DOI: https://doi.org/10.26565/2304-6201-2025-68-09

УДК (UDC) 004.8:342.9

Трусов Михайло Андрійович *аспірант ННІ «Комп'ютерних наук та штучного інтелекту», Харківський національний університет імені В.Н. Каразіна, майдан Свободи, 4, Харків, Україна, 61022*
e-mail: mykhailo.trusov@karazin.ua
https://orcid.org/0009-0001-4390-5307

Турута Олексій Петрович *к.т.н., доцент, доцент кафедри програмної інженерії, Харківський національний університет радіоелектроніки, просп. Науки 14, Харків, Україна, 61166*
e-mail: oleksii.turuta@gmail.com
https://orcid.org/0000-0002-0970-8617

Узлов Дмитро Юрійович *к.т.н., доцент, в.о. декана ННІ «Комп'ютерних наук та штучного інтелекту», Харківський національний університет імені В.Н. Каразіна, майдан Свободи, 4, Харків, Україна, 61022*
*e-mail:*dmytro.uzlov@karazin.ua
https://orcid.org/0000-00Q3-3308-424X

Аналіз ефективності бібліотеки Resemblyzer для короткокомандної голосової автентифікації

Актуальність. Голосова взаємодія широко використовується в системах Інтернету речей та автономних вбудованих пристроях, однак її застосування обмежується вимогами до безпеки, захисту приватності та обмеженими обчислювальними ресурсами периферійних платформ. Це зумовлює потребу у повністю локальних рішеннях голосової автентифікації, здатних працювати без залучення хмарних сервісів.
Метою роботи є оцінка можливостей відкритої Python-бібліотеки Resemblyzer для реалізації автономної голосової автентифікації користувачів за короткими голосовими командами в умовах відсутності доступу до хмарних обчислень та обмеженої апаратної потужності.
Методи дослідження. Дослідження виконано на основі декількох наборів аудіоданих із варіацією тривалості, якості та розміру файлів. Для формування ознак використовувалися голосові ембеддинги, згенеровані бібліотекою Resemblyzer. Кількісна оцінка подібності між записами здійснювалася за допомогою метрики косинусної подібності у сценаріях порівняння голосу одного мовця та різних мовців.
Результати. Показано, що надійна голосова автентифікація досягається для аудіозаписів тривалістю не менше 2.63 секунди та розміром файлу від 495 КБ. Короткі фрагменти тривалістю 1-1.5 секунди виявилися недостатньо інформативними для стабільного розрізнення мовців, особливо при зіставленні з високоякісним еталонним записом. Виявлено чітку залежність якості автентифікації від обсягу акустичної інформації, що міститься у голосовому сигналі.
Висновки. Отримані результати підтверджують доцільність використання Resemblyzer для побудови повністю автономних систем голосової біометричної автентифікації в реальному часі. Сформульовано практичні вимоги до мінімальної тривалості та інформаційної насиченості голосових команд, які можуть бути інтерпретовані як технічні обмеження на ентропію голосових паролів у захищених IoT-застосуваннях.
Ключові слова: *голосове керування, автентифікація користувача, Resemblyzer, короткі голосові команди, Інтернет речей (IoT), пристрої з обмеженими ресурсами, голосова верифікація, голосовий відбиток, косинусна подібність*.

Як цитувати: Трусов М. А., Турута О. П., Узлов Д. Ю. Аналіз ефективності бібліотеки Resemblyzer для короткокомандної голосової автентифікації. *Вісник Харківського національного університету імені В. Н. Каразіна, серія Математичне моделювання. Інформаційні технології. Автоматизовані системи управління*. 2025. вип. 68. С.84-97. https://doi.org/10.26565/2304-6201-2025-68-09

How to quote: M. Trusov, O. Turuta, D. Uzlov "Analysis of the effectiveness of the Resemblyzer library for short-command voice authentication", *Bulletin of V. N. Karazin Kharkiv National University, series Mathematical modelling. Information technology. Automated control systems*, vol. 68, pp. 84-97, 2025. https://doi.org/10.26565/2304-6201-2025-68-09 [in Ukrainian]

1. Вступ

Стрімке поширення пристроїв Інтернету речей (IoT) суттєво трансформувало характер взаємодії між людиною та технологічними системами. У цьому контексті голосові інтерфейси

(Voice User Interfaces, VUI) дедалі частіше розглядаються як природний та ефективний механізм керування підключеними пристроями [1-3]. На відміну від традиційних ручних методів введення, таких як клавіатури чи сенсорні панелі, голосові команди забезпечують безконтактну, інтуїтивну комунікацію, що є особливо важливим у мобільних та розподілених обчислювальних середовищах [4]. Однак, попри очевидні переваги, системи автоматичного розпізнавання мовлення (Automatic Speaker Verification, ASV) залишаються вразливими до акустичної варіабельності, фонового шуму, акцентів, темпу та стилю мовлення — чинників, що суттєво знижують точність у реальних умовах [5]. Поряд із технічними обмеженнями важливими є й ризики безпеки. Зокрема, голосові інтерфейси легко піддаються атакам типу spoofing, replay та синтетичного голосового клонування [6-8]. Останні дослідження показали, що навіть сучасні біометричні системи можуть бути скомпрометовані за допомогою синтетичної мови, згенерованої нейронними моделями, що підкреслює нагальну потребу у надійних методах автентифікації користувача [9]. Описані недоліки роблять розробку стійких механізмів верифікації пріоритетним завданням для індустрії IoT.

