Vol. 6, No. 1, 2024

DOI: https://doi.org/10.23939/csn2024.01.054

УДК 004.93

ВИКОРИСТАННЯ ЕМБЕДІНГІВ ГОЛОСУ В ІНТЕГРОВАНИХ СИСТЕМАХ ДЛЯ ДІАРИЗАЦІЇ МОВЦІВ ТА ВИЯВЛЕННЯ ЗЛОВМИСНИКІВ

І. С. Заєць[2], В. А. Бридінський[1], Д. В. Сабодашко[2], Ю. В. Хома[1], Х. С. Руда[2], М. Є. Швед[2]

Національний університет "Львівська політехніка",
кафедра інформаційно-вимірювальних технологій[1]
кафедра захисту інформації[2]
E-mail: illia.zaiets.mkbas.2022@lpnu.ua, vitalii.a.brydinskyi@lpnu.ua, dmytro.v.sabodashko@lpnu.ua, yurii.v.khoma@lpnu.ua, khrystyna.s.ruda@lpnu.ua, mariia.y.shved@lpnu.ua

У цій роботі досліджується використання систем діаризації, які застосовують передові алгоритми машинного навчання для точного виявлення та розділення різних спікерів в аудіозаписах для реалізації системи виявлення зловмисників. Порівнюються декілька передових моделей діаризації, зокрема NeMo від Nvidia, Pyannote та SpeechBrain. Ефективність цих моделей оцінюється за допомогою типових метрик, що використовуються для систем діаризації, таких як коефіцієнт помилки діаризації (DER) та коефіцієнт помилки Жакара (JER). Система діаризації була протестована в різних аудіоумовах. Це, зокрема, зашумлене середовище, чисте середовище, мала кількість спікерів та велика кількість спікерів. Встановлено, що Pyannote проявляє найкращу продуктивність з погляду точності діаризації, тому саме її було обрано для реалізації системи виявлення зловмисників. Цю систему додатково оцінено на власному наборі даних, що ґрунтується на українських подкастах, і було встановлено, що система працює з показником чутливості 100 % та точністю 93,75 %. Це означає, що система не пропустила жодного злочинця з набору даних, але іноді неправильно ідентифікувала особу, яка не є злочинцем. Ця система виявилася ефективною та гнучкою для завдань виявлення зловмисників в аудіофайлах із різним розміром та кількістю присутніх спікерів.

Ключові слова: глибинне навчання, діаризація, ембедінги голосу, розпізнавання мовців, кібербезпека.

Вступ

Останнім часом цифрові технології швидко трансформують сучасний світ, а великі обсяги інформації, що надходять до нас щодня, створюють серйозні виклики у багатьох галузях. Це особливо актуально у сфері кібербезпеки та оброблення великих обсягів аудіоданих, де точний та вчасний аналіз є важливим фактором успіху. У зв'язку з цим виникає потреба у розробленні ефективних систем для обробки аудіоданих, зокрема систем діаризації мовлення, що можуть виявляти та відокремлювати голоси у складних аудіозаписах. Такі системи стають основним інструментом для виявлення зловмисників за їхнім голосом та ідентифікації осіб у різних сценаріях.

У цій статті пропонуємо розглянути розробку системи діаризації мовлення, що ґрунтується на сучасних методах машинного навчання і використовує набір даних VoxConverse для забезпечення

ефективності та точності алгоритмів. VoxConverse містить різноманітні аудіозаписи, від окремих промов до складних обговорень з накладенням голосів, що робить його ідеальним середовищем для тестування систем діаризації в різних умовах [1].

Особливу увагу присвячено аналізу того, як системи впораються з найпоширенішими проблемами у процесі роботи з аудіо, такими як шум, накладення голосів і різний рівень гучності мовців.

Розроблено методологію тестування та аналізу даних для порівняння бібліотек діаризації. Це дало можливість не лише здійснити оцінку точності процесу розпізнавання, а й окреслити сильні та слабкі сторони кожної системи, а також виявити способи їх ефективного використання.

Для оцінки бібліотек діаризації використано показники, що допомагають об'єктивно оцінити точність і надійність кожної бібліотеки: коефіцієнт помилки діаризації (DER) і коефіцієнт помилки Жакара (JER).

Результати оцінки бібліотек діаризації використовувалися для вибору найкращої для побудови системи, здатної точно та ефективно ідентифікувати й розділяти дикторів в аудіозаписах для виявлення зловмисників. Врешті, була обрана бібліотека Pyannote як основний елемент системи, модифікації якої визначаються багаторівневою системою комплексної безпеки інформаційних технологій [2, 3].

1. Огляд літературних джерел

Технологія діаризації дає змогу виявляти та розрізняти окремих мовців в аудіозаписах. Вона ґрунтується на складному аналізі голосових даних, використовуючи алгоритми машинного та глибокого навчання для розпізнавання голосових характеристик кожного мовця з метою визначення осіб, які беруть участь у розмові. Це вміщає аналіз тембру голосу, швидкості мовлення, акцентів та інших унікальних особливостей, що відрізняють одного мовця від іншого. Основне завдання полягає в тому, щоб розділити аудіопотік на окремі сегменти, щоб кожен сегмент відповідав моменту, коли говорить одна людина або відбувається зміна мовців.

Цей процес прагне відповісти на запитання: "Хто говорить і коли?" впродовж усього [4]. Завдяки цьому аналіз аудіоматеріалів значно спрощується, особливо в ситуаціях, коли в розмові бере участь багато учасників, і їхні голоси часто перекриваються або змінюють один одного. Отже, діаризація стає незамінним інструментом для розуміння та аналізу складних аудіозаписів, особливо в контексті кібербезпеки та в інших сферах, де точність ідентифікації мовців є критичною [5].

