КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 3 (31), 2025
ISSN 2663 - 4023
DOI 10.28925/2663-4023.2025.31.1042
УДК 004.932.72:004.421.2

Руда Христина Степанівна
доктор філософії, доцент кафедри захисту інформації
Національний Університет "Львівська політехніка", Львів, Україна
ORCID: 0000-0001-8644-411X
khrystyna.s.ruda@lpnu.ua

Кос Ігор Іванович
студент кафедри захисту інформації
Національний Університет "Львівська політехніка", Львів, Україна
ORCID: 0009-0001-4558-5036
ihor.kos.mkbas.2025@lpnu.ua

Ахмедова Аліна Самедівна
студентка кафедри захисту інформації
Національний Університет "Львівська політехніка", Львів, Україна
ORCID: 0009-0005-4743-7140
alina.akhmedova.mkbas.2025@lpnu.ua

ОЦІНЮВАННЯ МАСШТАБОВАНОСТІ ГОЛОСОВИХ ЕМБЕДДІНГ-МОДЕЛЕЙ У БІОМЕТРИЧНИХ СИСТЕМАХ ВЕРИФІКАЦІЇ МОВЦЯ

Анотація. Стрімке розгортання цифрових платформ у фінансовому секторі, державному управлінні, електронній комерції та сервісних системах створює потребу у високонадійних та масштабованих засобах автентифікації користувачів. На цьому тлі біометричні технології, зокрема системи голосової автентифікації, демонструють значний потенціал завдяки поєднанню природної зручності взаємодії, мінімальних вимог до обладнання та можливості безперервної інтеграції у голосові інтерфейси. Однак швидке зростання числа користувачів і різноманіття сценаріїв використання формують нові виклики для дослідників і розробників. Сучасні системи мають забезпечувати високу точність у режимі реального часу, підтримувати стабільність роботи при збільшенні обсягів даних і гарантувати стійкість до кібератак, включаючи атаки із застосуванням синтетичної або модифікованої мови. Особливе значення набуває здатність моделей формувати компактні, інваріантні та робастні голосові ембеддінги, які забезпечують ефективне порівняння та класифікацію у великих базах даних. У статті проведено порівняльний аналіз масштабованості сучасних нейронних архітектур для задачі верифікації мовця, з акцентом на їх продуктивності, обчислювальній складності та поведінці при збільшенні кількості зареєстрованих користувачів. Розглянуто підходи до оптимізації моделей, методи індексованого пошуку ембеддінгів, а також роль репрезентативних багатомовних корпусів у підвищенні точності в умовах акустичної та мовної варіативності. Окрему увагу приділено питанням захисту від spoofing-атак та використанню спеціалізованих методів детекції синтетичної мови як невід'ємної складової масштабованих систем голосової біометрії. Отримані результати підкреслюють необхідність комплексного підходу до побудови сучасних систем голосової автентифікації, де інженерні рішення щодо архітектури поєднуються з вимогами інформаційної безпеки, високої продуктивності та здатності до адаптації в умовах динамічного зростання цифрових сервісів.

Ключові слова: голосова біометрія; масштабованість; верифікація мовця; ембеддінги; автентифікація; ECAPA-TDNN; Pyannote; WavLM.

ВСТУП

Масштабованість біометричної системи визначається її здатністю зберігати ефективність зі зростанням кількості користувачів та обсягу даних. Біометричні технології, зокрема розпізнавання відбитків пальців, обличчя, голосу та райдужки ока,

повинні забезпечувати високу точність і швидкодію навіть у випадках, коли система оперує мільйонами зареєстрованих шаблонів. Це особливо актуально для національних систем ідентифікації чи голосових асистентів, які щоденно обслуговують великі масиви користувачів, і продуктивність їхньої роботи не повинна знижуватися зі зростанням навантаження. Масштабованість у цьому контексті передбачає можливість безперервного додавання нових користувачів без необхідності перебудови системи та збереження стабільного часу відповіді.

Реалізація подібних вимог стає можливою завдяки використанню оптимізованих алгоритмів пошуку та індексування, методів розподілених обчислень і хмарної інфраструктури. Водночас розгортання біометрії на великі масиви даних супроводжується низкою викликів. Передусім, зі збільшенням бази користувачів суттєво зростають обчислювальні витрати, оскільки задача ідентифікації у форматі 1:N вимагає мільйонів порівнянь, що без оптимізації робить час відповіді неприйнятним. Не менш критичною є проблема затримок і навантаження, адже автентифікація повинна здійснюватися в реальному часі навіть у масових сценаріях, таких як робота кол-центрів чи прикордонний контроль. Додатково зростають вимоги до зберігання й передачі даних: необхідно ефективно оперувати великими обсягами біометричних шаблонів (відбитків, образів, ембеддінгів), забезпечуючи їхню компактність і швидкий доступ. Нарешті, важливою залишається проблема збереження точності в умовах збільшення кількості користувачів та появи схожих біометричних зразків, а також за наявності варіацій у даних одного й того ж користувача, зумовлених змінами середовища запису, шумом чи віковими факторами.

Таким чином, масштабованість біометричних систем є комплексною характеристикою, що поєднує продуктивність, точність та безпеку, і потребує застосування стійких до варіацій моделей, здатних функціонувати в умовах реального часу при дотриманні високих вимог до захисту даних [1].

Постановка проблеми. У цій частині статті описується проблема, розгляду якої присвячене дослідження, у загальному вигляді та її зв'язок із важливими науковими чи практичними завданнями.