Критичною проблемою залишається надійна верифікація користувача за короткими мовленнєвими фрагментами. У реальних умовах експлуатації IoT голосові команди часто тривають менше двох секунд, що суттєво обмежує обсяг фонетичної інформації, необхідної для вилучення стійких ознак мовця. Недостатня тривалість сигналу традиційно призводить до зниження точності систем ASV через високу внутрішньокласову варіативність коротких сегментів [10]. Еволюція методів розпізнавання мовця включає шлях від імовірнісних моделей на основі суміші ґаусових розподілів (GMM-UBM) та і-векторів [11] до методів глибокого навчання. Сучасний рівень технологій визначають архітектури, що генерують дискримінативні векторні представлення – ембединги, такі як x-vectors [12] та моделі на базі трансформерів, зокрема wav2vec 2.0 [13]. Попри високу точність на еталонних наборах даних, ці підходи часто вимагають значних обчислювальних ресурсів, що ускладнює їх імплементацію на периферійних пристроях. Це обмеження стимулювало розвиток легких (легковагових) архітектур, оптимізованих для роботи в реальному часі. Одним із перспективних інструментів є Resemblyzer – програмна реалізація нейромережевого енкодера, що базується на рекурентній архітектурі LSTM та навчається з використанням функції втрат Generalized End-to-End (GE2E) [14]. Цей метод, розроблений для максимізації косинусної схожості між векторами (d-vectors) одного користувача, забезпечує ефективну кластеризацію навіть за умов обмеженої довжини вхідних даних. На відміну від комплексних дослідницьких фреймворків, таких як SpeechBrain [15] або pyannote.audio [16], які орієнтовані на побудову складних конвеєрів обробки, Resemblyzer пропонує попередньо навчену модель для швидкого вилучення високорівневих ознак, що робить її привабливою для інтеграції в IoT-системи. Попри теоретичну обґрунтованість методу GE2E, систематичні дослідження ефективності бібліотеки Resemblyzer саме для задач верифікації на надкоротких аудіофрагментах залишаються фрагментарними. Більшість існуючих праць зосереджені на довших записах або завданнях діаризації (розділення користувачів) [17,18]. Питання щодо визначення порогового значення тривалості сигналу, за якого зберігається прийнятний рівень помилок, потребує детального вивчення.

Метою цієї роботи є експериментальна оцінка стійкості методу Resemblyzer при автентифікації користувачів за короткими голосовими командами. Дослідження спрямоване на встановлення залежності точності верифікації від тривалості аудіосигналу та визначення мінімальних технічних вимог для впровадження цього підходу в захищені голосові інтерфейси.

2. Методологія

2.1. Архітектура системи та вибір моделі кодування

Для вилучення голосових ознак застосовано нейромережевий енкодер, що базується на функції втрат Generalized End-to-End (GE2E). Модель реалізована у бібліотеці Resemblyzer, яка генерує 256-вимірні d-vectors — компактні векторні представлення мовця, адаптовані для задач верифікації. Вибір Resemblyzer зумовлений вимогою забезпечити оптимальний баланс між точністю та обчислювальною ефективністю в умовах обмежених ресурсів периферійних пристроїв (Edge-AI). GE2E-енкодер має невеликий обсяг (~15 МБ), що істотно менше, ніж у сучасних трансформерних моделей обробки мовлення (Табл. 1), і може виконувати інференс локально на пристрої без використання обчислювальних прискорювачів.

Таблиця 1. Порівняльна характеристика різних моделей обробки мовлення
Table 1. Comparative characteristics of different speech processing models

Назва моделі/архітектура	Розмір моделі, МБ	Основна сфера застосування	Примітки
Resemblyzer (GE2E)	~15	Верифікація користувача	Оптимізована для Edge AI, підтримує повністю локальний інференс
SpeakerNet	~30–70	Розпізнавання користувача	Середні обчислювальні вимоги, підтримує використання без GPU
Whisper (base)	~74	Розпізнавання мовлення (ASR)	Універсальна ASR-модель, яка забезпечує високу точність за рахунок більших ресурсів
ECAPA-TDNN (SpeechBrain)	~50-120	Верифікація користувача	Стандарт у сучасних системах верифікації користувача, який потребує серверної інференції
Wav2Vec 2.0	~95-320	Універсальні моделі для обробки мовлення	Модель широкого застосування, яка створена для задач великого масштабу

Такий підхід підвищує рівень конфіденційності та забезпечує автономність роботи системи, що особливо важливо для застосувань, де передавання голосових даних у мережу є небажаним або ризикованим. Крім того, мала обчислювальна складність моделі дозволяє розгортати підсистему автентифікації на ресурсно-обмежених IoT-платформах, що значно розширює можливості її практичного використання в автономних і мобільних середовищах.

2.2. Формування експериментального масиву даних

Для оцінювання точності верифікації сформовано набір аудіозаписів, що моделює реальні умови голосового керування автономними роботизованими або IoT-системами. Запис здійснювався за допомогою стандартного мікрофона ОС Windows без студійної обробки, що дозволяє відтворити низькоякісні умови, характерні для польового використання.

До масиву експериментальних даних було включено записи двох дикторів, чоловічого та жіночого голосів, що дозволило охопити основні відмінності спектральних характеристик мовлення. Структура набору даних охоплює три функціонально різні типи аудіоматеріалів (Табл. 2). Повний опис усіх даних подано у Додатку А. Перший тип представлений короткими командними словами тривалістю близько 1-2 секунд, які відповідають типовим однословним інструкціям, що використовуються у голосових інтерфейсах реального часу (наприклад, «Attack», «Destroy», «Autopilot»). Другий тип становлять фразові команди тривалістю приблизно 3 секунди, які поєднують звертання та основну команду (наприклад, «Ginger, listen to me, analyze»).

Таблиця 2. Структура та характеристики експериментального набору даних
Table 2. Structure and characteristics of the experimental datasets

ID	Тип голосу	Тип контенту	Кількість записів	Діапазон тривалості (сек)	Опис
Dataset #1	Чоловічий	Короткі команди	17	1.02-1.59	Однослівні команди («Attack», «Stop» тощо)
Dataset #2	Чоловічий	Зв’язне мовлення	1	15.78	Читання уривку прози для створення профілю
Dataset #3	Жіночий	Короткі команди	17	0.55-1.84	Однослівні команди («Attack», «Stop» тощо)
Dataset #4	Жіночий	Зв’язне мовлення	1	12.88	Читання уривку прози для створення профілю
Dataset #5	Чоловічий	Фразові команди	17	2.63-3.89	Команди зі звертанням («Ginger, listen to me...»)

