

М. Е. Бондаренко, Г. С. Іващенко

Харківський національний університет радіоелектроніки, Харків, Україна

ГІБРИДИЗАЦІЯ МЕТОДІВ ВИЛУЧЕННЯ ОЗНАК У СИСТЕМІ ГОЛОСОВОЇ ІДЕНТИФІКАЦІЇ КОРИСТУВАЧА

Анотація. Актуальність. Практичне застосування систем голосової ідентифікації обумовлює необхідність високої точності розпізнавання, обчислювальної ефективності та стійкості до синтезованих підрізок голосу. Статистичні підходи, такі як акустичний аналіз та спектральне порівняння, є невимогливими до апаратних ресурсів, проте поступаються за точністю сучасним глибинним підходам на основі машинного навчання. Штучні нейронні мережі, попри вищу точність, характеризуються значними обчислювальними витратами та чутливістю до викривлень. Одним із можливих шляхів вирішення є гібридизація методів на етапі вилучення ознак, що передбачає поєднання статистичного акустичного аналізу та нейромережевого підходу в єдиному векторному просторі. **Об'єкт дослідження:** гібридизація методів вилучення ознак у системі голосової ідентифікації на основі поєднання статистичного акустичного аналізу та штучної нейронної мережі. **Предмет дослідження:** метод гібридного вилучення ознак, що поєднує MFCC-вектори та вектори ознак нейронної мережі в єдиному векторному просторі, із порівняльним аналізом способів їх поєднання. **Мета дослідження:** розробка та дослідження методу гібридного вилучення ознак для підвищення точності голосової ідентифікації. **Завдання дослідження.** Реалізувати запропоновані підходи гібридизації статистичних та нейромережевих методів вилучення ознак у єдиному векторному просторі; експериментально оцінити точність ідентифікації у режимах closed-set та open-set; дослідити обчислювальну ефективність і масштабованість запропонованих варіантів. **Методи дослідження.** Статистичний акустичний аналіз на основі мел-частотних кепстральних коефіцієнтів, згортова нейронна мережа на основі архітектури VGGish, метод головних компонент для зниження розмірності, навчання проєкційного шару за критерієм кутового відступу, експериментальне порівняння отриманих результатів на наборі даних VoxCeleb1. **Результати дослідження.** Запропоновано та досліджено три варіанти гібридизації на рівні вилучення ознак, такі як конкатенація нормованих векторів, зважена лінійна комбінація після зниження розмірності та злиття через навчений проєкційний шар. Отримані результати демонструють скорочення EER та підвищення точності ідентифікації порівняно з використанням кожного методу окремо. **Висновки.** Результати дослідження підтверджують доцільність застосування гібридизації методів на етапі вилучення ознак для підвищення точності ідентифікації мовця. Встановлено, що гібридизація шляхом злиття через навчений проєкційний шар забезпечує найвищу точність, тоді як конкатенація нормованих векторів не потребує додаткового навчання.

Ключові слова: біометричні системи голосової ідентифікації; гібридизація методів; вилучення ознак; MFCC; штучні нейронні мережі; згортова мережа VGGish; навчання проєкційного шару; злиття ознак.

1. Вступ

1.1 Актуальність тематики

Голосова ідентифікація є актуальним способом контролю доступу, що використовується при створенні персоналізованих інтерфейсів, зокрема для розблокування авто, пристроїв, доступу до приміщень чи комп'ютерних систем [1–3].

Практичні умови застосування таких систем, зокрема мінливе акустичне середовище, фізіологічні варіації голосу, наявність фонового шуму та загрози підробки за допомогою синтезованого аудіо, передбачають суперечливі вимоги щодо точності ідентифікації, стійкості до міжсесійних варіацій та обчислювальної ефективності [3, 4].

Спосіб отримання векторного представлення мовця є основним компонентом архітектури системи біометричної ідентифікації, оскільки саме від якості вилучення ознак залежить подальша точність розрізнення користувачів [5]. Статистичні методи акустичного аналізу, такі як мел-частотні кепстральні коефіцієнти (MFCC), забезпечують інтерпретовані, відтворювані та обчислювально ефективні представлення, але усереднюють характеристики в часі та не відображають складну нелінійну динаміку сигналу [3, 6]. Глибокі нейронні мережі здатні автоматично виявляти складні закономірності, проте є вимогливими до апаратних

ресурсів та потребують великих навчальних вибірок і мають обмежену інтерпретованість [7–9].

Головними проблемними питаннями при цьому стають такі: необхідність поєднання переваг актуальних підходів до вилучення ознак без суттєвого ускладнення обчислювальної архітектури та відсутність універсальних методів їхньої інтеграції [10, 11].

Крім того, слід зазначити, що обчислювальні ресурси пристроїв для біометричної ідентифікації є обмеженими [12]. Це визначає необхідність пошуку методів ранньої інтеграції ознак, які б дозволили формувати компактні та інформативні вектори представлення мовця без суттєвого збільшення обчислювального навантаження.

Тому дослідження, присвячені формуванню ефективних гібридних представлень мовленнєвого сигналу шляхом інтеграції статистичних та глибинних ознак, є актуальним [13, 14]. Вирішення цього питання дозволить підвищити точність розпізнавання мовців в системах біометричної голосової ідентифікації за умов обмежених обчислювальних ресурсів.

1.2 Аналіз літературних джерел і визначення проблеми

У сучасних системах голосової ідентифікації дедалі частіше використовується поєднання різних

методів вилучення ознак з метою компенсації обмежень кожного з них [15-21].

Дослідження [15] розглядає голосову ідентифікацію як багатоетапний процес, що починається з перетворення акустичного сигналу шляхом екстракції 25 статистичних ознак. Цей етап охоплює розрахунок параметрів Гйорта, енергії Тігера та центральних моментів для детального опису морфології голосу мовця. Наступний рівень передбачає селекцію ознак на основі коефіцієнта кореляції Спірмена, що завершується етапом класифікації з використанням моделей на основі методів Decision Tree, Support Vector Machine (SVM) або Ensemble Bagged Tree для прийняття остаточного рішення. Такий підхід до обробки даних дозволив авторам досягти точності 90.16% для моделі SVM та 95.14% для алгоритму Decision Tree. Однак переважання морфологічних параметрів у часовій області обмежує здатність системи враховувати спектрально-динамічні властивості сигналу, що ускладнює часову сегментацію.

