УДК 004.89
DOI https://doi.org/10.32782/IT/2022-2-1

Віталій БРИДІНСЬКИЙ
аспірант, Національний університет «Львівська політехніка», вул. Степана Бандери, 12, м. Львів, Україна, 79000, vbrydinskyi@gmail.com
***ORCID:** 0000-0001-8583-9785*
***Spopus-Author ID:** 57456880300*

Бібліографічний опис статті: Бридінський, В. (2022). Побудова системи ідентифікації мовців на основі бібліотеки аудіообробки PyAnnote. *Information Technology: Computer Science, Software Engineering and Cyber Security*, 2, 3–11, doi: https://doi.org/10.32782/IT/2022-2-1

ПОБУДОВА СИСТЕМИ ІДЕНТИФІКАЦІЇ МОВЦІВ НА ОСНОВІ БІБЛІОТЕКИ АУДІООБРОБКИ PYANNOTE

У галузі машинного навчання одним із основних напрямків є опрацювання та розпізнавання мовлення. Серед важливих завдань роботи з аудіоданими є діаризація. Діаризація визначає часові межі в аудіозаписі, що належать окремим мовцям, тобто образно кажучи, вирішує задачу «коли хто говорить?». Проте відомі комерційні та відкриті засоби діаризації використовують кластеризацію сегментів, але не відповідають на питання «хто саме зараз говорить?». Існують системи, які ідентифікують мовця, але такі системи розраховані на те, що у аудіозапису присутній є лише один мовець. Тому актуальним завданням є створення системи діаризації, яка уможливлює ідентифікацію багатьох мовців, які довільним чином змінюються у аудіозаписах. У даному дослідженні запропоновано дві архітектури систем ідентифікації мовців на основі діаризації, які працюють відповідно на засадах по-сегментного та по-кластерного аналізу. Для побудови системи використано бібліотеку PyAnnote, що є у відкритому доступі. Верифікація роботи системи ідентифікації мовців здійснювалася на відкритій базі аудіозаписів AMI Corpus, у якому зібрано 100 годин анотованих та транскрибованих аудіо- та відеоданих. Розглянуто різні метрики оцінювання точності діаризації та, враховуючи специфіку розробленої системи, обґрунтовано доцільність застосування такої оцінки як F-Міра ідентифікації. Описано методику проведення досліджень, яка передбачала проведення трьох експериментів. Перший експеримент націлений на дослідження архітектури системи ідентифікації, що базується на по-сегментному аналізі, а другий експеримент – на дослідження архітектури, що застосовує по-кластерний аналіз. Третій експеримент стосується визначення оптимальної тривалості навчальної вибірки для класифікаторів системи ідентифікації. Результати експериментів показали, що по-кластерний підхід показав кращі результати ідентифікації порівняно із по-сегментним підходом. Також встановлено, що оптимальною тривалістю вибірки аудіоданих для тренування класифікатора під кожного конкретного мовця становить 20 секунд.

***Ключові слова:** система діаризації, бібліотека PyAnnote, машинне навчання, кластеризація, класифікація, аналіз аудіо, ідентифікація мовця.*

Vitalii BRYDINSKYI
Postgraduate Student, Lviv Polytechnic National University, Stepana Bandery str., 12, Lviv, Ukraine, 79000, vbrydinskyi@gmail.com
***ORCID:** 0000-0001-8583-9785*
***Spopus-Author ID:** 57456880300*

To cite this article: Brydinskyi V. (2022). Pobudova systemy identyfikatsii movtsiv na osnovi biblioteky audioobrobky PyAnnote [Implementation of the speaker identification system based on the PyAnnote audio processing library]. *Information Technology: Computer Science, Software Engineering and Cyber Security*, 2, 3–11, doi: https://doi.org/10.32782/IT/2022-2-1

IMPLEMENTATION OF THE SPEAKER IDENTIFICATION SYSTEM BASED ON THE PYANNOTE AUDIO PROCESSING LIBRARY

In the field of machine learning, one of the main areas is speech processing and recognition. One of the important tasks of working with audio data is diarization. Diarization determines the time boundaries in the audio recording belonging to individual speakers, that is, figuratively speaking, solves the problem of «who speaks when?». However, known commercial and open source diarization tools use segment clustering, but do not answer the question «who exactly is speaking now?». There are systems that identify the speaker, but such systems are designed for the fact

that there is only one speaker in the audio recording. Therefore, a relevant task is to create a diarization system that allows the identification of many speakers that arbitrarily change in audio recordings. In this study, we propose two architectures of speaker identification systems based on diarization, which work respectively on the per segment basis and per cluster analysis. To implement the system, we used the PyAnnote library, which is open source. The evaluation of the speaker identification system was carried out on the open audio database AMI Corpus, which contains 100 hours of annotated and transcribed audio and video data. Various metrics for assessing the accuracy of diarization are considered and, taking into account the specifics of the developed system, the expediency of using such an assessment as the F-Measure of identification is substantiated. The methodology of research is described, which included three experiments. The first experiment is aimed at studying the architecture of the identification system based on per segment analysis, and the second experiment is aimed at studying the architecture that uses per cluster analysis. The third experiment concerns the determination of the optimal training sample duration for the classifiers of the identification system. The experimental results showed that the cluster-based approach showed better identification results compared to the segment-based approach. It was also found that the optimal duration of audio data sampling for training the classifier for each specific speaker is 20 seconds.