Третій тип містить довші еталонні записи тривалістю 12-16 секунд, що складаються зі зв'язного читання й слугують для побудови стабільного референсного представлення голосу, необхідного для подальшої верифікації (наприклад, читання уривка з «The Adventures of Sherlock Holmes»).

2.3. Вилучення ознак та формування векторних представлень

Перетворення аудіосигналу у векторне представлення здійснювалося засобами Python із використанням бібліотек NumPy та Pandas, що забезпечують ефективну роботу з числовими даними. Кожен аудіозапис спочатку проходив стандартну процедуру попередньої обробки за допомогою функції preprocess_wav, яка виконує нормалізацію сигналу та підготовку аудіо до подальшого кодування відповідно до вимог моделі Resemblyzer. Після цього оброблений сигнал подавався до нейронного енкодера VoiceEncoder, що реалізує архітектуру GE2E. Енкодер формував 256-вимірний d-vector, який є компактним числовим представленням індивідуальних голосових характеристик користувача. Таке представлення є стійким до локальних варіацій сигналу та зберігає ключові ознаки, необхідні для подальшої верифікації. Згенеровані вектори зберігалися у форматі .csv, що забезпечує мінімальний розмір файлу (близько 1 КБ на один запис) і дозволяє проводити подальший аналіз без повторної обробки аудіоданих.

2.4. Метрика схожості та критерії прийняття рішень

Для порівняння векторних представлень мовця у даному дослідженні застосовано косинусну подібність. Формально косинусну подібність між векторами A (еталонний запис) та B (тестовий запис) визначають як нормалізований скалярний добуток:

$$S_C(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|} \qquad (2.1)$$

Отримане значення належить інтервалу $[-1,1]$, де значення, близькі до +1, вказують на високу подібність між векторними представленнями двох мовних зразків, а значення, близькі до нуля або від'ємні, свідчать про відсутність відповідності. У практичних системах верифікації мовця для справжніх пар (коли обидва зразки належать одному користувачу) характерні значення подібності у межах приблизно 0.80-0.95. У нашому дослідженні встановлено порогове значення $\tau = 0.75$, що забезпечує збалансований компроміс між пропускною здатністю та захищеністю системи.

3. Результати та обговорення

3.1. Аналіз внутрішньокласової варіативності

Перший етап дослідження було спрямовано на кількісну оцінку того, наскільки стабільно модель Resemblyzer відтворює голосові ембедінги для коротких команд, записаних одним і тим самим користувачем. Для цього використано набір Dataset #1, який містить 17 однословних англомовних команд (наприклад, analyze, attack, authentication, autopilot, check, defence, …, watch, див. Додаток А) тривалістю близько 1-1.5 с, промовлених одним диктором у подібних акустичних умовах.

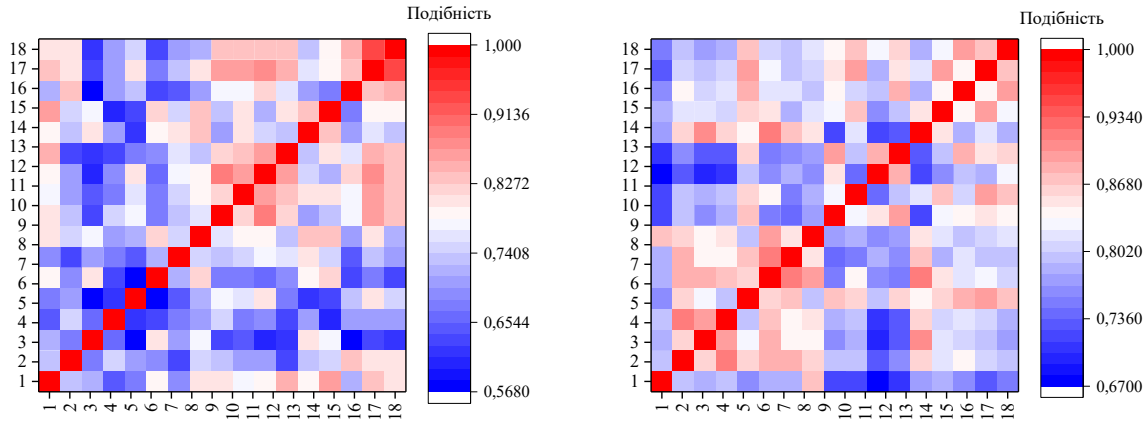

Рис. 3.1. Матриця косинусної подібності для коротких команд. Тип голосу: чоловічий – ліва панель, жіночий – права панель. Номерами позначені команди, повна розшифровка яких представлена у Додатку Б

Fig. 3.1. Cosine similarity matrix for short commands. Voice type: male – left panel, female – right panel. Command indices correspond to the full list provided in Appendix B.

Для кожного аудіофайла було згенеровано 256-вимірний голосовий ембеддинг за допомогою моделі Resemblyzer, після чого виконано попарне порівняння всіх ембеддингів між собою. У результаті отримано повну матрицю косинусної подібності розміром 17×17, де кожен елемент S_C (i,j) відображає схожість між двома командами того самого мовця. Сформована матриця подана в текстовому вигляді в Додатку Б (Табл. Б1). На основі цієї матриці побудовано теплову карту, яку наведено на Рис. 3.1, ліва панель, де по осях відкладено номери команд, назви яких наведено у Табл. Б1, а колірна шкала відображає значення косинусної подібності.