У дослідженні [16] пропонується багаторівнева гібридна модель розпізнавання голосу, що поєднує природно-інспіровані алгоритми оптимізації (такі як оптимізація роєм часток (PSO) та генетичні алгоритми (GA)) із системами нечіткої логіки [9]. Рішення ґрунтується на первинному вилученні акустичних характеристик сигналу з подальшим їх використанням у текстонезалежному режимі. У підході використовуються такі ознаки мовного сигналу, як енергія сигналу (EE), частота перетину нуля (ZCR) та короткочасна енергія (STE), які формують вхідний вектор ознак для класифікатора. Це дозволяє ефективно ідентифікувати особу незалежно від змісту вимовлюваних фраз. Гібридизація методів дозволила досягти точності розпізнавання до 88% при середньоквадратичній похибці (MSE) на рівні 10%. Проте розширення простору ознак ускладнює оптимізацію та збіжність моделі [16].

У роботі [17] продемонстровано модель для виявлення голосового спуфінгу, яка поєднує попередньо навчену модель ШНМ глибокого навчання (VGGish) для автоматичного вилучення ознак та довготривалу короткочасну пам'ять (LSTM) для подальшої класифікації. Архітектура VGGish-LSTM подана як гібридна, оскільки вона поєднує різні архітектури ШНМ, що застосовуються послідовно для розв'язання окремих підзадач – вилучення ознак і класифікації. Запропонована гібридна модель продемонструвала високу ефективність, досягнувши точності 99.11% та рівня рівної помилки (EER) 1.18% під час тестування на наборі даних ASV Spoof 2019 LA. Водночас в роботі зазначено, що обробка великих обсягів даних потребує використання складних попередньо навчених моделей (YAMNet), що спричиняє дефіцит обчислювальних ресурсів ще на етапі вилучення ознак, а застосування рекурентних нейронних мереж для довгих послідовностей ускладнює процес навчання через проблему зникнення градієнта.

У дослідженні [18] розглядається гібридна екстракція ознак шляхом поєднання статистичних

MFCC-параметрів із векторами ознак, відібраними згортковою мережею (CNN). Автори проводять порівняльний аналіз кількох конфігурацій гібридизації. Однак оцінка ефективності підходу обмежена єдиним невеликим набором даних і не враховує стійкості системи до міжсесійних варіацій голосу та її обчислювальної складності при масштабуванні.

Робота [19] пропонує інтеграцію механізму уваги для динамічного зважування кількох груп ознак у задачах верифікації мовця та виявлення піддробленого голосу (SASV). Запропонована архітектура здійснює спільне навчання ASV- та антиспуфінг-гілок для підвищення точності розпізнавання [22, 23]. Однак висока складність моделі та необхідність сумісного тренування обох компонентів ускладнюють інтерпретацію внеску окремих джерел ознак і суттєво підвищують обчислювальні витрати.

У дослідженні [20] пропонується злиття без залучення нейромережових екстракторів фізичних часових і частотних параметрів (спектральний потік та нахил, тон, інтенсивність) з їхніми статистичними варіаціями. Такий підхід орієнтований на реалізацію текстонезалежної ідентифікації мовця на рівні ознак низького порядку. Однак відсутність порівняння з сучасними глибокими архітектурами (VGGish, x-vector, ECAPA-TDNN) звукує застосовність отриманих висновків для сценаріїв із високою міжсесійною варіативністю голосового сигналу.

Дослідження [21] розглядає архітектуру AMDF-CNN, яка забезпечує увага-кероване злиття (attention-guided feature fusion) часових, спектральних та кепстральних вимірів в єдиній згортковій структурі. Застосування багатовимірної моделі дозволило знизити частку помилкової класифікації мовця на 2.52% порівняно з базовими підходами. Оскільки автори не наводять порівняння обчислювальних витрат із простішими методами злиття, оцінити практичну ефективність такого ускладнення архітектури наразі неможливо.

Аналіз розглянутих актуальних досліджень свідчить про доцільність переходу від послідовного використання окремих методів до їх гібридизації. Це пов'язано з тим, що ізолювані алгоритми вже не забезпечують достатнього приросту ефективності, тому все частіше застосовуються їх поєднання в межах єдиної системи. У сучасних рішеннях статистичні методи та ШНМ використовуються разом як частини одного підходу. Класичні алгоритми з виділенням ознак і традиційними класифікаторами (SVM, дерева рішень) ефективні при однорідності даних і не потребують складних обчислень [23], але не підтримують роботу зі складними варіаціями мовлення та великими наборами даних. Натомість моделі на основі штучних нейронних мереж (LSTM, CNN) виявляють складні закономірності в аудіосигналі та враховують його часову структуру, але характеризуються високими вимогами до ресурсів, мають проблему зникаючого градієнта на довгих послідовностях та потребують великі масиви даних для навчання. Принципова

різниця між їх використанням полягає в тому, на якому етапі відбувається обмін інформацією між компонентами системи і чи формується в результаті єдине спільне представлення сигналу.

Отже на основі проведеного аналізу робіт [15-21] є підстави стверджувати, що доцільним є проведення дослідження, спрямованого на підвищення точності текстонезалежної ідентифікації мовця шляхом гібридизації методів вилучення акустичних ознак голосового сигналу.

1.3 Мета і завдання дослідження

Метою роботи є розробка та дослідження методу гібридного вилучення ознак для задач голосової ідентифікації, що поєднує MFCC-представлення та векторні ознаки нейромережі на основі VGGish. Передбачається формування спільного гібридного векторного опису мовленнєвого сигналу з подальшим порівнянням трьох підходів до злиття ознак: конкатенації, зваженої лінійної комбінації та злиття через навчений проєкційний шар [24]. Для досягнення мети були поставлені такі завдання:

- розробити підходи гібридизації методів вилучення статистичних та нейромережових ознак у єдиному векторному просторі;
- визначити критерій та показники якості процесу ідентифікації мовця;
- оцінити точність ідентифікації мовця за допомогою запропонованих способів гібридизації у режимах closed-set та open-set;
- дослідити обчислювальну ефективність і масштабованість запропонованих підходів при збільшенні банку користувачів.

2. Матеріали і методи дослідження

Об'єктом дослідження є гібридизація методів вилучення ознак у системі голосової ідентифікації на основі поєднання статистичного акустичного аналізу та штучної нейронної мережі. Реалізовано ідентифікацію мовця шляхом порівняння з еталонними профілями гібридного вектора ознак, який формується злиттям (через конкатенацію, зважену суму або використання проєкційного шару) MFCC- та VGGish-представлень мовленнєвого сигналу.

Основна гіпотеза дослідження: інтеграція статистичних та нейромережових ознак у єдиному векторному просторі на етапі вилучення ознак дозволить зменшити рівень рівної помилки (ERR) та підвищити точність ідентифікації порівняно з ізольованим застосуванням кожного з підходів. Гіпотеза ґрунтується на припущенні про взаємодоповнювальний характер цих представлень: ознаки, що вилучені за допомогою статистичних методів, описують явні спектральні відмінності між мовцями, тоді як ознаки, вилучені за допомогою ШНМ, відображають нелінійні закономірності, виявлені в процесі навчання [25].