Key words: *diarization system, PyAnnote library, machine learning, clustering, audio analysis, speaker identification.*

Актуальність проблеми. Діаризація (англ. diarization) є важливим етапом у процесі розпізнавання усної мови, а її завданням є це поділ вхідного аудіозапису на однорідні сегменти відповідно до особи мовця (англ. speaker). Традиційно діарізація мовців поєднує в собі сегментацію аудіозапису та кластеризацію одержаних сегментів. Сегментація спрямована на пошук точок зміни мовця в аудіозаписі, натомість кластеризація – на групування сегментів мовлення на основі характеристик мовця.

Застосування діаризації завдяки структуризації аудіозапису за чергами мовців дає змогу підвищити читабельність автоматичної транскрипції мовлення на текст, а її застосування у системах біометричної ідентифікації за голосом дає змогу встановити справжню особу мовця (Juang B., 2005; Homayoon Beigi, 2011). Іншим прикладом застосування діаризації є її використання для розпізнавання мовлення у довгих аудіозаписах із декількома мовцями, де діаризація використовується для розділення довгого аудіозапису на коротші, які, в свою чергу, подаються на модель розпізнавання мовлення (Mao Huanru, 2020). Також схожа система може бути використана для систем перекладу, де аудіозапис розділяється на сегменти мовлення за допомогою діаризації, і система перекладу виконує переклад для кожного із сегментів, таким чином збільшується швидкодія та точність таких систем (Inaguma Hirofumi, 2021; Ueda Yushi, 2022).

Хоча розроблення методів діаризації почалося понад десятиліття тому (Anguera Xavier, 2012), на цей час продовжуються інтенсивні дослідження, метою яких є підвищення достовірності та обчислювальної ефективності алгоритмів діаризації. Варто зазначити, що для діаризації переважно застосовуються алгоритми машинного навчання без вчителя (unsupervised). Проте у деяких випадках для вирішення задачі діаризації виникає потреба в застосуванні машинного навчання з вчителем (supervised).

Основною відмінністю діаризації з вчителем є інший спосіб у індексації сегментів, який полягає не в групуванні схожих сегментів в окремі категорії (кластеризація), а співвіднесені семента до конкретного мовця на основі зразків його голосу (класифікація). Відтак останній етап перетворень для діаризації з вчителем здійснюватиметься за допомогою додаткового вихідного модуля – класифікатора мовців. Діаризація з учителем застосовується в задачах, коли потрібно ідентифікувати сегменти в аудіозаписі, які належать конкретному мовцю (тобто фактично здійснити пошук за голосом реплік розмови).

Натомість в діаризації без вчителя, сегменти розділяються між мовцями, але невідомо, яка із міток діаризації відноситься до конкретного мовця. Також може бути доцільно в умовах малих даних, коли треба додати підтримку нового мовця лише на основі одного короткого аудіозапису в кілька секунд. Наприклад, це може бути застосовано у відео конференції, для визначення мовця, котрий щойно доєднався до конференції, його голос буде записано та використано для підлаштування моделі діаризації, щоб у подальшому на цій конференції, його голос було визначено та ідентифіковано правильно.

Аналіз останніх досліджень і публікацій. Огляд літературних джерел дає підстави стверджувати, що системи діаризації традиційно будуються на основі підходу без учителя (Bredin Herve, 2022; Jin Qin, 2004; Tanveer Md, 2022; Le Lan Gaël, 2016; Dawalatabad Nauman, 2020). Є також кілька досліджень присвячених діаризації, що спирається на методи з учите-

лем (Zhang Aonan, 2019; Fini Enrico, 2020; Xie Weidi, 2019).

На цей час відомо декілька проектів, які вирішують завдання діаризації, зокрема це комерційні продукти IBM Watson, Google STT (Herchonvicz Andrey L, 2019) та бібліотеки з відкритим кодом pyAudioAnalysis (Giannakopoulos Theodoros, 2015), SpeechBrain (Ravanelli Mirco, 2021), PyAnnote (Bredin Herve, 2020). Комерційні продукти насамперед є рішеннями для розпізнавання мови, діаризація у цих рішеннях є додатковою функцією. Перевагою цих рішень є простота використання, швидкість розгортання та налаштування системи, хороша точність. Але є й недоліки: відсутність можливості налаштування параметрів системи або окремих її компонент, ціна, прив'язка до інфраструктури надавача послуг. Тому актуальним завданням є пошук рішення з відкритим кодом, з модульною структурою і можливістю налаштовування параметрів системи, та яке б не поступало за точністю комерційним засобам. На основі проведеного літературного огляду було встановлено, що найпопулярнішими системами, які задовольняють вищезгаданим вимогам, є фреймворки PyAnnote, SpeechBrain, pyAudioAnalysis.

