослідження впливу таких характеристик мовленнєвих даних на точність роботи моделей розпізнавання мовців.

Мета роботи – виявити закономірності впливу акустичних та кількісних характеристик мовленнєвих даних на якість роботи систем автоматичного розпізнавання мовців на матеріалі українськомовних корпусів.

У цій роботі було поставлено такі **завдання**:

- огляд підходів до створення систем автоматичного розпізнавання мовця та принципів їх роботи;
- визначення акустичних та кількісних характеристик мовленнєвих даних, що найбільше впливають на точність роботи таких систем;
- розгляд актуальних наукових праць та вибір найсучасніших моделей розпізнавання мовця для їх подальшого тестування та порівняння;
- створення системи автоматичного розпізнавання мовців шляхом кластеризації;

- формування та опис наборів даних для оцінки ступеня впливу кількісних та акустичних характеристик даних на результат роботи системи;
- проведення експерименту із використанням сформованих датасетів та створеної системи;
- проведення лінгвістичного аналізу наборів мовленнєвих даних для виявлення їх акустичних особливостей;
- аналіз впливу виявлених акустичних особливостей на точність роботи системи та виявлення закономірностей впливу цих характеристик на кожну з моделей.

Об’єктом дослідження є системи автоматичного розпізнавання мовців.

Предметом є вплив акустичних і кількісних характеристик мовленнєвих даних на точність роботи таких систем.

Методи, застосовані в цьому дослідженні для аналізу результатів, належать до традиційних лінгвістичних методів (перцептивний аналіз, аналіз формантної структури та паузації, показник зашумленості) та статистичних методів (оцінка ступеня впливу аналізованого явища на значення метрик, порівняння розбіжностей всередині груп та між групами).

Новизна полягає в тому, що у цій роботі було вперше проведено комплексне дослідження впливу акустичних та кількісних характеристик на роботу найсучасніших моделей розпізнавання мовця на матеріалі українськомовних записів.

Матеріалом дослідження є корпуси аудіозаписів українського мовлення, які було зібрано із відкритих джерел, взято з доступних наборів даних чи записано у природних умовах.

Практичне значення цієї роботи полягає в тому, що на зараз існує потреба у створенні та тестуванні систем автоматичного розпізнавання мовця саме для українськомовних даних. З накопиченням неструктурованих та нерозмічених наборів аудіозаписів, таких як телефонні перехоплення або записи систем безпеки,

зростає потреба в інструментах, які можуть виявляти, скільки унікальних голосів мовців присутні у всьому наборі, і який запис голосу належить кожному з них. З'ясування особливостей впливу кількісних та акустичних характеристик на роботу систем автоматичного розпізнавання мовця допоможе надалі вдосконалити ці системи та враховувати особливості мовленнєвих даних при застосуванні систем в реальних умовах.

Робота складається з трьох розділів: «Теоретичні основи автоматичного розпізнавання мовців», «Проведення експерименту», «Аналіз результатів експерименту».

У першому розділі міститься роз'яснення основних термінів та розгляд принципів роботи систем розпізнавання голосів. Також розділ містить огляд наукових джерел задля визначення акустичних та кількісних параметрів, які впливають на результати систем, і обґрунтування вибору моделей розпізнавання мовця для тестування.

Другий розділ містить опис створеної системи та її інтерфейсу, а також описи сформованих для тестування датасетів. У цьому розділі описано хід експерименту та здійснено попередній аналіз отриманих результатів.

У третьому розділі міститься лінгвістичний аналіз використаних для експерименту даних, розгорнутий аналіз та порівняння отриманих значень метрик та перцептивний аналіз помилок, яких припустилася кожна модель при її використанні у створеній системі.

Повний обсяг роботи – 91 сторінка. В кінці міститься загальний висновок та список використаних наукових джерел, який містить 41 посилання. До роботи докладено Додаток, де розміщено посилання на github-репозиторій зі створеною системою.

СПИСОК СКОРОЧЕНЬ

ARI (Adjusted Rand Index) – скоригований індекс Ренда

ASR (Automatic Speaker Recognition) – автоматичне розпізнавання мовця

DBSCAN (Density-based spatial clustering of applications with noise) – просторова кластеризація на основі щільності з врахуванням шуму

DER (Diarization Error Rate) – діаризаційна помилка

ECAPA-TDNN (Emphasized Channel Attention, Propagation, and Aggregation Time Delay Neural Network) – нейронна мережа з часовими затримками, підсиленою увагою до каналів, поширенням та агрегацією ознак

HDBSCAN (Hierarchical density-based spatial clustering of applications with noise) – ієрархічна просторова кластеризація на основі щільності з врахуванням шуму

MFCC (Mel-frequency cepstral coefficients) – мел-частотні кепстральні коефіцієнти

NMI (Normalized Mutual Information) – нормалізована взаємна інформація

ResNet (Residual neural network) – залишкова нейронна мережа

SNR (signal-to-noise ratio) – співвідношення сигналу та шуму

t-SNE (t-distributed Stochastic Neighbor Embedding) – T-розподілене вкладення стохастичної близькості

VAD (voice activity detection) – виявлення голосової активності

ЧОТ – частота основного тону

РОЗДІЛ 1. Теоретичні основи автоматичного розпізнавання мовців

У цьому розділі ми розглянемо ключові поняття, роз'яснення яких необхідне для проведення подальшого дослідження. Буде оглянуто підходи до створення та принципи роботи систем, які виконують розпізнавання мовця за допомогою моделей та кластеризації ембедингів. Далі на основі аналізу наукових робіт буде визначено кількісні та акустичні характеристики даних, що найбільше впливають на точність роботи систем. Також буде проаналізовано праці, що стосуються теми та поставленого завдання, з метою вибору моделей для подальшого дослідження.

1.1. **Огляд основних понять**

На початку варто роз'яснити основні поняття, якими ми будемо оперувати в цьому дослідженні: автоматична ідентифікація за голосом, верифікація мовця, діаризація та кластеризація. Крім того, варто пояснити різницю між ідентифікацією та верифікацією мовця, адже ця відмінність є досить важливою для поставленого завдання та нашого подальшого дослідження.

Для пошуку дефініцій варто звернутися до наукових джерел цієї сфери. Зокрема, для поняття «автоматична ідентифікація мовця» у праці М.М.Кабіра (2021, с. 2) подається таке визначення:

«**Ідентифікація** визначає особу анонімного мовця на основі його висловлювань. Вона дозволяє знайти конкретного мовця серед набору розпізнаних голосів. Це спосіб виявлення людини за її висловлюваннями, що зберігаються в базі даних. Такий підхід є зіставленням за схемою 1:N, де конкретне висловлювання порівнюється з N шаблонами.» (тут і далі переклад наш. – Ю. К.)

Голосова верифікація дещо відрізняється від попередньої задачі. Для неї М.М.Кабір (2021, с. 2) наводить таку дефініцію:

«**Верифікація** мовця використовує голос для підтвердження особи. У системах ідентифікації отримані характеристики голосу порівнюються з характеристиками всіх мовців, зібраними в базі. Натомість у системах верифікації

вони порівнюються лише з особливостями голосу мовця, за якого себе видає людина. Це зіставлення за схемою 1:1, де голос одного мовця порівнюється лише з одним зразком.»

Також варто привести визначення поняття «діаризація», яке теж є дотичним до завдань розробленої системи й визначається у роботі М.М.Кабіра (2021, с. 2) так:

«**Діаризація** – це поділ звукового запису, у якому присутні голоси кількох осіб, на сегменти, що відповідають кожному з мовців.»

Всі три наведені поняття є підзавданнями ширшої галузі – автоматичного розпізнавання мовця (англ. Automatic Speaker Recognition, ASR), яка охоплює всі перелічені задачі.

Далі варто розглянути інше поняття, яке є важливим для цього дослідження. Протас (2021, с. 7) наводить таке визначення кластеризації:

«**Кластеризація** – це задача розбиття множини об'єктів на підмножини (кластери) таким чином, щоб об'єкти з одного кластера були більш схожі один на одного, ніж на об'єкти з інших кластерів, за певним критерієм. Кластеризацію можна визначити як задачу групування об'єктів за певною мірою подібності».

Задача кластеризації є ключовою для створеної системи, адже її точність буде залежати не тільки від роботи моделі верифікації мовця, але й від вибраного алгоритму кластеризації. Навіть отримавши за допомогою моделі дуже якісні векторні представлення записів, які демонструють велику розбіжність між різними голосами, неможливо досягти високої точності без правильно підібраного та коректно налаштованого алгоритму кластеризації.

Поставлені задачі поєднують елементи діаризації, ідентифікації та верифікації мовця, а також залучають кластеризацію. Алгоритм кластеризації виконує роль класифікатора у нашій системі, використовуючи оцінку відстані між **ембедингами** (векторними представленнями записів, отриманими за допомогою моделі). Роблячи це, він зіставляє певний ембединг одразу із кількома сусідніми, що робить цей процес близьким до ідентифікації. Крім того, це завдання є одним з

етапів діаризації, адже вона включає кластеризацію відрізків записів з метою розподілу висловлювань за мовцями.

Врешті, варто дати роз'яснення тому принципу, за яким працює система, створена у цьому дослідженні для вимірювання точності моделей розпізнавання мовця.

Автоматичне розпізнавання мовців шляхом кластеризації – це групування звукозаписів за належністю голосу одній людині. У цьому завданні заздалегідь не відома кількість голосів мовців чи їх індивідуальні ознаки, за якими їх можна відрізнити. Завдання охоплює виявлення кількості мовців та визначення, які відрізки мовлення їм належать.

За класифікацією систем розпізнавання мовця це завдання «закритого набору», адже система працює лише з конкретним набором голосів, які містяться у вхідних даних (Kabir, 2021, p. 3).

За поділом на текстово-незалежні та текстово-залежні завдання, така система є текстово-незалежною. Використані підходи, а саме текстово-незалежне розпізнавання голосів за допомогою моделей верифікації мовця, передбачають що результати роботи системи не будуть залежати від конкретних висловлювань мовця у вхідних даних (Борисов і Трапезон, 2024, с. 2; Kabir, 2021, p. 3), що є дуже корисним для подальшого порівняння результатів роботи цієї системи на датасетах читаного та спонтанного мовлення.

1.2. **Основні принципи роботи систем автоматичного розпізнавання мовця**

Далі буде наведено огляд принципів роботи систем автоматичного розпізнавання мовця та методів, що використовуються на різних етапах їх роботи.

У сучасних системах автоматичного розпізнавання мовця можна виокремити два основних підходи: **наскрізні моделі** (англ. end-to-end), які опрацьовують запис за один етап, та **модульні конвеєри** (англ. modular pipelines), які передбачають послідовне виконання окремих етапів, до яких в тому числі належить кластеризація

(Fujita et al, 2019, p. 2). Також виділяють і **гібридні рішення**, що поєднують елементи обох підходів задля об’єднання їх переваг та подолання їхніх обмежень. Такі системи зазвичай створюються для конкретної задачі із врахуванням її специфіки (Lam, 2024, p. 3).

1.2.1. **Модульні конвеєри**

Один із традиційних підходів до створення систем автоматичного розпізнавання мовця, який був особливо поширеним до 2019–2020 років (Fujita et al, 2019, p. 2; Tranter and Reynolds, 2006, p. 2), ґрунтується на **модульній архітектурі з використанням кластеризації**. Він вважався найпоширенішим до того, як наскрізні моделі набули популярності (Fujita et al, 2019, p. 2). Цей метод передбачає побудову конвеєра з кількох незалежних етапів, кожен з яких виконується за допомогою окремої моделі. Один з варіантів будови такого конвеєра зображено на рис. 1.1.

Типова реалізація включає такі компоненти:

1. Виявлення мовленнєвої активності (VAD) – на цьому етапі система визначає сегменти запису, де присутній голос, відсіюючи тишу, шум або інші нерелевантні ділянки сигналу.
2. Виявлення зміни мовця – ідентифікуються моменти, коли голос одного мовця змінюється голосом іншого. Цей етап особливо важливий у випадках накладання голосів у записі.
3. Сегментація – аудіозапис розділяється на короткі ділянки, у межах яких, за припущенням, говорить лише один мовець.
4. Отримання векторних представлень – з кожного сегмента отримуються ембедінги, що представляють ознаки голосу. Перед подачею на вхід до моделі, сегменти аудіозаписів можуть проходити попередню обробку у вигляді перетворення на спектрограми чи інші ознаки, а отримані ембедінги можуть проходити постобробку у вигляді нормалізації.

5. Кластеризація – ембединги сегментів групуються за подібністю за допомогою алгоритмів кластеризації, дозволяючи виявити, які сегменти належать одному мовцю.

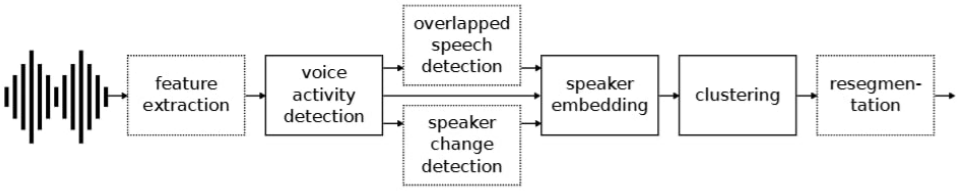

Рис. 1.1. Набір модулів, які можуть бути поєднані для побудови системи автоматичного розпізнавання мовця. Загальна послідовність етапів у традиційному конвеєрі (*Yin et al, 2020, p. 2).

Цей підхід включає комбінацію мінімум трьох різних моделей: для виявлення мовлення, для отримання векторних представлень, а також для кластеризації. Така архітектура дозволяє замінювати окремі компоненти системи, що робить її більш гнучкою.

Попри свою поширеність, цей традиційний підхід має низку обмежень. Наприклад, під час етапу тренування компоненти системи навчаються незалежно та не оптимізуються спільно під одну задачу. Також алгоритм чи модель, що виконує кластеризацію ембедингів звукозаписів, зазвичай не навчається для такої задачі або взагалі не піддається навчанню.

Водночас у цього підходу є і переваги, які забезпечують його популярність. Зокрема, якщо використовується алгоритм кластеризації, який не потребує вказування кількості кластерів, система може успішно працювати із набором даних з невизначеною кількістю мовців.

1.2.2. **Наскрізні моделі**

З огляду на обмеження традиційного підходу до побудови систем розпізнавання мовця, у подальших дослідженнях було запропоновано наскрізні нейронні моделі (англ. end-to-end neural speaker diarization) (Fujita et al, 2019, p. 2).

На відміну від традиційних систем, наскрізна модель не має окремого етапу кластеризації. Будову такої системи представлено на рис. 1.2: замість модульної архітектури вона має одну цілісну нейронну мережу. Система отримує на вхід повний аудіозапис, включно із моментами з перекриттям голосів та тишею, у вигляді необробленого сигналу або перетвореним на набір необхідних ознак (залежно від типу нейронної мережі). Далі система ділить аудіозапис на сегменти за мовцями безпосередньо під час проходження сигналу через мережу. Під час навчання така модель оптимізується за допомогою функції втрати, орієнтованої на завдання діаризації.

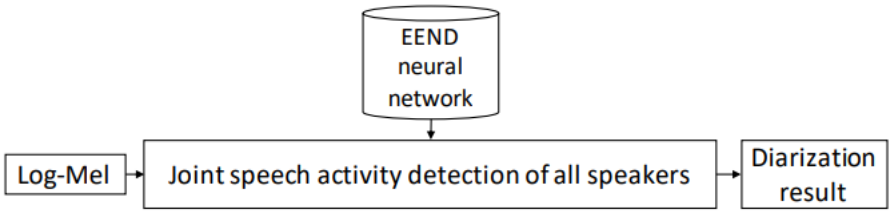

Рис. 1.2. Наскрізний метод. Будова системи, заснованої на підході з використанням наскрізної мережі (Fujita et al, 2019, p. 2).

Серед ключових переваг наскрізної моделі слід виокремити її вищу точність у порівнянні з підходами на основі кластеризації. За результатами експериментів (Fujita et al, 2019, p. 5), наскрізні системи демонструють значно нижчий рівень діаризаційної помилки (англ. DER), що свідчить про їхню ефективність у складних акустичних умовах.

Проте, попри високу точність, цей підхід має і недоліки. У більшості наскрізних моделей кількість мовців, яких система здатна розрізнити, задається фіксовано під час тренування. Це створює обмеження у застосуванні таких моделей до даних із невідомою або варіативною кількістю мовців (Horiguchi et al, 2020, p. 1).

1.2.3. **Гібридні рішення**

Одним із подальших напрямків розвитку систем автоматичного розпізнавання мовця стала поява гібридних систем, що поєднують переваги наскрізних підходів із більшою гнучкістю щодо кількості мовців. Зокрема, було запропоновано новий варіант наскрізної архітектури, який визначає кількість мовців за допомогою спеціального механізму кластеризації – атракторів (Horiguchi et al, 2020, p. 2).

Така архітектура пропонує додатковий модуль, що приймає на вхід послідовність ембедингів і генерує так звані атракторні точки – вектори, які виступають у ролі центрів кластерів і допомагають розподіляти сегменти мовлення за належністю різним мовцям. Таким чином, замість традиційної кластеризації або фіксованої кількості вихідних каналів, система динамічно приймає рішення про кількість мовців у записі. Модуль атракторів оптимізується разом з рештою мережі під час навчання.

Однак модель з атракторами має і певні обмеження, пов’язані зі способом формування цих атракторних точок. Зокрема, атрактори можуть бути чутливими до порядку ембедингів, поданих на вхід модуля. Результати кластеризації можуть змінюватися залежно від того, чи ембединги подаються в хронологічному чи випадковому порядку. У розглянутій роботі для уникнення прив’язки до послідовності використовується перемішаний порядок (Horiguchi et al, 2020, p. 4).

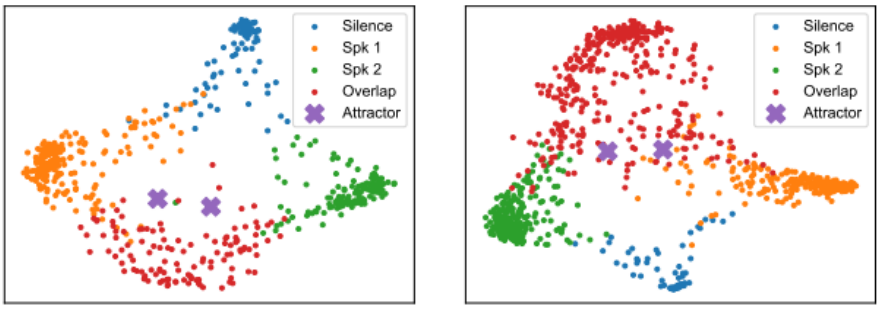

Рис. 1.3. Візуалізація ембедингів та атракторів для двох мовців (Horiguchi et al, 2020, p. 4).

На рис. 1.3 зображено візуалізацію ембедингів різних сегментів мовлення (голос першого мовця, голос другого мовця, накладання голосів, тиша) та точки-атрактори, за допомогою яких виконується розподіл.

Було розглянуто й інші гібридні підходи: зокрема, у роботі Кіношіти, Делькруа і Тавари (2021, p. 2) було запропоновано гібридний підхід, що теж поєднує переваги наскрізних архітектур та традиційних методів на основі кластеризації.

Суть методу полягає в тому, що аудіозапис розбивається на коротші частини, кожна з яких передбачає присутність щонайбільше двох мовців (у статті зазначено, що кількість мовців на сегмент є змінюваним параметром). Далі кожна така частина поділяється на відрізки, що містять голос лише одного мовця, після чого з кожного відрізку отримується його представлення за допомогою моделі. Наступний етап – це кластеризація всіх ембедингів, отриманих із всіх частин повного аудіозапису. Цей процес схематично зображено на рис. 1.4.

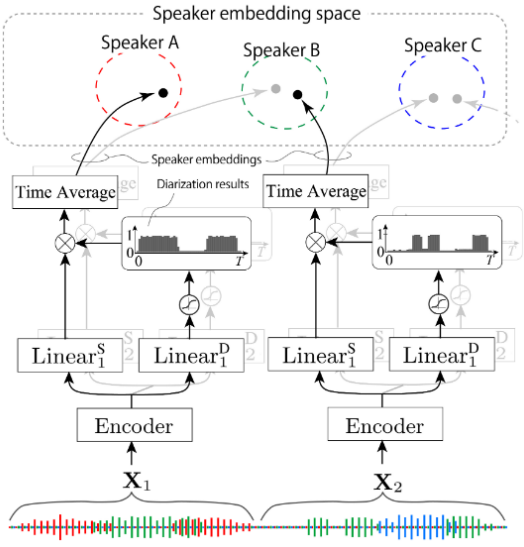

Рис. 1.4. Діаграма запропонованої системи. Вхідні дані містять голоси трьох мовців всього (позначені червоним, зеленим та синім), але лише два з них присутні на кожному сегменті (Kinoshita, Delcroix and Tawara, 2021, p. 2).