Головними цілями діаризації є:

- ідентифікація різних спікерів в аудіофайлі. Це важливо для розслідувань у сфері кібербезпеки, де потрібно розуміти, хто саме брав участь у розмові;
- аналіз комунікацій, таких як перехоплені телефонні дзвінки або нотатки з нарад, може допомогти виявити підозрілу або злочинну діяльність;
- виявлення спроб шахрайства у телефонних дзвінках, наприклад, виявленням непослідовностей у голосах або спроб маніпуляції;
- автоматизація процесу розподілу та аналізу аудіофайлів, що спрощує роботу аналітиків;
- захист конфіденційності інформації за допомогою "моніторингу" аудіокомунікацій у великих організаціях, щоб гарантувати, що конфіденційна інформація не розголошується.

Загалом аудіодіаризація відіграє важливу роль у кібербезпеці, допомагаючи виявляти та запобігати шахрайству, злочинам та іншим кіберзагрозам.

Такий підхід важливий для захисту інформації та виявлення потенційних загроз, що особливо актуально в контексті зростання кількості кібератак та шахрайської діяльності.

Впровадження діаризації в межах кібербезпеки має свої виклики, і одним з них є проблема наявності фонового шуму в аудіозаписах, що може істотно ускладнити процес ідентифікації мовця. Для розв'язання цієї проблеми застосовуються різні методи фільтрації та очищення аудіосигналу. Іншим важливим аспектом є мінливість голосових характеристик, таких як акценти, інтонації та

швидкість мовлення. Це потребує від систем діаризації високої гнучкості та здатності адаптуватися до різних умов.

Діаризація передбачає декілька критичних етапів для точного розпізнавання та розділення мовців в аудіозаписах. Ці етапи містять виявлення голосової активності, виявлення перекритої мови, виявлення зміни мовця, сегментацію, виділення ембедінгів спікерів, кластеризацію та нейронний діаризатор.

Процес діаризації зображено на рис. 1.

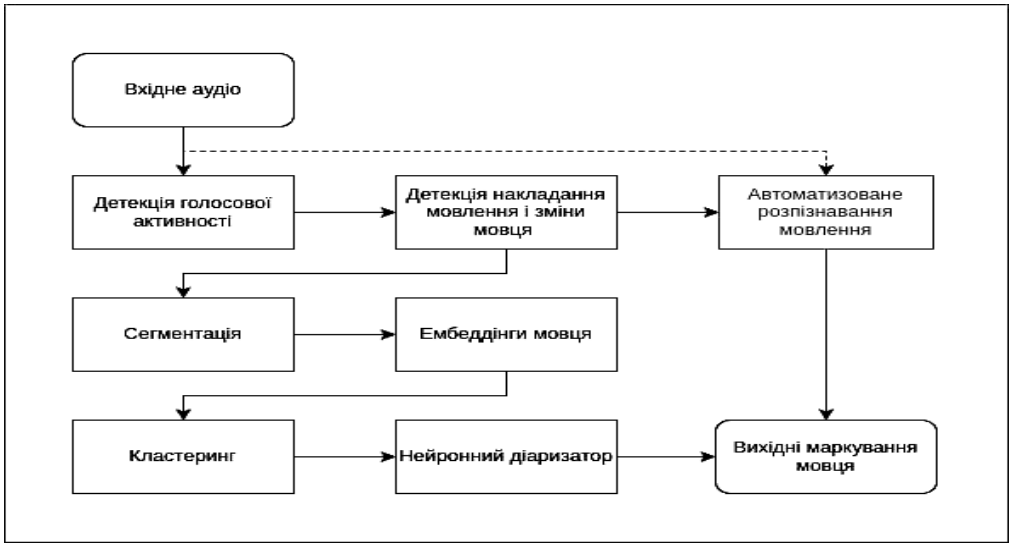

Рис. 1. Структура діаризації з основними кроками

Автоматизоване розпізнавання мовлення (Automatic Speech Recognition, ASR) – рис. 1 демонструє, що за потреби процес ASR можна використовувати паралельно з діаризацією [6].

Детекція голосової активності (Voice Activity Detection) – перший етап – виявлення голосової активності, на якому система визначає, чи є голосовий сигнал у певному аудіосегменті. Це дає можливість користувачам фільтрувати тихі ділянки або шум, зосереджуючись лише на сегментах з голосовою активністю, наприклад, мовою [7].

Детекція накладання мовлення і зміни мовця (Speaker Change Detection & Overlapped Speech Detection) – процес охоплює виявлення зміни диктора та моментів, коли одночасно говорить декілька осіб. Система аналізує аудіопотік, щоб визначити моменти, коли один диктор закінчує говорити, а інший починає. Це допомагає розділити аудіо на сегменти, кожен з яких відображає мову певного диктора. На цьому етапі процес діаризації стикається з низкою проблем. По-перше, якість запису звуку може значно відрізнятися, а шум, відлуння та інші звукові перешкоди можуть ускладнити ідентифікацію дикторів. По-друге, важливим аспектом забезпечення точної діаризації є врахування різноманітних мовних особливостей, а також акценти, діалекти та інтонації. По-третє, накладення голосів, коли одночасно говорить декілька людей, створює проблему для точної сегментації та ідентифікації дикторів [8].

Сегментація. На цьому етапі аудіозапис розділяється на менші фрагменти для детального аналізу. Кожен фрагмент перевіряється на наявність унікальних характеристик мовця [9].

Ембедінги мовця – один з найважливіших етапів. Система ідентифікує індивідуальні голосові атрибути, такі як тембр, темп та інтонація, що дає змогу створити унікальний “відбиток” кожного мовця [10].

Кластеринг. На цьому етапі відбувається кластеризація, де фрагменти з подібними характеристиками об’єднуються разом. Це дає можливість системі розпізнавати та групувати фрагменти, що належать одному й тому самому мовцеві.

Нейронний діаризатор. Останній етап передбачає використання нейронних діаризаторів – систем, що ґрунтуються на глибокому навчанні і можуть автоматично ідентифікувати різних мовців у складних аудіозаписах. Вони використовують потужні нейронні мережі для аналізу аудіосигналів, здатні вловлювати тонкі відмінності між різними голосами та ефективно кластеризувати аудіо-фрагменти за принципом належності певному мовцю.

Завдяки нейронним мережам процес діаризації стає більш точним та гнучким. Нейронні діаризатори здатні ефективно обробляти великі обсяги аудіоданих, а також краще справляються з такими проблемами, як накладення голосів або зміна умов запису [11].