Візуальний аналіз теплової карти демонструє очікувано високі значення на головній діагоналі (самопорівняння, S_C≈1.0) та помітну варіабельність позадіагональних елементів. Для частини пар коротких команд значення подібності досягають 0.80-0.89, тоді як для інших падають до 0.58-0.65. Це означає, що навіть у межах одного користувача короткі висловлювання можуть давати доволі різні голосові відбитки, що особливо критично для систем автентифікації, які покладаються на одиничні короткі команди.

Для формалізованої оцінки якості внутрішньої подібності всі ненульові позадіагональні елементи матриці було розподілено на три діапазони:

$0 \leq SC < 0.6$ – невдала автентифікація,

$0.6 \leq SC < 0.8$ – проміжний діапазон невизначеності,

$0.8 \leq SC \leq 1$ – успішна автентифікація.

Підрахунок часток елементів у кожному діапазоні показав, що лише 30% позадіагональних значень потрапляють у зону успішної автентифікації, тоді як 67% належать до проміжної зони невизначеності, а 3% мають значення нижче 0.6. Такий розподіл свідчить про те, що короткі команди тривалістю близько однієї секунди містять недостатній обсяг стійкої фонетичної інформації для гарантованої верифікації користувача за один-єдиний запит.

Отримані дані добре узгоджується з природою d-vector представлень, які покладаються на усереднення послідовностей прихованих станів нейронної мережі: чим коротший сигнал, тим сильніше на ембединг впливають локальні варіації артикуляції, темпу, інтонації та шуму [14]. Далі у даній роботі буде продемонстровано, що використання іншого диктора з тим самим набором команд та додавання довшого запису істотно змінює статистику подібності, ілюструючи можливість часткової компенсації цих ефектів.

Для більш детального аналізу внутрішньокласової варіативності було розглянуто ще один голосовий набір даних, Dataset #3 (Додаток А), що містить ті самі 17 коротких команд, проте записаних жіночим диктором. Формат запису (1-1.5 с), акустичні умови та словниковий склад повністю відповідають Dataset #1, що дозволяє безпосередньо порівнювати внутрішню структуру ембедингів між двома дикторами в межах одного й того самого експерименту. Аналогічно до Dataset #1, для кожного аудіофайла з Dataset #3 було сформовано 256-вимірні голосові ембединги, після чого виконано повне попарне порівняння всіх векторів. У результаті отримано матрицю косинусної подібності розміром 17×17, текстове подання якої наведене в Додатку Б (Табл. Б2). Ця матриця становить основу для теплової карти (heatmap), що відображатиме внутрішньокласову схожість коротких команд саме для жіночого голосу.

Порівняно з Dataset #1, результати показали значно вищу стабільність ембедингів. Усі позадіагональні значення матриці для жіночого голосу лежали вище 0.60, що вказує на повну відсутність невдалої автентифікації. Загальна структура подібності описується домінуванням високих значень у діапазоні 0.80-0.92, що відповідає 67.7% усіх пар коротких команд. Частка значень у межах 0.60-0.80 становила 32.3%, а значень <0.60 виявлено не було. Такий розподіл суттєво контрастує з Dataset #1, де більшість значень (≈67%) потрапили в невизначену зону, а частина – нижче 0.60.

Таким чином, внутрішньокласова варіативність для жіночого голосу помітно нижча, ніж для чоловічого, навіть за умов використання одного й того самого набору команд та однакової тривалості сигналів. Це свідчить про те, що у Resemblyzer голосові відбитки можуть суттєво залежати від індивідуальних артикуляційних особливостей диктора, темпу мовлення та спектральної структури голосу. Оскільки модель формує d-vector шляхом усереднення прихованих станів LSTM, короткі зміни в артикуляції чи інтонації можуть впливати на якість ембединга по-різному для різних користувачів.

Для кількісного порівняння внутрішньокласової варіативності двох користувачів було побудовано розподіли значень косинусної подібності, отриманих з відповідних матриць для чоловічого (Dataset #1) та жіночого (Dataset #3) голосів. Таке порівняння агрегованих розподілів

дозволяє оцінити загальну структуру варіативності, що є інформативним у контексті короткокомандної голосової автентифікації. Для цього всі позадіагональні значення подібності було об'єднано у дві вибірки, після чого було побудовано розподіл значень косинусної подібності для чоловічого та жіночого голосів (Рис. 3.2).

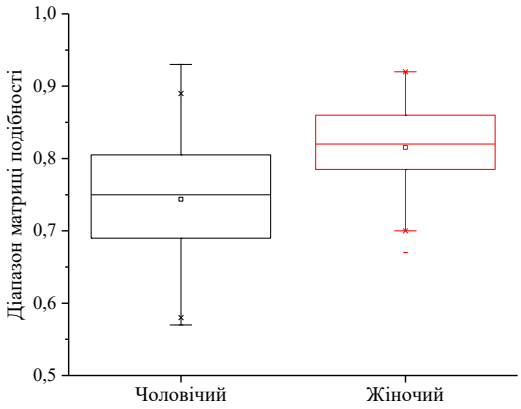

Рис. 3.2. Розподіл значень косинусної подібності, отриманих з відповідних матриць для чоловічого (Dataset #1) та жіночого (Dataset #3) голосів
Fig. 3.2. Distribution of cosine similarity values derived from the corresponding matrices for male (Dataset #1) and female (Dataset #3) voices

Аналіз відмінностей у медіані, інтерквартильному розмаху, щільності розподілу та наявності потенційних викидів показав, що жіночий голос характеризується більш компактним розподілом подібностей: медіанні значення вищі, а варіативність – суттєво менша. Натомість для чоловічого голосу спостерігається ширший інтервал значень (із зсувом до нижчих подібностей), що вказує на меншу відтворюваність коротких команд. Отримані відмінності у формі та ширині розподілів підтверджують, що внутрішньокласова стабільність ембедингів може істотно залежати від мовця, навіть за однакових умов запису. На практиці це означає, що для деяких користувачів короткі командні висловлювання можуть бути менш надійними для автентифікації.