Цей підхід може працювати з накладанням мовлення та обробляти довгі записи з довільною кількістю мовців без суттєвого погіршення точності. Проведені експерименти показали, що у випадку коротких аудіо (до 5 хвилин), гібридний

підхід демонструє порівнянну якість із наскрізною архітектурою, але при цьому перевершує її на довших записах (Kinoshita, Delcroix and Tawara, 2021, p. 4).

Проте в описаного підходу є обмеження – така система працює лише з визначеною кількістю мовців. Однак автори зазначають, що в поєднанні з архітектурою на зразок модуля атракторів, описаного вище, можлива модифікація моделі для роботи з невідомою кількістю мовців. Цей метод є спробою об'єднати переваги обох підходів: точність наскрізних систем на локальному рівні та узгодження на глобальному рівні за допомогою кластеризації.

Також варто розглянути конкретний приклад створення гібридної системи із поєднанням традиційних та новітніх методів, адаптованої під особливості певного завдання. Такий підхід використовується у дослідженні Ф. Лама (2024, p. 3). Метод, запропонований у цій статті, заснований на поєднанні діаризації та кластеризації. Етапи роботи системи включають виявлення сегментів із мовленням та отримання їх спектрограм, які подаються до моделі розпізнавання мовлення Whisper (Radford et al, 2023, p. 3). Отримані з неї ембедінги подаються в модель кластеризації Mix-SAE, представлену в дослідженні. Архітектуру системи та послідовність етапів оброблення даних зображено на рис. 1.5.

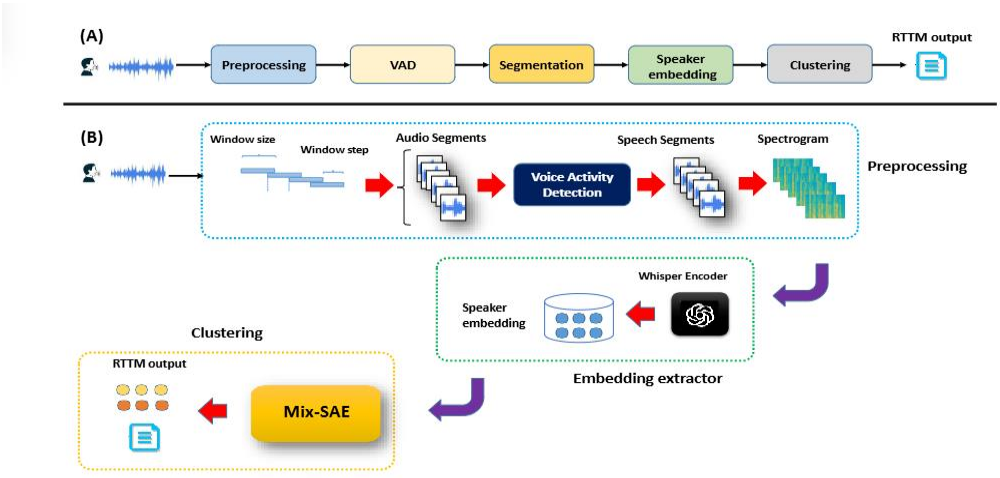

Рис. 1.5. Архітектури: (A) традиційної системи на основі кластеризації (B) запропонованої системи діаризації (Lam, 2024, p. 3).

1.2.4. **Теоретичні основи запропонованого підходу**

Система автоматичного розпізнавання мовця, створення якої є одним із поставлених завдань, належить до модульних конвеєрів та базується на підході на основі кластеризації ембедингів.

Завдання нашого дослідження включають оцінку та порівняння різних моделей, отже система повинна мати модульну архітектуру для легкої заміни однієї моделі на іншу. Також вона матиме єдиний алгоритм кластеризації, який оброблятиме отримані з моделі ембединги.

Запропонована система має ряд переваг, подібних до тих, якими володіють гібридні підходи. Наприклад, алгоритм кластеризації, що групуватиме ембединги за схожістю, буде динамічно визначати кількість різних голосів у наборі звукозаписів, що дозволить системі працювати з аудіокорпусами з невідомою кількістю мовців. Іншою перевагою є здатність системи працювати зі значною кількістю звукозаписів – на вхід може подаватися великий обсяг даних, розмір якого обмежений лише потужністю обчислювальних ресурсів.

Те, що в системі використовуватимуться претреновані моделі для етапу створення ембедингів, робить її ближчою до гібридних підходів, ніж до модульних архітектур. Запропонований метод є схожим на ті, що були описані у дослідженні Ф. Лама (2024, p. 3) та роботі Кіношіти, Делькруа і Тавари (2021, p. 2). Він так само є спробою сумістити переваги всіх підходів, об’єднуючи точність претренованих моделей на локальному рівні та узгодження результатів на глобальному рівні за допомогою кластеризації.

1.3. **Характеристики мовленнєвих даних, що можуть впливати на точність системи**

Оскільки метою цього дослідження є виявлення впливу акустичних та кількісних ознак даних на точність роботи моделей, далі варто виокремити конкретні ознаки, які будуть досліджуватися у подальших розділах. Для цього слід

звернутися до наукових досліджень, у яких було поставлено схожу мету або виокремлено чинники, що впливали на точність роботи моделей автоматичного розпізнавання мовця.

Наприклад, у дослідженні Борисова і Трапезона (2024, с. 5) було виявлено розбіжності у точності роботи моделей при розпізнаванні жіночих та чоловічих голосів: створені авторами нейронні мережі мали кращі показники для жіночого голосу, ніж для чоловічого, у завданні верифікації. У цьому дослідженні було створено та випробувано систему голосової верифікації, стійку до атак за допомогою клонування голосу. Щоправда, експеримент проводився на англомовних записах з відкритого набору даних LibriSpeech ASR corpus (Panayotov, Chen and Khudanpur, 2015, p. 1), в той час, як в нашому дослідженні тестування буде проводитися на україномовних даних.

У роботі Х.С. Рудої (2025) «Дослідження масштабованості систем біометричної автентифікації на основі ембеддінгів голосу» в ході експерименту було виявлено, що зі збільшенням числа зареєстрованих мовців точність системи суттєво падає, оскільки зростає ймовірність збігу ознак різних голосів. З п'яти систем верифікації з різним числом користувачів найкращі результати були отримані при середній кількості (16–20 осіб). Щоправда, ми припускаємо що спад точності при масштабуванні до більшої кількості користувачів міг відбутися через використання звукозаписів з іншого набору даних, які могли мати іншу якість. Крім того, серед рекомендацій, поданих в дослідженні, було тестування системи із використанням різних моделей, адже в ході всіх експериментів використовувалася лише модель TitaNet (Руда, 2025, с. 7). Цю рекомендацію було враховано у нашій роботі.

У дослідженні Зайця та ін. (2024) «Використання ембедінгів голосу в інтегрованих системах для діаризації мовців та виявлення зловмисників» порівнювалося три різні моделі для діаризації, і Pyannote продемонструвала найкращий результат із показником DER 0.09 (Заєць та ін., 2024, с. 8), тому

більшість експериментів проводилася саме з цією моделлю. Експеримент, що був присвячений окремому тестуванню Pyannote в завданні діаризації, показав, що при збільшенні кількості мовців модель демонструє гірші результати (зростає показник DER), а також її точність сильно залежить від рівня зашумленості середовища (результати цього дослідження подано в таблиці 1.1).

Аудіоумови	DER	JER
Зашумлене середовище; два мовці	0,19	0,19
Зашумлене середовище; 11 мовців	0,76	0,75
Чисте середовище; два мовці	0,07	0,07

Таблиця 1.1. Результати тестування моделі Pyannote в різних акустичних умовах та з різною кількістю мовців. (Заєць та ін., 2024, с. 7)

Інший етап тестування був проведений на українськомовних записах подкастів із використанням Pyannote для отримання ембедингів, і створена система досягла точності (accuracy) 93,75 % та повноти (recall) 100%, зробивши лише одну помилку (Заєць та ін., 2024, с. 11). Настільки високі результати могли бути зумовленими не тільки тим, що записи були зроблені в студійному середовищі, а й особливостями інтонацій учасників подкасту, адже це було підготовлене мовлення.

На основі проаналізованих досліджень можна зробити висновки, які саме характеристики звукозаписів мають значний вплив на точність роботи систем автоматичного розпізнавання мовця. В результаті проведеного аналізу було сформовано перелік кількісних та якісних особливостей мовленнєвих даних, вплив яких досліджуватиметься в практичній частині:

- стать мовця, що впливає на схожість голосів (Борисов і Трапезон, 2024, с. 5);
- кількість голосів, збільшення якої може призводити до помилок (Заєць та ін., 2024, с. 8; Руда, 2025, с. 7);
- зашумленість даних, яка може погіршувати точність, що є особливо актуальним для записів спонтанного мовлення в природних умовах (Заєць та ін., 2024, с. 7);

- більш підготовлене мовлення зі стабільними інтонаціями, близьке до читаного, яке може бути легшим для розпізнавання, ніж спонтанне (Заєць та ін., 2024, с. 11).

1.4. **Огляд актуальних досліджень**

Через актуальність досліджуваної нами сфери, останнім часом зростає кількість як вітчизняних, так і іноземних досліджень на цю тему. Зокрема, у 2023–2025 роках з'явилося кілька українських робіт, що досліджували особливості роботи систем автоматичного розпізнавання мовця.

Одна з цих робіт фокусувалася на тестуванні чотирьох моделей: Pyannote (Yin et al, 2020, p. 3), WavLM (Chen et al, 2022, p. 3), TitaNet (Koluguri, Park and Ginsburg, 2022, p. 2) та ECAPA-TDNN (Desplanques, Thienpondt and Demuynck, 2020, p. 1) з метою визначення найкращої моделі для роботи з українськомовними записами (Brydinskyi, 2024, p. 4). Модель ECAPA-TDNN продемонструвала найкращі результати (Brydinskyi, 2024, p. 8), що свідчить про доцільність її подальшого використання при роботі з українськомовними даними. Відкритий датасет, зібраний для експерименту в цій статті, є цінним інструментом для досліджень, тому буде використаний також і в нашій роботі разом із даними з інших відкритих джерел.

Іншим дослідженням, важливим для нашої роботи, є «Порівняння методів цифрової обробки сигналів та моделей глибинного навчання у голосовій аутентифікації» Х.С. Рудої та ін. (2024, с. 1). Ця робота представляє результати порівняння сучасних та традиційних підходів до верифікації мовця, а також детальний огляд моделей ResNet (Zeinali et al, 2019, p. 2) та вже згаданої вище ECAPA-TDNN. Моделі продемонстрували набагато кращі результати у завданні верифікації мовця, ніж традиційні методи. Найкращою виявилася ResNet із точністю 98,3% (Руда та ін., 2024, с. 17), що дозволяє стверджувати, що ця модель може мати кращі результати, ніж ECAPA-TDNN, або бути порівнянною з нею за

точністю. Експеримент був проведений на англомовних даних, тож є потреба у подальшому дослідженні ефективності цих моделей для українських записів.

Крім того, у 2024 р. було розроблено модель ReDimNet із новою архітектурою, що дозволяє їй досягати найкращих результатів у галузі розпізнавання мовця (Yakovlev et al, 2024, p. 1). У дослідженні було представлено порівняння варіантів ReDimNet із варіантами моделей з іншими архітектурами (наведено на рис. 1.6). Позаяк модель ReDimNet відносно нова, поки що немає наукових досліджень із її тестуванням на україномовних записах.

Рис. 1.6. Результати тестування на датасеті VoxCeleb моделей з архітектурою ReDimNet (позначено червоним), TDNN (позначено зеленим), ResNet (позначено блакитним) та іншими (позначено жовтим). Кількість параметрів (розмір) моделей позначено величиною кола, яке представляє цю модель. По горизонталі відзначено кількість затрачених обчислювальних ресурсів, а по вертикалі – метрику EER (equal error rate), менше значення якої відповідає кращому результату (Yakovlev et al, 2024, p. 1).

Огляд наведених вище наукових робіт дозволив визначити моделі розпізнавання мовця, які продемонстрували найкращі результати в усіх

дослідженнях: ResNet, ReDimNet та ECAPA-TDNN. Ці моделі було обрано нами для подальшого тестування. Інші моделі, такі як Pyannote, TitaNet та WavLM, мали гірші показники у більшості тестів.

Висновки до Розділу 1

На початку розділу ми роз'яснили основні поняття цієї роботи – діаризацію, ідентифікацію та верифікацію мовця, які є підгалузями автоматичного розпізнавання мовця (ASR).

Після цього було оглянуто основні принципи роботи та будови систем автоматичного розпізнавання мовця. Зокрема, виділено підходи до їх створення: модульний конвеєр із кількох незалежних компонентів, наскрізні моделі та гібридні підходи. Найбільш ефективними та поширеними є наскрізні моделі, але гібридні підходи, що поєднують наскрізну модель з кластеризацією або атракторами, також мають популярність. Визначено, що система, представлена у цьому дослідженні, є наближеною до традиційного модульного конвеєра, адже має модульну будову та використовує алгоритм кластеризації, не здатний до навчання. Проте використання претренованих моделей для отримання ембедингів та здатність системи працювати із великою кількістю записів робить її більш подібною до описаних гібридних підходів.

Після проведення аналізу сучасних наукових праць, що стосуються поставленого завдання, було визначено основні кількісні та якісні характеристики мовленнєвих даних, вплив яких досліджуватиметься в практичній частині: стать мовця, кількість різних голосів у тестувальному наборі даних, зашумленість середовища у поєднанні зі спонтанним мовленням (для записів, зроблених в природних умовах), ступінь підготовленості мовлення. Ці особливості виявилися основними чинниками, що впливали на точність роботи систем автоматичного розпізнавання мовця у розглянутих нами дослідженнях.

Розгляд найактуальніших сучасних досліджень, сфокусованих на створенні та тестуванні систем автоматичного розпізнавання мовця, дозволив визначити

моделі, які продемонстрували найкращі результати: ECAPA-TDNN, ResNet та ReDimNet. Ці моделі було обрано для подальшого тестування з метою оцінки впливу характеристик, перелічених вище, на точність їх роботи.

РОЗДІЛ 2. Проведення експерименту

У цьому розділі буде розглянуто моделі розпізнавання мовця, робота яких оцінюватиметься в ході дослідження, а також представлено будову, принципи роботи та інтерактивний вебінтерфейс системи, у складі якої вони будуть тестуватися. Далі буде описано якісний та кількісний склад датасетів, сформованих для експерименту із виявлення впливу характеристик даних на точність моделей. Після цього буде коротко оглянуто отримані результати, які будуть докладніше проаналізовані в наступній частині.

2.1. **Огляд моделей, вибраних для тестування**

ResNet

Модель, яка буде оцінюватися в межах цього експерименту, має один із варіантів популярної архітектури ResNet та була навчена на багатомовних корпусах даних VoxCeleb1 та VoxCeleb2 із мовленням відомих людей (Nagrani, Chung and Zisserman, 2017, p. 1; Nagrani, Chung and Zisserman, 2018, p. 1). Вона є одною з відкритих моделей для автоматичного розпізнавання мовця, що надаються бібліотекою SpeechBrain (Ravanelli et al, 2021, p. 8).

Архітектура цієї моделі використовувалась у завданнях комп'ютерного зору, але пізніше стала застосовуватись і в системах голосової верифікації (Villalba et al, 2020, p. 1; Zeinali et al, 2019, p. 2). Особливість роботи таких моделей полягає в тому, що перед поданням до мережі аудіосигнали перетворюються на спектрограми або мел-спектрограми, які відображають мел-частотні кепстральні коефіцієнти (MFCC). Вони представляють частотний склад аудіозапису на основі шкали Мела, яка відображає людське сприйняття звуку. Спектрограми є подібними до зображень (рис. 2.1.), тому модель вивчає їх особливості, як в задачі розпізнавання образів (Zeinali et al, 2019, p. 2). ResNet враховує такі ознаки як особливості тембру, спектральні характеристики, патерни в коливаннях інтонації. При її тренуванні використовувалася функція втрати Additive Margin Loss, яка

визнана найкращою для тренування моделей голосової верифікації (Coria et al, 2020, p. 8).

Така мережа має хорошу здатність до узагальнення та здатна навчатися на великих наборах даних. Розмірність ембедингів, отриманих з цієї моделі – 256 вимірів (Brydinskyi, 2024, p. 5).

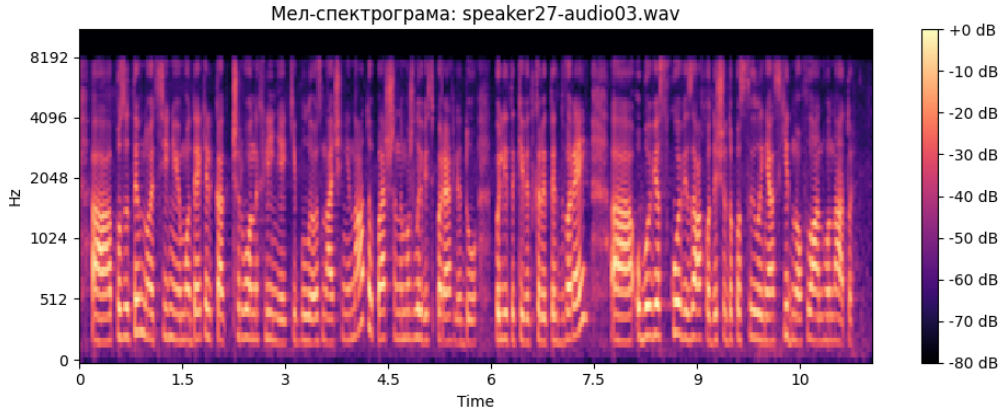

Рис. 2.1. Приклад мел-спектрограми, що подається до моделі ResNet.

ECAPA-TDNN

Модель ECAPA-TDNN, вибрана для тестування, теж надається бібліотекою для обробки мовлення SpeechBrain (Ravanelli et al, 2021, p. 8). Вона тренувалася на наборах даних Voxceleb1 та Voxceleb2, як і модель ResNet (Ravanelli et al, 2021, p. 8). Проте, на відміну від ResNet, архітектура TDNN була створена безпосередньо для завдання розпізнавання мовця. Вхідними даними, які ця модель використовує для створення представлень, теж є мел-спектрограми, для створення яких використовується 80 MFCC (Desplanques, Thienpondt and Demuynck, 2020, p. 3), як і у ResNet. Модель використовує вдосконалену версію архітектури TDNN із додатковими блоками (Hu, Shen and Sun, 2018, p. 2). При її навчанні використовувалася та сама функція втрати Additive Margin Loss, що й для ResNet.

Перевагою такої архітектури є те, що модель не просто враховує особливості тембру, а може вивчати та запам'ятовувати складні тривалі залежності та контексти, що допомагає при роботі з довгими записами (Руда та ін., 2024, с. 14).

Також вона здатна визначати найбільш репрезентативні особливості спектра за допомогою механізмів уваги (Desplanques, Thienpondt and Demuynck, 2020, p. 3).

Іншою перевагою є швидкість та невелика кількість параметрів, що робить її зручною у використанні навіть з обмеженими ресурсами (Руда та ін., 2024, с. 14, Brydinskyi, 2024, p. 5). Розмірність її ембедингів складає всього 192 виміри – це суттєво менше за розмірність ембедингів інших моделей (Brydinskyi, 2024, p. 5). ECAPA-TDNN використовувалась в кількох дослідженнях (Руда та ін., 2024, с. 13; Холєв і Барковська, 2023, с. 4; Brydinskyi, 2024, p. 6), де продемонструвала високі результати, а в одному з досліджень (Brydinskyi, 2024, p. 10) виявилася найкращою для українськомовних аудіозаписів.

ReDimNet

Архітектура ReDimNet є новітньою розробкою, представленою кілька років тому. За даними її розробників, вона суттєво перевершує інші моделі за точністю (Yakovlev et al, 2024, p. 1). У цьому експерименті буде використовуватись один із варіантів моделі, представлених розробниками, який має середній розмір (M), був навчений на датасетах VoxCeleb2, Voxblink2 (Lin, Y. et al, 2024, p. 1) і cn-celeb (Fan, Y. et al, 2019, p. 1) та використовував функцію втрати Additive Angular Margin Loss при тренуванні. Варто зауважити, що Voxblink2 є найбільшим на цей час відкритим датасетом, що містить голоси 111 тисяч мовців та мовлення 18 різними мовами (Lin, Y. et al, 2024, p. 1), і сильно перевищує за обсягом датасети VoxCeleb, які традиційно використовуються для тренування моделей розпізнавання мовця (зокрема й для двох попередньо розглянутих моделей). Датасет cn-celeb містить китайське мовлення, що сприяло більшій мовній різноманітності даних при навчанні моделі (Fan, Y. et al, 2019, p. 1).