Аналіз останніх досліджень і публікацій. Аналіз наукових джерел останніх років засвідчує значне зростання інтересу до масштабованих рішень у сфері голосової біометрії. Дослідження зосереджені на розвитку ембеддінг-моделей, оптимізації архітектур і прикладному використанні технологій у банківських сервісах, кол-центрах та голосових асистентах.

У статті [2] розглянуто масштабованість системи голосової автентифікації на базі архітектури TitaNet. Авторка встановила, що ефективність моделі зберігається на високому рівні при невеликих базах даних, проте поступово знижується зі збільшенням кількості користувачів. Це свідчить про необхідність адаптивного налаштування порогових значень та використання додаткових методів підвищення точності. Таким чином, навіть передові архітектури демонструють певні обмеження масштабованості, що визначає потребу у їх подальшій оптимізації.

Більш комплексний порівняльний підхід представили Brydinskyi et al. (2024) [3], які проаналізували низку сучасних моделей для верифікації мовця, серед яких ECAPA, TitaNet, WavLM та PyAnnote. Автори дійшли висновку, що технологія ембеддінгів мовця має високу масштабованість, оскільки нових користувачів можна додавати без повторного навчання моделі. Особливу увагу дослідники звернули на архітектуру ECAPA, яка продемонструвала збалансоване поєднання точності (EER близько 1,7 %) та

швидкодії (інференс ~69 мс), що робить її перспективною для використання у великомасштабних системах автентифікації.

У продовження теми оптимізації, Thienpondt і Demuynck (2023) [4] запропонували гібридну архітектуру ECAPA2, яка забезпечує підвищену стійкість ембеддінгів мовця до шуму та коротких уривків мовлення. Модель продемонструвала state-of-the-art результати на наборі VoxCeleb1 за меншої кількості параметрів порівняно з попередніми рішеннями. Це свідчить про ефективність підходів, спрямованих на зменшення обчислювальної складності без погіршення продуктивності, що є важливим чинником при масштабуванні систем.

Ще одним актуальним завданням є ідентифікації мовця за короткими фрагментами мовлення. Deng et al. (2025) [5] представили архітектуру Dense-Fusion2Net з модулем TFCA (Time-Frequency Channel Attention), яка дозволяє ефективно працювати з обмеженою тривалістю сигналів. Модель довела свою результативність на наборах VoxCeleb, зберігаючи високу точність і низьке обчислювальне навантаження. Це забезпечує перспективність її застосування у банківських та сервісних системах, де користувачі взаємодіють із системою протягом лічених секунд.

Більш ширший історичний контекст дослідження розвитку технологій подали Sharma et al. (2024) [6]. Автори простежили еволюцію від класичних підходів на основі i-vector до сучасних нейромережевих ембеддінг-моделей, підкресливши, що саме поява x-vector та подальших глибоких архітектур забезпечила перехід до масштабованих систем. В огляді відзначено, що сучасні методи відкрили можливість роботи в умовах відкритих множин користувачів, де база постійно розширюється, що відповідає потребам реальних сценаріїв застосування.

Проблема безпеки масштабованих систем досліджувалася у роботі Chen et al. (2023) [7], де вивчали спроби обману системи розпізнавання мовця спеціально змінених аудіосигналів, так званих аудіоадверсаріальних прикладів. Автори показали, що поєднання спеціальних перетворень звукового сигналу з методами навчання на прикладах атак (adversarial training) дозволяє підвищити точність приблизно на 13,6 % і суттєво ускладнює проведення атак. Це створює підґрунтя для розробки більш захищених і масштабованих біометричних систем.

Окремим напрямом безпеки масштабованих систем є приватність та гетерогенні умови навчання. Chen & Xu (2023) [8] запропонували підхід personalized federated learning (PFL) для задач верифікації та ідентифікації мовця. Їхня система дала змогу одночасно зберегти приватність даних користувачів, забезпечити адаптацію до різних акустичних доменів (кімнати, шуми, мови) та уникнути катастрофічного забування у безперервному навчанні (C-PFL). У 12 змодельованих сценаріях такий підхід показав нижчий середній EER і кращу збіжність, ніж централізоване навчання, що підкреслює його перспективність для масштабованих реальних систем.

Практичну площину масштабування демонструють результати індустріальних досліджень. Так, у проєкті RudderAnalytics (2024) [9] було впроваджено систему на базі SpeechBrain із використанням набору VoxCeleb2. Розробники показали, що система зберігає високу точність при масштабуванні, успішно витримавши зростання кількості зареєстрованих користувачів на 115 % без втрати продуктивності.

Водночас навіть класичні архітектури залишаються предметом оптимізації. Sharif-Noughabi et al. (2025) [10] показали, що модифікована VGG-CNN у поєднанні з методами розширення даних здатна суттєво підвищити точність ідентифікації на наборі VoxCeleb1 (з ~84 % до понад 91 %). Це свідчить про ефективність інтеграції класичних підходів з

сучасними техніками навчання для забезпечення продуктивності у великих базах користувачів.

Отже, аналіз сучасних публікацій свідчить, що масштабованість голосових біометричних систем спирається на компактні та робастні ембеддінги, легкі архітектури, методи квантування й рішення для коротких записів, доповнені механізмами безпеки та приватності. Сучасні нейромережеві моделі рухаються до синтезу високої точності, оптимальної швидкодії та стійкості до атак, відкриваючи шлях до широкого впровадження у банківських сервісах, кол-центрах і голосових асистентах.

Мета статті. Порівняльно оцінити масштабованість сучасних моделей голосової верифікації (ECAPA-TDNN, Pyannote, WavLM base-sv, WavLM base-plus-sv) у міру зростання кількості користувачів (10–70), вимірявши точність, FAR, FRR та EER, і на основі результатів дати практичні рекомендації для побудови масштабованих систем (компактні ембеддінги, індексований пошук, квантування, дистиляція знань).