PyAudioAnalysis – це бібліотека з відкритим кодом, призначена для опрацювання звуку, зокрема вилучення ознак (feature extraction) з аудіосигналу, з подальшою класифікацією та сегментацією аудіозаписів. Відтак цю бібліотеку можна використати для цілей діаризації з вчителем. Ця бібліотека має низку інструментів, призначених передовсім для аналізу даних та вилучення ознак з аудіозапису. У цієї бібліотеки також є інструменти для сегментування та класифікації аудіо, проте точність помітно поступається комерційним рішенням та іншим рішенням з відкритим кодом (Bredin Herve, 2020).

SpeechBrain – це набір програмних інструментів з відкритим кодом, які призначені для створення систем для аналізу аудіо, зокрема в таких задачах, як розпізнавання мови (automatic speech recognition), розпізнавання мовця (speaker identification), ідентифікація мови (language identification), тощо. Наразі ця бібліотека не містить у собі інструменту для сегментування аудіо, що ускладнює її застосування для задач діаризації (Ravanelli Mirco, 2021).

PyAnnote – бібліотека з відкритим кодом, розроблена безпосередньо для вирішення задачі діаризації (Bredin Herve, 2020; Bredin Herve, 2021; Mao Huanru 2020). Вона складається з навчених нейромережевих модулів, на основі спільного використання яких, можна імплементувати систему для діаризації без вчителя. Згадані модулі можна додаткового переналаштовувати, замінити на інші модулі, або ж додавати нові. Така відкрита модульна структура забезпечує високу гнучкість і можливості для покращення характеристик системи. Окрім цього ця бібліотека містить інструменти для опрацювання та маніпуляції сегментами аудіо (Bredin Herve, 2020).

Таким чином, для досліджень автори обрали PyAnnote, оскільки серед розглянутих рішень потенційно забезпечує найвищу точність (зіставиму з існуючими комерційними рішеннями), а модульна структура і доступність налаштування гіперпараметрів відкриває можливості адаптації під завдання діаризації з учителем.

Визначення мети дослідження. Метою роботи є удосконалення відомої системи діаризації без вчителя на основі PyAnnote через додавання нових модулів для розпізнавання мовців, що сприятиме не лише розширенню функціональності системи, але також покращенню результатів діаризації шляхом налаштування параметрів окремих модулів.

Виклад основного матеріалу дослідження. Опис стандартної системи діаризації на основі PyAnnote. У даному дослідженні за базову прийнято структуру системи діаризації без вчителя (з використанням кластеризації) на основі PyAnnote (Bredin Herve, 2020). До складу цієї системи діаризації входять такі модулі:

Виявлення голосової активності (voice activity detection) – модуль, який виконує функцію виявлення проміжків часу у аудіозаписі, де є присутній голос людини.

Виявлення зміни мовця (speaker change detection) – модуль, який виконує функцію виявлення моментів часу у аудіозаписі, де мовлення одного мовця завершилося і розпочалася мовлення іншого.

Виявлення накладеного мовлення (overlapped speech detection) – модуль, який виконує функцію виявлення проміжків часу у аудіозаписі, де два або більше мовців говорять одночасно. Такі сегменти зазвичай вилучають з аналізу, оскільки у таких випадках для виявлення конкретного мовця потрібно здійснювати додаткові перетворення пов'язані з розділення двох накладених в часі сигналів.

Векторне представлення мовця (Романюк Андрій, 2019) (speaker embedding) – модуль, який на основі фізіологічних параметрів голосу створює вектор із певною заданою розмірністю, числові значення якого, репрезентують характерні ознаки для кожного мовця. Таким чином відстань між векторами, які відповідають

фразам, що належать одній людині, буде менша ніж відстань до векторів, утворених із фраз сказаних іншими мовцями (Snyder David, 2018).

Кластеризація (clustering) – це модуль, який відповідає за групування сегментів, що відповідають мовцям, використовуючи їхнє векторне представлення (embeddings).

Ресегментація (resegmentation) – це модуль, який виконує завдання з уточнення/покращення меж та міток сегментації, передовсім для сегментів в межах яких спостерігається накладання голосів кількох мовців.

Таким чином, результатом роботи системи діаризації без вчителя є голосові сегменти згруповані в те чи інше число кластерів.

Пропонований підхід. Автори досліджували ідею адаптації PyAnnote під задачу діаризації з учителем шляхом заміни у базовій структурі системи (рис. 1) модуля кластеризації на модуль класифікації, а також можливість поєднання спільної роботи цих модулів. Структури цих двох варіантів систем діаризації з учителем наведено на рис. 1.