Особливостями архітектури ReDimNet є те, що вона об’єднує 1D- та 2D-згорткові блоки шляхом зміни розмірності між представленнями (Yakovlev et al, 2024, p. 2). Дослідники припускають, що завдяки поєднанню цих двох блоків модель зможе краще відображати складні часові та спектральні коливання,

притаманні спонтанному чи зашумленому мовленню. На відміну від ReDimNet, архітектура ResNet використовує лише 2D-згорткові блоки, що могло зумовити ті гірші результати, що були представлені у роботі розробників нової моделі ReDimNet. Розмірність ембедингів, отриманих з цієї моделі – 192 виміри, як і в ECAPA-TDNN.

2.2. **Архітектура створеної системи**

Розробка системи автоматичного розпізнавання мовців проходила у середовищі Google Colab, адже ця платформа підтримує виконання коду мовою програмування Python та надає безплатний доступ до графічного процесора на віддаленому сервері. Вибір цього середовища зумовлений тим, що оброблення великих обсягів даних за допомогою моделей зі значною кількістю параметрів потребує обчислювальних ресурсів, яких можуть не мати звичайні персональні комп'ютери (Pessoa et al, 2018, p. 2). Крім того, Google Colab надає 100 гігабайтів пам'яті у середовищі, що дозволяє працювати з великими наборами даних, не використовуючи пам'ять комп'ютера. Код програми доступний за посиланням на репозиторій GitHub, розміщеним у Додатку.

Система складається з двох основних частин: модуля отримання ембедингів і модуля кластеризації, а також має додаткові функції для візуалізації ембедингів, автоматичного обрахування значень метрик (якщо працює з розміченим набором даних) та швидкого відтворення аудіозаписів (використовується при проведенні перцептивного аналізу записів, які були помилково класифіковані системою).

Система може працювати у двох режимах: з розміченими даними, де кожен запис має відповідну мітку мовця (режим тестування) і з неанотованими даними (режим розпізнавання). У режимі тестування результати можуть оцінюватися за допомогою метрик, що дозволяє встановити точність роботи системи.

Гнучка архітектура та модульність дозволяють легко змінювати моделі та алгоритми кластеризації при кожному новому запуску коду, адже модель, яку

система має використовувати для отримання ембедингів, передається в програму разом з іншими аргументами. Перевагою такої будови є легкість додавання інших моделей для тестування, що спрощує розширення системи в майбутньому. Використання іншого алгоритму для кластеризації може бути імплементовано так само легко, як і додавання іншої моделі.

Система розрахована на роботу з великими наборами даних, що складаються із записів, кожен з яких містить мовлення лише однієї особи. Вхідний набір аудіозаписів проходить попереднє оброблення, отримання ембедингів із вибраної претренованої моделі, нормалізацію та кластеризацію цих ембедингів, після чого звукозаписи розподіляються за директоріями згідно з визначеними системою мітками кластерів. Всі описані етапи роботи системи послідовно представлено на рис. 2.2. та буде детально розглянуто далі.

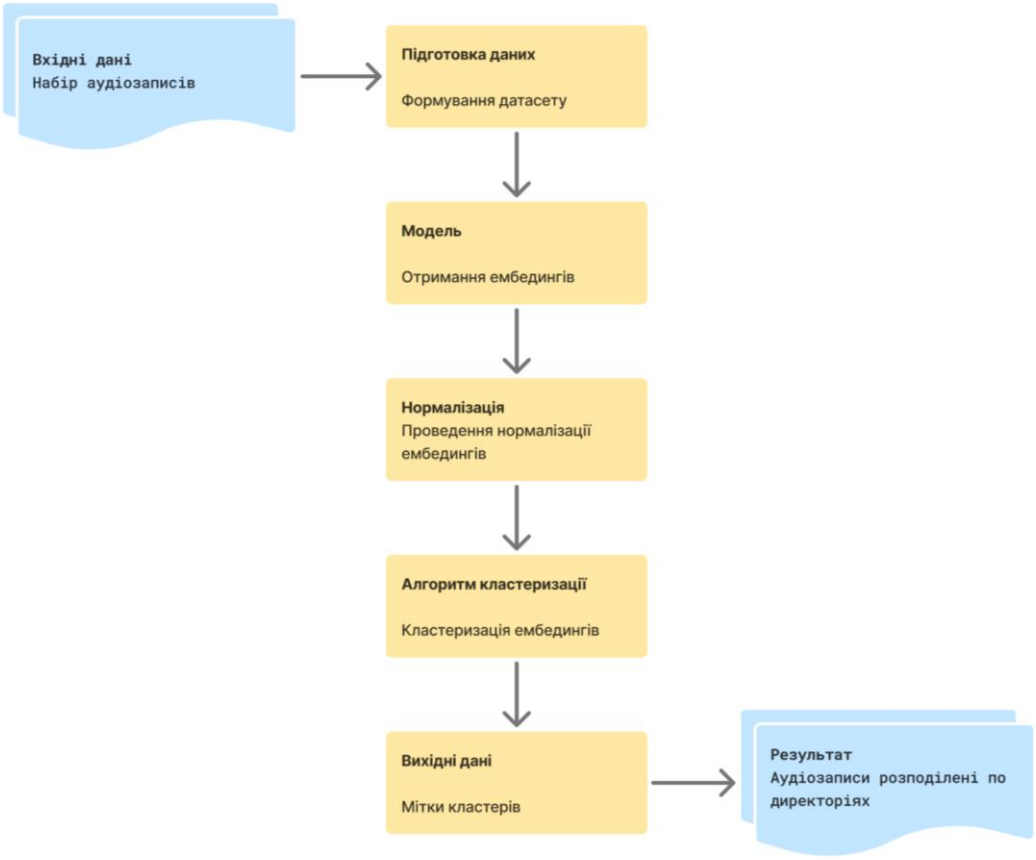

Рис. 2.2. Етапи оброблення даних.

Підготовка даних

Першим етапом у роботі системи є підготовка вхідних даних. На цій стадії відбувається формування датасету та приведення даних до необхідного формату. Система підтримує різні формати вхідних даних та обробляє їх залежно від типу та структури. Клас *Data_Preparation*, що виконує необхідні перетворення даних, підтримує такі формати датасетів:

- папка: директорія, в якій містяться вкладені папки із записами голосів мовців (назва папки використовується як мітка для мовця);
- папка без міток: директорія із записами голосів без будь-якої анотації.
- файл формату .pkl: файл, що містить об'єкт *DataFrame* (Mckinney, 2011, p. 2), який має колонку *path* зі шляхами до аудіофайлів та колонку *label* з мітками мовців.

Попереднє оброблення відбувається таким чином: клас *Data_Preparation* формує словник (dictionary), в якому ключем є індекс запису, а значенням – словник зі шляхом до аудіофайлу та міткою мовця. У разі відсутності анотації замість міток використовуються назви файлів для спрощення подальшого перегляду результатів кластеризації. Клас *Audio* приймає шлях до файлу та зчитує звуковий сигнал у потрібному форматі (частота дискретизації: 16 кГц; канал: моно) за допомогою функції *load_audio_resampled*, яка використовує бібліотеку torchaudio (Yang et al, 2022, p. 2). За допомогою класу *Audio* з кожного шляху до файлу отримується звуковий сигнал.

Модуль отримання та нормалізації ембедингів

Цей модуль програмного коду складається із класу *Embed* та класів-обгорток для моделей, які приймають звуковий сигнал та повертають його ембединг. Клас *Embed* приймає модель та датасет і виконує основну обробку набору даних: за допомогою моделі, кожен аудіосигнал перетворюється в ембединг, який потім нормалізується за допомогою функції normalize з бібліотеки scikit-learn (Pedregosa et al, 2011, p. 2).

Модуль кластеризації

Нормалізовані ембедінги подаються до алгоритму кластеризації. Архітектура дозволяє використовувати будь-який з реалізованих алгоритмів без необхідності заздалегідь вказувати кількість кластерів, що відповідає завданню із невідомою кількістю мовців. Наразі система підтримує два алгоритми кластеризації – DBSCAN та HDBSCAN. Їх реалізовано у відповідних класах, які приймають набір ембедингів та повертають мітки, що позначають належність запису до певного кластера.

У представленому тут експерименті було вирішено використовувати алгоритм **HDBSCAN**, адже він демонстрував кращі результати у всіх наших попередніх дослідженнях. HDBSCAN є вдосконаленою версією DBSCAN (Malzer and Baum, 2020, p. 2), яка будує ієрархію кластерів і автоматично визначає їх оптимальну кількість без необхідності явно задавати максимальну відстань між двома сусідніми точками (Malzer and Baum, 2020, p. 2). Він краще працює з даними нерівномірної щільності та складнішої структури, що дозволяє йому краще справлятися із завданням кластеризації ембедингів.

Вивід результатів та додаткові функції

Після кластеризації аудіозаписи розподіляються по окремих директоріях згідно з присвоєними кластерами. Система має кілька додаткових функцій, що полегшують інтерпретацію її результатів. Відповідний код реалізовано у таких класах:

Visualisation: приймає набір ембедингів та налаштування вигляду візуалізації, опціонально може приймати відповідні мітки мовців та мітки належності до кластерів. Залежно від вибраних параметрів налаштувань, може відображати двовимірну або тривимірну візуалізацію ембедингів у вигляді точок на графіку за допомогою алгоритмів зменшення розмірності t-SNE або PCA.

Metrics: приймає наявні мітки мовців та визначені системою мітки кластерів та розраховує значення метрик. Клас підтримує додавання нових метрик як методів.

Для оцінки якості кластеризації у цьому експерименті були використані дві метрики: скоригований індекс Ренда (ARI) та нормалізована взаємна інформація (NMI). Вибір цих метрик був обумовлений їхньою поширеністю у задачах кластеризації (Протас, 2021, с. 14; Mahmoudi and Jemielniak, 2024, p. 3; Santos and Embrechts, 2009, p. 1). Нижче наведений короткий опис кожної з використаних метрик та інтерпретацію їх значень.

ARI оцінює схожість між двома розбиттями вибірки та не залежить від значень та перестановок міток. Значення ARI варіюється від -1 до 1, де значення -1 та 0 відповідають випадковій кластеризації, а 1 – ідеальній відповідності правильному розбиттю (Протас, 2021, с. 14). Метрика вважається досить надійною та репрезентативною (Santos and Embrechts, 2009, p. 9).

NMI вимірює ступінь узгодженості між двома розбиттями на кластери з нормалізацією відносно розміру вибірки. Значення NMI лежить у діапазоні [0, 1], де 0 означає відсутність спільної інформації, а 1 – повну відповідність. Однією з переваг метрики є можливість порівняти два розбиття із різною кількістю кластерів, що підходить для нашої задачі. Щоправда, метрика має і певні недоліки, наприклад зниження репрезентативності при великій кількості кластерів (Mahmoudi and Jemielniak, 2024, p. 3; McCarthy et al, 2019, p. 3).

2.3. **Створення інтерактивного вебінтерфейсу системи**

Для пришвидшення виконання експериментів, легшого налаштування параметрів, перегляду візуалізації ембедингів, прослуховування аудіофайлів, аналізу помилок та динамічного завантаження датасетів, було вирішено об’єднати весь перелічений функціонал в межах одного інтерфейсу.

Для створення інтерфейсу використовувалась бібліотека *gradio* (Abid et al, 2019, p. 1), яка дозволяє легко створювати інтерактивні вебінтерфейси, що активуються при запуску програмного середовища і забезпечують інтеграцію з моделями та корпусами даних.

Для полегшення процесу тестування моделей, у створеному інтерфейсі було реалізовано всі елементи взаємодії з програмою, описаною в частині 2.2 «Архітектура створеної системи». Інтерфейс мав дві вкладки: «Кластеризація» та «Аналіз результатів», які містили відповідний функціонал (рис. 2.3).

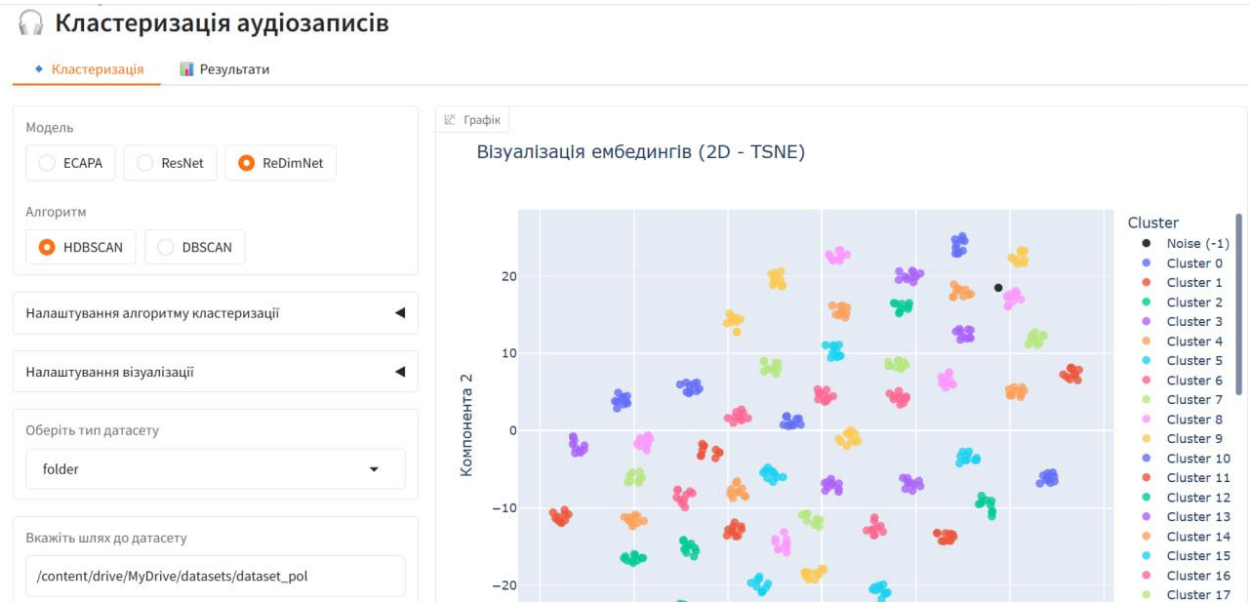

Рис. 2.3. Демонстрація інтерфейсу створеної системи, вкладка «Кластеризація».

Вкладка «Кластеризація» об’єднує в собі всі налаштування, які проводяться перед початком експерименту. Більш детальний огляд її елементів представлено нижче.

Вибір моделі

Як показано на рис. 2.3, першим з усіх налаштувань йде параметр «модель», де користувачу дається на вибір три опції: ECAPA, ResNet та ReDimNet. Експеримент буде проведено із використанням вибраної моделі для отримання ембедингів звукозаписів.

Вибір та налаштування алгоритму кластеризації

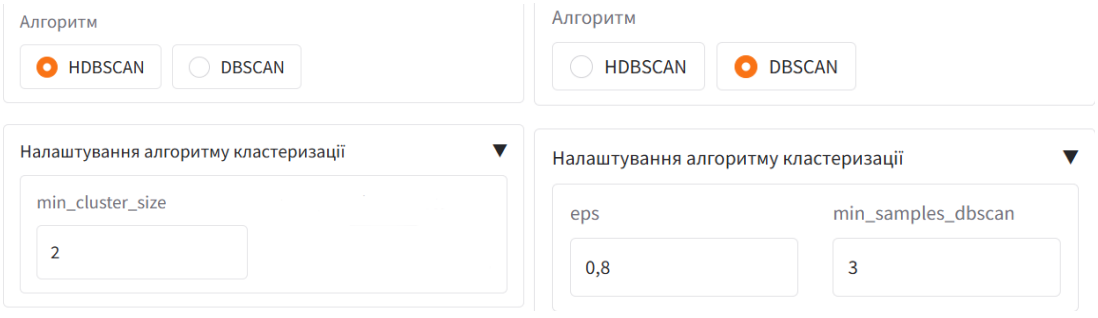

Рис. 2.4. Налаштування параметрів для алгоритмів кластеризації HDBSCAN та DBSCAN.

Наступним йде параметр «алгоритм», який надає вибір між двома варіантами алгоритму кластеризації: HDBSCAN та DBSCAN. Попри те, що було прийнято рішення провести всі експерименти із використанням алгоритму HDBSCAN, у нашій системі реалізовано кілька алгоритмів із можливістю легкого додавання нових для подальших досліджень.

Крім того, на рис. 2.4. показано, що алгоритми мають параметри налаштування, які користувач теж може задавати. Наприклад, для алгоритму HDBSCAN можна задати мінімальний розмір кластера: цей параметр за замовчуванням має значення 2, адже якщо в датасеті є принаймні два записи одного і того ж голосу, їх має бути виділено в окремий кластер.

Налаштування візуалізації ембедингів

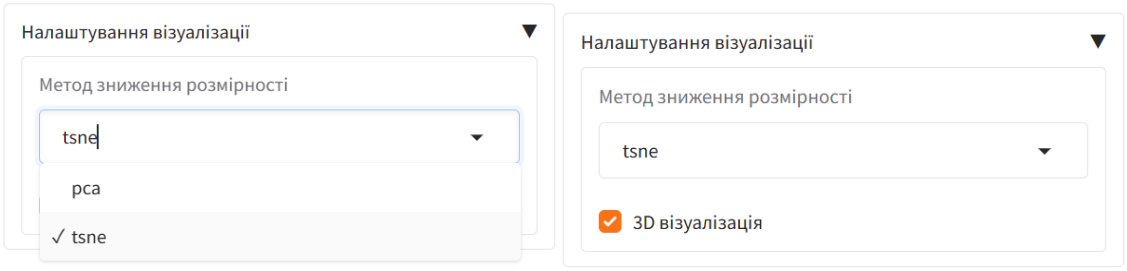

Рис. 2.5. Налаштування параметрів візуалізації ембедингів: вибір алгоритму для зниження розмірності ембедингів та можливість візуалізувати ембединги в 3D-просторі.

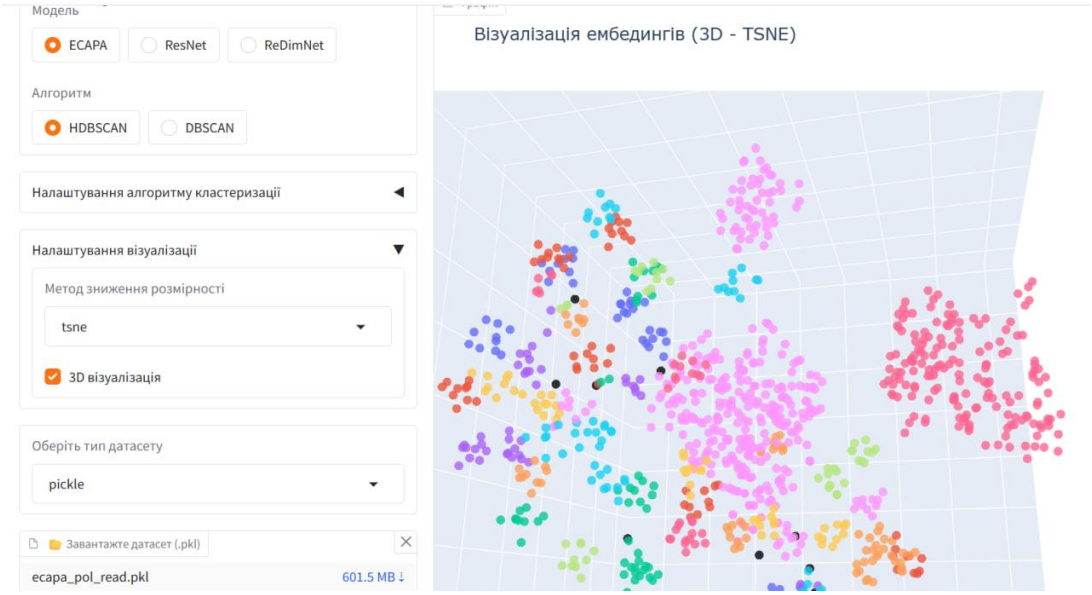

Рис. 2.6. Візуалізація ембедингів у 3D-просторі.

На рис. 2.5. продемонстровано, що візуалізацію ембедингів може бути здійснено за допомогою двох алгоритмів на вибір: t-SNE, що добре підходить для відображення великої кількості ембедингів, та PCA, який працює краще при дуже малій кількості звукозаписів у датасеті. Також користувач може обрати 3D-візуалізацію, приклад якої показано на рис. 2.6. Такий спосіб візуалізації ембедингів може знадобитися при дуже великій кількості мовців у датасеті чи значній кількості аудіозаписів, щоб детальніше проаналізувати розташування ембедингів у просторі. За замовчуванням, у системі використовується алгоритм t-SNE та 2D-візуалізація, адже такий тип відображення ембедингів обрано як найбільш зручний та репрезентативний.

Ембединги, які були об’єднані в один кластер, позначаються одним кольором, а мітки мовців, якщо вони є, відображаються у вигляді підписів поруч з точками ембедингів, які стають помітними при наведенні курсора на точку. Таким чином, можна помітити випадки помилкового об’єднання та неправильної атрибуції, коли точки одного кластеру мають різні мітки мовців.