Для досягнення даної мети потрібно:

- Проаналізувати сучасні нейронні архітектури верифікації мовця з точки зору їхньої масштабованості.
- Сформувати багатомовний датасет та підготувати еталонні ембеддінги для кожної моделі.
- Реалізувати процедуру порогової верифікації та визначити оптимальні пороги для різних масштабів системи.
- Провести експериментальні вимірювання точності, FAR, FRR та EER для наборів 10–70 користувачів.
- Порівняти поведінку моделей при зростанні кількості користувачів та визначити їхню стійкість до масштабування.
- Сформувати рекомендації щодо побудови масштабованих систем голосової автентифікації на основі отриманих результатів.

ТЕОРЕТИЧНІ ОСНОВИ ДОСЛІДЖЕННЯ

Найточніші нейромережі часто дуже великі, що ускладнює їх застосування на пристроях з обмеженими ресурсами або при масовому доступі в хмарі. Для подолання обмежень, пов’язаних із великою обчислювальною складністю моделей, застосовують методи оптимізації, зокрема квантизацію — зменшення розмірності нейронних мереж. Наприклад, компанія Amazon зменшувала модель голосового помічника Alexa до менш ніж 1% початкового розміру без істотної втрати точності [11]. Це стало можливим завдяки квантуванню параметрів та дистиляції знань, коли невелика «учнівська» модель навчається на виходах великої «вчительської». Такі оптимізовані моделі можуть запускатись на смартфонах, автомобільних системах та мікроконтролерах. Водночас для ідентифікації мовця важливий швидкий пошук по базі ембеддінгів – замість повного перебору використовують індекси, що дозволяє масштабувати бази до десятків мільйонів шаблонів [12].

Одним із критичних аспектів у розробці та розгортанні масштабованих систем, особливо у сфері штучного інтелекту та машинного навчання, є дані для навчання. Якість, обсяг та репрезентативність цих даних безпосередньо впливають на ефективність і адаптивність моделей. Щоб забезпечити високу точність та надійність функціонування моделей у різноманітних умовах, необхідно використовувати значні за обсягом і різноманітні за складом навчальні набори.

Для того, щоб моделі могли ефективно функціонувати з різними акцентами, мовами та серед широких груп користувачів, вони мають бути навчені на репрезентативних вибірках даних. Це означає, що навчальний набір даних повинен відображати все розмаїття вхідних даних, з якими модель стикатиметься у реальному світі. Недостатня репрезентативність може призвести до упереджень у роботі моделі, зниження точності для певних груп користувачів або варіацій у мовленні. Наприклад, у голосових службах, якщо модель навчена переважно на даних однієї демографічної групи, вона може демонструвати значно гірші показники розпізнавання мови для інших груп, що матимуть відмінні акценти, тембр голосу або словниковий запас. Тому інвестування у збір та анотування великих, збалансованих і різноманітних навчальних даних є фундаментальним для досягнення масштабованості та універсальності систем штучного інтелекту.

Окрім даних для навчання, аспект масштабованості нерозривно пов'язаний із надійністю системи. У сучасних високодоступних сервісах, таких як банківські голосові служби або побутові голосові асистенти, моделі штучного інтелекту функціонують у режимі 24 години та 7 діб на тиждень. Це вимагає розгортання моделей у кластерних середовищах, де навантаження ефективно розподіляється між численними вузлами. Така архітектура забезпечує неперервну роботу системи та стабільну відповідь навіть у періоди пікових навантажень.

Розподіл навантаження між вузлами є одним із ключових механізмів ефективного функціонування сучасних інформаційних систем. Він дозволяє уникати перевантаження окремих компонентів, рівномірно розподіляючи запити та обчислювальні задачі між усіма доступними ресурсами. Завдяки цьому забезпечується не лише стабільність, але й оптимальна продуктивність системи в умовах високих навантажень. Особливо важливою є здатність системи автоматично реагувати на відмови окремих вузлів: у разі їх виходу з ладу, інші вузли можуть оперативно перебрати на себе виконання необхідних функцій. Такий підхід гарантує безперервність надання послуг, підвищує загальну відмовостійкість архітектури та мінімізує ризики, пов’язані з простоєм. Це критично важливо для онлайн-сервісів, фінансових платформ, медичних систем та інших сфер, де навіть короткочасна затримка або технічний збій може призвести до значних фінансових втрат або втрати довіри з боку користувачів.Таким чином, поняття масштабованості системи охоплює не лише технічну здатність обробляти дедалі більший обсяг даних і зростаючу кількість запитів, але й включає в себе забезпечення високої надійності, стійкості до збоїв, а також підтримку постійної доступності сервісів у реальних умовах експлуатації.

ТЕОРЕТИЧНІ ОСНОВИ ДОСЛІДЖЕННЯ

Для проведення експериментального дослідження масштабованості нейронних моделей у системах голосової біометричної автентифікації було сформовано багатомовний аудіодатасет, що включає україномовних та англомовних мовців. Україномовна частина бази даних складалася із записів голосів 50 публічних осіб (переважно політиків, урядовців та комунікаторів), отриманих із відкритих джерел. Як основне джерело використано публічний датасет на платформі Hugging Face [13], який містив структуровані аудіофайли українською мовою належної якості для проведення дослідження. Кожен мовець був представлений 10 записами тривалістю близько 10 секунд, що забезпечило достатній обсяг даних для побудови еталонних ембеддінгів.