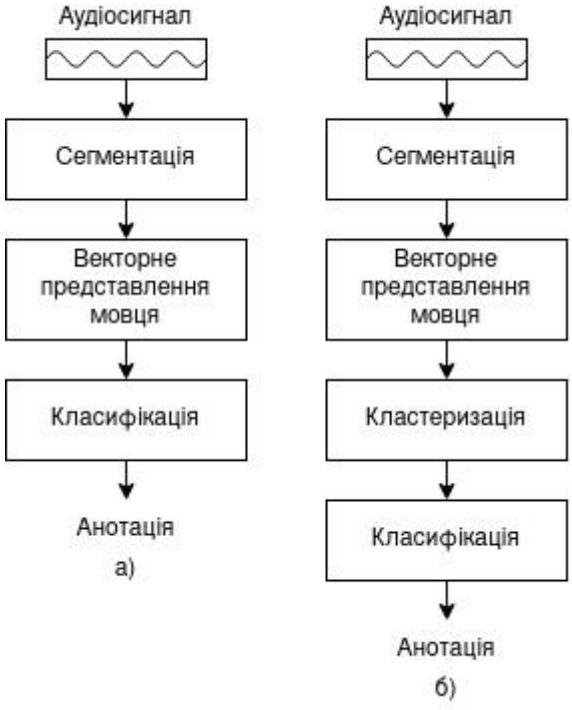

Рис. 1. Структури системи діаризації з учителем на основі PyAnnote: а – ідентифікація за окремими сегментами; б – ідентифікація на основі групи (кластеру) сегментів

Як зазначалося вище, на виході системи діаризації без учителя вхідний аудіозапис розділений на сегменти та згрупований на кластери, проте невідомо, до якого конкретного мовця належить конкретний кластер. Застосування класифікатора дасть змогу розпізнавати мовців, або іншими словами, присвоювати певні групи сегментів аудіо до конкретного мовця. Для цього, до базової системи діаризації, додається новий модуль натренований на розпізнавання мовця, а існуючі модулі системи та їхні параметри додатково підлаштовуються для кращого результату діаризації.

Досліджено чотири методи для розпізнавання мовця, зокрема такі поширені у машинному навчанні як випадковий ліс (random forest) і k-найближчих сусідів (k-nearest neighbors). Крім того, автори розробили два спеціалізовані методи класифікації сегментів мови:

відстань до векторного представлення цільового мовця (target speaker embedding) – сегменти аудіо, які відносяться до якогось конкретного мовця, вибираються залежно від відстані між векторним представленням цільового мовця та векторним представленням певного сегменту аудіо. Залежно від наперед заданого порогу, визначається, чи сегмент належить до мовця, чи ні.

Відстань до групи – перед тим як вибрати, які сегменти відносяться до певного мовця, спершу всі сегменти аудіо групуються, тобто по суті виконується діаризація без учителя. Після того, як ембедінги сегментів аудіо згруповано, для кожної групи визначається її центроїд (середнє значення з всіх ембедінгів у групі), а далі, обчислюється відстань від ембедінгу цільового мовця (target speaker) до кожного із центроїдів, тоді обирається найближчий центроїд, і сегменти аудіо його відповідної групи позначаються як сегменти, які належать до цільового мовця.

Принцип проведення експериментів. Кожен з експериментів досліджує вплив кожного із модулів або його параметрів на кінцеву точність діаризації. Тобто, для різних модулів системи відбувається незалежна оптимізація їх параметрів, або навіть заміна методу модуля як такого, причому параметри інших модулів зафіксовано значеннями за замовчуванням. Таким чином, можна оцінити вагу кожного з модулів та вплив його параметрів на точність системи діаризації. Експерименти проводяться як для діаризації без учителя, так і для діаризації з учителем.

Всі експерименти проводяться на одному і тому ж наборі даних – AMI Corpus. AMI Corpus, відкритий набір даних, який містить у собі аудіо- та відеозаписи робочих нарад та засідань, де мовлення є анотовано, і для кожного сегменту анотації присвоєно ідентифікаційний номер мовця, тому ці дані підходять для

діаризації як з вчителем, так і без вчителя. Аудіодані зберігаються у форматі WAV, а анотація – у форматі XML, який перетворюється в формат RTTM для роботи з бібліотекою pyAnnote. Формат RTTM (Rich Transcription Time Marked) – це текстовий формат, який містить інформацію про сегменти та ідентифікацію мовця для кожного з аудіофайлів. Розмір датасету становить 100 годин. Середня тривалість одного аудіо файлу – 2104,48 секунд. Середня тривалість одного анотованого сегменту – 4,1 секунд. Число унікальних мовців – 189 (серед них 65,1 % чоловіки). Середня тривалість аудіо на кожного мовця – 546,96 секунд.

Також для діаризації з учителем, експерименти проводяться для кожного мовця у кожному із аудіозаписів. Наприклад, аудіозапис з 4 мовцями буде аналізуватися почергово 4 рази і точність оцінюватиметься для кожного учасника окремо.

Метрики для оцінювання точності. Для коректності точність методів з учителем та без учителя обчислювалась на одних і тих самих даних. Було розглянуто метрики які традиційно використовуються в задачах діаризації (Bredin Herve, 2017): F-Міра виявлення (Detection F-Score), F-Міра сегментування (Segmentation F-Score), Похибка діаризації (Diarization Error Rate), F-Міра ідентифікації (Identification F-Score).

Основною метрикою обрано F-Міру ідентифікації, як таку, що відображає правильність присвоєння ідентифікаторів мовців для кожного із сегментів у аудіозаписі, на відміну від похибки діаризації, що використовує автоматичне зіставлення сегментів, які не прив'язані до якогось із мовців.