Завантаження датасету

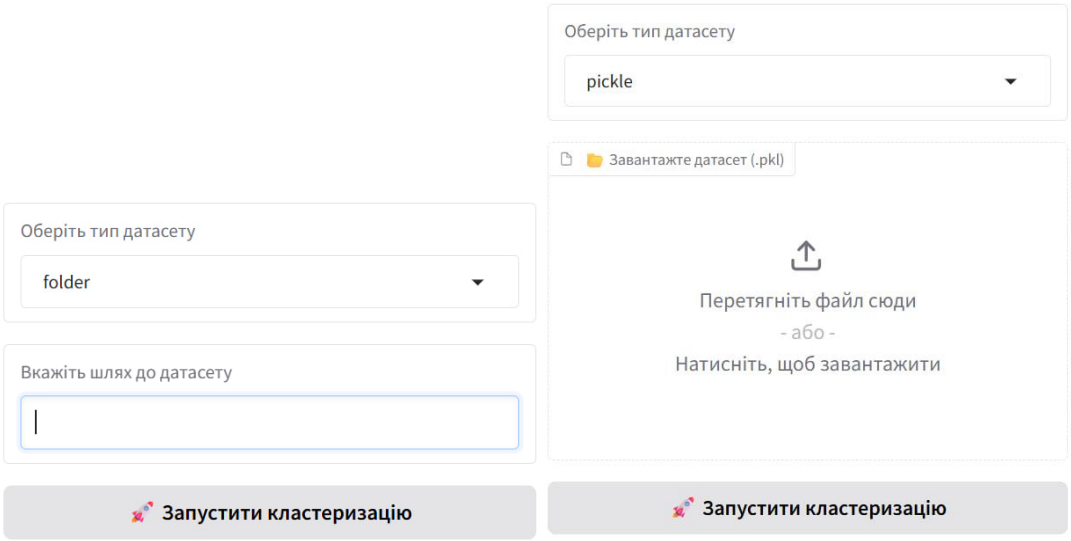

Рис. 2.7. Способи завантаження набору даних.

Як показано на рис. 2.7, система має два способи завантаження даних: «folder» (папка) та «pickle» (pkl-файл). При виборі першого способу, дані завантажуються за вказаним шляхом до папки на гугл-диску, де знаходиться набір даних. Також користувач може завантажити у відповідне поле файл .pkl, де містяться шляхи до звукозаписів та мітки мовців.

Завантаження отриманих ембедингів

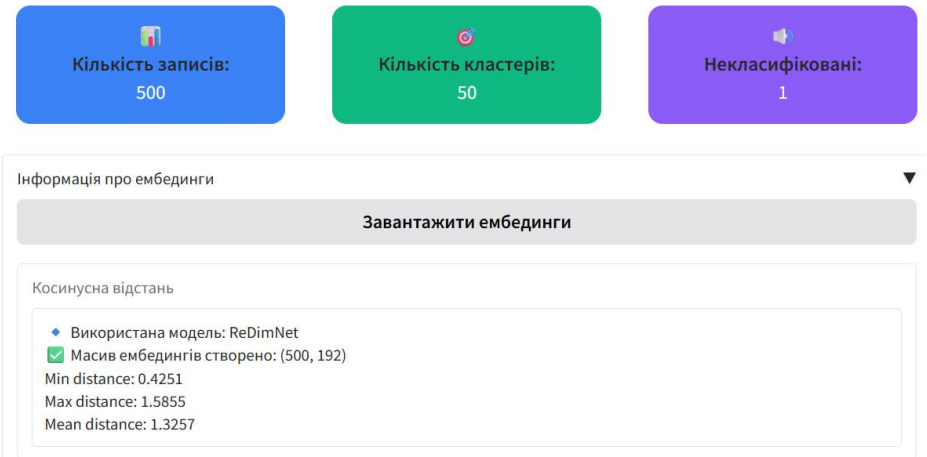

Рис. 2.8. Загальна інформація про результати кластеризації та можливість завантаження ембедингів.

Після закінчення обробки даних, під графіком із візуалізацією ембедингів розміщується загальна інформація про результати кластеризації (рис. 2.8): кількість записів у датасеті, кількість отриманих кластерів та кількість некласифікованих

записів, які було віднесено до категорії “шум”. Також можна переглянути значення мінімальної, максимальної та середньої косинусної відстані між ембедингами.

Крім того, тут реалізовано можливість завантажити отриманий датасет з ембедингами, згенерованими моделлю на цьому етапі, у форматі pkl-файлу, і використовувати його в подальших експериментах. Такий датасет можна повторно завантажити у систему – в такому разі програма працює з готовими ембедингами, присутніми в датасеті, і не використовує модель. Такий підхід дозволяє зберігати результати проведених експериментів, не виконувати повторних тестів для одних і тих самих даних та економити обчислювальні ресурси, потрібні для роботи моделей. Також використання готових ембедингів дозволяє спробувати інший алгоритм їх кластеризації чи іншу візуалізацію – зокрема, для 3D-візуалізації на рис. 2.6 було використано датасет із готовими ембедингами моделі ECAPA-TDNN.

Друга вкладка, «Результати», надає можливість провести аналіз результатів експерименту (рис. 2.9). Нижче представлено докладний опис її функцій та елементів.

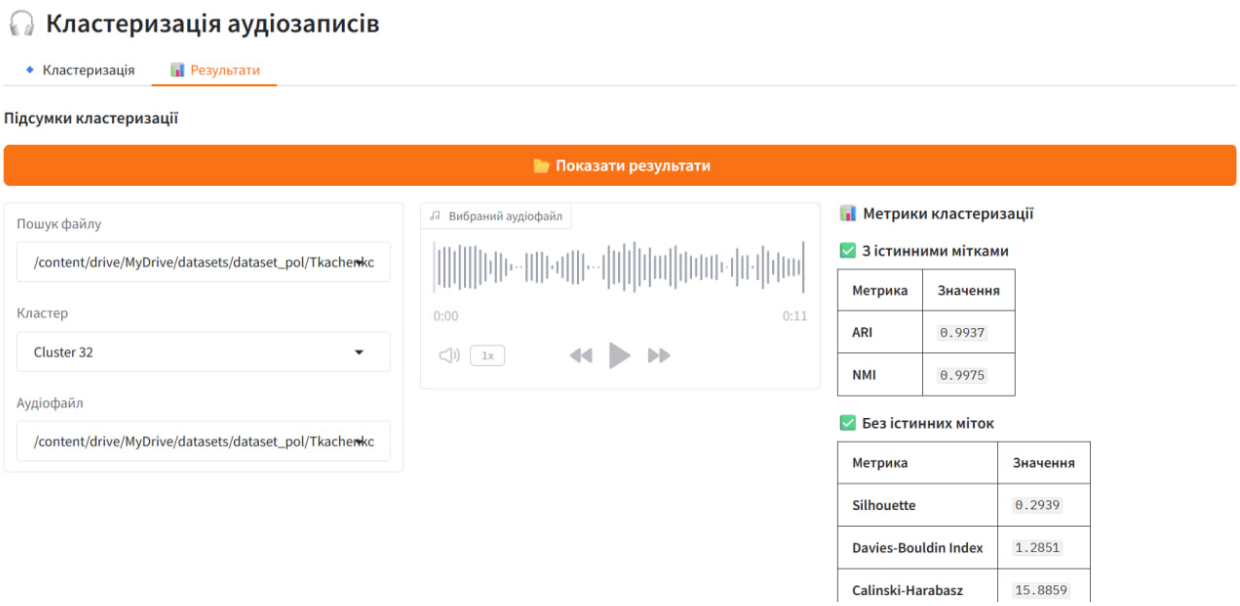

Рис. 2.9. Демонстрація інтерфейсу створеної системи, вкладка «Результати»

Відображення таблиці із метриками

При натисканні кнопки «Показати результати» відбувається відображення розрахованих метрик у таблицях зліва (рис. 2.9). Якщо система працює в режимі «без міток мовців», відображається лише друга таблиця, адже метрики в цій таблиці розраховуються тільки на основі відстані між ембедингами та не потребують міток.

Вибір кластера та файлу для прослуховування

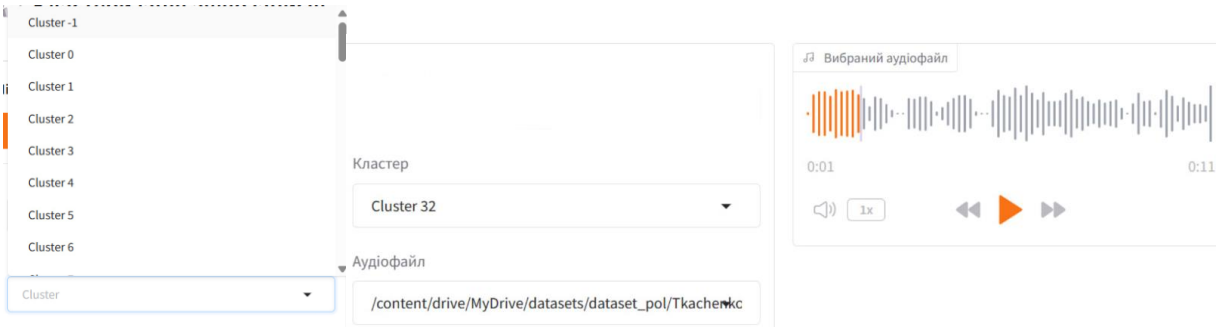

Рис. 2.10. Можливість обрати кластер та прослухати його аудіозаписи.

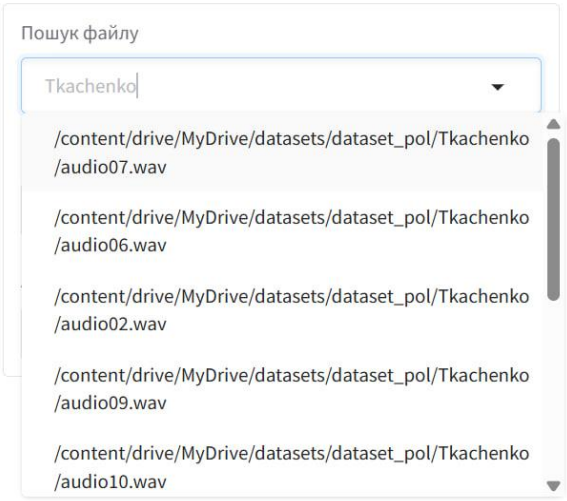

Рис. 2.11. Пошук запису за назвою файлу.

Для полегшення перцептивного аналізу, було реалізовано можливість вибору та прослуховування аудіозаписів із датасету прямо у веббраузері. Наприклад, користувач може обрати певний кластер та прослухати записи, які туди потрапили (рис. 2.10), або знайти конкретний запис за його назвою, щоб побачити, до якого кластеру його було віднесено (рис. 2.11). Такі функції значно полегшують аналіз помилково класифікованих даних та визначення мовленнєвих характеристик, типових для мовця.

Діаграма розподілу даних за кластерами

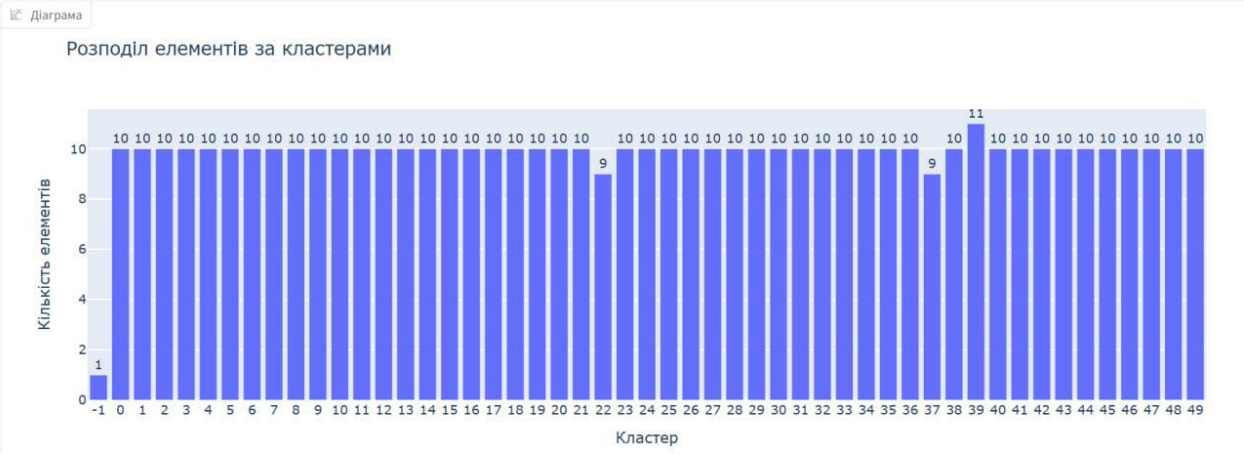

Рис. 2.12. Діаграма, що відображає кількість записів у кожному кластері.

Діаграма на цій вкладці демонструє розподіл даних за кластерами – цей графік дозволяє побачити, скільки записів потрапило до кожного кластера. Якщо датасет містить однакове число записів для кожного мовця, як на рис. 2.12, це значно полегшує виявлення та аналіз всіх помилок, яких припустилася система.

2.4. **Опис датасетів**

Для оцінки ступеня впливу кількісних та акустичних характеристик даних на точність моделей було створено шість наборів даних. Перша пара датасетів була створена для перевірки впливу різниці між читаним та спонтанним мовленням на результати кластеризації, друга – для виявлення відмінностей у результатах для чоловічих та жіночих голосів, а третя – для перевірки залежності результату від кількості мовців. Аудіозаписи було зібрано із відкритих джерел, взято з доступних наборів даних чи записано у природних умовах. Усі записи містять мовлення тільки українською мовою. Далі буде наведено більш детальний огляд кожного з цих датасетів.

Датасет читаного мовлення

Цей набір даних було створено для тестування системи на записах читаного мовлення, зроблених у відносно тихих умовах. Датасет містить 309 записів по 9

секунд кожен, та містить 16 різних голосів. Загальна тривалість становить 2781 секунду, або 46,4 хвилини. Він є не дуже збалансованим за статтю та за віком, оскільки містить переважно жіноче мовлення, а вік всіх мовців становить приблизно 18-20 років. Склад датасету відображено на рис. 2.13 нижче.

Рис. 2.13. Склад датасету читаного мовлення.

Датасет спонтанного мовлення

Цей датасет було створено для оцінки впливу спонтанного мовлення на результат роботи моделей. Він містить записи непідготовленого мовлення в зашумлених умовах. Кількість записів становить 348 (по 9 секунд кожен), а кількість мовців – 16, як і в попередньому датасеті. Загальна тривалість становить 3132 секунд, тобто 52 хвилини. Цей датасет, як і попередній, містить переважно жіноче мовлення мовців віком 18-20 років, тому такі параметри як стать та вік мовців, тривалість записів, загальний обсяг датасету та кількість мовців не вплинуть на різницю в результатах. Склад датасету відображено на рис. 2.14 нижче.

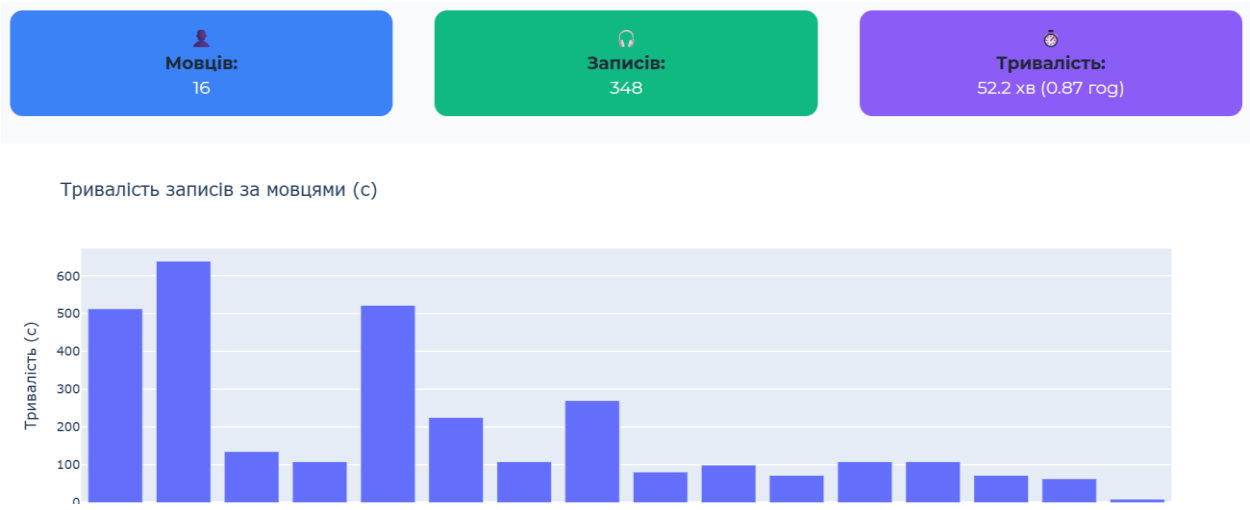

Рис. 2.14. Склад датасету спонтанного мовлення.

Датасети читаного та спонтанного мовлення не є повністю порівнянними, адже датасет зі спонтанним мовленням є дещо більшим за обсягом (на 5,8 хвилини). Ця різниця не є надто суттєвою, адже всі інші параметри здебільшого збігаються (кількість мовців, їх вік та стать). До того ж ця невелика різниця в обсягах датасетів дозволить перевірити взаємодію між різними факторами (кількість записів та підготовленість мовлення).

Датасет чоловічого мовлення

Цей набір даних було сформовано для виявлення особливостей роботи системи для чоловічого мовлення. Він містить 330 записів добре підготовленого мовлення в умовах різного ступеня зашумленості, тривалість кожного запису – 10-11 секунд, кількість мовців становить 33 (для кожного мовця – 10 записів). Загальний обсяг набору даних – 3378 секунд, тобто 56,3 хвилини. Варто відзначити, що датасет є збалансованим за віком та представляє мовлення різних вікових груп (віковий діапазон становить від 32 до 70 років). Склад датасету відображено на рис. 2.15 нижче.

Рис. 2.15. Склад датасету чоловічого мовлення.

Датасет жіночого мовлення

Датасет було створено для аналізу особливостей роботи системи для жіночого мовлення. Він містить 170 записів підготовленого мовлення з різною зашумленістю, тривалість кожного запису становить 10-11 секунд, кількість мовців – 17 (для кожної дикторки по 10 записів). Загальний обсяг набору даних – 1746 секунд (29,1 хвилини). Цей набір даних, як і датасет чоловічого мовлення, є збалансованим за віком (віковий діапазон – від 27 до 58 років). Склад датасету відображено на рис. 2.16 нижче.

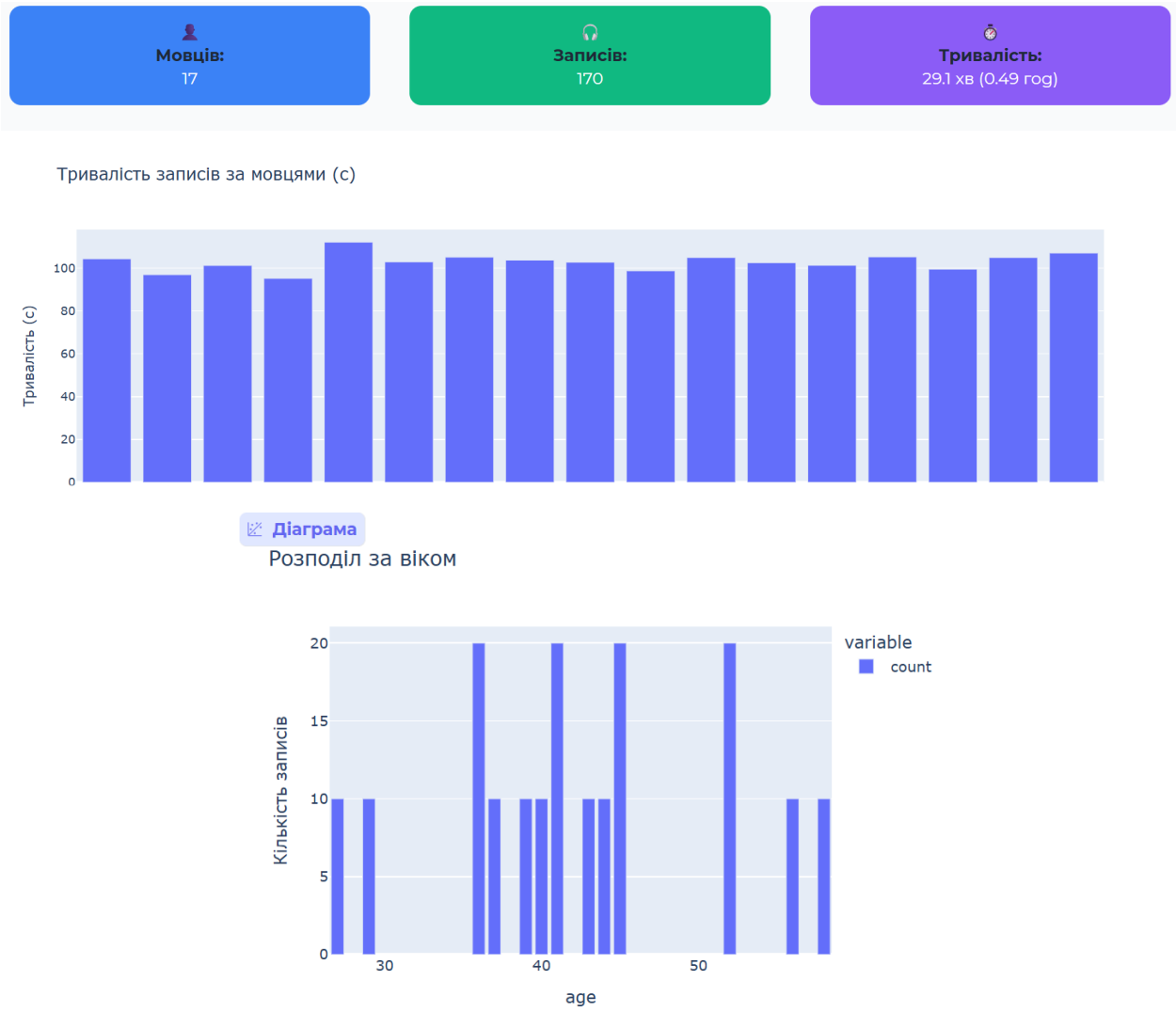

Рис. 2.16. Склад датасету жіночого мовлення.