З метою оцінки масштабованості моделей в умовах мовної варіативності датасет було доповнено записами 20 англомовних мовців, серед яких відомі актори, журналісти та публічні особи. Аудіоматеріали збиралися вручну з відкритих джерел, насамперед із відеоінтерв'ю, опублікованих на платформі YouTube.

Для уникнення перетину навчальних і тестових прикладів було сформовано розширений набір даних. Зокрема, для кожного з 70 мовців було додатково підготовлено по 20 унікальних аудіозаписів, які не дублювали зразки, використані для формування еталонних ембеддінгів. Це дозволило забезпечити коректність оцінки точності верифікації та обґрунтованість експериментальних результатів.

На рисунку 1 зображено схему роботи використаної системи автентифікації. Для першого етапу дослідження роботи системи усі файли одного мовця опрацьовувались кожною з обраних нейронних мереж: Pyannote, WavLM base-sv, WavLM base-plus-sv та ECAPA. У результаті генерувались 10 ембеддінгів, які усереднювались в один вектор – еталонний ембеддінг мовця. Це усереднення дозволяє зменшити вплив варіацій голосу (інтонації, темпу, шумів). Сформовані еталонні ембеддінги зберігались у базі даних разом з унікальним ідентифікатором мовця.

Для наступної частини дослідження необхідно було визначити оптимальний поріг верифікації. Для цього було змодельовано повноцінний процес автентифікації, який дозволяє оцінити поведінку системи в умовах реального використання. Метою було досягнення збалансованого співвідношення між помилками першого та другого роду (хибно-негативні та хибно-позитивні). У рамках моделювання формувалися пари записів, що охоплювали два сценарії: перевірку валідних користувачів, де тестові зразки порівнювалися з еталонами тих самих осіб, та перевірку невалідних користувачів, у якій тестові ембеддінги зіставлялися з еталонними векторами інших користувачів.

Після створення еталонної бази система переходила до верифікації. Тією ж нейронною моделлю, що й при побудові еталону, опрацьовувався розширений набір голосових даних мовця задля екстракції тестового ембеддінга. Отриманий вектор порівнювався з відповідним еталонним ембеддінгом користувача за допомогою косинусної відстані. Якщо значення косинусної відстані було меншим за заздалегідь встановлений поріг, верифікація вважалась успішною. Інакше – зразок не вважався таким, що належить користувачу.

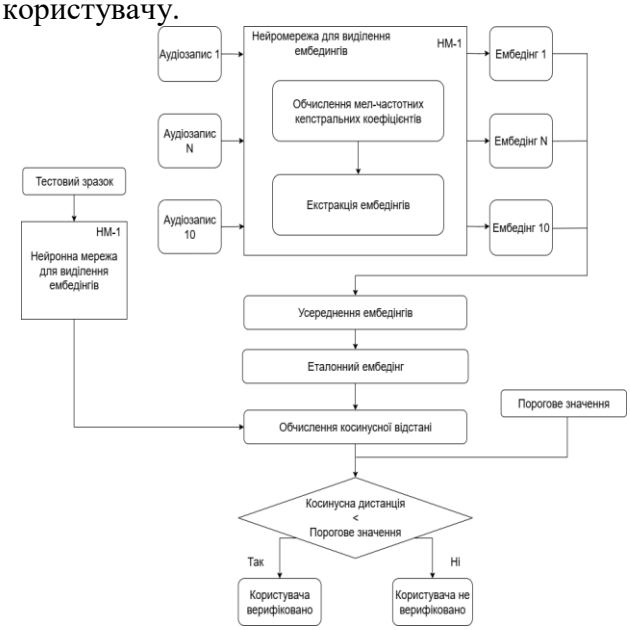

Рис. 1. Схема роботи системи біометричної автентифікації за голосом

Для оцінювання результатів експериментального дослідження було застосовано низку загальноприйнятих метрик, які традиційно використовуються у сфері біометричної верифікації. Найчастіше для кількісного аналізу ефективності систем розглядають показники Accuracy, False Acceptance Rate (FAR), False Rejection Rate (FRR) та інтегральний критерій Equal Error Rate (EER).

Accuracy (точність) є базовим показником, що відображає загальну частку правильно класифікованих випадків у задачі бінарної верифікації мовця. Вона визначається як відношення кількості коректно прийнятих рішень (успішних авторизацій та правильних відмов) до загальної кількості перевірок. Незважаючи на інтуїтивну зрозумілість, дана метрика є чутливою до дисбалансу класів і тому не вважається визначальною у верифікаційних дослідженнях.

False Acceptance Rate (FAR) характеризує рівень безпеки системи та визначає частку випадків, коли неавторизований користувач помилково ідентифікується як легітимний. Формально, FAR обчислюється як відношення кількості хибних прийнять (False Acceptances) до загальної кількості спроб доступу, здійснених сторонніми мовцями. Зниження цього показника є критично важливим для підвищення стійкості системи до атак та забезпечення її надійності.

False Rejection Rate (FRR), навпаки, характеризує зручність користування системою та відображає частку випадків, коли зареєстрований користувач був помилково відхилений як неавторизований. Показник визначається як відношення кількості хибних відмов (False Rejections) до загальної кількості спроб доступу, здійснених легітимними мовцями. Низьке значення FRR гарантує кращий користувацький досвід та вищу доступність системи. Очевидно, що одночасно мінімізувати обидва показники — FAR і FRR — неможливо. Оптимальний баланс між ними досягається шляхом вибору відповідного порогового значення подібності ембеддінгів, яке визначає пріоритет: мінімізацію хибних прийнять чи хибних відмов.