Основними аспектами діаризації є використання спеціальних алгоритмів машинного навчання. Наприклад, нейронні мережі ефективно працюють із завданням класифікації аудіофрагментів за мовцями. Кластеризація допомагає групувати схожі мовні характеристики, що полегшує ідентифікацію окремих мовців. Розпізнавання мовця є важливим для визначення того, хто говорить у конкретний момент запису.

Опрацювання даних також є невід'ємною частиною цього процесу. Для точності алгоритмів критично важливим є виділення з аудіо ознак, таких як тон, швидкість мовлення та інші характеристики. Нормалізація даних забезпечує однорідність вхідних даних, що допомагає підвищити точність моделей машинного навчання.

Під час оптимізації та налаштування моделей важливо обрати правильні гіперпараметри для максимізації ефективності діаризації. Стратегії оптимізації містять вибір архітектури моделі, налаштування швидкості навчання та інших параметрів, які можуть впливати на результат.

Використані у роботі інструменти PyAnnote [12], NVIDIA NeMo [13] та SpeechBrain [14] мають широкий спектр функціональних можливостей для глибокого аналізу мовлення, ідентифікації дикторів та діаризації. PyAnnote чудово підходить для автоматичної анотації аудіо та ідентифікації дикторів. NVIDIA NeMo пропонує потужні інструменти для роботи з нейронними мережами, що ідеально підходить для складних задач діаризації. SpeechBrain, завдяки своїй гнучкості та відкритості, є ще одним чудовим інструментом для обробки мовлення та діаризації.

Загалом використання алгоритмів машинного навчання в задачах діаризації аудіо є важливим кроком на шляху розвитку технологій обробки мовлення, що пропонує більш точні та ефективні рішення.

2. Постановка завдання

Основне завдання цього дослідження полягає у вивченні можливостей використання систем діаризації, що ґрунтуються на передових алгоритмах машинного навчання, для розділення та ідентифікації спікерів у аудіозаписах. Для цього проводиться порівняння ефективності кількох передових моделей діаризації, таких як NeMo від Nvidia, Pyannote та SpeechBrain, за допомогою стандартних метрик, включно з коефіцієнтом помилки діаризації (DER) та коефіцієнтом помилки Жакара (JER). У дослідження також входить тестування системи діаризації в різних аудіоумовах, зокрема зашумлене та чисте середовища, з різною кількістю спікерів. Після вибору найефективнішої моделі діаризації вона застосовується для реалізації системи виявлення зловмисників. Завершальним етапом є оцінка ефективності цієї системи на власному наборі даних, що ґрунтується на українських подкастах.

3. Мета дослідження

Основною метою цієї статті є реалізація системи, яка зможе точно розпізнавати та розділяти голоси людей в аудіозаписах, а також виявляти несанкціонованих користувачів. Це має велике значення не лише для інформаційної безпеки, а й для інших сфер, де важливий точний аналіз мовлення. Для досягнення цієї мети ми проаналізували можливості бібліотек, таких як Pyannote, NVIDIA NeMo та SpeechBrain, і на основі цього аналізу побудували систему діаризації, здатну виявляти сторонніх осіб.

4. Огляд моделей

Pyannote. Головною перевагою Pyannote є його гнучкість та висока точність, що забезпечується використанням алгоритмів глибокого навчання. Він застосовує нейронні мережі для аналізу аудіозаписів, виявлення особливостей голосу (embeddings) для ідентифікації різних мовців, їх розділення та класифікації. Цей підхід дає можливість Pyannote ефективно розбивати аудіозаписи на сегменти, кожен з яких відповідає конкретному спікеру, навіть за складних умов, коли голоси перекриваються або наявний фоновий шум. Крім діаризації, Pyannote також пропонує інструменти для інших завдань обробки аудіо, таких як визначення активності голосу, розпізнавання віку та статі. Це робить його багатофункціональним рішенням для аналізу аудіоданих.

Іншою важливою особливістю Pyannote є його відкрита спільнота розробників. Розробники та дослідники можуть пропонувати власні покращення та адаптації, що сприяє постійному оновленню та вдосконаленню інструменту. Це також означає, що бібліотека постійно адаптується до нових проблем та технологічних проривів в обробці аудіоданих.

Загалом Pyannote є інноваційним рішенням для діаризації аудіозаписів, що постійно еволюціонує та застосовується у різних галузях, зокрема дослідженні та комерційному використанні. Ця бібліотека відзначається високою ефективністю у виконанні складних завдань обробки аудіо, що призводить до отримання надійних та точних результатів.

NVIDIA NeMo. Ця бібліотека, розроблена компанією NVIDIA, спеціалізується на використанні глибокого навчання для різноманітних завдань обробки мовлення, включно з діаризацією аудіозаписів. Варто відзначити, що особливістю NeMo є його модульна архітектура, яка дає можливість дослідникам та розробникам легко створювати, налаштовувати та оптимізувати різноманітні компоненти нейронних мереж для розв’язання конкретних завдань. Це перетворює NeMo на засіб, який не лише корисний для професіоналів у галузі машинного навчання, але й доступний для широкого кола користувачів, навіть тих, які не мають глибоких знань у цій сфері.

Отже, NVIDIA NeMo відіграє основну роль у сучасних дослідженнях та застосуваннях діаризації аудіо, пропонуючи гнучкі, масштабовані та передові рішення для різноманітних сфер, включно з кібербезпекою.

SpeechBrain є ще одним важливим гравцем у сфері алгоритмів машинного навчання для аудіодіаризації. Цей відкритий інструмент був спроєктований як універсальне рішення для різноманітних завдань обробки мовлення, зокрема авдіодіаризації, розпізнавання мовлення та синтезу мовлення. Гнучкість є основним аспектом SpeechBrain, що дає можливість йому легко адаптуватися до потреб користувачів за допомогою налаштування та оптимізації системи. Використання алгоритмів глибокого навчання дає змогу ефективно опрацьовувати складні аудіозаписи та точно визначати голоси різних дикторів. Однією з основних переваг SpeechBrain є його здатність ефективно опрацьовувати великі обсяги даних, що робить його ідеальним для обробки аудіозаписів у великому масштабі. Крім того, важливою є його адаптивність до різних умов запису, таких як різні мови, акценти та якість звуку.