Для реалізації підходу передбачається формування спільного векторного опису голосового сигналу з використанням різних стратегій злиття

ознак, що впливають на структуру фінального представлення.

Розробка та експериментальне оцінювання підходів гібридного вилучення ознак здійснювалися за дотримання таких крайових умов та обмежень:

Умова 1. Мовленнєвий сигнал розглядається як одноканальний із частотою дискретизації 16 кГц, що відповідає формату використаного набору даних.

Умова 2. Профіль користувача формується з фіксованої кількості висловлень, а еталонний вектор профілю обчислюється шляхом усереднення векторів цих висловлень.

Умова 3. Розмірності вихідних представлень є фіксованими: 240 ознак для статистичного методу та 128 ознак для нейромережового.

Умова 4. При навчанні шару злиття ваги базової нейромережової архітектури залишаються сталими, що запобігає деградації попередньо сформованих представлень.

Умова 5. Вплив адитивного акустичного шуму та багатомовності мовлення в межах даного дослідження не моделюється, а міжсесійна варіативність враховується лише через кількість висловлень у профілі користувача.

Процес голосової ідентифікації складається з кількох послідовних етапів, таких як попередня обробка вхідного аудіо сигналу $x(t)$, вилучення ознак, формування векторного представлення мовленнєвого сигналу, а також етап порівняння з еталонними профілями користувачів і прийняття рішення щодо ідентифікації.

На етапі вилучення ознак виконується перетворення мовленнєвого сигналу у компактний числовий вектор, що зберігає індивідуальні характеристики голосу та є стійким до акустичних варіацій між сесіями.

Вибір MFCC як статистичного методу вилучення ознак зумовлений обчислювальною ефективністю, можливостями інтерпретації та відтворення представлення мовного сигналу, не потребує навчання і враховує особливості слухового сприйняття людини [26].

На MFCC реалізовані класичні системи ідентифікації мовця на основі гауссових сумішей, а їх властивості та обмеження докладно проаналізовані в оглядових роботах [27]. Це дозволяє застосовувати вилучення ознак MFCC в умовах обмежених ресурсів та використовувати результати для порівняльного аналізу.

Сигнал розбивається на короткі перекривні кадри тривалістю 25 мс із кроком 10 мс [28]. Для кожного кадру застосовується вікно Геммінга для зменшення спектральних викривлень. Далі виконується швидке перетворення Фур'є та фільтрація через банк із 40 мел-фільтрів у діапазоні 80-7600 Гц.

Після цього обчислюється логарифм енергій фільтрів і застосовується дискретне косинусне перетворення, у результаті чого отримуються MFCC-коефіцієнти. Додатково обчислюються їх перші та другі похідні, які описують зміну ознак у часі.

$$D_t = \frac{\sum_{n=1}^N n \cdot (c_{t+n} - c_{t-n})}{2 \sum_{n=1}^N n^2} t, \quad (1)$$

де c_t – значення статичного MFCC-коефіцієнта у кадри t , а N – розмір часового вікна ($N = 2$).

Для кожного кадру формується вектор ознак, а для отримання фіксованого представлення всього сигналу, обчислюються середнє значення та стандартне відхилення по всіх кадрах:

$$\phi M(x) = [\mu_M; \sigma_M] \in \mathbb{R}^{240}, \quad (2)$$

де μ_M та σ_M – вектори середнього значення та стандартного відхилення ознак кадру по всій тривалості сигналу.

Головним обмеженням підходу є втрата часової структури сигналу внаслідок усереднення кадрів. Як наслідок, порядок появи локальних динамічних змін не враховується, а різні за часовою структурою сигнали отримують подібні представлення, що погіршує розрізнення мовців [16, 29].

Сучасною альтернативою опису сигналу через усереднення спектральних характеристик (зокрема MFCC) є використання згорткових нейронних мереж, які виконують автоматичне вилучення ознак безпосередньо з мел-спектрограми. У роботі у якості нейромережевого екстрактору застосовано модель на основі архітектури VGGish [30]. Завдяки попередньому навчанню на наборі аудіоподій, ця модель формує інформативні високорівневі представлення. Додатковими перевагами є компактність вихідного вектора розмірністю 128. Нейронну мережу було донаведено на цільовому наборі даних для її специфічної адаптації до задачі ідентифікації мовців.

Вхідні дані для мережі формуються у вигляді мел-спектрограми за схемою конвеєра попередньої обробки, наведеною на рис. 1. Аудіосигнал спочатку розбивається на короткі перекривні фрейми з застосуванням віконної функції, після чого для кожного фрейму обчислюється спектр за допомогою швидкого перетворення Фур'є.

На наступному етапі виконується проекція спектральної енергії на мел-шкалу з використанням набору фільтрів, після чого застосовується логарифмування отриманих значень. У результаті формується двовимірне представлення сигналу у вигляді матриці «час/частота». Для подальшої обробки мел-спектрограма фреймується на патчі розміром 96 часових кадрів на 64 мел-смуги, що відповідає інтервалу приблизно 0.96 с.

Архітектура мережі включає чотири згорткові блоки з ядрами 3×3 та кількістю фільтрів 64, 128, 256 і 256 відповідно. Після кожного згорткового блоку застосовується операція максимального об'єднання 2×2 для поступового зменшення просторової розмірності ознак.

Далі використовуються два повнозв'язні шари розмірністю 4096 і 128 нейронів.

Вихід останнього шару формує 128-вимірний вектор ознак:

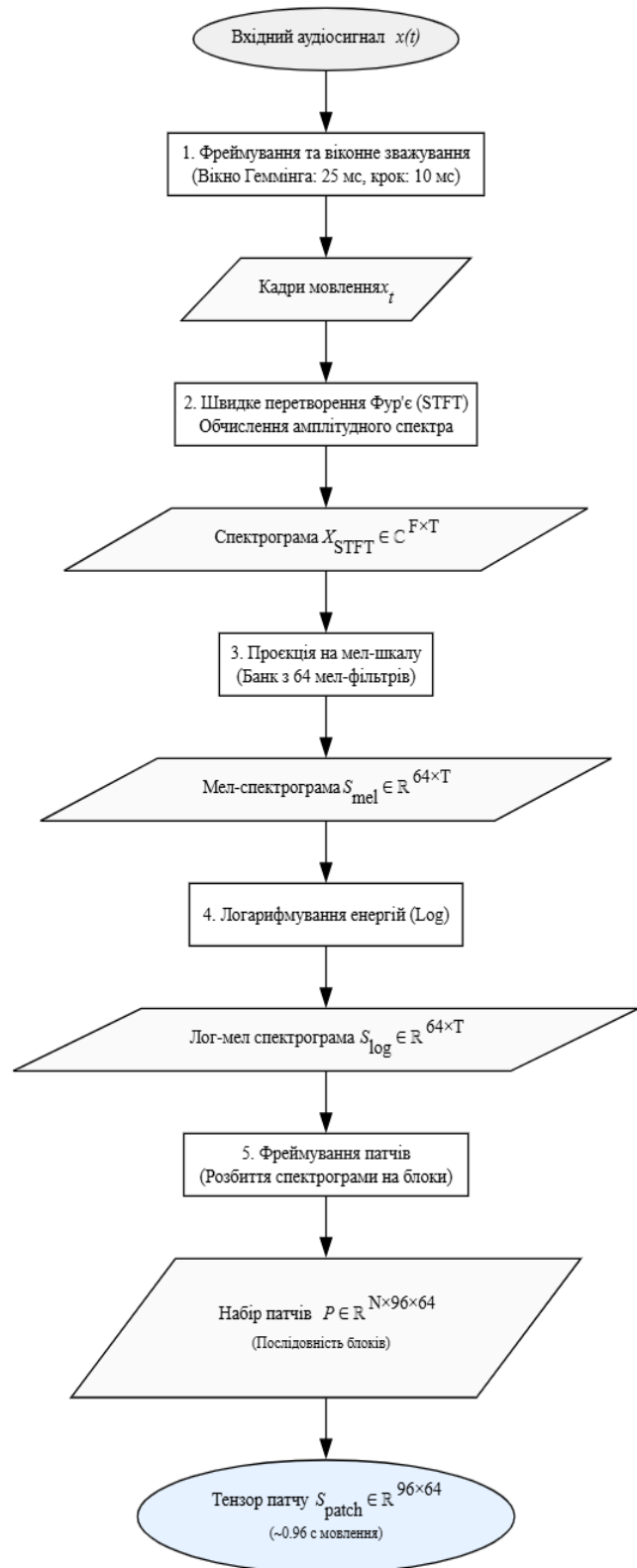


Рис. 1. Блок-схема конвеєра попередньої обробки аудіосигналу та формування мел-спектрограмних патчів

$$\phi V(x) = fVGGish(x; \theta) \in \mathbb{R}^{128}, \quad (3)$$

де $fVGGish$ – нелінійне перетворення вхідного сигналу x нейронною мережею з параметрами θ .

Навчання мережі виконується з використанням функції втрат на основі кутового відступу (Angular

Margin Loss), що обґрунтовано специфікою задачі ідентифікації, у якій рішення приймається за мірою подібності між векторними представленнями, а не за належністю до фіксованого класу. Це забезпечує формування розділених кластерів у просторі ознак для різних мовців і збільшує міжкласову відстань, що безпосередньо сприяє зниженню EER [10, 12]. Для сигналів довільної тривалості вихідний аудіопотік розбивається на патчі з перекриттям, після чого для кожного з них обчислюється вектор ознак, а фінальне представлення отримується шляхом усереднення по часовій осі [11].

На відміну від статичного методу вилучення ознак, запропонований метод на основі ШНМ VGGish дозволяє моделювати складні нелінійні залежності між спектральними компонентами та враховувати часову динаміку сигналу. Водночас такий підхід є обчислювально більш затратним і потребує достатнього обсягу навчальних даних для стабільної роботи моделі.

Методи вилучення ознак формують вектори різного типу: класичний метод описує локальні спектральні характеристики та їхню динаміку, нейромережовий моделює складні закономірності, виявлені під час навчання. Завдяки цьому описані методи є взаємодоповнюваними: на основі MFCC чутливий до явних спектральних відмінностей між мовцями, а на основі VGGish здатний виявляти малопомітні особливості мовлення та артикуляції, які складно формалізувати аналітично [3, 25, 29].

Для приведення представлень вилучених ознак різної розмірності до єдиного простору застосовано метод головних компонент, що виконує лінійне перетворення з мінімальною втратою дисперсії вихідних даних і не потребує навчання з учителем, завдяки чому вплив зниження розмірності можна відокремити від впливу самого способу злиття ознак у спільний простір.

Експериментальну реалізацію гібридної системи виконано за допомогою фреймворку PyTorch в інтерактивному середовищі Apache Zeppelin для структурованого виконання сценаріїв та забезпечення відтворюваності результатів експериментів. Обчислення проводилися на серверній платформі на основі процесора AMD EPYC 9354, графічного прискорювача NVIDIA A100 (80 ГБ VRAM) та 256 ГБ DDR5 ECC.

Як експериментальний набір даних обрано VoxCeleb1 (1251 мовець, понад 145 000 висловлювань) [31, 32]. Вибір обумовлено наявністю записів мовців у неконтрольованих акустичних умовах, що дає змогу оцінити стійкість методів до міжсесійних варіацій голосу, а також його ефективність у задачах ідентифікації мовця для забезпечення порівняльності результатів. Дані розділено на дві вибірки: 1151 мовець для навчання та формування профілів (по 3 висловлювання на користувача), і 100 мовців – для тестування.

Вибір показників FAR, FRR та EER у якості критеріїв обґрунтований тим, що FAR і FRR характеризують два протилежні типи помилок системи залежно від обраного порога прийняття

рішення, тоді як EER є інваріантним критерієм для порівняння конфігурацій.

Оцінювання у режимах closed-set (оцінка точності розпізнавання) та open-set (оцінка здатності відхиляти невідомих мовців) зумовлене необхідністю аналізу роботи системи у різних практичних сценаріях застосування. Додатково вимірювалися експлуатаційні характеристики: середній час обробки запиту, тривалість виконання окремих модулів, обсяг використаної VRAM, залежність часу ідентифікації від розміру банку користувачів та вплив кількості векторів у профілі на підсумкові показники.

3. Результати дослідження

3.1 Підходи гібридизації методів вилучення ознак

Об'єднання традиційних акустичних (MFCC) та нейромережових (VGGish) представлень у єдиний гібридний вектор $\phi H(x)$ дає змогу ефективно використати переваги обох підходів, поєднуючи класичні спектрально-часові характеристики з високорівневими абстрактними ознаками сигналу, вилученими за допомогою ШНМ [17, 30].

Запропоновано три підходи інтеграції векторів ознак сигналу, що відрізняються обчислювальною складністю: конкатенація нормованих векторів, зважена лінійна комбінація та злиття через навчений проєкційний шар.

Конкатенація нормованих векторів передбачає об'єднання векторів після їх нормування. Перед конкатенацією обидва вектори приводяться до одиничної норми за евклідовою метрикою, що дає змогу сформувати цілісний вектор розширеної розмірності без втрати первинної інформації з кожного джерела:

$$\begin{aligned}\hat{\phi}M &= \frac{\phi M}{\|\phi M\|_2}, \\ \hat{\phi}V &= \frac{\phi V}{\|\phi V\|_2},\end{aligned}\quad (4)$$

У подальшому конкатенуються в вектор:

$$\hat{\phi}H = [\hat{\phi}M; \hat{\phi}V] \in R^{368}. \quad (5)$$

Нормування перед конкатенацією запобігає домінуванню вектора з більшою нормою при обчисленні міри подібності та пригніченню внеску іншого джерела ознак. Після нормування обидва вектори мають рівноцінний вплив на формування гібридного представлення. Підсумковий вектор містить 240 статистичних та 128 нейромережових ознак, що забезпечує формування цілісного опису сигналу без втрати інформації з жодного із джерел. Запропонований підхід не потребує навчання та не вносить додаткових апроксимаційних похибок [33].