Опис виконаних експериментів та представлення їх результатів. У дослідженні було виконано 3 експерименти, які мають на меті дослідити архітектурні питання діаризації з учителем. Перший експеримент мав на меті дослідити посегментний класифікатор (рис. 2, а) Другий експеримент мав на меті дослідити класифікатор для кластеризованих сегментів (рис. 2, б). Третій експеримент мав на меті оцінити оптимальну навчальну вибірку для класифікаторів. На етапі попередніх досліджень встановлено можливість використання у експериментах меншого набору аудіоданих (тривалістю 10 годин), оскільки різниця у результатах моделювання у порівнянні із повним набором (100 год.) не перевищувала відсотка. Це дало змогу заощадити обчислювальні ресурси.

Експеримент 1: Дослідження впливу функцій відстані між векторними представленнями на точність ідентифікації мовця. Ціль цього експерименту полягає у визначенні, яка із функцій відстані дає найкращий результат за використання діаризації з учителем. Досліджувалися такі функції відстані: Евклідова відстань, відстань Манхетена (Manhattan), відстань Пірсона (Pearsons) та відстань Спірмена (Spearmans).

Функція відстані використовується для ідентифікації мовця, шляхом обчислення відстані між зразковим векторним представлення цього мовця та векторним представленням певного сегменту аудіозапису. Якщо ця відстань є меншою за попередньо встановлений поріг, тоді це означає, що голос на даному сегменті відповідає голосу цього мовця. Оскільки функцій відстані є кілька, є сенс оцінити, яку з них найкраще використати для діаризації з учителем. Архітектура діаризації з учителем у цьому експерименті відповідає ілюстрації на рис. 2, а.

У таблиці 1 наведено результати експерименту 1.

Таблиця 1

Результати дослідження функцій відстані на рівні сегментів

Функція відстані	F-Міра ідентифікації, %
Euclidean (поріг=7)	58.30 %
Manhattan (поріг=7)	57.95 %
Pearsons (поріг=0.5)	24.41 %
Spearmans (поріг=0.5)	23.64 %

Результати цього експерименту показали, що евклідова відстань та відстань Манхетена краще підходять для діаризації з учителем. Проте обчислення Евклідової відстані є простішим порівняно із обчисленням відстані Манхетена.

Експеримент 2: Дослідження методів групування сегментів (кластеризація). Ціль цього експерименту є дослідження методу вибору кластера, відповідає архітектурі діаризації з учителем, яку зображено на ілюстрації б) на рис. 2.

Цей метод передбачає використання ще одного модуля - кластеризації (clustering), який виконується після створення векторного представлення (embedding) для кожного із сегментів аудіозапису і перед класифікацією створених векторних представлень. Модуль кластеризації об'єднує векторні представлення в групи, кожна з яких належить до якогось мовця на аудіозаписі. На стадії ідентифікації голосу мовця, обчислюється відстань між векторним представленням цього мовця на певному сегменті та

центроїдом кожного із кластерів, і тоді сегмент з голосом цього мовця відповідає кластеру, до якого відстань є найменшою.

Отже, перший крок цього експерименту полягає у визначенні оптимального алгоритму кластеризації/групування. Було досліджено такі алгоритми кластеризації: Ієрархічна (Hierarchical), K-середніх (K-Means) та Спектральна (Spectral) кластеризація.

У таблиці 2 наведено результати дослідження алгоритмів кластеризації у методі вибору кластеру.

Таблиця 2

Результати дослідження алгоритмів кластеризації

Алгоритм кластеризації	F-Міра ідентифікації, %
Без кластеризації	58.30 %
Hierarchical (поріг=2)	62.35 %
K-Means	63.01 %
Spectral	47.90 %

З результатів цього експерименту можна дійти висновку, що для діаризації з учителем за методом вибору кластерів, ієрархічний агоритм та алгоритми кластеризації K-Means показують кращий результат ніж спектральний алгоритм кластеризації. Проте для кластеризації доцільно обрати K-Means алгоритм, оскільки його імплементація є значно простішою.

Другий крок цього експерименту полягає у визначенні, яка із функцій відстані дає найкращий результат діаризації з учителем, за методом вибору кластера. Цей крок аналогічний експерименту 1, але його було проведено на рівні кластерів, а не окремих сегментів. Тобто, у методі вибору кластера відстань обчислюється між всіма центроїдами кластерів, які відповідають кожному із голосів на аудіозапису, та векторним представленням певного сегменту аудіозапису.

У таблиці 3 наведені результати дослідження функцій відстані у методі вибору кластеру.

Таблиця 3

Результати дослідження функцій відстані на рівні кластерів

Функція відстані	F-Міра ідентифікації, %
Euclidean	63.01 %
Manhattan	62.75 %
Pearsons	51.43 %
Spearmans	53.03 %

Подібно як і у експерименті 1 Евклідова відстань забезпечує кращий результат діаризації і до того не є вимогливою до обчислювальних ресурсів.