Ці два набори даних теж не є повністю порівнянними, адже датасет чоловічого мовлення є значно більшим за обсягом та за кількістю мовців. Проте інші параметри, такі як підготовленість мовлення, кількість записів для мовця, рівень зашумленості та варіативність віку, збігаються. Різна кількість мовців у цьому випадку є додатковим параметром, вплив якого також буде проаналізовано разом із впливом статі мовців.

Датасет з великою кількістю мовців

Для перевірки впливу кількісних характеристик даних на точність роботи моделей було створено датасет зі значною кількістю мовців. Він містить 500 записів тривалістю близько 10-12 секунд кожен та 50 різних голосів (по 10 записів для

мовця). Загальний обсяг цього набору даних – 5127 секунд (85,5 хвилин). Датасет є збалансованим як за статтю, так і за віком мовців, і містить записи різного ступеня зашумленості.

Рис. 2.17. Склад датасету з великою кількістю мовців.

Датасет із малою кількістю мовців

Також для перевірки впливу кількості мовців на роботу системи було створено датасет із незначною кількістю мовців, але великою кількістю записів для кожного диктора. Цей набір даних містить 523 записи тривалістю по 9 секунд для 4 мовців. Кількість записів для різних мовців значно відрізняється. Загальний обсяг датасету – 4710 секунд (78,5 хвилин). Таким чином, датасет є порівнянним із попереднім набором даних за обсягом, але значно відрізняється за кількістю представлених мовців. Різниця в загальній тривалості становить лише ~7 хвилин, що не є значною відмінністю для датасетів тривалістю понад годину. Попри

невелику кількість, мовці є різними за статтю та за віком, як і в попередньому датасеті. Умови, в яких було зроблено записи, є природними (мають різний ступінь зашумленості).

Попри те, що кількість мовців у цьому датасеті невелика, його обсяг є досить значним, адже диктори мають багато записів голосу в різноманітних умовах.

Порівняння результатів для цього і для попереднього датасету допоможе оцінити вплив саме кількості мовців на точність роботи: якщо результати залежать лише від кількості різних голосів, а не від числа їх записів та різноманітності акустичних умов, система має продемонструвати кращі результати для датасету з чотирма мовцями. Натомість якщо такі параметри як різноманітність акустичних умов, велика кількість записів для одного голосу, та незбалансованість за мовцями сильніше вплинуть на точність, система продемонструє кращі результати для датасету із великою кількістю мовців, де для кожного мовця в датасеті міститься всього 10 записів.

Рис. 2.18. Склад датасету із малою кількістю мовців.

2.5. **Хід експерименту та попередній огляд результатів**

Основний експеримент складався з декількох етапів, що повторювалися для кожної із трьох моделей розпізнавання мовця: **ECAPA-TDNN**, **ResNet** та **ReDimNet**. Хід експерименту зображено за допомогою блок-схеми на рис. 2.19. Кожен із шести сформованих датасетів був послідовно оброблений системою спочатку із використанням ECAPA-TDNN, потім – ResNet, а потім – ReDimNet. Результати роботи кожної моделі з кожним із датасетів було збережено та детально проаналізовано.

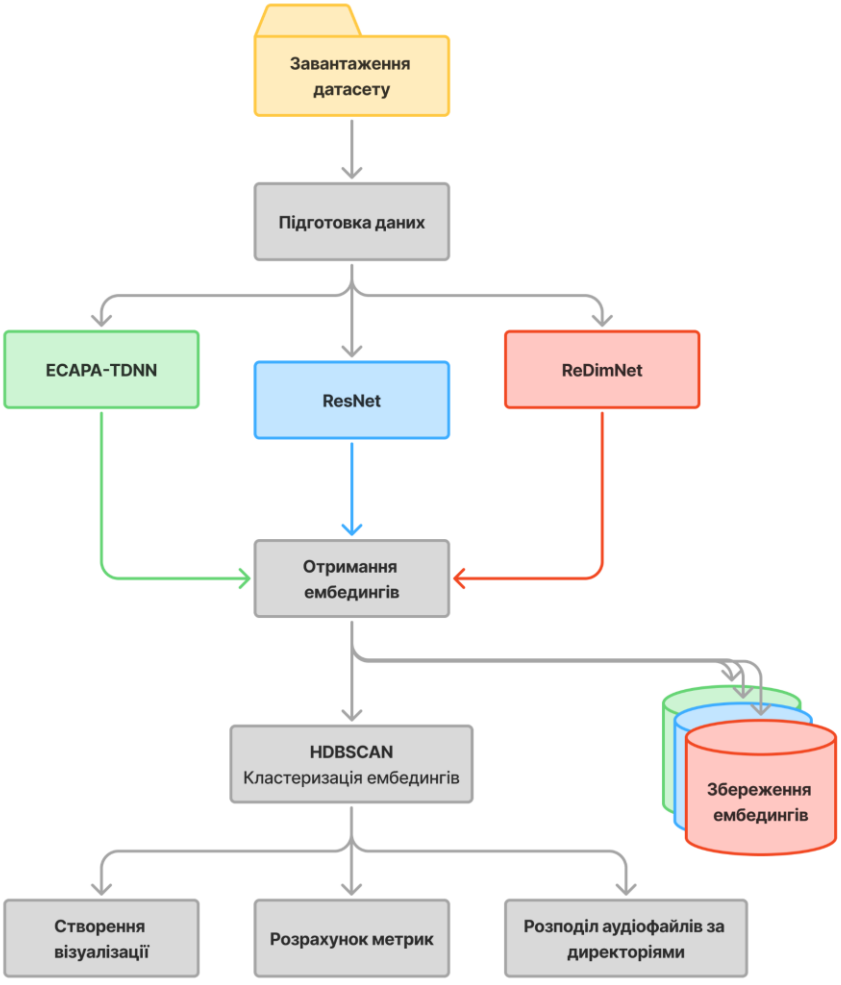

Рис. 2.19. Блок-схема, що демонструє етапи експерименту для одного датасету.

Послідовність виконання етапів обробки кожного датасету за допомогою кожної моделі була такою:

1. Отримання ембедингів аудіозаписів за допомогою поточної моделі.
2. Кластеризація ембедингів за допомогою алгоритму HDBSCAN.
3. Візуалізація результатів кластеризації.
4. Автоматичний розрахунок метрик точності кластеризації.
5. Оцінка якості кластеризації за візуалізацією, перцептивний аналіз помилково класифікованих записів.
6. Збереження результатів поточного етапу тестування.

Під час експерименту в алгоритмі кластеризації HDBSCAN використовувалась евклідова відстань для розрахунку дистанції між ембедингами. Цей спосіб визначення подібності широко застосовується для порівняння ембедингів у завданні голосової верифікації – наприклад, у метриці потрійної втрати під час навчання (Ren, Chen and Xu, 2019, p. 2).

Попередній огляд результатів

В результаті проведення експерименту було попередньо оцінено роботу кожної моделі із різними датасетами. ReDimNet продемонструвала найкращий результат у всіх експериментах. ResNet посіла друге місце за точністю: для датасетів читаного мовлення та невеликої кількості мовців вона показала таку ж точність, як і ReDimNet, для інших датасетів результати були дещо гіршими за результати ReDimNet, і лише для датасету чоловічого мовлення вона спрацювала найгірше з усіх моделей. ECAPA-TDNN опинилася на третьому місці за якістю ембедингів, хоча її результати теж були досить високими та близькими до результатів ResNet, а для датасету чоловічого мовлення вона спрацювала краще за ResNet.

Щодо відмінностей у роботі моделей для різних датасетів, було виявлено ряд закономірностей, які підтверджують наші припущення, висловлені в Розділі 1 та пункті 2.4 «Опис датасетів». Зокрема, спонтанне мовлення у зашумлених умовах дійсно гірше розпізнається усіма моделями: для цього датасету було отримано найнижчі значення метрик, в той час, як для читаного мовлення результати були

близькими до ідеальних. Також було виявлено відмінності у роботі моделей для чоловічих та жіночих голосів: ResNet продемонструвала кращі результати для жіночих, ECAPA-TDNN – для чоловічих, а результати моделі ReDimNet не залежали від статі мовця та перевершували показники обох інших моделей. Також всі моделі ідеально впоралися із датасетом з невеликою кількістю мовців, в той час, як значна кількість різноманітних голосів виявилася складнішою задачею. Більш докладний аналіз усіх отриманих результатів буде представлено у наступному розділі.

Висновки до Розділу 2

Оглянувши вибрані для тестування моделі (ResNet, ECAPA-TDNN, ReDimNet), ми встановили, що всі вони використовують як вхідні дані мел-спектрограми, що представляють частотний склад запису на основі шкали Мела, яка відображає людське сприйняття звуку. З огляду на це, при розгляді помилок системи з цими моделями є доцільним аналіз формант, тембру, паузації та ступеня зашумленості мовленнєвих даних.

Далі було представлено процес створення системи та її будову. Вона складається з двох основних модулів – моделі для отримання ембедингів і алгоритму, що їх кластеризує, а також додаткових модулів для візуалізації результатів та автоматичного обрахування метрик.

Інтеграція всього створеного програмного забезпечення в один інтерактивний вебінтерфейс в рази пришвидшує як проведення експериментів, так і аналіз їх результатів. Перевагами створеного інтерфейсу є візуальна демонстрація результатів, швидкість та відтворюваність проведення експериментів у реальному часі, полегшення ідентифікації та аналізу помилково розпізнаних даних, можливість проведення нових експериментів з іншими датасетами та легка інтеграція нових моделей, метрик і алгоритмів.

Також був представлений опис датасетів, сформованих для оцінки ступеня впливу характеристик мовленнєвих даних на точність роботи створеної системи.

Для перевірки впливу підготовленості мовлення було створено датасети із читаним та спонтанним мовленням, для аналізу впливу статі – датасети жіночого та чоловічого мовлення, а для перевірки впливу такої кількісної характеристики як число мовців – датасети із 50 та 4 голосами відповідно.

Після формування датасетів було проведено основний експеримент цього дослідження. Під час попереднього аналізу всіх його етапів було встановлено, що моделі краще працюють із читаним мовленням та мають труднощі зі спонтанним; ResNet та ECAPA-TDNN демонструють різну точність залежно від статі мовця; всі моделі демонструють вищу точність для датасету із малою кількістю мовців попри його великий обсяг. В результаті порівняння моделей найкращою визнана ReDimNet, другою за точністю – ResNet, а третьою – ECAPA-TDNN, хоча всі моделі продемонстрували досить високі результати.

РОЗДІЛ 3. Аналіз результатів експерименту

У цій частині буде проведено докладний лінгвістичний аналіз звукозаписів кожного датасету для встановлення того, якою мірою акустичні характеристики проявлялися в аудіозаписах. Після цього буде проаналізовано отримані значення метрик для оцінки ступеня впливу кожної характеристики на точність роботи моделей. Також будуть розглянуті помилки кожної з моделей та проведено перцептивний аналіз неправильно класифікованих записів для встановлення акустичних особливостей, що могли їх спричинити. Таким чином, буде проведено триступеневий аналіз результатів експерименту.

3.1. **Вплив відмінностей читаного та спонтанного мовлення**

Для записів у датасетах читаного та спонтанного мовлення було виміряно такі параметри як темп, середня кількість коротких, середніх та довгих пауз у записі, SNR (singnal-to-noise ratio).

Через велику кількість мовців в обох датасетах, було вирішено обрати для вимірювань та аналізу тільки перших 10 мовців, що становить приблизно половину датасету. На нашу думку, такий обсяг матеріалу буде достатнім для того, щоб виявити та проаналізувати розбіжності читаного та спонтанного мовлення, а також дослідити інші можливі закономірності у показниках параметрів.

Читане мовлення

№ мовця у датасеті	Темп (скл/с)	Короткі паузи (сер. кількість)	Середні паузи (сер. кількість)	Довгі паузи (сер. кількість)	SNR, dB
#1	3.2	22	26	20	31.2
#2	3.8	2	4	23	26.0
#3	4	8	6	1	23.8
#4	3.6	9	3	16	23.9
#5	4.7	2	14	22	7.5
#6	2.8	9	13	11	27.7
#7	3.5	6	5	6	11.3
#8	3.4	2	3	14	21.0
#9	5.1	9	13	17	27.6
#10	3.2	14	25	39	17.6

Таблиця 3.1. Темп, середня частота пауз різної довжини та показник SNR для записів читаного мовлення.

На основі отриманих значень для читаного мовлення перших десяти мовців у датасеті, можна зробити висновки про особливості цих даних. Темп мовлення є переважно середнім – у межах 3-4 складів на секунду, що є характерним для читання: більшість записів мають стабільний темп, який не змінюється протягом більшої частини запису. Кілька мовців читають текст у швидкому (5 скл/с) чи повільному (2,8 скл/с) темпі, але загальна тенденція залишається незмінною: темп читаного мовлення є помірним, розміреним та стабільним протягом запису.

Щодо паузації у проаналізованих записах читаного мовлення можна сказати, що вона значною мірою залежить від мовця, а не тільки від ступеня підготовленості мовлення: для деяких мовців характерна однаково велика кількість пауз будь-якої

довжини (наприклад, мовці №1 та №10), для інших – мала кількість пауз (мовці №3, №6 та №7).

Але все ж у результатах вимірювання простежується тенденція кількісного переважання довгих пауз: зазвичай середніх пауз більше, ніж коротких, а довгих найбільше серед усіх. Така паузація є характерною для читаного мовлення, де синтагми часто збігаються з реченнями та є довгими. Вимовляючи довгі синтагми під час читання тексту, мовець робить між ними тривалі паузи, заповнені мовленнєвим диханням.

Показники SNR знаходяться переважно в діапазоні 20-30 дБ, що свідчить про значне переважання сигналу над шумом. Значення у 20 дБ зазвичай відповідає нормальній якості запису, а 30 дБ – дуже чистому звуку. Отже, можна зробити висновок, що більшість мовців у цьому датасеті мають якісні, незашумлені записи. Кілька мовців мають трохи нижчі показники (7 дБ та 11 дБ), що свідчить про те, що запис мовлення було зроблено у зашумленому середовищі чи на неякісне обладнання. Також можливою причиною таких показників є те, що мовець знаходився надто далеко від мікрофона.

Спонтанне мовлення

№ мовця у датасеті	Темп (скл/с)	Короткі паузи (сер. кількість)	Середні паузи (сер. кількість)	Довгі паузи (сер. кількість)	SNR, dB
#1	3.5	2.7	2.4	3.7	16.7
#2	4	0.2	0.8	0.2	21.9
#3	4.6	0.8	1.6	1.8	22.4
#4	2.4	7	7	7	22.7
#5	3.8	1.5	4.3	8.9	13.6
#6	6	6.25	7	6.5	16.4
#7	6	0.2	1	2.4	14.7
#8	4	0.04	0.2	0.3	20.4
#9	5	0.57	0.7	3	12.5
#10	5.7	1.2	2.4	1.7	31.8

Таблиця 3.2. Темп, середня частота пауз різної довжини та показник SNR для записів спонтанного мовлення.

На відміну від читаного мовлення, спонтанне має набагато вищий темп – більшість записів містять швидке мовлення (або таке, яке можна схарактеризувати як близьке до швидкого). Декілька записів мають помірний темп. Варто зауважити, що наведені значення є лише усередненими показниками, адже зазвичай темп варіювався протягом запису. Деякі записи, навпаки, містили мовлення із повільним темпом. Оскільки датасет спонтанного мовлення був сформований переважно із голосових повідомлень, показник темпу сильно залежав від комунікативної ситуації та змісту самого повідомлення.

Кількість пауз на один запис у датасеті спонтанного мовлення теж є приблизною, адже обчислювалась як середнє значення для дуже багатьох записів мовця. Саме тому показники середньої кількості пауз у таблиці 3.2 подані в середніх

значеннях у вигляді десяткових дробів. Показники середньої кількості пауз для читаного мовлення були виміряні не для такої великої кількості записів, як для спонтанного, тому їх середні значення були подані у вигляді цілих чисел.

Отримані значення сильно відрізняються від показників для читаного мовлення. Для спонтанного мовлення характерна набагато менша кількість пауз, а також довгі паузи майже не переважають над короткими.

Показники SNR переважно знаходяться в межах 12-22 дБ, за винятком записів мовця №10. Таке значення SNR відповідає середній зашумленості записів: 20-22 дБ означає нормальну якість звукозапису, а значення ближче до 10 дБ свідчить про наявність дуже помітного шуму на фоні. На основі отриманих значень SNR можна зробити висновок, що записи спонтанного мовлення є більш зашумленими, ніж записи читаного, хоча трапляються й винятки – наприклад, записи мовця № 10 мають значне переважання сигналу над шумом.

Аналіз результатів тестування моделей: читане мовлення

Читане мовлення		
Моделі	ARI	NMI
ECAPA-TDNN	0.988	0.993
ResNet	**1**	**1**
ReDimNet	**1**	**1**

Таблиця 3.3. Результати тестування моделей на датасеті читаного мовлення. Жирним виділено найвищі значення метрик.

Як помітно зі значень метрик, всі три моделі продемонструвати майже ідеальну точність на записах читаного мовлення – ResNet та ReDimNet взагалі не зробили жодної помилки. Розпізнавання читаного мовлення виявилося легкою задачею для протестованих моделей.

Далі буде приведено більш докладний огляд їх результатів та перцептивний аналіз помилково класифікованих записів.

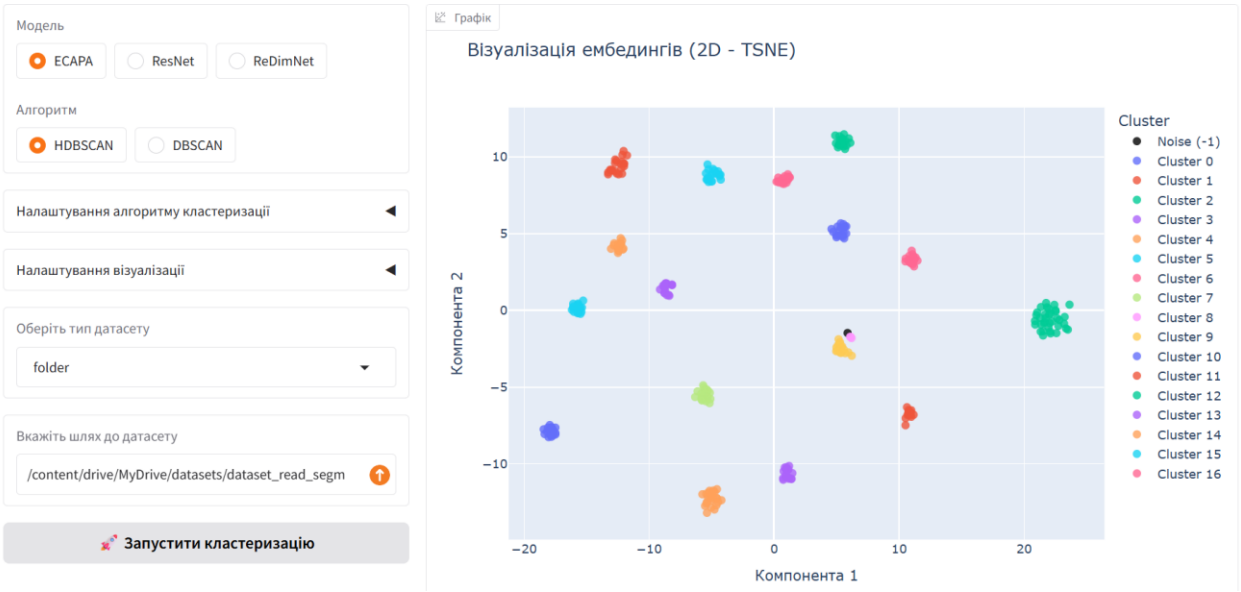

Рис. 3.1. Візуалізація ембедингів читаного мовлення (результати роботи моделі ECAPA-TDNN). Всі ембединги утворюють компактні кластери, що відповідають мовленню однієї особи. Ембединги помилково класифікованих записів знаходяться біля кластера №9 (позначено жовтим): один некласифікований запис (позначено чорним) та кілька записів, помилково відділених в окремий кластер (позначено рожевим).