Спосіб зваженої лінійної комбінації передбачає приведення обох векторів до спільного простору однакової розмірності з подальшим їх зваженим поєднанням. Статистичний вектор має розмірність 240, тоді як нейромережовий 128. Для приведення їх

до спільного простору виконується лінійне перетворення статистичного вектора до 128 вимірів за допомогою матриці головних компонент, побудованої на навчальній вибірці:

$$\varphi' M = P \cdot \varphi M \in R^{128}, \quad (6)$$

де P – матриця проєкції розмірністю 128×240 , стовпці якої відповідають 128 головним компонентам з найбільшою дисперсією.

Після цього обидва вектори нормуються, і формується їх зважена лінійна комбінація:

$$\varphi H = \alpha \cdot \varphi' M + (1 - \alpha) \cdot \varphi V \in R^{128}, \quad (7)$$

де $\alpha \in [0, 1]$ ваговий коефіцієнт (від тільки нейромережевого при $\alpha = 0$ до проєктованого статистичного при $\alpha = 1$), який визначає відносний внесок джерел ознак і підбирається на валідаційній вибірці за критерієм мінімізації EER.

Спосіб злиття через навчений проєкційний шар передбачає нелінійне поєднання ознак за допомогою двошарової штучної нейронної мережі з навченими параметрами. На відміну від попередніх підходів, система самостійно навчається оптимальному способу інтеграції двох джерел ознак без використання припущень щодо їх відносної ваги або структури.

Обидва вектори спочатку конкатенуються у проміжний вектор розмірністю 368, який подається на вхід першого повнозв'язного шару:

$$\begin{aligned} h &= \tanh(W_1 \cdot [\varphi M; \varphi V] + b_1), \\ W_1 &\in R^{256 \times 368}, \\ b_1 &\in R^{256}. \end{aligned} \quad (8)$$

Ваги базової архітектури залишаються сталими, що запобігає деградації попередньо сформованих представлень.

Отриманий у кожному з підходів гібридний вектор $\varphi H(x)$ використовується для формування профілів користувачів та їх порівняння з вхідним сигналом у банку ознак. На основі цих представлень виконується обчислення міри подібності між тестовими та еталонними векторами, що дозволяє здійснювати ідентифікацію мовця.

3.2 Критерій та показники якості процесу ідентифікації мовця

Для виконання оцінки ефективності розпізнавання сформовано систему кількісних показників якості, що ґрунтуються на вимірюванні кількості помилок першого та другого роду. Основними критеріями оцінки обрано метрики помилка хибного прийняття (FAR) та помилка хибного відхилення (FRR), які відображають залежність помилок системи ідентифікації від обраного порога прийняття рішення. Для забезпечення порогово-інваріантного порівняння різних підходів та їхніх конфігурацій визначено інтегральний показник ERR, який відповідає точці

рівноймовірної помилки ($FAR = FRR$). Система показників якості додатково враховує режими closed-set (для визначення загальної точності ідентифікації) та open-set (для аналізу стійкості до міжсесійних варіацій голосу та оцінки здатності системи відхиляти незареєстрованих мовців).

Окрім метрик точності, до системи оцінювання залучено експлуатаційні характеристики: середній час обробки одного запиту $T_{проц}$, тривалість роботи окремих модулів (статистичного та нейромережевого вилучення ознак), обсяг використаної відеопам'яті (VRAM), а також залежність часу ідентифікації від розміру банку користувачів. На основі виділених показників сформовано загальний критерій оцінки процесу ідентифікації мовця, який полягає у мінімізації сукупного рівня помилок EER за умови дотримання часового регламенту обробки сигналу:

$$\min_{\theta} EER(\theta) \text{ при } T_{проц}(\theta) \leq T_{макс}, \quad (9)$$

де θ – вектор параметрів конфігурації системи (метод гібридизації ознак, розмірність векторного представлення, поріг ухвалення рішення), $T_{проц}$ – середній час обробки одного запиту, $T_{макс}$ – гранично допустимий час відповіді системи.

3.3 Оцінка точності ідентифікації в режимах closed-set та open-set

Для оцінки впливу способу гібридизації класичного статичного та нейромережевого підходів на точність ідентифікації користувачів проведено експериментальне порівняння (табл. 1) розглянутих конфігурацій в умовах closed-set постановки. Оцінювання здійснювалося шляхом порівняння записаних тестових висловлювань мовців із заздалегідь сформованими еталонними профілями користувачів у закритій множині мовців.

Таблиця 1 – Порівняння методів вилучення ознак в умовах closed-set постановки

Конфігурація	FAR, %	FRR, %	ERR, %	Точність ідентифікації, %
Статистичний метод (MFCC)	9.2	11.2	10.5	87.6
Нейромережевий метод (VGGish)	4.1	5.1	4.8	93.8
Гібридизація (конкатенація)	3.6	4.8	4.2	94.7
Гібридизація (при $\alpha=0.35$)	3.8	3.1	4.4	94.3
Гібридизація (FC-fusion)	2.9	3.8	3.3	95.9

Основними критеріями порівняння є точність розпізнавання та рівень рівноймовірної помилки

EER. Додатково аналізувалася стабільність роботи системи ідентифікації при кожній з конфігурацій при зміні тривалості тестових записів, що дозволило встановити архітектуру гібридного вектора для практичного застосування.

За результатами експерименту найнижчі значення FAR (2.9%) та EER (3.3%), а також найвищу точність ідентифікації (95.9%) зафіксовано для гібридизації через злиття за допомогою навченого проєкційного шару.

Найнижче значення FRR (3.1%) отримано для зваженої лінійної комбінації при ваговому коефіцієнті $\alpha = 0.35$.

Для статистичного методу зафіксовано найвищі значення FAR (9.2%), FRR (11.2%) та EER (10.5%).

Експеримент (табл. 2) дозволяє оцінити стабільність роботи конфігурацій системи біометричної голосової ідентифікації в умовах open-set постановки при змінах голосових характеристик користувачів між сесіями запису, а також за появи незареєстрованих мовців.