Отже, SpeechBrain надає комплексне рішення для завдань аудіодіаризації, що характеризується високою точністю, гнучкістю та масштабованістю. Це важливо для різноманітних сфер застосування, від наукових досліджень до комерційних проєктів обробки аудіо.

5. Хід експерименту

5.1. Метрики

Помилка оцінки діаризації (DER) – це помилка визначення меж та перекриття сегментів у записі звуку з урахуванням правильного або неправильного призначення ідентифікатора мовця сегменту запису звуку. Ця помилка є основною для діаризації та є загальноприйнятою метрикою в комерційних системах.

Помилку оцінки діаризації можна обчислити за таким рівнянням (1):

$$DER = \frac{T_a+T_m+T_c}{T}, \quad (1)$$

де: T – загальна тривалість аудіофайлу; T_a – тривалість немовлення, помилково визначеного як мовлення, в аудіофайлі, T_m – тривалість мовлення, помилково визначеного як немовлення, в аудіофайлі, T_c – тривалість помилкової ідентифікації мовця в аудіофайлі.

Коефіцієнт помилки Жакара (JER) – помилка, яка визначає, як часто мовці помилково ідентифікуються як інші мовці. Ця метрика основана на індексі Жакара, який вимірює схожість між множинами.

Коефіцієнт помилки Жакара можна обчислити за формулою (2):

$$JER = 1 - \frac{\sum_{i=0}^{N}|S_i \cap D_i|}{\sum_{i=0}^{N}|S_i \cup D_i|}, \quad (2)$$

де: S_i – множина сегментів для мовця в тестовому наборі даних; D_i – множина сегментів, де було передбачено мовця; N – загальна кількість мовців.

5.2. Початковий аналіз

Перед початком масштабного тестування на великих обсягах даних із різних бібліотек Python ми провели попередні експерименти з алгоритмом Pyannote, використовуючи аудіозаписи з різноманітними умовами. Ці аудіозаписи охоплювали файли з різною кількістю учасників, змінним рівнем шуму та різним ступенем перекриття мовлення. Такий підхід дав можливість нам глибше зрозуміти продуктивність та надійність Pyannote в різних акустичних сценаріях, що має велике значення для подальшого аналізу великих обсягів даних.

Спершу для аналізу було вибрано двохвилинний аудіофайл, що містив запис новин. Особливістю цього запису був високий рівень фонового шуму, хоча не було перекриття аудіодоріжок. Такий вибір дав змогу оцінити ефективність алгоритму в обробці аудіо зі складними умовами звукового середовища, які не супроводжуються одночасним мовленням декількох дикторів.

Результат діаризації із двома дикторами та фоновим шумом візуалізовано на рис. 2.

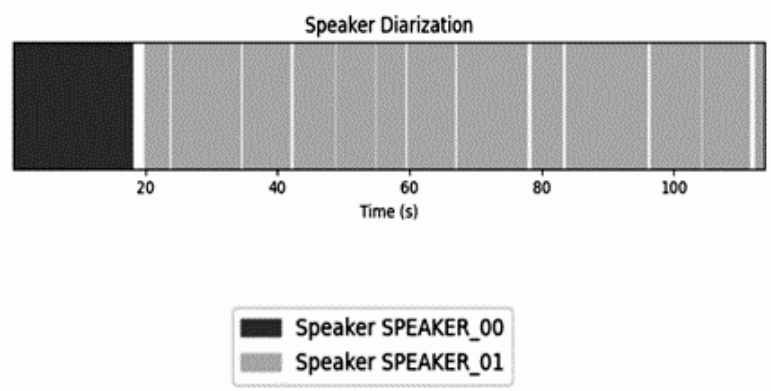

Рис. 2. Шкала часу аудіофайлу, що містить двох дикторів із фоновим шумом

Наступним етапом нашого дослідження стало складніше випробування для бібліотеки Pyannote. Для цього ми обрали 16-хвилинний аудіозапис конференції, на якій одночасно брали участь 11 осіб. Цей запис відрізнявся високим рівнем шуму, змінами гучності мовлення та частими перебиваннями між спікерами. Цей вибір дав можливість нам оцінити, наскільки ефективно Pyannote справляється з високим рівнем складності в розпізнаванні мовлення та ідентифікації дикторів у складних аудіозаписах із багатьма голосами.

Результат діаризації з одинадцятьма спікерами та фоновим шумом візуалізовано на рис. 3.

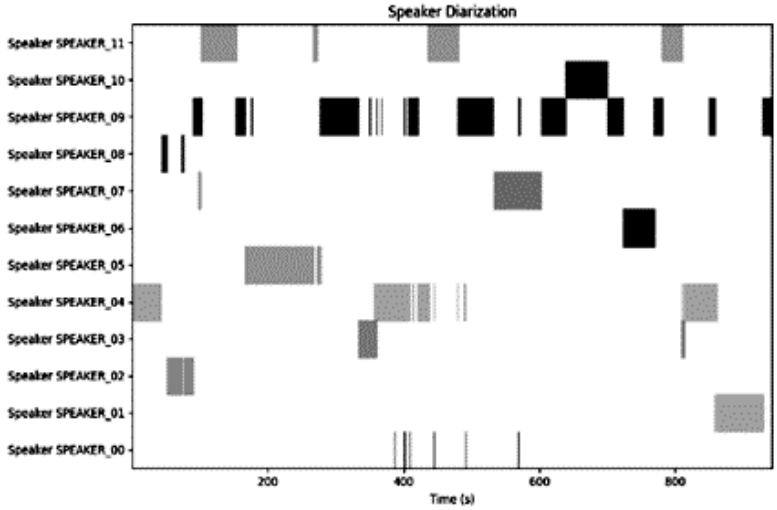

Рис. 3. Таймлайн аудіофайлу з одинадцятьма спікерами

Дослідження було продовжено перевіркою системи на аудіозаписі, що складається лише з чистого мовлення, без наявності фонового шуму, накладених доріжок або перерв у мовленні. Цей запис має тривалість приблизно 15 хвилин і надає ідеальні умови для аналізу, що дає можливість оцінити роботу алгоритму в оптимальних умовах. Цей підхід дає змогу встановити базовий рівень точності роботи системи діаризації за ідеальних умов, без зовнішнього впливу.