3.2. Порівняння коротких голосових команд із довгим еталонним голосовим відбитком

На наступному етапі дослідження було проаналізовано, наскільки ефективно короткі голосові команди можуть бути співвіднесені з еталонним голосовим відбитком, сформованим на основі довгих та акустично стабільних записів мовлення. Такий сценарій відтворює типову конфігурацію систем голосової автентифікації, де процес enrollment передбачає використання тривалого та якісного аудіосигналу, тоді як щоденні запити користувача представлені короткими командами тривалістю 1-1.5 секунди. У нашому випадку довгі еталонні записи було отримано з Dataset #2 (чоловічий голос, 15.78 с, Додаток А) та Dataset #4 (жіночий голос, 12.88 с, Додаток А), тоді як короткі команди взято відповідно з Dataset #1 і Dataset #3 (Додаток А).

Для кожного диктора було сформовано еталонний d-vector, після чого всі короткі командні висловлювання порівнювалися з цим відбитком за допомогою стандартної метрики косинусної подібності. Таким чином отримано по 17 значень для кожного диктора, що відображають ступінь відповідності коротких команд довгому «золотому стандарту» голосової ідентичності. Такий підхід дозволяє оцінити, наскільки надійно Resemblyzer може зіставляти короткий запит зі стійким, багатофреймовим enrollment-профілем, який містить значно більше акустичної інформації. Результати порівняння представлені у вигляді діапазонів подібності і часток значень у кожному діапазоні (Табл. 3, 4, Рис. 3.3).

Як видно з цих таблиць та рисунку, в обох випадках жоден із коротких аудіосигналів не досяг порога успішної автентифікації при порівнянні з високоякісним еталонним відбитком. Переважна більшість значень перебувала у проміжному інтервалі (0.60-0.80), а значна частина – у зоні однозначної невдалої автентифікації (<0.60).

Таблиця 3. Порівняння даних Dataset #1 з Dataset #2 (чоловічий голос, короткі команди ↔ довгий еталонний відбиток)
Table 3. Comparison of data from Dataset #1 and Dataset #2 (male voice, short commands ↔ long reference embedding)

Діапазон	Частка (%)	Інтерпретація
[0;0.6]	17.65	Невдала автентифікація
(0.6;0.8)	82.35	Проміжний діапазон невизначеності
[0.8;1]	0.0	Успішна автентифікація
Всього зразків: 17		

Таблиця 4. Порівняння даних Dataset #3 з Dataset #4 (жіночий голос, короткі команди ↔ довгий еталонний відбиток)
Table 4. Comparison of data from Dataset #3 and Dataset #4 (female voice, short commands ↔ long reference embedding)

Діапазон	Частка (%)	Інтерпретація
[0;0.6]	35.29	Невдала автентифікація
(0.6;0.8)	64.71	Проміжний діапазон невизначеності
[0.8;1]	0.0	Успішна автентифікація
Всього зразків: 17		

Ці результати свідчать, що короткі команди тривалістю 1-1.5 секунди не містять достатнього обсягу стабільних індивідуальних акустичних ознак, необхідних для коректної верифікації при порівнянні з довгим записом.

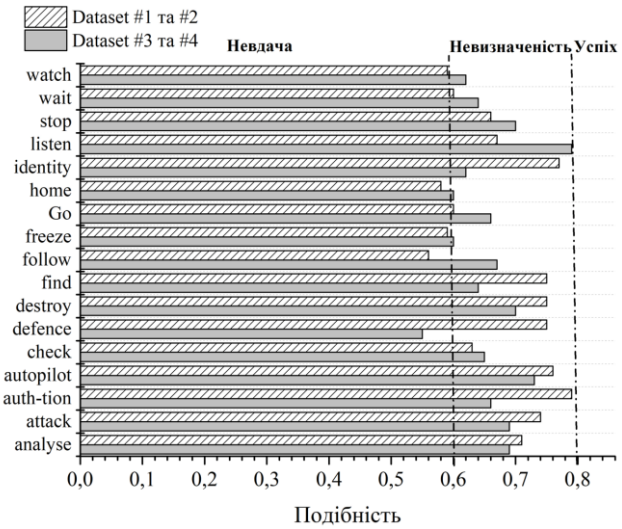

Рис. 3.3. Порівняння подібності коротких команд із довгим еталонним записом (чоловічий та жіночий голоси. Невдача, невизначеність та успіх означають діапазони подібностей, які відповідають невдалій автентифікації, проміжному діапазону невизначеності та успішній автентифікації, відповідно
Fig. 3.3. Comparison of similarity between short commands and a long reference recording (male and female voices. Failure, uncertainty, and success denote similarity ranges corresponding to unsuccessful authentication, an intermediate uncertainty region, and successful authentication, respectively)

Отримані результати можуть бути пояснені архітектурою Resemblyzer, що використовує LSTM-енкодер з усередненням прихованих станів на часовому горизонті всього запису. Короткі аудіофрагменти містять недостатню кількість інформаційно насичених фреймів, що призводить до нестабільності ембеддингу, підвищеної чутливості до випадкових інтонаційних варіацій та фонетичних коливань і, відповідно, – до низької подібності з довгим еталонним записом. У контексті IoT-застосувань це означає, що короткі командні висловлювання не можуть бути використані як самостійний матеріал для автентифікації, якщо enrollment здійснено на довгому аудіосигналі.

Описаний аналіз демонструє фундаментальну обмеженість Resemblyzer у сценаріях короткокомандної автентифікації: короткі команди тривалістю близько однієї секунди не є

достатньо інформативними для відтворення структури еталонного d-vector. Для практичного застосування такі результати вказують на необхідність використання або довших команд, або багатократної агрегації коротких фрагментів, або застосування спеціальних технік нормалізації та підсилення голосових ознак.