Таблиця 2 – Результати оцінки стійкості системи до міжсесійних варіацій у open-set постановці

К-сть сесій	Метод	EER, %	FRR, %	Точність ідентифікації, %
1	Статистичний	5.5	89.8	90.2
1	Нейромережевий	4.9	91.5	92.1
1	Гібрид (оцінки)	4.6	92.3	92.8
1	Гібрид (рішення)	5.0	91.0	91.7
3	Статистичний	4.8	91.2	91.6
3	Нейромережевий	3.5	94.0	94.8
3	Гібрид (оцінки)	3.1	95.4	95.9
3	Гібрид (рішення)	3.4	94.6	95.1
5	Статистичний	4.5	92.6	92.9
5	Нейромережевий	3.0	95.8	96.3
5	Гібрид (оцінки)	2.7	96.9	97.2
5	Гібрид (рішення)	2.9	96.2	96.7

Найвищу точність ідентифікації (97.2%) зафіксовано для гібридизації на рівні оцінок при п'яти сесіях у профілі.

3.4 Оцінка обчислювальної ефективності та масштабованості

В процесі дослідження (табл. 3) аналізується вплив розміру банку користувачів на час обробки запиту та використання GPU-пам'яті для нейромережевого та гібридних методів.

За результатами експерименту при збільшенні банку користувачів з 50 до 1000 осіб час обробки

одного запиту зростає з 42 до 341 мс для нейромережевого методу, з 55 до 450 мс для гібридизації на рівні ознак та з 50 до 390 мс для гібридизації на рівні оцінок.

Обсяг використаної відеопам'яті змінювався в межах від 812 до 910 МБ.

Таблиця 3 – Порівняння обчислювальної ефективності залежно від розміру банку користувачів

Кількість користувачів	ШНМ VGGish, мс	Гібрид (ознаки), мс	Гібрид (оцінки), мс	GPU memory, МБ
50	42	55	50	812
100	57	75	68	820
200	89	120	105	835
500	176	230	205	860
1000	341	450	390	910

4. Обговорення результатів дослідження

Реалізовано підходи гібридизації на рівні вилучення ознак для задач голосової ідентифікації, що ґрунтуються на інтеграції статистичного акустичного аналізу та нейромережевого підходу в єдиному векторному просторі.

Запропонований гібридний підхід передбачає формування спільного представлення, в якому обидва джерела вилучених ознак присутні одночасно та рівноправно впливають на результат ідентифікації.

Порівняльний аналіз показав, що статистичний метод на основі MFCC продемонстрував найнижчу точність ідентифікації (87.6%) та найвищі значення FAR (9.2%), FRR (11.2%) і EER (10.5%) (табл. 1).

Застосування нейромережевої архітектури VGGish забезпечило підвищення точності до 93.8% та істотне зниження рівня EER до 4.8%, тоді як використання гібридизації дозволило додатково покращити результати: злиття через проєкційний шар (FC-fusion) забезпечило максимальну точність 95.9% та мінімальний EER 3.3% (табл. 1).

Експериментальні дослідження стійкості методів до міжсесійних варіацій показали, що збільшення кількості сесій у профілі з 1 до 5 забезпечує стабільність розглянутих конфігурацій (табл. 2).

При 1 сесії використання статистичного підходу вилучення ознак демонструє точність 90.2%, а використання нейромережевого методу на основі VGGish – 92.1%, гібридизація підвищує цей показник до 92.8%.

При розширенні профілю до 5 сесій гібридний метод дозволяє досягти мінімального рівня EER 2.7% та максимальної точності ідентифікації 97.2% (табл. 2), що є найкращими показниками серед усіх

розглянутих методів у контексті open-set задач. Оцінка обчислювальної ефективності свідчить, що зі збільшенням кількості профілів у банку з 50 до 1000 осіб час обробки зростає майже лінійно для всіх конфігурацій системи голосової ідентифікації (табл. 3).

Мінімальний час відповіді забезпечує нейромережовий метод (від 42 до 341 мс) завдяки відсутності додаткових етапів обробки. Гібридизація на рівні ознак вимагає найбільших обчислювальних ресурсів (55-450 мс) через раннє об'єднання векторів, тоді як гібридизація на рівні оцінок поступається нейромережовому підходу за часом обробки на 15-25% (50-390 мс) (табл. 3), проте є обчислювально ефективнішою при збереженні високої точності. Найбільшу частку ресурсів використовує нейромережовий компонент, про що свідчить незначне коливання GPU-пам'яті в межах 812-910 МБ при розширенні бази з 50 до 1000 користувачів (табл. 3).

Отримані результати підтвердили, що гібридизація на рівні вилучення ознак перевершує окреме використання будь-якого з методів (табл. 1, 2). Це пояснюється взаємодоповненням статистичних та нейромережових представлень: явні спектральні характеристики та глибинні закономірності разом формують більш інформативний опис голосового сигналу мовця, ніж кожен підхід окремо. Перевага гібридних методів зростає зі збільшенням міжсесійної варіативності, що є важливим для практичного застосування систем голосової ідентифікації в біометричних комплексах аутентифікації та контролю доступу.

Застосування запропонованих рішень має обмеження: нейромережовий компонент потребує GPU-прискорення, а при розширенні бази до 1000 користувачів час обробки зростає до 390-450 мс (табл. 3), що важливе для високонавантажених сервісів реального часу.

Крім того, у роботі не досліджено вплив викривлень, зокрема, каналних завад (наприклад, при стисненні сигналу в телефонних мережах). Подальші дослідження передбачають впровадження алгоритмів швидкого векторного пошуку для масштабування бази мовців та стисненні ШНМ для її використання без GPU.

5. Висновки

1. Розроблено три підходи гібридизації статистичного (MFCC) та нейромережового (VGGish) методів вилучення ознак: на основі

конкатенації, зваженої комбінації (з PCA) та злиття через навчений шар (FC-fusion). На відміну від послідовного аналізу мовного сигналу, обидва джерела ознак використовуються одночасно у спільному векторному просторі. Гібридизація через конкатенацію не потребує навчання, зважена комбінація зберігає компактність вектора, а злиття через проєкційний шар забезпечує адаптацію для врахування нелінійних взаємозв'язків.

2. Визначено критерій та показники якості ідентифікації мовця на основі метрик FAR, FRR та інтегрального показника EER у режимах closed-set й open-set. За критерій якості обрано мінімізацію сукупного рівня помилок EER за умови дотримання часового регламенту обробки ($T_{\max} \leq T_{\text{проц}}$).

3. Оцінено точність ідентифікації в режимах роботи closed-set та open-set. Гібридизація методів вилучення ознак демонструє вищу точність, ніж при ізольованому використанні через взаємодоповнення спектральних та глибинних ознак. У режимі роботи closed-set найкращий результат ідентифікації забезпечує злиття через проєкційний шар (EER 3.3%, точність 95.9%), а в режимі open-set при 5 сесіях у профілі – гібридизація на рівні оцінок (EER 2.7%, точність 97.2%).