ECAPA-TDNN

Модель продемонструвала майже ідеальний результат, хоча й допустила декілька незначних помилок. Кластери, утворені ембедингами читаного мовлення, виглядали надзвичайно компактними при візуалізації у 2D-просторі, на відміну від кластерів ембедингів спонтанного мовлення.

Серед результатів був лише один некласифікований запис. Його ембединг розташувався досить близько до кластера, але все ж не був віднесений до цього кластеру алгоритмом, що свідчить про те, що дистанція між ним та іншими ембедингами перевищила порогове значення. При перцептивному аналізі цього звукозапису було встановлено, що він містить дуже довгу паузу (1.63 секунди), численні хезитаційні явища та досить нетипову для мовця інтонацію.

Для цього того ж самого мовця моделлю було допущено ще одну помилку: три записи відділено від основної групи та виділено в окремий кластер (хоча вони теж знаходилися дуже близько до основного кластера). Можлива причина полягала в хезитаційних явищах, які загалом були нехарактерними для читаного мовлення цього мовця. Також в одному із помилково класифікованих записів темп був значно швидшим, ніж на інших.

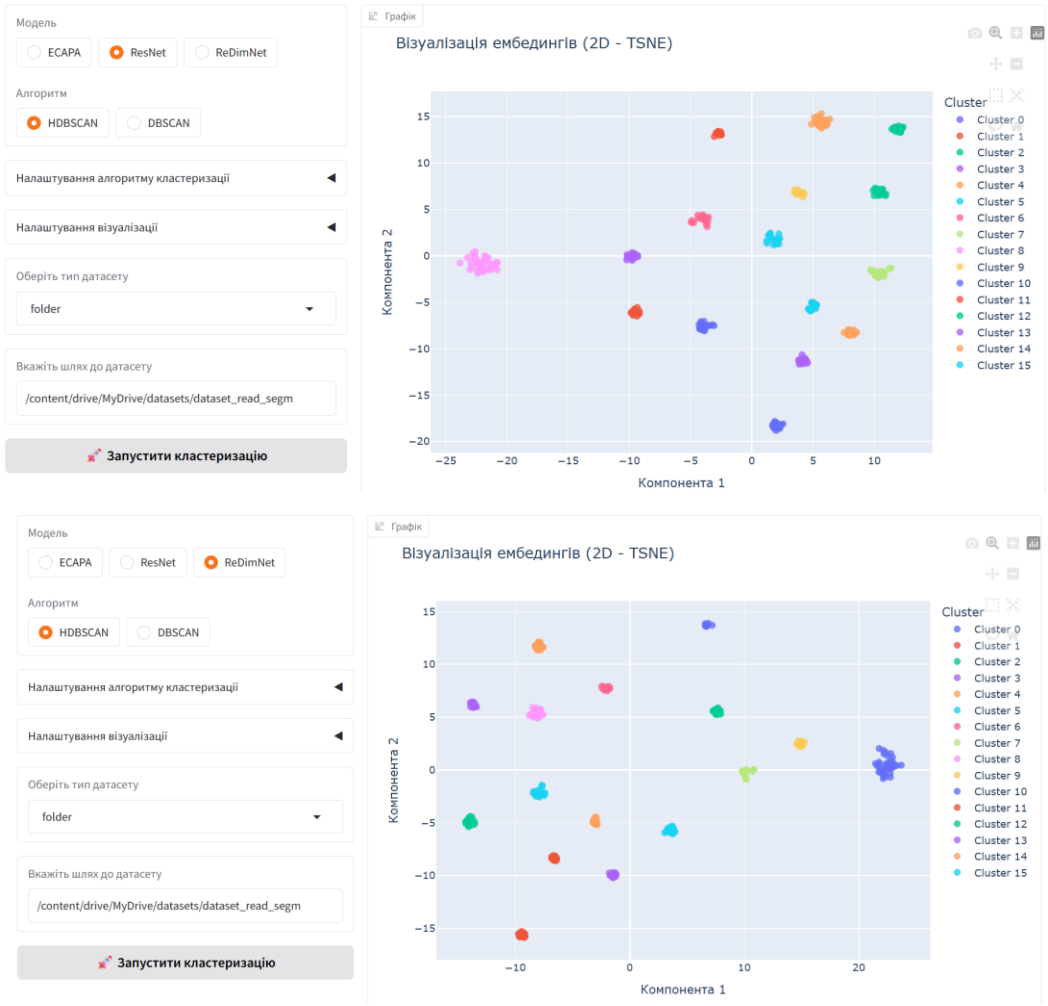

Рис. 3.2. Візуалізація ембедингів читаного мовлення: результати роботи моделей ResNet (вгорі) та ReDimNet (знизу). Кластери ембедингів моделі ReDimNet виглядають компактнішими.

ResNet та ReDimNet

Ці дві моделі продемонстрували ідеальний результат без нерозпізнаних записів та помилкових класифікацій: всі записи було класифіковано правильно. Як

помітно з візуалізації, кластери ембедингів обох моделей є дуже компактними, що свідчить про те, що записи читаного мовлення були дуже подібними між собою для цих моделей.

Щоправда, один кластер завжди був більш розрідженим за інші, адже цей мовець мав більше записів. Велика кількість записів могла зумовити їх більше різноманіття, але проте жодна із моделей не припустилася помилок для записів цього мовця, адже його голос досить сильно відрізнявся за тембром від усіх інших голосів у цьому датасеті.

Аналіз результатів тестування моделей: спонтанне мовлення

Спонтанне мовлення		
Моделі	ARI	NMI
ECAPA-TDNN	0.319	0.655
ResNet	0.349	0.677
ReDimNet	**0.451**	**0.712**

Таблиця 3.4. Результати тестування моделей на датасеті спонтанного мовлення. Жирним виділено найвищі значення метрик.

Зі значень метрик, отриманих в результаті тестування моделей на датасеті спонтанного мовлення, можна зробити висновок, що цей тип мовлення розпізнається моделями набагато гірше. Цей набір даних виявився надзвичайно складною задачею для всіх представлених моделей. Отже, підготовленість мовлення значною мірою впливає на точність його розпізнавання.

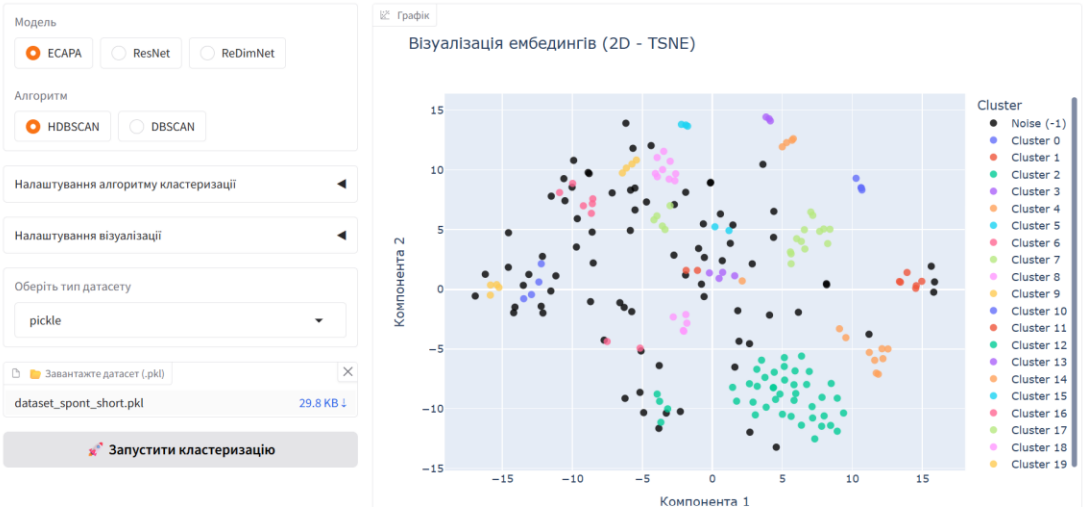

Рис. 3.3. Візуалізація ембедингів моделі ECAPA-TDNN (датасет із записами спонтанного мовлення).

ECAPA-TDNN

Модель продемонструвала низьку точність, ембединги були дуже розрідженими та майже не згрупованими. Багато записів було віднесено до категорії “шум”, що свідчить про те, що модель не змогла зробити ближчими ембединги одного голосу. До того ж було дуже багато некласифікованих записів (79 із 212), які не були віднесені до жодної з груп ембедингів. Для кількох мовців звукозаписи все ж були правильно об’єднані в кластери, але таких мовців було небагато.

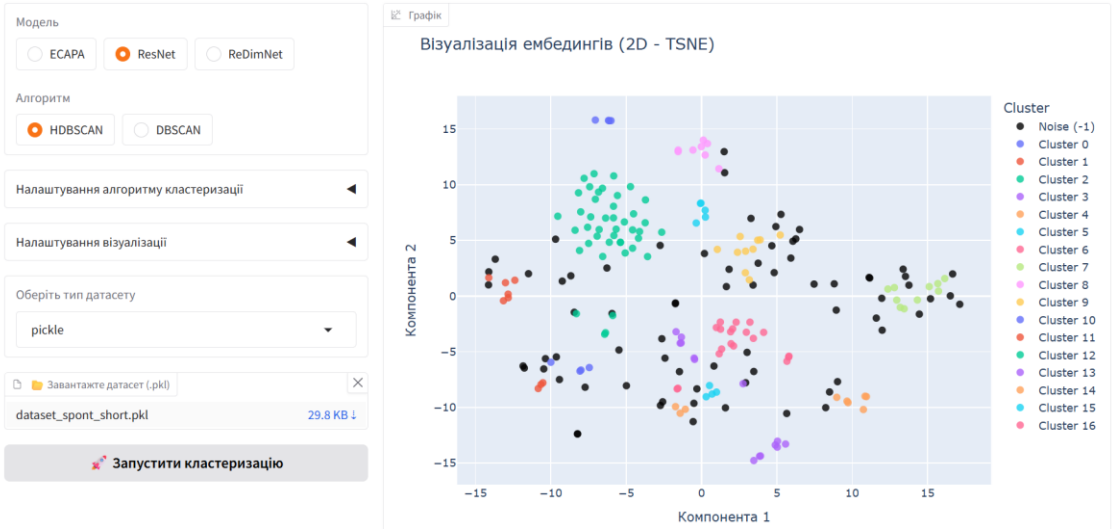

Рис. 3.4. Візуалізація ембедингів моделі ResNet (датасет із записами спонтанного мовлення).

ResNet та ReDimNet

Ці моделі зіткнулися з тою самою проблемою, що й ECAPA-TDNN: ембединги були надто розріджені, що добре видно із їх візуалізації. Щоправда, вони були трохи краще згруповані: результати містили більше правильних кластерів, ніж у попередньої моделі, і в ці кластери потрапило більше записів. Також було менше некласифікованих записів (73 із 212 для ResNet та 61 із 212 для ReDimNet). Модель ReDimNet найкраще справилася із поставленим завданням.

Під час перцептивного аналізу було виявлено, що більшість записів містить сильно помітні шуми, переривання мовлення, а всі акустичні параметри голосу сильно варіюються, адже голосові повідомлення записані в різних середовищах та на різній відстані від мікрофона. Записи спонтанного мовлення мали надзвичайно різноманітні інтонації, великий діапазон ЧОТ та нестабільність темпу.

3.2. **Вплив відмінностей чоловічих та жіночих голосів**

Для обох датасетів (чоловічого та жіночого мовлення) було проведено однакові вимірювання – ЧОТ (частоти основного тону), F1 (першої форманти) та F2 (другої форманти). Для всіх отриманих результатів було проведено аналіз, що кількісно підтвердив акустичну відмінність між мовленням цих двох груп. Через велику кількість мовців в обох датасетах, вимірювання проводилися лише для перших десяти мовців з кожного набору даних.

Чоловіче мовлення

№ мовця у датасеті	F0 (ЧОТ), ГЦ	F1, ГЦ	F2, ГЦ
#1	105	560	1680
#2	100	500	1270
#3	125	535	1167
#4	110	350	1545
#5	130	574	1373
#6	127	413	1794
#7	96	494	1264
#8	104	302	1173
#9	108	496	1210
#10	145	619	1393

Таблиця 3.5. Значення ЧОТ, першої та другої формант для чоловічих голосів

На основі отриманих значень ЧОТ, F1 та F2 для чоловічих голосів можна зробити висновки про їх варіативність.

Більшість значень ЧОТ знаходяться в діапазоні 100-125 Гц; голоси можна схарактеризувати як низькі та не дуже відмінні за частотою основного тону.

Однак у значеннях формант спостерігається набагато більша різноманітність. Для F1, деякі голоси мають значення близько 300 Гц (мовці #4 та #8), а верхня межа діапазону значень першої форманти перевищує 500 Гц і в одного мовця (позначеного #10) доходить до 600 Гц. Щоправда, для цього мовця характерні також високі значення ЧОТ, які значно відрізняються від інших у цій вибірці. Можливо, для цього вищого голосу характерні також вищі значення формант.

У значеннях другої форманти теж простежується значна різноманітність – значення варіюються від близько 1200 Гц до близько 1700 Гц, що є досить широким

діапазоном і свідчить про відмінність голосів. Отже, хоч чоловічі голоси не надто відрізняються за частотою основного тону, вони мають відмінності у формантах.

Жіноче мовлення

№ мовця у датасеті	F0 (ЧОТ), ГЦ	F1, ГЦ	F2, ГЦ
#1	210	683	2135
#2	255	551	1660
#3	223	636	1810
#4	203	580	2025
#5	294	594	1748
#6	220	690	1858
#7	241	639	1789
#8	188	473	1820
#9	235	639	1636
#10	239	574	1778

Таблиця 3.6. Значення ЧОТ, першої та другої формант для жіночих голосів

Для жіночих голосів було виміряно ті самі показники, що й для чоловічих (значення представлені в таблиці 3.6). Отримані значення досить суттєво відрізняються від значень для чоловічих голосів. Показники ЧОТ майже вдвічі вищі; більшість значень знаходиться в діапазоні 200 - 250 Гц, що свідчить про те, що жінки мають значно вищі голоси. Показники першої форманти відрізняються від чоловічих і є вищими (в середньому 550 - 650 Гц). Це підтверджує наше попереднє спостереження щодо того, що для вищих голосів характерні вищі значення F1 та F2.

Показники другої форманти у жіночих голосів теж відрізняються від чоловічих. Хоча значення цієї форманти досить сильно варіюються як в першій, так

і в другій групі мовців, загалом можна зробити висновок, що значення F2 у жінок в середньому вищі.

Окремо варто відзначити варіативність всередині групи. Для жінок характерна більша варіативність у значеннях ЧОТ – діапазон у 200 - 250 Гц значно ширший за діапазон 100 - 125 Гц у чоловіків. Проте для жінок характерна менша варіативність у значеннях другої форманти – всі значення є вищими за 1600 Гц та рідко перевищують 2000 Гц.

Загалом, на основі проведеного попереднього аналізу можна зробити висновок, що для чоловічих голосів у цьому датасеті характерна стабільна, майже однакова ЧОТ і варіативні значення формант, в той час, як для жіночих голосів навпаки – більша розбіжність у значеннях ЧОТ між мовцями та менша розбіжність між формантами у різних осіб.

Аналіз результатів тестування моделей: чоловіче мовлення

Чоловіче мовлення				
Моделі	ARI	NMI	mean intra-cluster distance	mean inter-cluster distance
ECAPA-TDNN	0.961	0.981	0.355	0.806
ResNet	0.951	0.977	**0.345**	0.809
ReDimNet	**0.993**	**0.997**	0.363	**0.885**

Таблиця 3.7. Результати тестування моделей на датасеті чоловічого мовлення із розрахованою середньою косинусною відстанню між ембедингами всередині кластера (mean intra-cluster distance) та між кластерами (mean inter-cluster distance)

Проаналізувавши отримані результати, ми дійшли висновку, що найкращі значення за більшістю метрик має модель ReDimNet (найвищі показники ARI та NMI, що означає точне розбиття на кластери, та найбільша середня косинусна відстань між кластерами). Модель ResNet продемонструвала найнижчий результат

за метриками ARI та NMI, проте її ембединги виявилися найбільш згрупованими всередині кластерів. Ембединги моделі ReDimNet виявилися найбільш розрідженими у просторі, адже були далі одне від одного у межах кластера, а самі кластери теж розташовувалися далеко одне від одного.

Також середня косинусна відстань між ембедингами всередині кластера не дуже відрізняється у всіх трьох моделей, в той час, як середня відстань між кластерами виявилася набагато більшою у ReDimNet, моделі з найкращими результатами.

З наведених значень помітно, що сильна згрупованість ембедингів у кластері не завжди означає вищу точність, проте значна відстань між кластерами може свідчити про якісну кластеризацію.

Крім того, було здійснено докладний перцептивний аналіз записів, помилково класифікованих моделями.

ECAPA-TDNN

Модель опинилася на другому місці за точністю кластеризації, продемонструвавши досить високий результат, проте все ж припустилася кількох помилок.

П’ять звукозаписів було класифіковано як “шум” та не віднесено до жодної групи, хоча всі вони знаходилися досить близько до тих кластерів, до яких мали належати. Більшість із цих аудіофайлів мали шуми на фоні або були зроблені не в тих самих умовах що й інші записи голосів мовців.

Також модель ECAPA-TDNN здійснила одне помилкове розбиття записів голосу мовця на два кластери: три записи було відділено в окремий кластер. Методом перцептивного аналізу було з’ясовано, що ці три записи було зроблено в інших умовах – можливо, на інший пристрій чи в іншому середовищі, що й спричинило різницю в якості.

Крім того, результати містили дві помилкові атрибуції. В першому випадку, ембединг було приписано до групи записів голосу іншого мовця, при цьому він

знаходився дуже далеко від свого справжнього кластера. Перцептивно, цей запис голосу був дуже схожий із записами того кластеру, до якого його було віднесено: мовлення мало схожу ЧОТ, а також запис мав подібний рівень зашумленості. В другому випадку, два кластери знаходилися поруч, і один запис, хоч і був досить близьким до свого кластера, виявився ближчим до іншого. На цьому записі була присутня незначна реверберація, а тембр голосу був подібний до тембру іншого мовця. Крім того, мовлення було швидшим та більш емоційним, ніж зазвичай характерно для цього мовця.

ResNet

Модель ResNet продемонструвала досить високу точність, проте якість її ембедингів виявилася нижчою ніж в інших моделей. Після проведення кластеризації її ембедингів було виявлено 9 некласифікованих записів. Всі вони знаходилися досить близько до своїх справжніх кластерів, проте їх подібність до інших ембедингів групи виявилася недостатньою. Деякі мовці мали по декілька некласифікованих записів. Можливими причинами цих помилок були несхожість тембру та шум середовища.

Також модель два рази здійснила помилкове розбиття записів мовця на два кластери. У першому випадку, це була та сама помилка, якої припустилася модель ECAPA-TDNN – записи цього мовця було розділено на два окремі кластери. У другому випадку, два записи голосу мовця теж було відділено від основної групи в окремий кластер. В обох випадках, можливою причиною помилок став шум середовища та варіації тембру голосу.

ReDimNet

Модель продемонструвала надзвичайно високі результати. Вона здійснила лише одну неправильну класифікацію – ембединг було помилково віднесено до іншого кластеру. Це була та сама помилка, якої припустилася модель ECAPA-TDNN; голос на цьому записі був дуже схожим на голос іншого мовця.

Аналіз результатів тестування моделей: жіноче мовлення

Жіноче мовлення				
Моделі	ARI	NMI	mean intra-cluster distance	mean inter-cluster distance
ECAPA-TDNN	0.928	0.978	0.330	0.732
ResNet	0.987	0.992	**0.320**	0.778
ReDimNet	**0.994**	**0.997**	0.337	**0.850**

Таблиця 3.8. Результати тестування моделей на датасеті жіночого мовлення із розрахованою середньою косинусною відстанню між ембедингами всередині кластера (mean intra-cluster distance) та між кластерами (mean inter-cluster distance)

Для жіночих голосів, так само як і для чоловічих, модель ReDimNet продемонструвала найкращі результати. Проте в показниках метрик інших двох моделей виявилися відмінності для різних датасетів: ECAPA-TDNN спрацювала краще для чоловічих голосів, а ResNet показала кращий результат для жіночих. Показники косинусної відстані між ембедингами свідчать про те, що модель ResNet забезпечує більшу компактність кластерів – як із чоловічими голосами, так і з жіночими; це можна вважати особливістю самої моделі.