Результат діаризації з двома мовцями та чистим фоном показано на рис. 4.

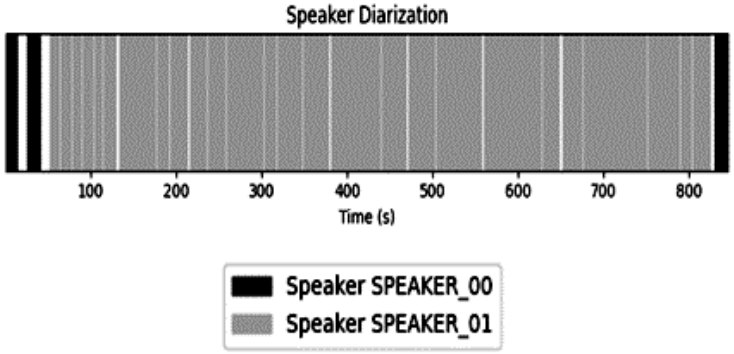

Рис. 4. Таймлайн аудіофайлу з двома спікерами з чистим звуком

У табл. 1 подано результати початкового дослідження стійкості бібліотеки діаризації Pyannote.

Таблиця 1

Результати початкового дослідження

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

Отримані результати експерименту свідчать, що система діаризації успішно справлялася з обробкою аудіозаписів з невеликою кількістю мовців, хоча її ефективність дещо зменшувалася в умовах наявності шумного оточення. У разі діаризації більшої кількості мовців у шумному

середовищі система виявилася менш ефективною і продемонструвала недостатню точність. Отже, не рекомендується використовувати цю систему для обробки даних, де присутня велика кількість мовців з потенційним перекриттям голосів та високим рівнем шуму у середовищі.

5.3. Тестовий набір даних для вибору моделі

Щоб оцінити ефективність Python-бібліотек для діаризації, було обрано набір даних VoxConverse [16]. Цей набір даних є важливим ресурсом для дослідників та розробників у галузі обробки мовлення. Він містить широкий спектр аудіозаписів, охоплюючи різні сценарії, такі як публічні виступи, інтерв'ю, новини та дебати. Особливістю VoxConverse є наявність аудіозаписів, де спостерігається перекриття мовлення, що є характерною рисою реальних ситуацій та становить велике зацікавлення для досліджень у цій галузі. Кожен аудіозапис у наборі VoxConverse містить детальні анотації з часовими інтервалами та ідентифікаторами дикторів. Ця інформація надзвичайно цінна, оскільки дає можливість точно оцінити ефективність різних алгоритмів та систем діаризації у розподілі мовлення між різними дикторами. Використання таких анотацій важливо для порівняння результатів різних систем діаризації з ідеальним станом та оцінки їхньої продуктивності.

З набору даних VoxConverse, який містить 464 записи, для проведення тестування було відібрано лише перші 50 записів для оптимізації часу виконання процесу діаризації та раціонального використання ресурсів. Обрані аудіозаписи відрізняються тривалістю, від 3 до 20 хвилин, що забезпечує широкий спектр сценаріїв мовлення для оцінки продуктивності обраних бібліотек Python для машинного навчання. Цей підхід не лише спрямований на аналіз деталей, а й має практичне значення, оскільки він відображає адаптивність систем до різноманітних сценаріїв мовлення у реальних умовах. Це дає можливість краще зрозуміти роботу кожної системи в умовах, що можуть виникнути у реальному житті, та визначити можливі напрями подальшого вдосконалення.

5.4. Вибір моделі для системи виявлення зловмисників

Для кожної з бібліотек ми розробили відповідний код, враховуючи їх унікальні особливості, з метою забезпечення відповідності кінцевого результату загальноприйнятому стандарту RTTM для позначок від різних систем діаризації.

Під час експерименту ми використовували моделі цих бібліотек, навчені на наборі даних VoxConverse, для оцінки їхньої ефективності в реальних умовах. Головною метою цього експерименту було визначити, яка з них найкраще підходить для кінцевого завдання виявлення зловмисників.

Для оцінки ефективності кожної бібліотеки були обрані такі показники, як середні значення DER та JER, розраховані на основі 50 вибраних тестових записів із VoxConverse. Також був врахований час діаризації для кожної системи, щоб оцінити найефективніший алгоритм за цим параметром. Цей аналіз допоможе визначити не лише найточнішу систему розпізнавання голосу, але й ту, яка забезпечить найкраще співвідношення між швидкодією та якістю обробки даних.

Результати експерименту з вибору моделі представлено в табл. 2.

Таблиця 2

Результат експерименту щодо вибору моделі

Модель	Час обробки	DER	JER
SpeechBrain	3 хв 53 с	0,31	0,41
NVIDIA NeMo	17 хв 32 с	0,14	0,39
Pyannote	20 хв 7 с	0,09	0,28

Бібліотека машинного навчання з відкритим кодом SpeechBrain відзначилася швидкістю обробки, зайнявши лише 3 хвилини 53 секунди. Загалом DER на рівні 31 % можна вважати задовільним, особливо враховуючи складність тестових даних.

Бібліотека NeMo від компанії NVIDIA потребувала більш тривалого часу для обробки даних – 17 хвилин і 32 секунди. Незважаючи на це, вона продемонструвала сильні результати в діаризації, особливо враховуючи складність аудіоданих. Зокрема, середній DER на рівні 14 % свідчить про високу ефективність алгоритму.

Для реалізації системи виявлення зловмисників було вибрано використання Pyannote, яка продемонструвала найвищі результати діаризації на цьому наборі даних. Середній DER на рівні 9 % свідчить про високу ефективність Pyannote в розв'язанні проблем, що виникають у цьому контексті. Незважаючи на те, що тривалість обробки даних системою Pyannote становила 20 хвилин і 7 секунд, це компенсується високою точністю та документацією високої якості, що сприяє швидкому початку роботи з бібліотекою. Хоча час обробки не має такого вирішального значення, як у NeMo, час впровадження та налаштування Pyannote був значно скороченим.