Equal Error Rate (EER) виступає інтегральним критерієм якості біометричної системи, що визначається у точці, де значення FAR та FRR збігаються. Чим нижче значення EER, тим більш ефективною вважається система. Завдяки незалежності від конкретного порогового значення, дана метрика є універсальним показником для порівняння різних алгоритмів біометричної автентифікації [14].

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

У ході дослідження було проведено серію експериментів для п’яти рівнів масштабованості системи біометричної автентифікації за голосом: 10, 20, 35, 50 та 70 зареєстрованих користувачів. Для кожного з рівнів було виконано порівняльне тестування чотирьох нейронних моделей – Pyannote, WavLM base-sv, WavLM base-plus-sv та ECAPA – із подальшим аналізом точності, кількості помилок першого (FAR) і другого (FRR) роду, а також показника EER (Equal Error Rate). У таблиці 1 подано детальні показники EER на усіх етапах експерименту.

Таблиця 1

Показники EER кожної з моделей на усіх етапах експерименту

	10, %	20, %	35, %	50, %	70, %
Pyannote	2.73	4.46	5.40	5.30	4.55
WavLM base-sv	7.42	10.13	11.72	10.90	8.09

WavLM base-plus-sv	9.77	10.12	11.27	9.92	7.47
ECAPA-TDNN	1.17	1.33	2.38	2.14	3.04

З рисунку 2 можна побачити, що незалежно від масштабу найкращі результати демонструє ECAPA-TDNN: EER тримається в межах приблизно 1.2-3.0% і залишається найнижчим на кожному етапі. Pyannote стабільно друга (приблизно 2.7-5.4%), із помірним зростанням помилки при збільшенні кількості користувачів і невеликим покращенням на 70. Обидві версії WavLM мають найвищі EER (приблизно 7-11%). Спостерігається тенденція – з ростом числа спікерів EER зростає, але на найбільшому наборі спостерігається часткова стабілізація.

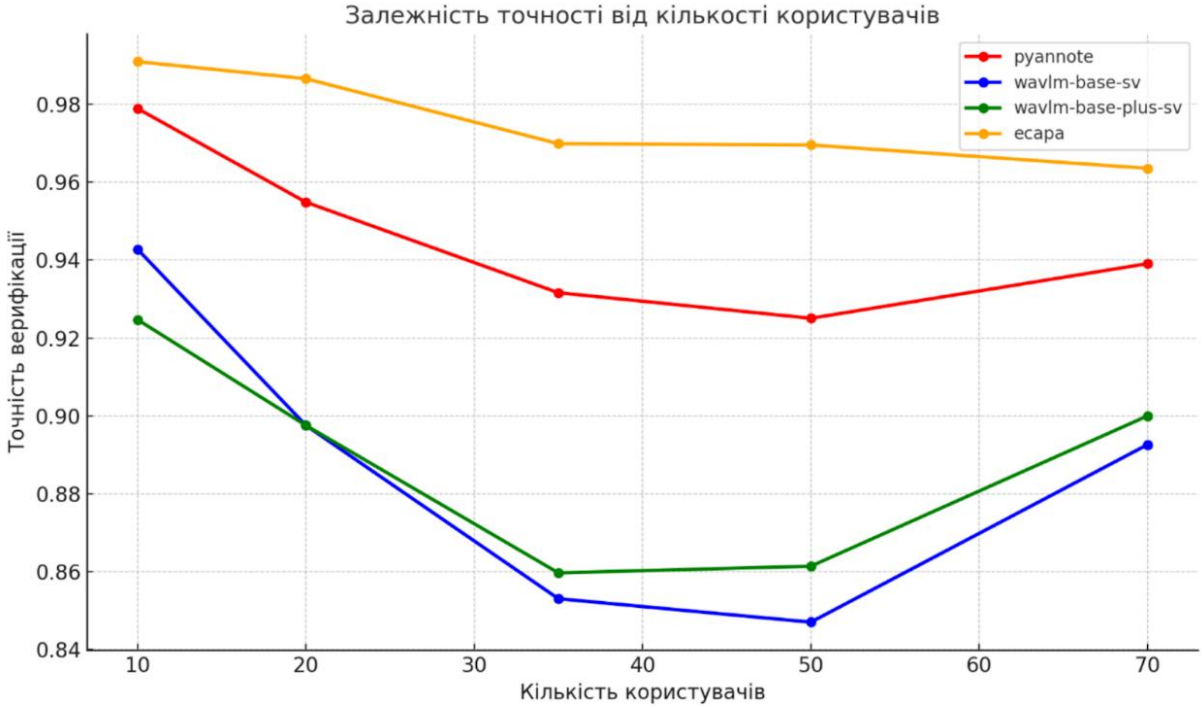

Рис. 2. Графік залежності точності від кількості користувачів

На основі побудованого графіка залежності точності (accuracy) від кількості користувачів встановлено, що модель ECAPA-TDNN демонструє найвищу стабільність та точність у всіх конфігураціях. Її точність коливається в межах 95–98%, при цьому навіть при максимальному масштабуванні (70 користувачів) зберігається високий рівень. Pyannote також показала стабільну точність, лише трохи поступаючись ECAPA-TDNN, з переважно рівномірною деградацією при зростанні кількості користувачів.

Моделі WavLM base-sv та WavLM base-plus-sv демонструють дещо нижчі показники точності. Особливо помітне зниження спостерігається після перевищення межі у 35 користувачів. Це може свідчити про підвищення складності відрізняти ознаки серед дедалі більшої кількості користувачів для цих моделей, що призводить до зростання кількості хибних рішень системи.