3.3. Аналіз якості автентифікації при збільшенні тривалості та розміру аудіозаписів

Останній етап дослідження було спрямовано на дослідження здатності Resemblyzer формувати стабільніші голосові відбитки та підвищувати точність автентифікації при збільшенні тривалості коротких команд та, відповідно, обсягу мовленнєвого матеріалу. Доцільність такого роду аналізу випливає з результатів попереднього підпункту, де короткі фрагменти тривалістю приблизно 1 секунду демонстрували недостатню інформативність для надійного зіставлення з довгими enrollment-записами. Для перевірки цієї гіпотези було сформовано Dataset #5 (Додаток А), у якому кожна команда з Dataset #1 та Dataset #3 була перезаписана у вигляді довших та якісніших фрагментів. Як показано в Табл. В1, Додаток В, середній розмір файлу та середня тривалість збільшились у ~2-2.4 рази. Це дозволило оцінити, чи є наявність додаткових мовленнєвих кадрів достатньою для формування більш стабільних голосових відбитків.

Спочатку було обчислено матрицю косинусної подібності для всіх аудіозразків Dataset #5 відносно один одного (Рис. 3.4). Результати виявили значне підвищення рівня внутрішньокласової стабільності порівняно з аналогічними матрицями для коротких команд: більшість значень лежала у діапазоні 0.85-0.95, що наближається до типових показників «високої подібності» для систем контролю доступу. Максимальні значення коливалися біля 0.94-0.95, а мінімальні – практично не виходили за межі 0.8.

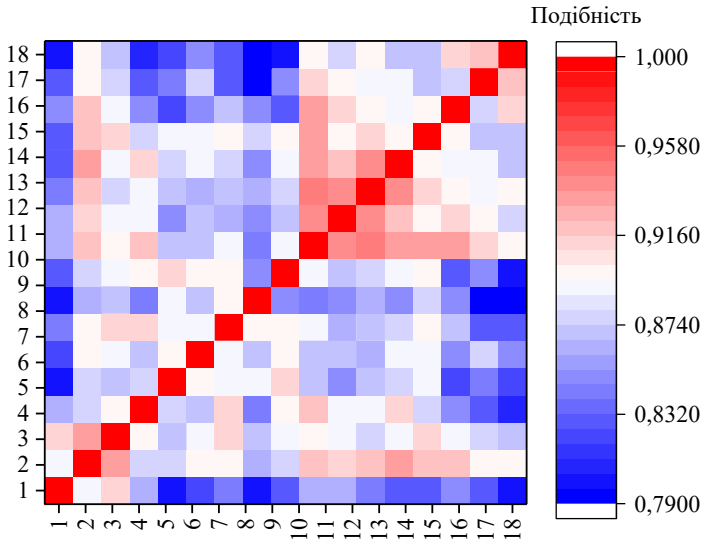

Рис. 3.4. Матриця косинусної подібності для всіх аудіозразків Dataset #5 відносно один одного. Номерами позначені команди, повна розшифровка яких представлена у Додатку В
Fig. 3.4. Cosine similarity matrix for all audio samples in Dataset #5 relative to each other. Command indices correspond to the full list provided in Appendix C.

Статистичний розподіл подібності всередині Dataset #5 показав, що 98.53% усіх пар лежать у діапазоні [0.8-1.0], тобто відповідають критерію «успішна автентифікація», 1.47% – у проміжному інтервалі (0.6-0.8) та 0% – у зоні невдалої автентифікації [0-0.6] (Рис. 3.5, штрихована заливка). Загальна кількість порівнянь становила 272, і серед них не було зафіксовано жодного хибно-позитивного результату. Це свідчить про те, що вже збільшені до 2.6–3.0 секунд аудіозаписи забезпечують утворення стабільних голосових відбитків, придатних для внутрішньокласового зіставлення.

Далі було оцінено здатність збільшених команд із Dataset #5 узгоджуватися з голосовим еталоном (d-vector), побудованим на основі високоякісного довгого аудіофрагмента із Dataset #2. Цей сценарій моделює реальну задачу автентифікації, коли enrollment виконується на основі

«еталонного» запису, а перевірка – за допомогою покращених, але все ще відносно коротких команд.

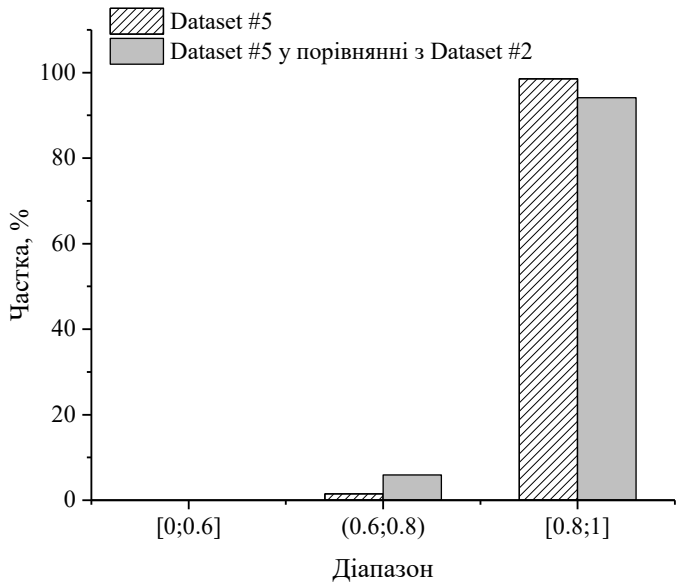

Рис. 3.5. Статистичний розподіл подібностей для Dataset #5 (штрихована заливка) та при порівнянні Dataset #5 з Dataset #2 (сіра заливка)
Fig. 3.5. Statistical distribution of similarity values for Dataset #5 (dashed shading) and for the comparison between Dataset #5 and Dataset #2 (gray shading)

Результати порівняння, представлені на Рис. 3.5 (сіра заливка) показали, що 94.12% команд успішно зіставилися з еталоном (подібність ≥0.8), 5.88% потрапили у зону невизначеності (0.6-0.8) та 0% не пройшли автентифікацію (≤0.6). Таким чином, покращені записи забезпечують рівень успішних збігів, який відповідає характеристикам повноцінних систем верифікації мовця, що використовують значно довші аудіосигнали.