Отже, з урахуванням швидкості, точності та простоти використання, Pyannote стала оптимальним вибором для розв'язання кінцевої задачі, показавши відмінний баланс між часом обробки та якістю результатів.

5.5. Діаризація для завдання виявлення зловмисника

5.5.1. Підготовка даних

Перед розробкою та аналізом системи виявлення зловмисників важливо провести етап підготовки та обробки вхідних даних. Цей етап охоплює вибір відповідних аудіозаписів і їх попередню обробку для ефективного навчання та тестування моделі. Після створення надійного та репрезентативного навчального набору даних наступним кроком є розробка методів виявлення та ідентифікації потенційних злочинців у базі даних. Використання різних методів діаризації дає можливість перевірити ефективність системи та забезпечити її практичне використання в реальних сценаріях.

Для проведення досліджень було обрано кілька епізодів відомого українського Youtube-подкасту, учасниками яких постійно є двоє ведучих, а також кожного епізоду приходять різні гості. У межах експерименту деяких гостей умовно було позначено як "зловмисників". Для кожного з цих гостей було створено п'ять окремих трихвилинних аудіозаписів з метою отримання їхнього голосового відбитку. Загальний розмір набору даних для ідентифікації "зловмисників" становив 56 записів, тривалість кожного з яких – 2–3 хвилини. З цієї кількості 15 записів містять голоси визначених "зловмисників", тоді як у решти – 41, таких голосів немає. Це дає унікальну можливість оцінити, наскільки добре розроблена система діаризації здатна розпізнавати та відрізняти голоси в реальних сценаріях, що є основним для практичного використання.

Набір даних, який використовувався для цього експерименту, можна знайти у [15].

Приклад підготовленого аудіофайлу для виявлення зловмисників наведено на рис. 5.

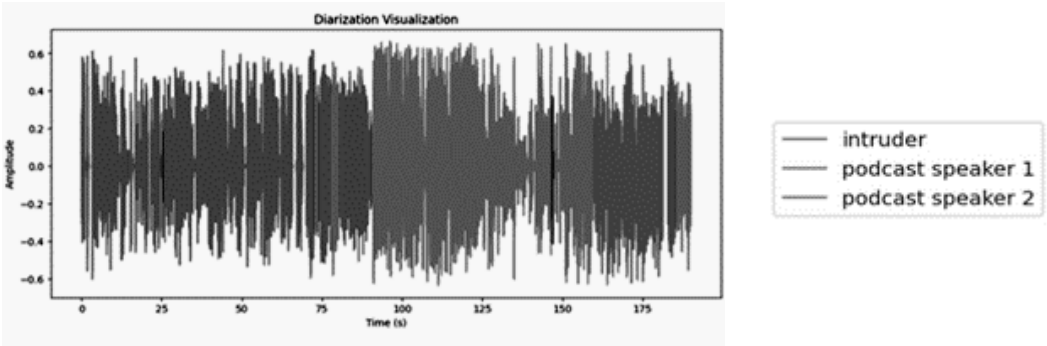

Рис. 5. Приклад аудіофайлу з голосом зловмисника

5.5.2. Впровадження системи виявлення зловмисника

Наступним етапом нашої дослідницької роботи була розробка спеціального алгоритму, призначеного для вилучення ембедінгів з аудіозаписів, що містять голосові зразки осіб, відзначених як "зловмисники". Цей алгоритм має за мету перетворення різноманітних голосових характеристик кожного мовця на складні, високорозмірні числові вектори.

Використання цих голосових ембедінгів має вагоме значення у нашому дослідженні, оскільки воно дає можливість проводити точніше і глибше порівняння та аналіз. Це досягається через вимірювання косинусних відстаней між цими числовими векторами. За допомогою аналізу цих відстаней ми можемо з високою точністю визначити, чи можна віднести певний фрагмент мови до конкретного "злочинця" або іншого мовця. Ця методологія виявляється дуже ефективною для розрізнення різних голосів у записах аудіо, що є основним для розробки систем, застосовуваних у судових експертизах та інших сферах, де потрібна точна ідентифікація особи за голосом.

Наступним кроком у дослідженні є застосування розробленого алгоритму для отримання векторних ембедінгів з усіх підозрілих записів у базі даних. Цей процес містить аналіз кожного аудіофайлу та вилучення відповідних ембедінгів. Після їх збору вони групуються, щоб об'єднати схожі голосові характеристики, що спрощує подальший процес ідентифікації. Групування допомагає зменшити потребу багаторазового порівняння кожного сегмента аудіо з усіма записами зловмисника, таким способом підвищуючи ефективність та точність ідентифікації. Крім того, групування сприяє визначенню спільних характеристик голосів "зловмисників", що може допомогти точно ідентифікувати потенційних підозрюваних.

Обробка подкастів – це завантаження кожного аудіофайлу, запуск діаризації на цьому файлі, а пізніше вибір лише сегментів, що перевищують 5 секунд. Це забезпечує детальний аналіз та точну ідентифікацію різних голосів у записі. Пізніше кожен сегмент перевіряється на відповідність мовним шаблонам зловмисника, що заздалегідь визначені та зберігаються в базі даних. Ця методика дає змогу точно визначити моменти присутності підозрюваного в аудіоматеріалі, а також можливість ідентифікувати зловмисників, які говорять лише в деяких фрагментах подкасту. Це потрібно для досягнення високої точності ідентифікації, водночас не втрачаючи швидкості обробки.

Перед порівнянням сегментів потрібно визначити параметр порогового значення, який істотно впливає на результати дослідження. Порогове значення відіграє вирішальну роль у визначенні того, чи голос у сегменті подкасту збігається з голосом відомого зловмисника. Воно слугує мірою для порівняння рівня схожості між ембедінгами голосів. Основна ідея полягає в тому, що якщо косинусна відстань між ембедінгом сегмента та найближчим ембедінгом зловмисника менша за цей поріг, система розпізнає присутність зловмисника в цьому сегменті.