Проте варто відзначити цікаву особливість: в останній ітерації (на 70 користувачах) точність моделей Pyannote, WavLM base-sv та WavLM base-plus-sv зросла порівняно з попереднім етапом, що порушує загальну тенденцію до деградації. Причиною цьому можна виокремити більшу кількість мовців, що дало змогу моделям побудувати кращі порогові значення та уникати помилок першого та другого родів.

Загальна тенденція підтверджує класичну закономірність: збільшення кількості користувачів у системі супроводжується складнішим завданням розмежування між голосовими ембеддінгами, що може призводити до зростання кількості помилкових рішень і зниження точності.

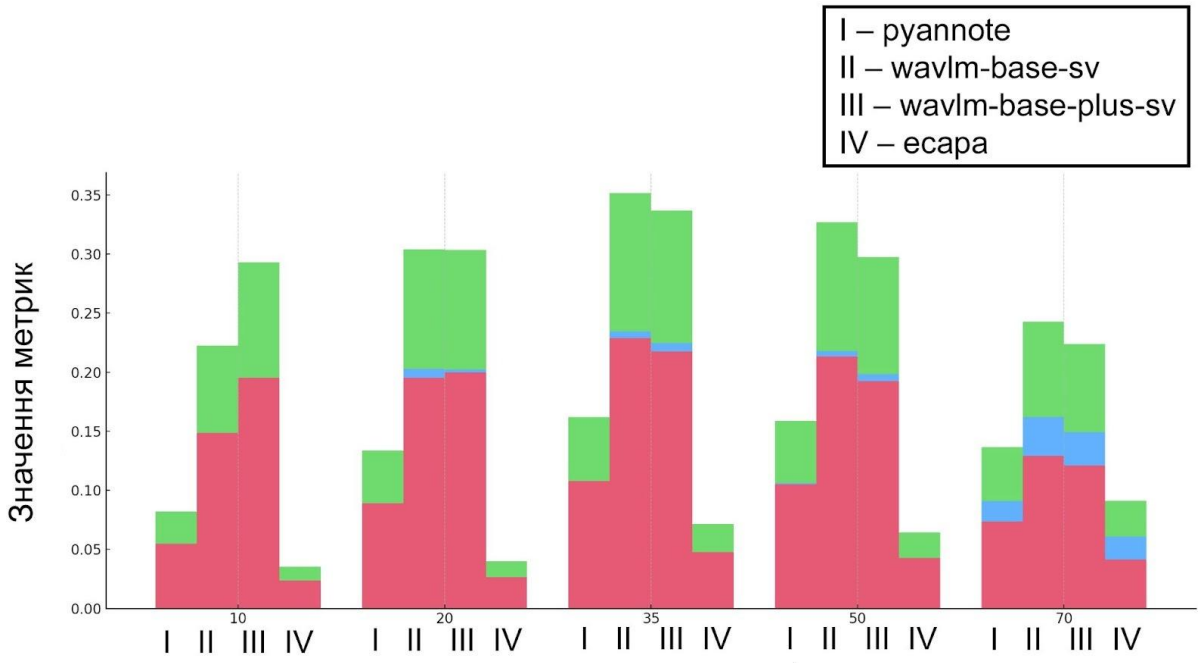

Рис. 3. Графік залежності FAR, FRR та EER від кількості користувачів

На рисунку 3 подано гістограму, побудовану за трьома ключовими метриками – FAR (False Acceptance Rate) – червоного кольору, FRR (False Rejection Rate) – синього кольору та EER (Equal Error Rate) – зеленого кольору. За її допомогою глибше проаналізовано поведінку кожної моделі в умовах масштабування. Отримані результати дозволяють всебічно оцінити масштабованість та робастність різних архітектур у завданні біометричної верифікації мовця.

1. Архітектура ECAPA-TDNN підтвердила свою високу ефективність: вона зберігає мінімальні значення EER на всіх рівнях масштабування, а також демонструє збалансованість між FAR і FRR. Це свідчить про оптимальне калібрування порогів та здатність моделі формувати дискримінативні ембеддінги, стійкі до зростання кількості користувачів. Такі характеристики роблять ECAPA-TDNN найбільш придатною для реального впровадження у системах із високими вимогами до продуктивності та безпеки.

2. Модель Pyannote також показала відносно низькі рівні помилок, однак на високих масштабах (50–70 користувачів) зафіксовано поступове зростання FRR. Це свідчить про певне зниження здатності моделі коректно розпізнавати зареєстрованих користувачів за умов збільшення бази даних. Попри це, загальний рівень помилок залишається прийнятним, що робить Pyannote придатною для застосувань середнього масштабу.

3. На противагу, WavLM base-sv продемонструвала значні обмеження: підвищені значення FAR вказують на підвищену ймовірність помилкового прийняття сторонніх користувачів. Подібна поведінка є критично небезпечною у безпеково-орієнтованих

сценаріях, таких як банківські сервіси чи державні реєстри, де навіть поодиноке хибне прийняття може мати суттєві наслідки.

4. Модифікація WavLM base-plus-sv дещо знижує EER порівняно з базовою версією, однак модель не демонструє стабільності на рівні ECAPA чи Pyannote. При зростанні кількості користувачів спостерігається значне підвищення FRR, що вказує на втрату здатності моделі надійно розпізнавати легітимних мовців. Це суттєво знижує зручність використання системи та може негативно впливати на користувацький досвід.

Загальна тенденція показує, що масштабування бази користувачів не є нейтральним чинником для всіх моделей: у найстійкіших архітектур (як-от ECAPA) EER залишається стабільним або навіть знижується за рахунок кращої узгодженості ембеддінгів, тоді як слабші моделі демонструють синхронне зростання як FAR, так і FRR.