Як і у випадку з чоловічими голосами, модель ReDimNet має найбільші відстані між ембедингами як всередині кластерів, так і між самими кластерами, тому більша відстань між ембедингами теж є особливістю роботи самої моделі. Косинусна відстань між кластерами є набагато більшою у моделі ReDimNet. Як і у попередньому датасеті, значення косинусної відстані всередині кластерів відрізняється менше, ніж між кластерами. Отже, можна зробити висновок, що велика відстань між різними кластерами більш важлива для точності, ніж компактність цих кластерів – і результати моделі ReDimNet демонструють цю перевагу.

Крім того, косинусна відстань як всередині кластерів, так і між ними виявилася набагато меншою для жіночих голосів, ніж для чоловічих – незалежно від моделі. Отже, для всіх моделей, незалежно від особливостей їх архітектури чи точності результатів, жіночі голоси були більш схожими між собою, ніж чоловічі.

Далі представлено результати перцептивного аналізу записів, щодо яких моделі припустилися помилок.

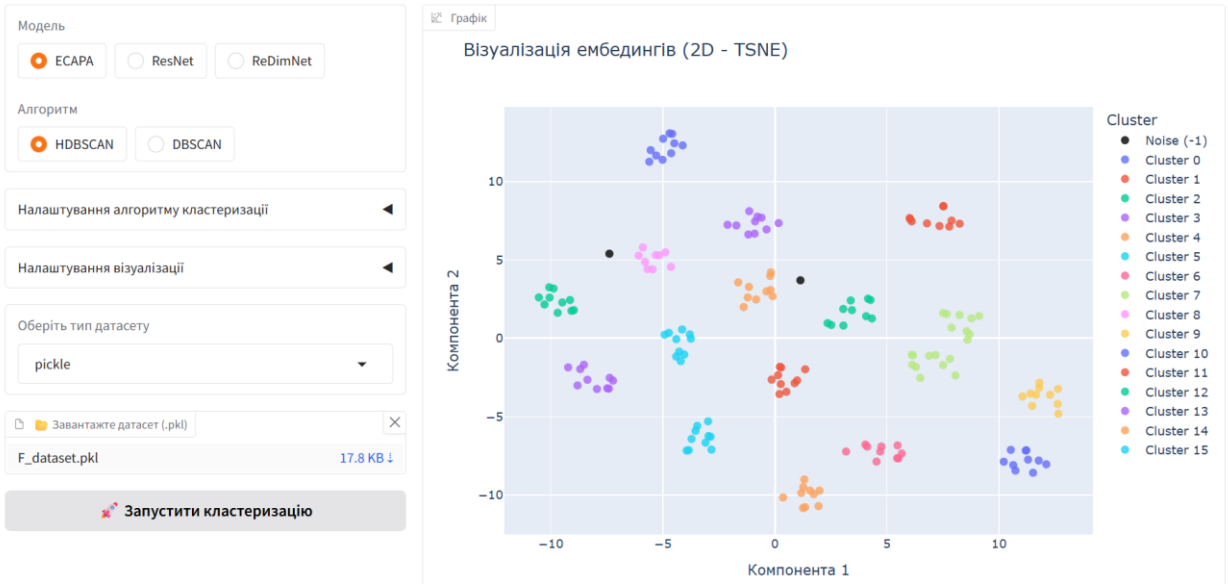

Рис. 3.5. Візуалізація ембедингів моделі ECAPA-TDNN (датасет жіночого мовлення). Чорним позначено ембединги, далекі від кластерів (категорія "шум"). Кластер 7 (позначено світло-зеленим) містить помилково об'єднані записи голосів різних мовців.

ECAPA-TDNN

Два записи не ввійшли до жодного кластера: один із них знаходився близько до інших ембедингів голосу цього мовця, але не був віднесений до них, в той час, як другий знаходився на великій відстані від свого кластера, що було добре помітно на діаграмі візуалізації. При перцептивному аналізі цього звукозапису було виявлено, що на фоні присутній дуже сильний білий шум – запис є надто зашумленим, і моделі не можуть створити якісне представлення голосу.

Також модель ECAPA-TDNN помилково об'єднала записи двох різних голосів в один кластер (кластер №7 на рис. 3.5). Перцептивно, ці два жіночі голоси

звучали дуже схоже: обидва мали швидкий темп, високу інтенсивність, схожу висоту голосу, подібні інтонаційні контури при емоційному мовленні. Обидві дикторки майже не робили пауз, а та невелика кількість пауз, що все ж була присутня в їх мовленні, утворювала схожі патерни.

ResNet та ReDimNet

Модель ResNet продемонструвала набагато кращі результати, ніж ECAPA-TDNN, не зробивши помилкових об'єднань кластерів. Її помилками були два некласифіковані записи, причому один із них був тим самим аудіофайлом, для якого зробила помилку ECAPA-TDNN. Також цей запис не віднесла до жодного кластера модель ReDimNet – це була єдина помилка у роботі цієї моделі на всьому датасеті. Цей запис не змогла правильно класифікувати жодна із протестованих моделей – можливо, він є надто зашумленим для завдання розпізнавання мовця за голосом, хоча перцептивно так не здається, і до того ж датасет містить набагато більш зашумлені аудіофайли із музикою чи шумом аудиторії на фоні.

3.3. **Вплив кількості голосів у датасеті та обсягу даних**

Аналіз результатів тестування моделей: велика кількість мовців

Кількість мовців: **50**; обсяг датасету: **~1,5** годин				
Моделі	ARI	NMI	mean intra-cluster distance	mean inter-cluster distance
ECAPA-TDNN	0.945	0.982	0.345	0.837
ResNet	0.962	0.984	**0.338**	0.834
ReDimNet	**0.994**	**0.998**	0.355	**0.893**

Таблиця 3.9. Результати тестування моделей на датасеті з великою кількістю мовців із розрахованою середньою косинусною відстанню між ембедингами всередині кластера (mean intra-cluster distance) та між кластерами (mean inter-cluster distance)

Акустичні характеристики мовлення не були релевантними для вимірювання впливу кількості мовців у тестувальному датасеті на точність роботи моделей, тому в цій частині увагу було приділено передусім порівнянню значень косинусних відстаней та аналізу помилок.

При вимірюванні середньої косинусної відстані між ембедингами всередині кластерів та між окремими кластерами, було отримано результати, схожі на попередні показники для датасетів чоловічого та жіночого мовлення. Зокрема, ResNet має найбільш компактні кластери серед усіх моделей, а також найменшу відстань між кластерами, що свідчить про близькість розташування всіх її ембедингів. Попри те, що ембединги ResNet так сильно схожі між собою, ця модель зазвичай демонструє кращі результати, ніж ECAPA-TDNN, ембединги якої менш подібні – наприклад, ResNet перевершувала іншу модель в тестуванні на датасетах читаного, спонтанного та жіночого мовлення.

Такий результат зумовлений тим, що близькість між ембедингами схожих голосів для ResNet має більше значення, ніж відстань між несхожими: наприклад, різниця в компактності кластерів між ResNet та ECAPA-TDNN дорівнює 0.007, а їх різниця у відстані між окремими кластерами – 0.003. Це означає, що ResNet демонструє кращі результати шляхом того, що робить однакові голоси якомога ближчими.

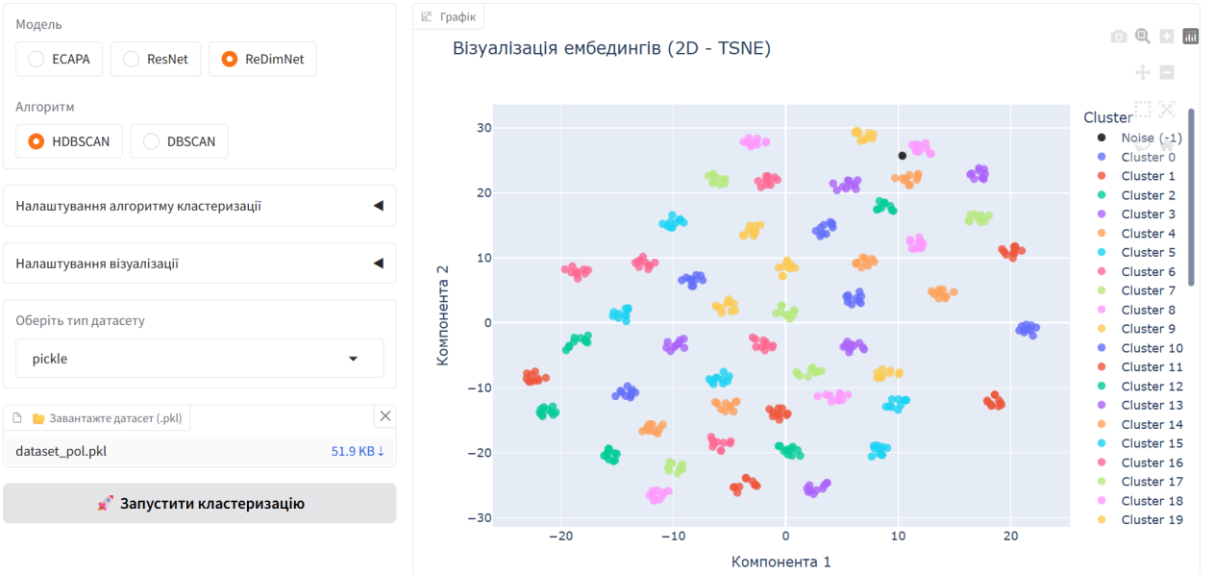

Рис. 3.6. Візуалізація ембедингів моделі ReDimNet для датасету з великою кількістю мовців.

На противагу їй, модель ReDimNet має протилежну стратегію. Її ембединги є далекими один від одного – як всередині кластерів, так і між кластерами, що помітно зі значень у таблиці 3.9. Для неї велика відстань між ембедингами несхожих голосів є більш значущою, ніж близькість між схожими: наприклад, різниця в компактності кластерів між ReDimNet та ECAPA-TDNN становить 0.01, а їх різниця у відстані між окремими кластерами – 0.056. Отже, модель ReDimNet стабільно демонструє кращі результати за ECAPA-TDNN через більшу відстань між кластерами, тобто між ембедингами несхожих голосів. Її стратегія полягає в тому, щоб робити ембединги різних голосів максимально далекими.

Крім аналізу середніх косинусних відстаней між ембедингами, було також проведено аналіз допущених помилок та перцептивне дослідження неправильно класифікованих записів.

ECAPA-TDNN

Велика кількість мовців спричинила певні проблеми в роботі моделі та вплинула на якість кластеризації: записи двох мовців було помилково об’єднано в один кластер, а записи іншого мовця було помилково розбито на два кластери (хоча ембединги виглядали досить близькими на візуалізації). Також є кілька помилкових класифікацій та сім нерозпізнаних записів – багато з них теж знаходилося досить близько до правильного кластера.

ResNet

Значна кількість різних голосів також виявилася складним завданням для цієї моделі: дев’ять записів не було віднесено до жодного кластера (“нерозпізнані”) – як і в результатах тестування ECAPA-TDNN, переважна більшість із них знаходилася досить близько до правильного кластера. Також модель кілька разів помилково розбивала на різні кластери записи одного голосу. Один з таких випадків було досліджено методом перцептивного аналізу: звукозаписи, що виділились в окремий

кластер, мали низьку якість (шум, реверберація), швидший темп та меншу кількість пауз. Проте інші записи голосу цього мовця також іноді містили реверберацію, тому найбільш вірогідним є те, що проблема полягала у темпі та паузації.

ReDimNet

Ця модель найкраще справилася із поставленим завданням, зробивши лише дві помилки: один із записів потрапив до категорії “шум”, другий було помилково приписано іншому мовцеві. Запис, що не був віднесений до жодного з кластерів, мав низьку якість, а неправильно класифікований запис голосу був дійсно схожий за тембром на мовлення тієї особи, якій його було помилково приписано. Така незначна кількість помилок дає підстави стверджувати, що модель чудово справляється із датасетами з великою кількістю мовців, і наявність надто схожих голосів чи мовлення з великою варіативністю не є для неї такою значною проблемою, як для інших моделей.

Аналіз результатів тестування моделей: мала кількість мовців

Кількість мовців: **4**; обсяг датасету: **~1,5** годин						
Моделі	ARI	NMI	Кластер 1 (284 записи)	Кластер 2 (170 записів)	Кластер 3 (54 записи)	Кластер 4 (15 записів)
ECAPA-TDNN	**1**	**1**	0.457	**0.299**	0.288	0.317
ResNet	**1**	**1**	0.468	0.328	0.298	0.300
ReDimNet	**1**	**1**	**0.403**	0.326	**0.224**	**0.158**

Таблиця 3.10. Результати тестування моделей на датасеті з невеликою кількістю мовців та середня косинусна відстань між ембедингами кожного кластера

Всі три моделі продемонстрували ідеальний результат на цьому датасеті, попри те, що він містив переважно спонтанне мовлення, і різні мовці мали неоднакову кількість записів, що робило датасет незбалансованим.

Оскільки цей набір даних містив малу кількість мовців із неоднаковим розподілом записів, було вирішено розрахувати середню косинусну відстань між ембедингами всередині кожного кластера. З результатів, представлених у таблиці 3.10, та візуалізації на рис. 3.7 добре помітно, що всі чотири кластери мають різну компактність.

Для двох мовців (№1 та №2) спостерігається менша компактність кластерів: ембединги більш розріджені та групуються у невеликі підкластери всередині основного. В результаті перцептивного аналізу було з’ясовано, що записи голосів цих мовців мають різну якість, були зроблені в різний час, в неоднакових умовах та в різних комунікативних ситуаціях. Всередині кластерів навіть помітно невеликі підгрупи, що відповідають різним акустичним умовам на записах. Проте жодна з моделей не припустилася помилок і не здійснила помилкового розбиття цих великих кластерів на кілька груп: всі записи цих голосів було класифіковано правильно. Можливо, свою роль тут відіграло те, що обидва голоси були жіночими

– а як показав попередній експеримент, записи жіночих голосів менш варіативні та мають меншу ймовірність бути помилково розбитими на різні кластери, на відміну від чоловічого мовлення.

Для інших двох мовців (№3 та №4) незалежно від моделі ембедінги розташовувались дуже близько одне до одного. Перцептивний аналіз аудіофайлів та їх змісту довів, що це результат того, що всі записи були зроблені в однакових умовах, на один і той самий пристрій в межах тієї самої комунікативної ситуації. Візуалізація ембедингів демонструє, наскільки сильно варіативність середовища, фону, комунікативної ситуації та змісту записів впливає на компактність кластерів.

Також варто порівняти показники компактності кластерів у цьому і в попередньому датасетах. У наборі даних, що містив п’ятдесят голосів, кожен мовець мав по 10 записів, зроблених в різних ситуаціях та неоднакових умовах середовища. Ці кластери були в середньому менш компактними, ніж кластери №2, №3 та №4 із цього датасету, які мали 170, 54 та 15 записів відповідно. Отже, можна зробити висновок, що сама лише кількість записів голосу мовця не впливає на компактність кластера – кластер може містити 170 записів та бути більш компактним, ніж кластер, що містить 10 записів. Чинники, що сприяють меншій подібності ембедингів всередині одного кластеру – це якість, акустичні умови запису та різноманітність мовленнєвих ситуацій, а не безпосередня кількість записів.

Крім того, при тестуванні на датасетах із великою та збалансованою кількістю мовців, модель ReDimNet мала в середньому найменш компактні кластери, в той час, як на цьому наборі даних три із чотирьох її кластерів виявилися найбільш компактними серед усіх моделей, що помітно не тільки зі значень в таблиці 3.10, а й на візуалізації ембедингів (рис. 3.7).

Рис. 3.7. Візуалізація ембедингів моделей ECAPA, ResNet та ReDimNet відповідно.

Отже, можна зробити висновок, що ReDimNet створює більш розріджені та далекі кластери (а ResNet – компактніші та ближчі) *в середньому*, а не кожного разу. Схильність створювати більш віддалені чи більш подібні ембединги починає простежуватись тільки при великій кількості утворених кластерів (тобто при великій кількості мовців у датасеті). Якщо ж датасет містить голоси лише кількох мовців, ця закономірність не простежується: кілька утворених кластерів можуть виявитися компактними у ReDimNet чи розрідженими у ResNet.

Цей факт варто враховувати при подальших застосуваннях цих моделей у різноманітних завданнях. Середня компактність кластерів і, відповідно, точність роботи моделі залежить більше від кількості різних мовців у датасеті та різноманітності акустичних умов, а не від обсягу цього датасету та кількості записів голосу кожного диктора.

Висновки до Розділу 3

В результаті перцептивного дослідження та вимірювання акустичних параметрів наборів даних, порівняння результатів роботи моделей та перцептивного аналізу помилково класифікованих записів, було сформовано висновки про вплив кількісних та акустичних характеристик мовленнєвих даних на точність роботи моделей.

Зокрема, попередній аналіз показав, що записи спонтанного мовлення в середньому мають більшу зашумленість (нижчу якість), більшу варіативність інтонацій, швидший темп та коротші паузи, ніж записи читаного мовлення. Для

читаного мовлення, на противагу, характерний сталий темп, монотонна інтонація, довгі паузи між довгими синтагмами та краща якість (переважання сигналу над шумом). Вочевидь, сталий темп та низька варіативність інтонацій у читаному мовленні зробили його легшим для розпізнавання: моделі ResNet та ReDimNet продемонстрували ідеальний результат, а ECAPA-TDNN зробила лише кілька помилок. Проте розпізнавання спонтанного мовлення виявилось набагато складнішою задачею, з якою всі моделі справилися гірше. З візуалізації ембедингів було помітно, що записи читаного мовлення утворили надзвичайно компактні, чітко розділені кластери, в той час ембединги записів спонтанного мовлення були менш згрупованими у кластери та більш далекими одне від одного, через що багато з них не було віднесено до жодного кластера та позначено як “шум”. Крім того, варто відзначити додатковий вплив такого фактора як обсяг набору даних – хоча обидва датасети містили однакову кількість голосів, датасет спонтанного мовлення мав більше записів, що теж могло погіршити точність для цього набору даних.

Вимірювання акустичних параметрів чоловічих та жіночих голосів показало їх якісні відмінності: жіночі голоси були більш подібними між собою за формантною структурою, а чоловічі мали більшу варіативність у значеннях формант. Результати тестування моделей підтвердили це: для жіночих голосів було більш характерним помилкове об’єднання записів різних голосів в один кластер, а для чоловічих – помилкове розбиття записів одного голосу на різні кластери. Таким чином, жіночі голоси є складними для розпізнавання через свою схожість, а чоловічі – через високу варіативність голосу кожного мовця. Це підтверджували й показники косинусної відстані всередині кластерів та між ними: ембединги жіночих голосів завжди були ближчими між собою, незалежно від моделі.

Також варто відзначити вплив додаткового фактора – різної кількості мовців. Датасет жіночого мовлення мав менше голосів, що могло спричинити меншу середню відстань між кластерами та всередині них. Проте подібність між жіночими голосами та їх нижча варіативність була спричинена не тільки меншою кількістю

мовців, що помітно із допущених системою помилок. Наприклад, одна з моделей помилково об’єднала записи двох дикторок в один кластер, визначивши їх як надто схожі, але ця ж модель не зробила такої помилки для датасету чоловічого мовлення. Це означає, що у датасеті чоловічого мовлення не знайшлося двох “надто схожих” голосів, попри вдвічі більшу кількість мовців. До того ж для жіночого мовлення жодного разу не було здійснено помилкового розбиття записів голосу одного мовця на різні кластери, на відміну від чоловічого. Отже, жіночі голоси у цьому випадку дійсно були більш схожими між собою, а чоловічі – більш варіативними, і вплив кількості мовців у датасеті виявився другорядним.

Варто відзначити й те, що обидва датасети містили добре підготовлене мовлення, що дуже позитивно вплинуло на точність всіх моделей. Як показав попередній експеримент, читане мовлення є легшим для розпізнавання, адже сталий темп та монотонний ритм полегшують задачу. Проте у випадку з жіночим мовленням цей фактор швидше спричинив ускладнення розпізнавання, адже зробив схожі жіночі голоси ще більш схожими через ритмічність читаного мовлення.

Також було виявлено відмінності в роботі самих моделей. ResNet демонструвала високі показники через те, що робила ембединги однакових голосів якомога ближчими між собою, через що спрацювала краще для жіночого мовлення, в той час, як ECAPA-TDNN показала вищі результати для чоловічого. ReDimNet продемонструвала найкращий результат на обох датасетах незалежно від статі мовця. В результаті аналізу компактності кластерів було з’ясовано, що ця модель робить ембединги різних голосів якомога більш несхожими, через що й демонструє таку високу точність.