Порівняння внутрішньокласових значень подібності для Dataset #5, а також їх зіставлення з еталонним записом із Dataset #2, демонструє узгоджену та чітко виражену тенденцію: збільшення тривалості аудіосигналу до приблизно 2.6-3.0 секунд істотно підвищує якість отриманих ембедингів. Довші записи забезпечують більшу кількість мовленнєвих кадрів, що, у свою чергу, призводить до формування ембедингів із вищою міжфреймовою узгодженістю та меншою варіативністю. Це дозволяє моделі Resemblyzer відтворювати індивідуальні голосові характеристики значно точніше, ніж у випадку коротких команд, зафіксованих у Dataset #1 та Dataset #3.

Внутрішньокласові значення косинусної подібності для Dataset #5 стабільно наближалися до діапазону високої відповідності, що традиційно вважається достатнім для надійної верифікації користувача. Така поведінка суттєво контрастує з результатами попередніх експериментів, де короткі записи не могли забезпечити порівнюваний рівень акустичної інформативності. Примітно, що збільшені командні записи зберігають високу подібність навіть у випадку співставлення з довгим еталонним голосовим профілем, що додатково підтверджує їхню придатність для практичних задач автентифікації.

Отримані результати дають підстави стверджувати, що короткі команди тривалістю близько однієї секунди є недостатніми для надійного розпізнавання користувача, тоді як фрагменти тривалістю близько трьох секунд вже формують d-vector ембединги достатньої якості для стабільного відтворення голосових характеристик. У прикладних сценаріях, пов’язаних з IoT та голосовим керуванням, це означає, що оптимальною стратегією є використання довгих enrollment-записів у поєднанні з командними аудіофрагментами середньої тривалості. Такий підхід забезпечує доцільний компроміс між зручністю користувача та вимогами до безпеки, зберігаючи при цьому можливість реалізації автентифікації без застосування високопродуктивних обчислювальних ресурсів.

4. Висновки

У даній роботі проведено комплексну оцінку можливостей бібліотеки Resemblyzer для задач голосової автентифікації на основі коротких аудіофрагментів у контексті автономних та ресурсно обмежених систем. Аналіз внутрішньокласової варіативності та порівняння коротких команд із довгими еталонними записами показали, що фрагменти тривалістю близько однієї секунди не забезпечують достатньої інформативності для надійної верифікації користувача. Це підтверджує обмеженість ультракоротких команд у біометричних системах та узгоджується з сучасними уявленнями про необхідність достатнього часово-спектрального покриття мовного сигналу для вилучення індивідуально-специфічних ознак. Натомість, збільшення тривалості команд до 2.6-3.0 секунд істотно покращує якість d-vector ембеддингів і забезпечує рівень подібності, який відповідає практичним вимогам систем контролю доступу та персональної ідентифікації. Отримані результати демонструють формування більш компактних і стабільних кластерів у векторному просторі ознак, що знижує внутрішньокласову дисперсію та підвищує надійність автентифікації.

Проведені дослідження також підтвердили, що Resemblyzer може ефективно працювати на пристроях із низькою обчислювальною потужністю без використання хмарних сервісів, що робить його перспективним інструментом для створення легких, автономних і безпечних голосових інтерфейсів. Подальші дослідження доцільно спрямувати на оцінку стійкості моделі в умовах акустичних завад, реверберації та фонового шуму, а також на вивчення впливу мовних факторів, таких як темп мовлення, емоційний стан та багатомовність. Окремий інтерес становить аналіз продуктивності підходу у багатокористувацьких IoT-сценаріях, де критичними є масштабованість, захист від спуфінгових атак і довготривала стабільність голосових шаблонів.

REFERENCES

1. A. Choudhary, Internet of Things: a comprehensive overview, architectures, applications, simulation tools, challenges and future directions. *Discov. Internet Things*. 2024. Vol. 4. P. 31. https://doi.org/10.1007/s43926-024-00084-3.

2. M. Lombardi, F. Pascale, D. Santaniello, Internet of Things: A General Overview between Architectures, Protocols and Applications. *Information*. 2021. Vol. 12. P. 87. https://doi.org/10.3390/info12020087.

3. L. Atzori, A. Iera, G. Morabito, The Internet of Things: A survey. *Computer Networks*. 2010. Vol. 54. P. 2787-2805. https://doi.org/10.1016/j.comnet.2010.05.010.

4. M. Hoy, Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. *Med. Ref. Serv. Q*. 2018. Vol. 37. P. 81-88. https://doi.org/10.1080/02763869.2018.1404391.

5. M. Benzeghiba, R. De Mori, O. Deroo et al., Automatic speech recognition and speech variability. *Speech Commun*. 2007. Vol. 49. P. 763-786. https://doi.org/10.1016/j.specom.2007.02.006.

6. A. Javed, K. Malik, H. Malik, A. Irtaza, Voice spoofing detector: a unified anti-spoofing framework. *Expert Systems Applic*. 2022. Vol. 198. P. 116770. https://doi.org/10.1016/j.eswa.2022.116770.

7. N. Ahmed, J. Khan, N. Sheta et al., Detecting Replay Attack on Voice-Controlled Systems using Small Neural Networks. *2022 IEEE 7th Forum on Research and Technologies for Society and Industry Innovation (RTSI)*, Paris, France. 2022. P. 50-54. https://doi.org/10.1109/RTSI55261.2022.9905158.

8. Z. Wu, N. Evans, T. Kinnunen et al., Spoofing and countermeasures for speaker verification: A survey. *Speech Commun*. 2015. Vol. 66. P. 130-153. https://doi.org/10.1016/j.specom.2014.10.005.