Таким чином, результати підтверджують, що масштабованість біометричних систем безпосередньо залежить від архітектури моделі. Потужні рішення на кшталт ECAPA забезпечують високу толерантність до розширення користувацької бази, тоді як інші архітектури потребують додаткових заходів — оптимізації ембеддінгів, адаптивного вибору порогових значень або навіть повторного донавчання на репрезентативних даних. Це вказує на необхідність не лише ретельного вибору моделі, але й застосування комплексних методів підтримки її масштабованості при розробці реальних систем голосової біометрії.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

У статті проведено експериментальне порівняння чотирьох нейронних моделей голосової автентифікації з метою оцінки їхньої масштабованості. Встановлено, що ECAPA-TDNN демонструє найвищу стабільність та мінімальні значення EER на всіх рівнях кількості користувачів, що підтверджує її придатність для великомасштабних систем реального часу. Модель Pyannote також зберігає прийнятну точність, хоча характеризується поступовим зростанням FRR при збільшенні бази. Архітектури WavLM виявили обмеження масштабованості: base-sv має підвищений рівень FAR, а base-plus-sv — нестабільність показників FRR.

Показано, що для підтримання високої продуктивності та надійності систем доцільним є використання компактних ембеддінгів, індексованого пошуку та методів оптимізації моделей (квантування, дистиляція знань). Практичне значення результатів полягає у можливості застосування отриманих висновків під час розробки масштабованих біометричних сервісів у сферах банківських послуг, кол-центрів та голосових асистентів.

Щодо перспектив подальших досліджень, вони можуть включати розширення масштабу експериментів за межі 70 користувачів та сценаріїв тривалих/шумних записів, щоб перевірити стійкість висновків; дослідити доменну гнучкість (різні мови, акценти, кодеки) і режими із сильним стисненням файлу; поглибити аналіз безпеки; оптимізувати моделі для периферійних пристроїв через дистиляцію та квантування); розширити україномовні бенчмарки й перевірки fairness (стать/вік/акцент); вивчити UX-стратегії зниження FRR без росту FAR у банківських та кол-центрних кейсах. Мотивацією слугують стабільність ECAPA-TDNN та обмеження WavLM, виявлені у роботі, а також поточний діапазон масштабування 10–70 користувачів, який варто розширити.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Biostatistics.io. (n.d.). *Implementing biometrics for large-scale applications: Overcoming 6 challenges*. https://biostatistics.io/qa/implementing-biometrics-for-large-scale-applications-overcoming-6-challenges
2. Ruda, K. (2025). Study of the scalability of biometric authentication systems based on voice embeddings. *Social Development and Security*, *15*(1), 161–170. https://doi.org/10.33445/sds.2025.15.1.15
3. Brydinskyi, V., Khoma, Y., Sabodashko, D., Podpora, M., Khoma, V., Konovalov, A., & Kostiak, M. (2024). Comparison of modern deep learning models for speaker verification. *Applied Sciences*, *14*(4), Article 1329. https://doi.org/10.3390/app14041329
4. Thienpondt, J., & Demuynck, K. (2023). ECAPA2: A hybrid neural network architecture and training strategy for robust speaker embeddings. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 1–8). IEEE. https://doi.org/10.1109/ASRU57964.2023.10389750
5. Deng, F., Huang, R., Jiang, P., & Deng, L. (2025). Dense-Fusion2Net: A more efficient and lightweight short speech speaker recognition system with time-frequency channel attention. *Scientific Reports*, *15*, 9601. https://doi.org/10.1038/s41598-025-93873-x
6. Sharma, R., Govind, D., Mishra, J., Dubey, A. K., Deepak, K. T., & Prasanna, S. R. M. (2024). Milestones in speaker recognition. *Artificial Intelligence Review*, *57*, Article 58. https://doi.org/10.1007/s10462-023-10688-w
7. Chen, G., et al. (2023). Towards understanding and mitigating audio adversarial examples for speaker recognition. *IEEE Transactions on Dependable and Secure Computing*, *20*(5), 3970–3987. https://doi.org/10.1109/TDSC.2022.3220673
8. Chen, Z., & Xu, S. (2023). Learning domain-heterogeneous speaker recognition systems with personalized continual federated learning. *EURASIP Journal on Audio, Speech, and Music Processing*, *2023*, Article 33. https://doi.org/10.1186/s13636-023-00299-2
9. RudderAnalytics. (n.d.). *Building a robust speaker verification system for secure voice authentication*. Medium. https://medium.com/@rudderanalytics/voice-based-security-implementing-a-robust-speaker-verification-system-12c5fd98f1c1
10. Sharif-Noughabi, M., Razavi, S. M., & Mohamadzadeh, S. (2025). Improving the performance of speaker recognition system using optimized VGG convolutional neural network and data augmentation. *International Journal of Engineering*, *38*(10), 2414–2425. https://doi.org/10.5829/ije.2025.38.10a.17
11. Amazon Science Blog. (n.d.). *On-device speech processing makes Alexa faster, lower bandwidth*. https://www.amazon.science/blog/on-device-speech-processing-makes-alexa-faster-lower-bandwidth
12. Google Research. (n.d.). *An overview of speech recognition techniques*. https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/42535.pdf
13. Hugging Face. (2023). *ua-polit-tiny* [Dataset]. https://huggingface.co/datasets/vbrydik/ua-polit-tiny
14. Alice Biometrics. (2023). *Defining the core accuracy metrics of biometric systems*. https://alicebiometrics.com/en/defining-the-core-accuracy-metrics-of-biometric-systems

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 3 (31), 2025
ISSN 2663 - 4023