Третій та останній крок цього експерименту полягає у перевірці потреби використання класифікатора на основі методу k-найближчих сусідів (k-nearest neighbours), замість функції відстані. Класифікатор k-найближчих сусідів заснований на такій ідеї: після отримання векторних представлень мовців на всіх сегментах аудіозапису та кластеризації отриманих векторних представлень, відбувається навчання класифікатора, де входом є векторне представлення, а виходом – ідентифікатор одного із кластерів. Тоді отриманий навчений класифікатор використовується для визначення відповідності голосу мовця на певному сегменті аудіозапису до певного кластера, який відповідає до голосу одного із мовців присутніх у цьому аудіозаписі.

У таблиці 4 наведено результати дослідження класифікатора K-найближчих сусідів у методі вибору кластеру.

Таблиця 4

Результати дослідження класифікатора k-NN на рівні кластерів

Алгоритм класифікатора	F-Міра ідентифікації, %
Euclidean (+K-Means clustering)	63.01 %
k-NN (+K-Means clustering)	60.74 %

Як видно із експерименту, поєднання класифікації за Евклідовою відстанню та кластеризації за методом K-Means clustering дає кращий результат.

Експеримент 3: Дослідження впливу тривалості зразкового аудіозапису мовця на точність ідентифікації. Ціль цього експерименту полягає у виявленні оптимальної тривалості аудіозапису для додавання в систему діаризації з учителем нового користувача. Фактично на основі цього аудіозапису формується зразкове векторне представлення для конкретного мовця, з яким буде проводитися порівняння вкладень з інших сегментів для задачі ідентифікації. У експерименті досліджувались аудіозаписи тривалістю 10, 20, 30 та 60 секунд. Такі значення обрано з міркувань зручності інтерфейсу для кінцевого користувача (тривалість фраз зручна для запису зразкових фраз і калібрування системи).

У таблиці 5 наведені результати експерименту.

Таблиця 5

Результати дослідження довжини зразкового аудіозапису мовця

Довжина аудіо зразкового мовця, с	F-Міра ідентифікації, %
10 (+ K-Means)	63.01 %
20 (+ K-Means)	64.81 %
30 (+ K-Means)	64.10 %
60 (+ K-Means)	63.49 %

Хоча всі тривалості показали схожі результати, перевагу слід віддати тривалості 20 сек. З результатів цього експерименту можна дійти висновку, що найбільш оптимальною довжиною аудіо зразкового мовця становить 20 секунд. Хоча загалом всі значення в діапазоні 20–60 сек показали прийнятні результати.

Висновки і перспективи подальших досліджень. PyAnnote – одне з найбільш поширених і функціональних рішень в галузі опрацювання аудіосигналів, зокрема діаризації. До основних її переваг над відомими альтерантивами можна віднести: імплементація на популярній мові Python, відкритий вихідний код під ліцензією MIT, модульна архітектура з можливістю додавання до системи нових модулів та налаштовування параметрів наявних компонент. Крім того, PyAnnote за точністю практично не поступається існуючим комерційним рішенням, таким як IBM Watson чи Google Speech.

Існуючі приклади імплементації бібліотеки PyAnnote заточені під задачу діаризації без учителя (unsupervised). Разом з тим існує можливість розширити функціональність системи і адаптувати PyAnnote до задач діаризації з учителем (supervised), що відкриває можливості ідентифікації мовців. Побудова такої системи і стала предметом дослідження, результати якого представлено в даній статті.

Пропонований в роботі підхід передбачає дві архітектурні реалізації системи ідентифікації мовців – із заміною модуля кластеризації на модуль класифікатора та з поєднанням цих модулів. Фактично в першому випадку здійснюється ідентифікація мовця на рівні окремих сегментів аудіо, а в другому – ідентифікація проводиться на основі груп схожих між собою сегментів об'єднаних в кластери.

З метою дослідження запропонованих архітектур системи ідентифікації мовців було проведено три експерименти, а для оцінювання точності вибрано F-Міру ідентифікації. Перший експеримент мав на меті дослідити точність діаризиції за різних варіантів реалізації посегментного класифікатора, а в рамках другого експерименту виконано дослідження оптимального поєднання класифікатора і кластеризатора. Третій експеримент націлений на вибір оптимальної тривалості зразкової вибірки з голосами мовців, необхідної для навчання моделей класифікаторів. Всі дослідження проводилися на відкритому наборі даних – AMI Corpus, який містить розмічені аудіо- та відеозаписи робочих нарад і засідань.

За результатами експериментів встановлено, що найвищу точність для по-сегментної класифікації забезпечує алгоритм на основі Евклідової відстані за порогового значення 7. У випадку архітектури з класифікатором згрупованих сегментів оптимальним алгоритмом кластеризації є K-Means, а оптимальним алгоритмом класифікатора – алгоритм на основі Евклідової відстані. Також встановлено, що оптимальна тривалість зразкового аудіозапису, за яким система зіставляє сегменти для ідентифікації мовця, становить 20 сек.

В подальшому виглядає доцільним провести додаткові дослідження з метою оптимізації інших модулів каналу перетворення PyAnnote (виявлення голосової активності, зміни мовця чи накладання мовлення, обчислення векторного представлення) під задачу діаризації з учителем.