Система виявлення зловмисників зображена на рис. 6.

5.5.3. Експерименти із виявлення системою зловмисника

Після завершення розробки основних компонентів системи наступним етапом є її запуск та тестування на обраному наборі даних. Цей процес спрямований на оцінку функціональності та ефективності розробленої системи в реальних умовах. Важливою частиною цього етапу є аналіз результатів, що допомагає ідентифікувати переваги і хиби програми, а також визначити можливі напрями для подальшого вдосконалення. У досягненні балансу між precision і recall з пропорцією 1:1 ми ставимо однаковий акцент на влучність і повноту класифікації. Це означає, що ми намагаємося максимізувати як відсоток правильно класифікованих позитивних випадків серед усіх випадків, які модель визначила як позитивні, так і відсоток позитивних випадків, що були правильно ідентифіковані моделлю. Це дає можливість створити модель, яка ефективно ідентифікує позитивні приклади, водночас мінімізуючи як помилкові визначення, так і пропущені позитивні приклади.

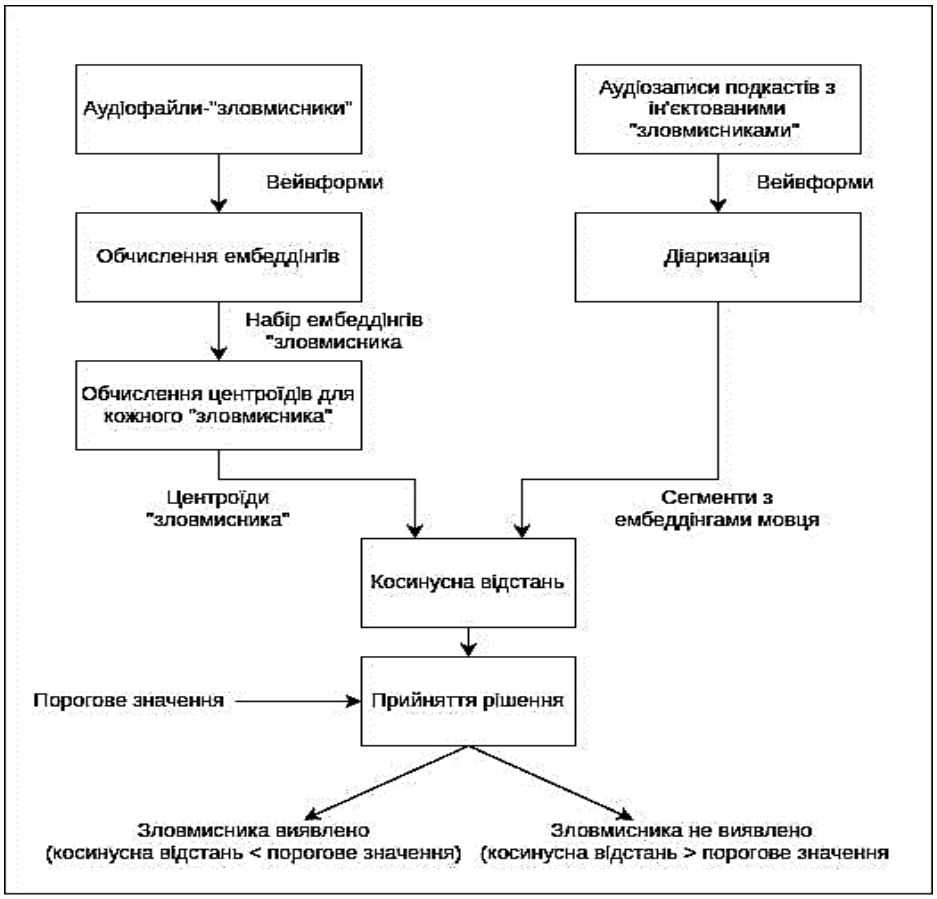

Рис. 6. Система виявлення зловмисника

6. Результати дослідження

Результати експерименту показали, що алгоритм продемонстрував високий рівень точності в завданнях ідентифікації. Серед усіх оброблених аудіофайлів лише один файл помилково класифікували як такий, що містить голос злочинця.

У табл. 3 подано результат експериментів з виявлення зловмисника.

Таблиця 3

Результат експериментів з виявлення зловмисника

Точність, %	Влучність, %	Повнота, %	F1-міра, %
98,21	93,75	100	96,77

Аналізуючи дані, представлені у табл. 3, можна зробити висновок, що досягнення високої точності виявлення порушників є досяжним, проте потребує ретельного налаштування системи відповідно до характеристик кожного окремого запису. Основними факторами, які впливають на успішність ідентифікації, є гіперпараметри “min_segment_duration” (мінімальна тривалість сегмента) та “similarity_threshold” (порогове значення схожості). Встановлення мінімальної тривалості сегмента допомагає уникнути помилкової ідентифікації порушників, але це може призвести до пропуску їхніх коротких висловлювань. З другого боку, налаштування порогового значення схожості для вбудованих елементів є важливим для точного розпізнавання голосів порушників, водночас уникаючи помилкових спрацювань.

Висновки

У цьому дослідженні ми провели ґрунтовний аналіз різних моделей глибокого навчання у сфері діаризації мовця, зосереджуючись на їх застосуванні для виявлення зловмисників. Ми оцінили стійкість та ефективність цих систем діаризації у різних умовах середовища.

Основною метою нашого дослідження було розроблення інноваційної системи виявлення зловмисників, що базується на принципах та технологіях діаризації аудіо. Головним результатом наших експериментальних досліджень є помітна перевага моделі діаризації Pyannote. Незважаючи на дещо тривалий час обробки порівняно з іншими моделями, ця модель продемонструвала високі показники якості діаризації, з найнижчими показниками помилок діаризації (DER) на рівні 14 % та помилок Жакара (JER) також на рівні 14 %.

Якість роботи розробленої системи виявлення зловмисників на основі бібліотеки Pyannote була надзвичайно високою, що підтверджується точністю на рівні 98,21 % та ідеальним показником повноти (recall) на рівні 100 %. Крім того, система продемонструвала влучність (precision) на рівні 93,75 %. F1-міра, яка є збалансованим показником точності та повноти системи, становила 96,77 %, підкреслюючи загальну ефективність системи.