4. Досліджено обчислювальну ефективність та масштабованість роботи при розширенні банку мовців до 1000 осіб. Гібридизація на рівні оцінок збільшила час обробки лише на 15-25% у порівнянні з методом на основі VGGish при сталому споживанні GPU-пам'яті (до 910 МБ).

Конфлікт інтересів

Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Фінансування

Дослідження проводилося без фінансової підтримки.

Доступність даних

Рукопис не має пов'язаних даних.

Використання засобів штучного інтелекту

Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

СПИСОК ЛІТЕРАТУРИ

1. Chung, J.S., Huh, J., Mun, S., Lee, M., Heo, H.-S., Choe, S., Ham, C., Jung, S., Lee, B.-J., and Han, I. (2020), "In Defence of Metric Learning for Speaker Recognition", *Interspeech 2020*, Shanghai, China, pp. 2977–2981, doi: <https://doi.org/10.21437/Interspeech.2020-1064>
2. Chen, Z. et al. (2022), "Large-Scale Self-Supervised Speech Representation Learning for Automatic Speaker Verification", *ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, pp. 6147–6151, doi: <https://doi.org/10.1109/ICASSP43922.2022.9747814>
3. Ramzan, H.A., Zahra, S., Ramzan, S., Musa, S.M.A., Kalsum, T., and Jalil, R. (2024), "Speaker Identification Using Convolutional Neural Networks and Mel-Frequency Cepstral Coefficients", *2024 5th International Conference on Innovative Computing (ICIC)*, Lahore, Pakistan, pp. 1–6, doi: <https://doi.org/10.1109/ICIC63915.2024.11116334>

4. Hershey, S., Chaudhuri, S., Ellis, D.P.W., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., Slaney, M., Weiss, R.J., and Wilson, K. (2017), "CNN architectures for large-scale audio classification", *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, pp. 131–135, doi: <https://doi.org/10.1109/ICASSP.2017.7952132>
5. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018), "X-Vectors: Robust DNN Embeddings for Speaker Recognition", *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, pp. 5329–5333, doi: <https://doi.org/10.1109/ICASSP.2018.8461375>
6. Aliaskar, M., Mazakov, T., Mazakova, A., Jomartova, S., and Shormanov, T. (2022), "Human voice identification based on the detection of fundamental harmonics", *2022 IEEE 7th International Energy Conference (ENERGYCON)*, Riga, Latvia, pp. 1–4, doi: <https://doi.org/10.1109/ENERGYCON53164.2022.9830471>
7. Okabe, K., Koshinaka, T., and Shinoda, K. (2018), "Attentive Statistics Pooling for Deep Speaker Embedding", *Interspeech 2018*, Hyderabad, India, pp. 2252–2256, doi: <https://doi.org/10.21437/Interspeech.2018-993>
8. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019), "PyTorch: An Imperative Style, High-Performance Deep Learning Library", *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pp. 8024–8035, URL: <https://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library>
9. Snyder, D., Garcia-Romero, D., Povey, D., and Khudanpur, S. (2017), "Deep Neural Network Embeddings for Text-Independent Speaker Verification", *Interspeech 2017*, Stockholm, Sweden, pp. 999–1003, doi: <https://doi.org/10.21437/Interspeech.2017-620>
10. Бондаренко М.Є., Іващенко Г.С. Багатовекторна голосова ідентифікація користувача з урахуванням фонетичних варіацій голосу. *Системи управління, навігації та зв'язку*. Полтава: ПНТУ, 2026. Вип. 1 (83). С. 40–44. doi: <https://doi.org/10.26906/SUNZ.2026.1.040>
11. Бондаренко М.Є., Іващенко Г.С. Методи оновлення профіля користувача в системах багатовекторної голосової ідентифікації. *Вісник Хмельницького національного університету. Технічні науки*. 2026. № 2 (363). С. 371–378. doi: <https://doi.org/10.31891/2307-5732-2026-363-51>
12. Бондаренко М.Є., Іващенко Г.С. Використання послідовності методів попередньої обробки в системах голосової ідентифікації. *Системи управління, навігації та зв'язку*. Полтава: ПНТУ, 2025. Вип. 2 (80). С. 90–96. doi: <https://doi.org/10.26906/SUNZ.2025.2.090>
13. Nassif, A.B., Shahin, I., Attili, I., Azzeh, M., and Shaalan, K. (2019), "Speech Recognition Using Deep Neural Networks: A Systematic Review", *IEEE Access*, vol. 7, pp. 19143–19165, doi: <https://doi.org/10.1109/ACCESS.2019.2896880>
14. Jahangir, R., Teh, Y.W., Memon, N.A., Mujtaba, G., Zareei, M., Ishtiaq, U., Akhtar, M.Z., and Ali, I. (2020), "Text-Independent Speaker Identification Through Feature Fusion and Deep Neural Network", *IEEE Access*, vol. 8, pp. 32187–32202, doi: <https://doi.org/10.1109/ACCESS.2020.2973541>
15. Bora, B., Emanet, A.E., Elmacı, E., Kandaz, D., and Uçar, M.K. (2023), "Hybrid AI-based Voice Authentication", *Turkish Journal of Forecasting*, vol. 7, no. 2, pp. 17–22, doi: <https://doi.org/10.34110/forecasting.1260073>
16. Al-Tekreeti, N., and Ibrahim, A. A. (2020), "Speaker Voice Recognition Using A Hybrid PSO/Fuzzy Logic System", *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, Istanbul, Turkey, pp. 1–5, doi: <https://doi.org/10.1109/ISMSIT50672.2020.9254309>
17. Shahzad, K., Farhan, S., Yasin-Ul-Haq, Sana, R., and Pathan, S. (2025), "Enhancing Voice Spoofing Detection: A Hybrid Approach With VGGish-LSTM Model for Improved Security in Automatic Speaker Verification Systems", *IEEE Access*, vol. 13, pp. 40682–40702, doi: <https://doi.org/10.1109/ACCESS.2025.3544639>
18. Hasan Abdulqader, A., Al-Haddad, S.A.R., Abdo, S., Abdulghani, A., and Natarajan, S. (2022), "Hybrid Feature Extraction MFCC and Feature Selection CNN for Speaker Identification Using CNN: A Comparative Study", *2022 2nd International Conference on Emerging Smart Technologies and Applications (eSmarTA)*, Marib, Yemen, doi: <https://doi.org/10.1109/eSmarTA56775.2022.9935422>
19. Shen, Q., Guo, M., and Huang, Y. et al. (2024), "Attentional multi-feature fusion for spoofing-aware speaker verification", *International Journal of Speech Technology*, vol. 27, pp. 377–387, doi: <https://doi.org/10.1007/s10772-024-10112-w>
20. Shah, Z., Jang, G., and Farooq, A. (2024), "Feature Fusion for Performance Enhancement of Text Independent Speaker Identification", *ICCK Transactions on Intelligent Systematics*, vol. 2, no. 1, pp. 27–37, doi: <https://doi.org/10.62762/TIS.2024.649374>
21. Karthikeyan, V., Priyadharsini, S.S., and Balamurugan, K. (2025), "Attention-based multi dimension fused-feature convolutional neural network framework for speaker recognition", *Multimedia Tools and Applications*, vol. 84, pp. 38927–38957, doi: <https://doi.org/10.1007/s11042-025-20694-5>
22. Variani, E., Lei, X., McDermott, E., Lopez-Moreno, I., and Gonzalez-Dominguez, J. (2014), "Deep neural networks for small footprint text-dependent speaker verification", *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, pp. 4052–4056, doi: <https://doi.org/10.1109/ICASSP.2014.6854363>
23. Todisco, M., Wang, X., Vestman, V., Sahidullah, Md., Delgado, H., Nautsch, A., Yamagishi, J., Evans, N., Kinnunen, T., and Lee, K. A. (2019), "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection", *Interspeech 2019*, Graz, Austria, pp. 1008–1012, doi: <https://doi.org/10.21437/Interspeech.2019-2249>
24. Tak, H., Patino, J., Todisco, M., Nautsch, A., Evans, N., and Larcher, A. (2021), "End-to-End anti-spoofing with RawNet2", *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, pp. 6369–6373, doi: <https://doi.org/10.1109/ICASSP39728.2021.9414234>
25. Semenov, S., Davydov, V., and Grynov, D. (2024), "Study of Machine Learning Methods for Voice Biometric Identification in Data Protection Systems", *2024 14th International Conference on Dependable Systems, Services and Technologies (DESSERT)*, Athens, Greece, pp. 1–6, doi: <https://doi.org/10.1109/DESSERT65323.2024.11122211>
26. Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. (2021), "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, doi: <https://doi.org/10.1109/TASLP.2021.3122291>

27. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., and Liu, W. (2018), "CosFace: Large Margin Cosine Loss for Deep Face Recognition", *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, pp. 5265–5274, doi: <https://doi.org/10.1109/CVPR.2018.00552>
28. Zeghidour, N., Teboul, O., de Chaumont Quitry, F., and Tagliasacchi, M. (2021), "LEAF: A Learnable Frontend for Audio Classification", *International Conference on Learning Representations (ICLR 2021)*, 16 p., available at: <https://openreview.net/pdf?id=jM76BCb6F9m>
29. Desplanques, B., Thienpondt, J., and Demuynck, K. (2020), "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification", *Interspeech 2020*, Shanghai, China, pp. 3830–3834, doi: <https://doi.org/10.21437/Interspeech.2020-2650>
30. Kurada, S., and Kurada, A. (2020), "Poster: VGGish Embeddings Based Audio Classifiers to Improve Parkinson's Disease Diagnosis", *2020 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, Crystal City, VA, USA, pp. 9–11, doi: <https://doi.org/10.1145/3384420.3431775>
31. Nagrani, A., Chung, J. S., and Zisserman, A. (2017), "VoxCeleb: A Large-Scale Speaker Identification Dataset", *Interspeech 2017*, Stockholm, Sweden, pp. 2616–2620, doi: <https://doi.org/10.21437/Interspeech.2017-950>
32. Chung, J.S., Nagrani, A., and Zisserman, A. (2018), "VoxCeleb2: Deep Speaker Recognition", *Interspeech 2018*, Hyderabad, India, pp. 1086–1090, doi: <https://doi.org/10.21437/Interspeech.2018-1929>

Received (Надійшла) 30.06.2026

Accepted for publication (Прийнята до друку) 26.08.2026

Publication date (Дата публікації) 25.09.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Івашенко Георгій Станіславович – кандидат технічних наук, доцент, доцент кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;

Heorhii Ivashchenko – Candidate of Technical Sciences, Associate Professor at the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;

e-mail: heorhii.ivashchenko@nure.ua; ORCID Author ID: <http://orcid.org/0000-0003-1027-5262>;

Scopus Author ID: <https://www.scopus.com/authid/detail.uri?authorId=57217030807>.

Бондаренко Максим Едуардович – аспірант кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;

Maksym Bondarenko – PhD student at the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;

e-mail: maksym.bondarenko@nure.ua; ORCID Author ID: <http://orcid.org/0000-0002-2500-7626>;

Scopus Author ID: <https://www.scopus.com/authid/detail.uri?authorId=57216159508>.

Hybridization of feature extraction methods in a user voice identification system

Maksym Bondarenko, Heorhii Ivashchenko

Abstract. Background. The practical application of voice identification systems necessitates meeting exclusive criteria: identification accuracy, computational efficiency and resistance to synthesised voice forgeries. Statistical approaches, such as acoustic analysis and spectral comparison, place low demands on hardware resources but are inferior in accuracy to modern deep learning architectures. At the same time, deep neural networks, despite their higher accuracy, significantly increase computational costs and sensitivity to distortions. One possible way of reconciling these requirements is to hybridise methods at the feature extraction stage, which involves combining statistical acoustic analysis and a neural network approach within a single vector space. **Object of the study:** the hybridisation of feature extraction methods in a voice identification system based on the combination of statistical acoustic analysis and a neural network. **Subject of the study:** the hybrid feature extraction method that combines MFCC vectors and neural network feature vectors within a single vector space, with a comparison of different ways of merging them. **Mera Objective:** to develop and investigate a hybrid feature extraction method that combines MFCC vectors and neural network feature vectors in a single vector space, with a view to improving the accuracy of voice identification. **Research tasks** are to implement the proposed approaches for hybridising statistical and neural network features within a single vector space; to experimentally evaluate the identification accuracy of the proposed variants in closed-set and open-set modes; to investigate the computational efficiency and scalability of the proposed variants. **Research methods.** Statistical acoustic analysis based on mel-frequency cepstral coefficients, a convolutional neural network based on the VGGish architecture, the principal component analysis method for dimensionality reduction, training of the projection layer using the angular deviation criterion, and an experimental comparison of the results obtained on the VoxCeleb1 dataset. **Results.** Three variants of hybridisation at the feature extraction stage were proposed and investigated, namely concatenation of normalised vectors, a weighted linear combination following dimensionality reduction, and merging via a trained projection layer. Results demonstrate a reduction in EER and an improvement in identification accuracy compared with the use of each method individually. **Conclusions.** The results of the study confirm the feasibility of applying method hybridisation at the feature extraction stage to improve speaker identification accuracy. It has been established that merging via a trained projection layer provides the highest accuracy, whilst concatenation of normalised vectors is a practically effective solution without the need for additional training.

Keywords: biometric systems of voice identification; hybridisation of methods; feature extraction; MFCC; artificial neural networks; VGG-style convolutional network; projection layer training; feature fusion.