ЛІТЕРАТУРА:

1. Juang B., Rabiner Lawrence. Automatic Speech Recognition – A Brief History of the Technology Development. 2005.

2. Homayoon Beigi. Fundamentals of Speaker Recognition. *New York: Springer*. 2011.

3. Anguera Xavier, Bozonnet Simon, Evans Nicholas, та ін. Speaker Diarization: A Review of Recent Research. 2012. IEEE Transactions on Audio, Speech & Language Processing. DOI: 10.1109/TASL.2011.2125954.

4. Li Runxin, Schultz Tanja, Jin Qin. Improving speaker segmentation via speaker identification and text segmentation. 2009.

5. Bredin Herve, Yin Ruiqing, Coria Juan, Gelly Gregory, та ін. Pyannote.Audio: Neural Building Blocks for Speaker Diarization. 2020. DOI: 10.1109/ICASSP40776.2020.9052974.

6. Jin Qin, Laskowski Kornel, Schultz Tanja, Waibel Alex. Speaker segmentation and clustering in meetings. 2004.

7. Tanveer Md, Casabuena Diego, Karlgren Jussi, Jones Rosie. Unsupervised Speaker Diarization that is Agnostic to Language, Overlap-Aware, and Tuning Free. 2022. DOI: 10.21437/Interspeech.2022-10605.

8. Le Lan Gaël, Meignier Sylvain, Charlet Delphine, Deléglise Paul. Speaker diarization with unsupervised training framework. 2016. DOI: 10.1109/ICASSP.2016.7472741.

9. Dawalatabad Nauman, Madikeri Srikanth, Sekhar Chandra, Murthy Hema. Novel Architectures for Unsupervised Information Bottleneck based Speaker Diarization of Meetings. 2020.

10. Zhang Aonan, Wang Quan, Zhu Zhenyao, Paisley John, Wang Chong. Fully Supervised Speaker Diarization. 2019. DOI: 10.1109/ICASSP.2019.8683892.

11. Fini Enrico, Brutti Alessio. Supervised Online Diarization with Sample Mean Loss for Multi-Domain Data. 2020. DOI: 10.1109/ICASSP40776.2020.9053477.

12. Xie Weidi, Nagrani Arsha, Chung Joon Son, Zisserman Andrew. Utterance-level Aggregation for Speaker Recognition in the Wild. 2019. DOI: 10.1109/ICASSP.2019.8683120.

13. Herchonvicz Andrey L., Franco Cristiano R., Jasinski Marcio G.. A comparison of cloud-based speech recognition engines. 2019. DOI: 10.14210/cotb.v0n0.p366-375.

14. Ravanelli Mirco, Parcollet Titouan, Plantinga Peter, Rouhe Aku, та ін. SpeechBrain: A General-Purpose Speech Toolkit. 2021.

15. Giannakopoulos Theodoros. pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. 2015. DOI: 10.1371/journal.pone.0144610.

16. Bredin Hervé, Laurent Antoine. End-to-end speaker segmentation for overlap-aware resegmentation. 2021.

17. Wang Keke, Mao Xudong, Wu Hao, Ding Chen, та ін. The ByteDance Speaker Diarization System for the VoxCeleb Speaker Recognition Challenge 2021. 2021.

18. Mao Huanru, McAuley Julian, Cottrell Garrison. Speech Recognition and Multi-Speaker Diarization of Long Conversations. 2020. DOI: 10.21437/Interspeech.2020-3039.

19. Inaguma Hirofumi, Yan Brian, Dalmia Siddharth, Guo Pengcheng, та ін. ESPnet-ST IWSLT 2021 Offline Speech Translation System. 2021.

20. Ueda Yushi, Maiti Soumi, Watanabe Shinji, Zhang Chunlei, та ін. EEND-SS: Joint End-to-End Neural Speaker Diarization and Speech Separation for Flexible Number of Speakers. 2022.

21. Bredin Hervé. pyannote.metrics: A Toolkit for Reproducible Evaluation, Diagnostic, and Error Analysis of Speaker Diarization Systems. 2017. DOI: 10.21437/Interspeech.2017-411.

22. Романюк Андрій. Векторні представлення слів для української мови. *Науковий* журнал «Україна Модерна». 2019. №27. DOI: 10.30970/uam.2019.27.1062

23. Snyder David, Garcia-Romero Daniel, Sell Gregory, Povey Daniel, Khudanpur Sanjeev. X-Vectors: Robust DNN Embeddings for Speaker Recognition. 2018. DOI: 10.1109/ICASSP.2018.8461375.

REFERENCES:

1. Juang, B. & Rabiner, Lawrence. (2005). Automatic Speech Recognition - A Brief History of the Technology Development.

2. Homayoon, Beigi. (2011) Fundamentals of Speaker Recognition. *New York: Springer*.

3. Anguera, Xavier & Bozonnet, Simon & Evans, Nicholas & Fredouille, Corinne & Friedland, Gerald & Vinyals, Oriol. (2012) Speaker Diarization: A Review of Recent Research. IEEE Transactions on Audio, Speech & Language Processing. DOI: 10.1109/TASL.2011.2125954.