Ruda Khrystyna
Doctor of Philosophy, Associate Professor, Department of Information Security
Lviv Polytechnic National University, Lviv, Ukraine
ORCID: 0000-0001-8644-411X
khrystyna.s.ruda@lpnu.ua

Kos Ihor
Student, Department of Information Security
Lviv Polytechnic National University, Lviv, Ukraine
ORCID: 0009-0001-4558-5036
ihor.kos.mkbas.2025@lpnu.ua

Akhmedova Alina
Student, Department of Information Security
Lviv Polytechnic National University, Lviv, Ukraine
ORCID: 0009-0005-4743-7140
alina.akhmedova.mkbas.2025@lpnu.ua

EVALUATION OF THE SCALABILITY OF VOICE EMBEDDING MODELS IN BIOMETRIC SPEAKER VERIFICATION SYSTEMS

Abstract. The rapid expansion of digital platforms in the financial sector, public administration, e-commerce, and service systems has created a growing demand for highly reliable and scalable user authentication technologies. In this context, biometric methods — particularly voice-based authentication systems — demonstrate significant potential due to their natural ease of interaction, minimal hardware requirements, and seamless integration into voice-driven interfaces. However, the increasing number of users and the diversity of usage scenarios introduce new challenges for researchers and developers. Modern systems must ensure high accuracy in real time, maintain stable performance as data volumes grow, and provide resilience against cyberattacks, including those involving synthetic or manipulated speech. A critical requirement is the ability of models to generate compact, invariant, and robust voice embeddings that enable efficient comparison and classification within large-scale databases. This paper presents a comparative analysis of the scalability of contemporary neural architectures for speaker verification, with emphasis on their performance, computational complexity, and behavior as the number of enrolled users increases. The study examines model optimization techniques, indexed embedding-based search methods, and the role of representative multilingual corpora in enhancing accuracy under conditions of acoustic and linguistic variability. Particular attention is given to protection against spoofing attacks and the use of specialized synthetic speech detection methods as an essential component of scalable voice biometric systems. The results highlight the need for a comprehensive approach to designing modern voice authentication systems, in which architectural engineering decisions are combined with requirements for information security, high performance, and adaptability in the rapidly evolving landscape of digital services.

Keywords: voice biometrics; scalability; speaker verification; embeddings; authentication; ECAPA-TDNN; Pyannote; WavLM.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Biostatistics.io. (n.d.). *Implementing biometrics for large-scale applications: Overcoming 6 challenges*. https://biostatistics.io/qa/implementing-biometrics-for-large-scale-applications-overcoming-6-challenges
2. Ruda, K. (2025). Study of the scalability of biometric authentication systems based on voice embeddings. *Social Development and Security, 15*(1), 161–170. https://doi.org/10.33445/sds.2025.15.1.15
3. Brydinskyi, V., Khoma, Y., Sabodashko, D., Podpora, M., Khoma, V., Konovalov, A., & Kostiak, M. (2024). Comparison of modern deep learning models for speaker verification. *Applied Sciences, 14*(4), Article 1329. https://doi.org/10.3390/app14041329

4. Thienpondt, J., & Demuynck, K. (2023). ECAPA2: A hybrid neural network architecture and training strategy for robust speaker embeddings. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 1–8). IEEE. https://doi.org/10.1109/ASRU57964.2023.10389750
5. Deng, F., Huang, R., Jiang, P., & Deng, L. (2025). Dense-Fusion2Net: A more efficient and lightweight short speech speaker recognition system with time-frequency channel attention. *Scientific Reports, 15*, 9601. https://doi.org/10.1038/s41598-025-93873-x
6. Sharma, R., Govind, D., Mishra, J., Dubey, A. K., Deepak, K. T., & Prasanna, S. R. M. (2024). Milestones in speaker recognition. *Artificial Intelligence Review, 57*, Article 58. https://doi.org/10.1007/s10462-023-10688-w
7. Chen, G., et al. (2023). Towards understanding and mitigating audio adversarial examples for speaker recognition. *IEEE Transactions on Dependable and Secure Computing*, *20*(5), 3970–3987. https://doi.org/10.1109/TDSC.2022.3220673
8. Chen, Z., & Xu, S. (2023). Learning domain-heterogeneous speaker recognition systems with personalized continual federated learning. *EURASIP Journal on Audio, Speech, and Music Processing, 2023*, Article 33. https://doi.org/10.1186/s13636-023-00299-2
9. RudderAnalytics. (n.d.). *Building a robust speaker verification system for secure voice authentication*. Medium. https://medium.com/@rudderanalytics/voice-based-security-implementing-a-robust-speaker-verification-system-12c5fd98f1c1
10. Sharif-Noughabi, M., Razavi, S. M., & Mohamadzadeh, S. (2025). Improving the performance of speaker recognition system using optimized VGG convolutional neural network and data augmentation. *International Journal of Engineering*, *38*(10), 2414–2425. https://doi.org/10.5829/ije.2025.38.10a.17
11. Amazon Science Blog. (n.d.). *On-device speech processing makes Alexa faster, lower bandwidth*. https://www.amazon.science/blog/on-device-speech-processing-makes-alexa-faster-lower-bandwidth
12. Google Research. (n.d.). *An overview of speech recognition techniques*. https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/42535.pdf
13. Hugging Face. (2023). *ua-polit-tiny* [Dataset]. https://huggingface.co/datasets/vbrydik/ua-polit-tiny
14. Alice Biometrics. (2023). *Defining the core accuracy metrics of biometric systems*. https://alicebiometrics.com/en/defining-the-core-accuracy-metrics-of-biometric-systems