Найважливішим досягненням системи є її безпомилкова здатність виявляти всіх зловмисників, долучених до набору даних. Це підкреслює цінність системи як надійного інструменту у сценаріях виявлення зловмисників.

Список літератури

1. *Landini F., Glembek O., Matejka P., Rohdin J., Burget L., Diez M., Silnova A. (2021). Analysis of the but Diarization System for Voxconverse Challenge. Conference: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). DOI: 10.1109/ICASSP39728.2021.9414315*
2. *Dudykevych V., Mykytyn H., Ruda K. (2022). The concept of a deepfake detection system of biometric image modifications based on neural networks, in: 2022 IEEE 3rd KhPI Week on Advanced Technology (KhPIWeek), IEEE. DOI: 10.1109/khpiweek57572.2022.9916378*
3. *Shtefaniuk Y. and Opirskyy I. (2021). Comparative Analysis of the Efficiency of Modern Fake Detection Algorithms in Scope of Information Warfare, 2021 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 207–211. DOI: 10.1109/IDAACS53288.2021.9660924.1*
4. *Anguera Miro X., Bozonnet S., Evans N., Fredouille C., Friedland G., Vinyals O. (2012). Speaker Diarization: A Review of Recent Research, IEEE Trans. Audio, Speech, Lang. Process, Vol. 20, 356–370. DOI:10.1109/tasl.2011.2125954*
5. *Khoma V., Khoma Y., Brydinskyi V., Konovalov A. (2023). Development of Supervised Speaker Diarization System Based on the PyAnnote Audio Processing Library, Sensors, Volume 23, 2082. DOI: 10.3390/s23042082*
6. *Hannun A., Case C., Casper J., Catanzaro B., Diamos G., Elsen E., Prenger R., Satheesh S., Sengupta Sh., Coates A., Ng A. Y. (2014). Deep Speech: Scaling up end-to-end speech recognition. Available at: https://doi.org/10.48550/arXiv.1412.5567 (Accessed: 15 February 2024).*
7. *Ball J. (2023). Voice Activity Detection (VAD) in Noisy Environments. Available at: https://arxiv.org/html/2312.05815v1 (Accessed: 15 February 2024).*
8. *Cornell S., Omologo M., Squartini S., Vincent E. (2022). Overlapped Speech Detection and speaker counting using distant microphone arrays, Comput. Speech & Lang, Volume 72, 101306. DOI: 10.1016/j.csl.2021.101306*
9. *Kotti M., Moschou V., Kotropoulos C. (2008). Speaker segmentation and clustering, Signal Process, Volume 88, 1091–1124. DOI: 10.1016/j.sigpro.2007.11.017*
10. *Dawalatabad N., Ravanelli M., Grondin F., Thienpondt J., Desplanques B., Na H. (2021). ECAPA-TDNN Embeddings for Speaker Diarization. Proc. Interspeech, 3560–3564. DOI: 10.21437/Interspeech.2021-941*
11. *Garcia-Romero D., Snyder D., Sell G., Povey D. and McCree A. (2017). Speaker diarization using deep neural network embeddings, 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 4930–4934. DOI: 10.1109/ICASSP.2017.7953094*
12. *Bredin H. (2023). Pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe, in: INTERSPEECH 2023, ISCA, ISCA. Doi:10.21437/interspeech.2023-105*

13. Harper E., Majumdar S., Kuchaiev O., Jason, et al. NeMo: a toolkit for Conversational AI and Large Language Models [Computer software]. https://github.com/NVIDIA/NeMo

14. Ravanelli M., Parcollet T., Plantinga P., et al. (2021). SpeechBrain: A General-Purpose Speech Toolkit. Available at: https://arxiv.org/abs/2106.04624 (Accessed: 15 February 2024).

15. Chung J. S., Huh J., Nagrani A., Afouras T., Zisserman A. (2020). Spot the Conversation: Speaker Diarisation in the Wild, in: Interspeech 2020, ISCA, ISCA. DOI:10.21437/interspeech.2020-2337

16. Zaiets I. (2024). Dataset of ukrainian podcasts for intruder detection by voice. DOI:10.57967/hf/0701

UTILIZATION OF VOICE EMBEDDINGS IN INTEGRATED SYSTEMS FOR SPEAKER DIARIZATION AND MALICIOUS ACTOR DETECTION

I. Zaiets[2], V. Brydinskyi[1], D. Sabodashko[2], Yu. Khoma[1], Kh. Ruda[2], M. Shved[2]

Lviv Polytechnic National University,
Department of Measuring Information Technologies[1]
Department of Information Security[2]
E-mail:illia.zaiets.mkbas.2022@lpnu.ua, vitalii.a.brydinskyi@lpnu.ua, dmytro.v.sabodashko@lpnu.ua, yurii.v.khoma@lpnu.ua, khrystyna.s.ruda@lpnu.ua, mariia.y.shved@lpnu.ua

This paper explores the use of diarization systems which employ advanced machine learning algorithms for the precise detection and separation of different speakers in audio recordings for the implementation of an intruder detection system. Several state-of-the-art diarization models including Nvidia's NeMo, Pyannote and SpeechBrain are compared. The performance of these models is evaluated using typical metrics used for the diarization systems, such as diarization error rate (DER) and Jaccard error rate (JER). The diarization system was tested on various audio conditions, including noisy environment, clean environment, small number of speakers and large number of speakers. The findings reveal that Pyannote delivers superior performance in terms of diarization accuracy, and thus was used for implementation of the intruder detection system. This system was further evaluated on a custom dataset based on Ukrainian podcasts, and it was found that the system performed with 100 % recall and 93,75 % precision, meaning that the system has not missed any criminal from the dataset, but could sometimes falsely detect a non-criminal as a criminal. This system proves to be effective and flexible in intruder detection tasks in audio files with different file sizes and different numbers of speakers which are present in these audio files.

Keywords: deep learning, diarization, speaker embeddings, speaker recognition, cyber security.