4. Li, Runxin & Schultz, Tanja & Jin, Qin. (2009). Improving speaker segmentation via speaker identification and text segmentation.

5. Bredin, Herve & Yin, Ruiqing & Coria, Juan & Gelly, Gregory & Korshunov, Pavel & Lavechin, Marvin & Fustes, Diego & Titeux, Hadrien & Bouaziz, Wassim & Gill, Marie-Philippe. (2020). Pyannote.Audio: Neural Building Blocks for Speaker Diarization. DOI: 10.1109/ICASSP40776.2020.9052974.

6. Jin, Qin & Laskowski, Kornel & Schultz, Tanja & Waibel, Alex. (2004). Speaker segmentation and clustering in meetings.

7. Tanveer, Md & Casabuena, Diego & Karlgren, Jussi & Jones, Rosie. (2022). Unsupervised Speaker Diarization that is Agnostic to Language, Overlap-Aware, and Tuning Free. DOI: 10.21437/Interspeech.2022-10605.

8. Le, Lan Gaël & Meignier, Sylvain & Charlet, Delphine & Deléglise, Paul. (2016). Speaker diarization with unsupervised training framework. DOI: 10.1109/ICASSP.2016.7472741.

9. Dawalatabad, Nauman & Madikeri, Srikanth & Sekhar, Chandra & Murthy, Hema. (2020). Novel Architectures for Unsupervised Information Bottleneck based Speaker Diarization of Meetings.

10. Zhang, Aonan & Wang, Quan & Zhu, Zhenyao & Paisley, John & Wang, Chong. (2019). Fully Supervised Speaker Diarization. DOI: 10.1109/ICASSP.2019.8683892.

11. Fini, Enrico & Brutti, Alessio. (2020). Supervised Online Diarization with Sample Mean Loss for Multi-Domain Data. DOI: 10.1109/ICASSP40776.2020.9053477.

12. Xie, Weidi & Nagrani, Arsha & Chung, Joon, Son & Zisserman, Andrew. (2019). Utterance-level Aggregation for Speaker Recognition in the Wild. DOI: 10.1109/ICASSP.2019.8683120.

13. Herchonvicz, Andrey L. & Franco, Cristiano R. & Jasinski, Marcio G.. (2019). A comparison of cloud-based speech recognition engines. DOI: 10.14210/cotb.v0n0.p366-375.

14. Ravanelli, Mirco & Parcollet, Titouan & Plantinga, Peter & Rouhe, Aku & Cornell, Samuele & Lugosch, Loren & Subakan, Cem & Dawalatabad, Nauman & Heba, Abdelwahab & Zhong, Jianyuan & Chou, Ju-Chieh & Yeh, Sung-Lin & Fu, Szu-Wei & Liao, Chien-Feng & Rastorgueva, Elena & Grondin, François & Aris, William & Na, Hwidong & Gao, Yan & Bengio, Y. (2021). SpeechBrain: A General-Purpose Speech Toolkit.

15. Giannakopoulos, Theodoros. (2015). pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. DOI: 10.1371/journal.pone.0144610.

16. Bredin, Hervé & Laurent, Antoine. (2021). End-to-end speaker segmentation for overlap-aware resegmentation.

17. Wang, Keke & Mao, Xudong & Wu, Hao & Ding, Chen & Shang, Chuxiang & Xia, Rui & Wang, Yuxuan. (2021). The ByteDance Speaker Diarization System for the VoxCeleb Speaker Recognition Challenge 2021.

18. Mao, Huanru & McAuley, Julian & Cottrell, Garrison. (2020). Speech Recognition and Multi-Speaker Diarization of Long Conversations. DOI: 10.21437/Interspeech.2020-3039.

19. Inaguma, Hirofumi & Yan, Brian & Dalmia, Siddharth & Guo, Pengcheng & Shi, Jiatong & Duh, Kevin & Watanabe, Shinji. (2021). ESPnet-ST IWSLT 2021 Offline Speech Translation System.

20. Ueda, Yushi & Maiti, Soumi & Watanabe, Shinji & Zhang, Chunlei & Yu, Meng & Zhang, Shi-Xiong & Xu, Yong. (2022). EEND-SS: Joint End-to-End Neural Speaker Diarization and Speech Separation for Flexible Number of Speakers.

21. Bredin, Hervé. (2017). pyannote.metrics: A Toolkit for Reproducible Evaluation, Diagnostic, and Error Analysis of Speaker Diarization Systems. DOI: 10.21437/Interspeech.2017-411.

22. Romanyuk, Andriy. (2019). Vektorni predstavlennia sliv dlia ukrainskoi movy [Vector representation of words for Ukrainian language]. *Naukovyi zhurnal «Ukraina Moderna»*. 27 [in Ukrainian]. DOI: 10.30970/uam.2019.27.1062

23. Snyder, David & Garcia-Romero, Daniel & Sell, Gregory & Povey, Daniel & Khudanpur, Sanjeev. (2018). X-Vectors: Robust DNN Embeddings for Speaker Recognition. DOI: 10.1109/ICASSP.2018.8461375.