Було проведено дослідження впливу кількості мовців у датасеті на точність роботи моделей. Було встановлено, що моделі набагато краще справляються із набором даних з невеликою кількістю мовців, навіть якщо обсяг датасету великий і ці кілька мовців мають багато записів. Всі моделі продемонстрували гірший результат на наборі із великою кількістю голосів – більшість помилок було

допущено через схожість різних голосів або велику варіативність в межах одного голосу. Це дві типові проблеми, з якими стикаються моделі, коли в тестувальному датасеті багато різних мовців.

Додатковим фактором впливу в цьому випадку була зашумленість записів. Великий за обсягом датасет, який містив лише чотири голоси, був досить зашумленим, адже записи були зроблені в природних умовах та у різних середовищах, проте це не вплинуло на точність системи. Натомість у датасеті зі значною кількістю мовців зашумлені записи спричинили ряд помилок системи. Це свідчить про те, що фактор зашумленості впливає на результат меншою мірою, ніж кількість голосів, але при збільшенні числа мовців його вплив теж зростає, і зашумленість починає призводити до неточностей. Також з порівняння компактності кластерів було виявлено, що вона більше залежить від різноманітності акустичних умов, ніж від кількості записів.

Висновки

У межах цієї роботи було проведено огляд методів та технологій в основі систем автоматичного розпізнавання мовців, створено систему розпізнавання мовців шляхом кластеризації та оцінено вплив кількісних та якісних характеристик мовленнєвих даних на точність роботи системи із використанням різних моделей, які також було порівняно за ефективністю.

Дослідження, проведене у першому розділі, надало теоретичну основу для створення системи. Визначивши та роз'яснивши ключові поняття, ми встановили, що завдання створення системи розпізнавання мовця поєднує елементи діаризації, ідентифікації та верифікації за голосом, а також включає задачу кластеризації. Проаналізувавши основні типи будови систем, ми дійшли висновку, що наскрізні моделі та гібридні підходи (такі як поєднання наскрізної моделі з кластеризацією) є найбільш поширеними та ефективними рішеннями, і створена система є прикладом саме гібридного підходу. Оглянувши сучасні праці, що стосувалися сфери автоматичного розпізнавання мовця, зокрема роботи В.Бридінського, Х.С.Рудої та інших, ми визначили кількісні та якісні характеристики мовленнєвих даних, які мають найбільший вплив на точність роботи досліджуваних систем. Такими характеристиками є стать мовця, кількість різних голосів у тестувальному наборі даних, зашумленість середовища у поєднанні зі спонтанним мовленням (для записів, зроблених в природних умовах) та ступінь підготовленості мовлення. Також в результаті огляду цих робіт було вибрано моделі розпізнавання мовця, які показали хороші результати при роботі з українськомовними даними: ResNet, ReDimNet та ECAPA-TDNN.

У другому розділі було розглянуто принципи роботи вибраних моделей та встановлено, що вони використовують мел-спектрограми як вхідні дані. Оскільки мел-спектрограми представляють частотний склад звукозапису, який відображає людське сприйняття звуку, при розгляді помилок системи було визнано доцільним здійснити аналіз формант, тембру, паузації та ступеня зашумленості мовленнєвих

даних. Далі було описано будову створеної системи. Її основними компонентами є два модулі: модель для отримання ембедингів звукозаписів та алгоритм, що їх кластеризує. Модульність створеної архітектури дозволяла динамічно замінювати використовувану в системі модель, що дало можливість оцінити ефективність та порівняти результати усіх трьох моделей, згаданих вище. Створене програмне забезпечення було інтегровано в один інтерактивний вебінтерфейс, що значно полегшило й пришвидшило проведення експериментів та аналіз їх результатів. Далі було представлено три пари датасетів, сформованих для аналізу впливу якісних та кількісних характеристик мовленнєвих даних на роботу системи. Для оцінки впливу підготовленості мовлення було створено датасети із читаним та спонтанним мовленням, для аналізу впливу статі мовця – датасети жіночого та чоловічого мовлення, а для оцінки впливу числа мовців – датасети із 50 та 4 голосами відповідно. Після цього було проведено головний експеримент дослідження. В результаті попереднього аналізу отриманих показників метрик було встановлено, що всі моделі демонструють кращі результати для читаного мовлення та меншого числа мовців, а також демонструють різну точність залежно від статі мовця. Найвищу точність, незалежно від датасету, демонструє модель ReDimNet.

У третьому розділі було здійснено вимірювання та аналіз акустичних характеристик звукозаписів у сформованих датасетах. Проведений аналіз показав, що записи спонтанного мовлення зазвичай більш зашумлені (нижчий показник SNR), мають швидший темп, коротші паузи та більшу варіативність інтонацій. Читаному мовленню притаманний стабільніший та повільніший темп, довгі паузи, менш варіативні інтонації та більше переважання сигналу над шумом (вищий SNR). Це робить читане мовлення легшим для розпізнавання голосів: дві з трьох моделей (ResNet та ReDimNet) продемонстрували ідеальну точність на датасеті читаного мовлення, а третя модель (ECAPA-TDNN) зробила лише кілька помилок. Спонтанне мовлення спричинило значне погіршення результату: всі моделі продемонстрували набагато нижчі значення метрик, проте їх результати

зіставлялися так само: найкращою була модель ReDimNet, трохи нижчу точність демонструвала ResNet, найнижчі показники мала ECAPA-TDNN. Саме записи спонтанного мовлення в умовах зашумленого середовища є найскладнішими для розпізнавання голосів, адже для цього датасету було отримано найнижчі показники всіх моделей.

Головними відмінностями в акустичних параметрах чоловічих та жіночих голосів були особливості формантної структури: жіночі голоси мали більшу подібність у значеннях формант, в той час, як чоловічі мали більшу варіативність. Ця особливість значним чином позначилася на результатах роботи моделей: жіночі голоси мали меншу косинусну відстань між ембедингами звукозаписів. Для датасету жіночого мовлення характерною помилкою було об'єднання різних голосів в один кластер, а для чоловічого – розбиття записів одного голосу на різні кластери. Таким чином було встановлено, що складність розпізнавання жіночих голосів полягає у їх схожості, а чоловічих – у надмірній варіативності кожного голосу. У цій частині експерименту було виявлено відмінності в роботі моделей для записів чоловічних та жіночих голосів. Зокрема, ResNet продемонструвала кращі показники для датасету жіночих голосів, ніж ECAPA-TDNN, адже особливості її роботи полягають в тому, що ця модель робить ембединги однакових голосів якомога ближчими між собою. Модель ECAPA-TDNN продемонструвала кращі показники для чоловічих голосів, ніж ResNet, адже виявилася стійкішою до їх варіацій, на відміну від іншої моделі. Модель ReDimNet виявилася найбільш ефективною з усіх: вона демонструвала високі показники незалежно від статі мовця. При аналізі компактності кластерів було встановлено, що модель ReDimNet робить ембединги несхожих голосів якомога більш далекими у просторі – вочевидь, ця стратегія спрацювала найкраще і для надто варіативного чоловічого мовлення, і для надто схожих жіночих голосів.

Також було встановлено, що всі моделі краще справляються із набором даних з меншою кількістю голосів, ніж із датасетом зі значною кількістю мовців. Коли

число різних голосів невелике, навіть значна кількість записів, зашумленість середовища та незбалансованість датасету не погіршують показників роботи системи. Всі моделі продемонстрували для цього датасету ідеальний результат, не зробивши жодної помилки. Натомість при значній кількості мовців зростає й кількість помилок через схожість різних голосів або варіативність в межах одного голосу. При роботі із датасетом зі значною кількістю мовців найкращий результат мала модель ReDimNet, трохи гірший – ResNet, а найнижчий – ECAPA-TDNN.

Слід зазначити, що різні фактори, проаналізовані в ході експерименту, значним чином взаємодіють між собою, підсилюючи чи послаблюючи вплив один одного. Наприклад, зашумленість записів тим сильніше погіршує результат, чим більше різних голосів у датасеті; при цьому вона майже не впливає на результат при невеликій кількості мовців. Читане мовлення може покращувати результати для чоловічих голосів, адже зменшує варіативність між записами одного голосу, яка є головною проблемою у їх розпізнаванні. Натомість для жіночих голосів читане мовлення навпаки може погіршити результат, зробивши схожі голоси ще більш схожими. Розгорнутий аналіз взаємодії цих факторів є одним із можливих напрямків для подальших досліджень.

Відмінності в ефективності моделей для різних датасетів дають підстави для формування рекомендацій із їх застосування залежно від типу вхідних даних. Модель ReDimNet продемонструвала найвищі показники метрик на всіх етапах експерименту й набагато краще справлялася із датасетами спонтанного мовлення та великою кількістю мовців – завданнями, що виявились найскладнішими. Вона підійде для будь-якої задачі, але особливо доречним є її використання у випадках, коли вхідні дані мають високий показник зашумленості та містять спонтанне мовлення значної кількості осіб. У задачах, що включають розпізнавання невеликого числа мовців чи роботу із читаним або добре підготовленим мовленням, достатнім є використання моделей ResNet та ECAPA-TDNN. Застосування ResNet

є більш ефективним, якщо датасет містить переважно жіноче мовлення, а ECAPA-TDNN – якщо переважають записи чоловічих голосів.

Отримані результати дозволили виявити закономірності впливу акустичних та кількісних характеристик мовленнєвих даних на точність роботи систем автоматичного розпізнавання мовця на матеріалі корпусів українських аудіозаписів. Можна зробити висновок, що врахування якісних та кількісних особливостей даних, а також особливостей роботи конкретної моделі може значно підвищити точність системи. Робота має як теоретичну, так і практичну цінність, зокрема для формування наборів вхідних даних при застосуванні систем автоматичного розпізнавання мовця в реальних умовах для діаризації, розпізнавання мовців у сфері безпеки чи аналізі великих неструктурованих наборів мовленнєвих даних різної якості.

Список використаних джерел

Abid, A., Abdalla, A., Abid, A., Khan, D., Alfozan, A. and Zou, J., 2019. Gradio: Hassle-free sharing and testing of ml models in the wild. [online] ICML Workshop on Human in the Loop Learning (HILL 2019). https://doi.org/10.48550/arXiv.1906.02569

Brydinskyi, V., Khoma, Y., Sabodashko, D., Podpora, M., Khoma, V., Konovalov, A. and Kostiak, M., 2024. Comparison of modern deep learning models for speaker verification. [online] Applied Sciences, 14(4), 1329. https://doi.org/10.3390/app14041329

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z. and Wei, F., 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. [online] IEEE Journal of Selected Topics in Signal Processing, 16(6), pp. 1505–1518. https://doi.org/10.48550/arXiv.2110.13900

Coria, J.M., Bredin, H., Ghannay, S. and Rosset, S., 2020. A comparison of metric learning loss functions for end-to-end speaker verification. [online] International Conference on Statistical Language and Speech Processing, pp. 137–148. Cham: Springer International Publishing. https://doi.org/10.48550/arXiv.2003.14021

Desplanques, B., Thienpondt, J. and Demuynck, K., 2020. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. [online] Available at: <https://arxiv.org/pdf/2005.07143> [Accessed 19 March 2025].

Fan, Y., Kang, J. W., Li, L. T., Li, K. C., Chen, H. L., Cheng, S. T. and Wang, D., 2020. Cn-celeb: a challenging chinese speaker recognition dataset. [online] International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7604–7608.

https://doi.org/10.48550/arXiv.1911.01799

Fortune Business Insights, 2024. Voice biometrics market size, share & industry analysis, by component, application, end-user, and regional forecast, 2024–2032. [online] Available at: <https://www.fortunebusinessinsights.com/industry-reports/voice-biometric-solutions-market-100509> [Accessed 25 March 2025].

Fujita, Y., Kanda, N., Horiguchi, S., Xue, Y., Nagamatsu, K. and Watanabe, S., 2019. End-to-end neural speaker diarization with self-attention. [online] Automatic Speech Recognition and Understanding Workshop (ASRU) 2019, pp. 296–303. https://doi.org/10.48550/arXiv.1909.06247

Ghaemmaghami, H., Dean, D., Sridharan, S., van Leeuwen, D., 2016. A study of speaker clustering for speaker attribution in large telephone conversation datasets. [online] Computer Speech & Language 40, pp. 23–45. https://doi.org/10.1016/j.csl.2016.03.005

Horiguchi, S., Fujita, Y., Watanabe, S., Xue, Y. and Nagamatsu, K., 2020. End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors. [online] INTERSPEECH 2020. https://doi.org/10.48550/arXiv.2005.09921

Hu, J., Shen, L. and Sun, G., 2018. Squeeze-and-Excitation Networks [online] Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132–7141. Available at: <https://openaccess.thecvf.com/content_cvpr_2018/papers/Hu_Squeeze-and-Excitation_Networks_CVPR_2018_paper.pdf> [Accessed 19 March 2025].

Kabir, M.M., Mridha, M.F., Shin, J., Jahan, I. and Ohi, A.Q., 2021. A Survey of Speaker Recognition: Fundamental Theories, Recognition Methods and Opportunities. [online] IEEE Access 9, pp. 79236–79263. 10.1109/ACCESS.2021.3084299

Kinoshita, K., Delcroix, M. and Tawara, N., 2021. Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds. [online] International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2021, pp. 7198–7202. https://doi.org/10.48550/arXiv.2010.13366

Koluguri, N.R., Park, T. and Ginsburg, B., 2022. Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context. [online] IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8102–8106. https://doi.org/10.48550/arXiv.2110.04410

Lam, P., Pham, L., Nguyen, T., Ngo, D., Pham, T., Nguyen, T. and Schindler, A., 2024. Towards Unsupervised Speaker Diarization System for Multilingual Telephone Calls Using Pre-trained Whisper Model and Mixture of Sparse Autoencoders. [online] International Symposium on Information and Communication Technology, pp. 39–53. Singapore: Springer Nature Singapore. Available at: <https://arxiv.org/pdf/2407.01963> [Accessed 20 March 2025].

Lin, Y., Cheng, M., Zhang, F., Gao, Y., Zhang, S. and Li, M., 2024. Voxblink2: A 100k+ speaker recognition corpus and the open-set speaker-identification benchmark. [online] InterSpeech2024.
https://doi.org/10.48550/arXiv.2407.11510

Mahmoudi, A. and Jemielniak, D., 2024. Proof of biased behavior of Normalized Mutual Information. [online] Scientific Reports, 14(1), 9021. https://doi.org/10.1038/s41598-024-59073-9

Malzer, C. and Baum, M., 2020. A Hybrid Approach To Hierarchical Density-based Cluster Selection. [online] 2020 IEEE international conference on multisensor fusion and integration for intelligent systems (MFI), pp. 223–228. https://doi.org/10.1109/MFI49285.2020.9235263

McCarthy, A.D., Chen, T., Rudinger, R. and Matula, D.W., 2019. Metrics matter in community detection. [online] International Conference on Complex Networks and Their Applications, pp. 164–175. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-36687-2_14

Mckinney, W., 2011. pandas: a Foundational Python Library for Data Analysis and Statistics. [online] Available at: <https://www.researchgate.net/publication/265194455_pandas_a_Foundational_Python_Library_for_Data_Analysis_and_Statistics> [Accessed 28 March 2025.

Nagrani, A., Chung, J.S. and Zisserman, A., 2017. Voxceleb: a large-scale speaker identification dataset. [online] https://doi.org/10.21437/Interspeech.2017-950

Nagrani, A., Chung J.S. and Zisserman, A., 2018. Voxceleb2: Deep speaker recognition. [online] https://doi.org/10.21437/Interspeech.2018-1929

Panayotov, V., Chen, G. and Khudanpur, S., 2015. Librispeech: An ASR corpus based on public domain audio books. [online] p. 1. Summary. Available at: <https://ieeexplore.ieee.org/document/7178964> [Accessed 28 March 2025].

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. and Duchesnay, É., 2011. Scikit-learn: Machine learning in Python. [online] The Journal of machine Learning research, 12, pp. 2825–2830. https://doi.org/10.48550/arXiv.1201.0490

Pessoa, T., Medeiros, R., Nepomuceno, T., Bian, G.B., Albuquerque, V.H.C. and Filho, P.P., 2018. Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications. [online] IEEE Access, pp. 1–8, 10.1109/ACCESS.2018.2874767

Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C. and Sutskever, I., 2023, Robust speech recognition via large-scale weak supervision. [online] International conference on machine learning, pp. 28492–28518. https://doi.org/10.48550/arXiv.2212.04356

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L. and Bengio, Y., 2021. SpeechBrain: A general-purpose speech toolkit. [online] Available at: <https://arxiv.org/pdf/2106.04624> [Accessed 25 March 2025].

Ren, Z., Chen, Z. and Xu, S., 2019. Triplet based embedding distance and similarity learning for text-independent speaker verification. [online] 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 558–562. https://doi.org/10.48550/arXiv.1908.02283

Santos, J.M. and Embrechts, M., 2009. On the use of the adjusted rand index as a metric for evaluating supervised classification. [online] International conference on artificial

neural networks, pp. 175–184. Berlin, Heidelberg: Springer Berlin Heidelberg. 10.1007/978-3-642-04277-5_18

Tranter, S.E. and Reynolds, D.A., 2006. An overview of automatic speaker diarization systems. [online] Transactions on audio, speech, and language processing, 14(5), pp. 1557–1565. https://doi.org/10.1109/TASL.2006.878256

Villalba, J., Chen, N., Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Borgstrom, J., García-Perera L.P., Richardson F., Dehak R., Torres-Carrasquillo P.A. and Dehak N., 2020. State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and Speakers in the Wild evaluations. [online] Computer Speech & Language, 60, p. 101026, 10.1016/j.csl.2019.101026

Yakovlev I. et al, 2024. Reshape dimensions network for speaker recognition [online] Intespeech 2024, pp. 1–5.
https://doi.org/10.21437/Interspeech.2024-2116

Yang, Y.Y., Hira, M., Ni, Z., Astafurov, A., Chen, C., Puhrsch, C. and Quenneville-Bélair, V., 2022. Torchaudio: Building blocks for audio and speech processing. [online] IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6982–6986. https://doi.org/10.48550/arXiv.2110.15018

Yin, R., Coria, J.M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W. and Gill, M.P., 2020. Pyannote.Audio: Neural Building Blocks for Speaker Diarization [online] pp. 7124–7128. Summary. International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2020, 10.1109/ICASSP40776.2020.9052974

Zeinali, H., Wang, S., Silnova, A., Matˇejka, P. and Plchot, O., 2019, BUT System Description to VoxCeleb Speaker Recognition Challenge 2019. [online] Available at: <https://arxiv.org/pdf/1910.12592> [Accessed 27 March 2025].

Борисов, Г. і Трапезон, К., 2024. Дослідження особливостей створення текстонезалежних голосових систем доступу із захистом від спуффінг-атак. Вісник КрНУ імені Михайла Остроградського. [онлайн] 1, с. 258–265. Режим доступу: <https://visnikkrnu.kdu.edu.ua/statti/2024_1_258.pdf> [Дата звернення: 26 Березня 2025].

Заєць, І.С., Бридінський, В.А., Сабодашко, Д.В., Хома, Ю.В., Руда, Х.С., і Швед, М.Є., 2024. Використання ембедінгів голосу в інтегрованих системах для діаризації мовців та виявлення зловмисників. Львів: НУ «ЛП» https://doi.org/10.23939/csn2024.01.054

Протас, О.М., 2021. Порівняльний аналіз якості методів кластеризації: задача кластеризації італійських вин. [онлайн] Суми: Сумський Державний Університет. Режим доступу: <https://essuir.sumdu.edu.ua/bitstream-download/123456789/86511/1/Protas_mag_rob.pdf> [Дата звернення: 26 Березня 2025].

Руда, Х.С., 2025. Дослідження масштабованості систем біометричної автентифікації на основі ембеддінгів голосу. Львів: НУ «ЛП» https://doi.org/10.33445/sds.2025.15.1.15

Руда, Х.С., Сабодашко, Д.В., Микитин, Г.В., Швед, М.Є., Бордуляк, С.М. і Коршун, Н.В., 2024. Порівняння методів цифрової обробки сигналів та моделей глибинного навчання у голосовій аутентифікації. Кібербезпека: освіта, наука,

техніка. [онлайн] 1 (25), с. 141–160. Режим доступу: <https://csecurity.kubg.edu.ua/index.php/journal/article/view/645/508> [Дата звернення: 26 Березня 2025].

Холєв, В. і Барковська, О., 2023. Порівняльний аналіз нейромережних моделей для розв'язання завдань розпізнавання спікера. [онлайн] Харків: ХНУРЕ, Сучасний стан наукових досліджень та технологій в промисловості, 2(24), с. 172–178. 10.30837/ITSSI.2023.24.172

Додаток

Посилання на відкритий репозиторій GitHub із повним кодом програми:

https://github.com/U-Karpcova/Voice-Recognition-System

У репозиторії містяться докладні інструкції щодо роботи зі створеною системою, частина використаних даних та посилання на інші відкриті датасети для відтворення експериментів, візуалізація результатів проведеного дослідження, а також файл із середовища Google Colab із програмним кодом, доступний для завантаження та запуску.