9. T. Kinnunen, Z. Wu, K. Lee et al., Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan. 2012. P. 4401-4404. https://doi.org/10.1109/ICASSP.2012.6288895.

10. A. Poddar, M. Sahidullah, G. Saha, Speaker verification with short utterances: a review of challenges, trends and opportunities. *IET Biometrics*. 2018. Vol. 7. P. 403-411. https://doi.org/10.1049/iet-bmt.2017.0065.

11.N. Dehak, P. Kenny, R. Dehak et al., Front-End Factor Analysis for Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing*. 2011. Vol. 19. P. 788-798. https://doi.org/10.1109/TASL.2010.2064307.

12. D. Snyder, D. Garcia-Romero, G. Sell et al., X-Vectors: Robust DNN Embeddings for Speaker Recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada. 2018. P. 5329-5333. https://doi.org/10.1109/ICASSP.2018.8461375.

13. A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. Neural Inform. Proces. Syst.* 2020. Vol. 33. P. 12449-12460. https://doi.org/10.48550/arXiv.2006.11477.
14. L. Wan, Q. Wang, A. Papir, I. Moreno, Generalized end-to-end loss for speaker verification. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018. P. 4879-4883. https://doi.org/10.48550/arXiv.1710.10467.
15. M. Ravanelli, T. Parcollet, P. Plantinga et al., SpeechBrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*. 2021. https://doi.org/10.48550/arXiv.2106.04624.
16. H. Bredin, R. Yin, J. Coria, Pyannote. audio: neural building blocks for speaker diarization. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020. P. 7124-7128). https://doi.org/10.48550/arXiv.1911.01255.
17. Y. Jia, Y. Zhang, R. Weiss et al., Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Adv. Neur. Inform. Proces. Systems*. 2018. arXiv:1806.04558. https://doi.org/10.48550/arXiv.1806.04558.
18. Q. Wang, C. Downey, L. Wan et al., Speaker diarization with LSTM. 2018 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018. P. 5239-5243. https://doi.org/10.48550/arXiv.1710.10468.

Додаток А

Повний опис експериментального масиву даних

Dataset #1. Тип голосу: чоловічий
Dataset #1. Voice type: male

File name	Text in audio	Length of audio (seconds)
Analyze.wav	Analyze	1,16
Attack.wav	Attack	1,34
Authentication.wav	Authentication	1,50
Autopilot.wav	Autopilot	1,42
Check.wav	Check	1,29
Defence.wav	Defence	1,48
Destroy.wav	Destroy	1,53
Find.wav	Find	1,46
Follow.wav	Follow	1,42
Freeze.wav	Freeze	1,40
Go.wav	Go	1,42
Home.wav	Home	1,34
Identity.wav	Identity	1,59
Listen.wav	Listen	1,20
Stop.wav	Stop	1,02
Wait.wav	Wait	1,18
Watch.wav	Watch	1,23

Dataset #2. Тип голосу: чоловічий
Dataset #2. Voice type: male

File name	Text in audio	Length of audio (seconds)
A1-m.waw	The Adventures of Sherlock Holmes" by Arthur Conan Doyle is a collection of detective stories written during the late 19th century. The book introduces the legendary detective Sherlock Holmes and his loyal companion, Dr. John Watson	15,78

Dataset #3. Тип голосу: жіночий
Dataset #3. Voice type: female

File name	Text in audio	Length of audio (seconds)
Analyze.wav	Analyze	1,49
Attack.wav	Attack	1,24
Authentication.wav	Authentication	1,84
Autopilot.wav	Autopilot	1,84
Check.wav	Check	1,19
Defence.wav	Defence	1,52
Destroy.wav	Destroy	1,33
Find.wav	Find	1,37
Follow.wav	Follow	1,38
Freeze.wav	Freeze	0,55
Go.wav	Go	1,2
Home.wav	Home	1,42
Identity.wav	Identity	1,48
Listen.wav	Listen	1,47
Stop.wav	Stop	1,15
Wait.wav	Wait	1,19
Watch.wav	Watch	1,46

Dataset #4. Тип голосу: жіночий
Dataset #4. Voice type: female

File name	Text in audio	Length of audio (seconds)
A1-f.wav	The Adventures of Sherlock Holmes" by Arthur Conan Doyle is a collection of detective stories written during the late 19th century. The book introduces the legendary detective Sherlock Holmes and his loyal companion, Dr. John Watson	12,88

Dataset #5. Тип голосу: чоловічий
Dataset #5. Voice type: male

File name	Text in audio	Length of audio (seconds)
Gltm-Analyze.wav	Ginger, listen to me, Analyze	3,16
Gltm-Attack.wav	Ginger, listen to me, Attack	3,07
Gltm-Authentication.wav	Ginger, listen to me, Authentication	3,64
Gltm-Autopilot.wav	Ginger, listen to me, Autopilot	3,02
Gltm-Check.wav	Ginger, listen to me, Check	3,21
Gltm-Defence.wav	Ginger, listen to me, Defence	3,03
Gltm-Destroy.wav	Ginger, listen to me, Destroy	3,16
Gltm-Find.wav	Ginger, listen to me, Find	2,68
Gltm-Follow.wav	Ginger, listen to me, Follow	3,89
Gltm-Freeze.wav	Ginger, listen to me, Freeze	2,86
Gltm-Go.wav	Ginger, listen to me, Go	2,63
Gltm-Home.wav	Ginger, listen to me, Home	2,73
Gltm-Identity.wav	Ginger, listen to me, Identity	2,92
Gltm-Listen.wav	Ginger, listen to me, Listen	2,95
Gltm-Stop.wav	Ginger, listen to me, Stop	2,78
Gltm-Wait.wav	Ginger, listen to me, Wait	2,68
Gltm-Watch.wav	Ginger, listen to me, Watch	2,86

Додаток Б

Таблиця Б1. Матриця косинусної подібності Dataset # 1
Table B1. Cosine similarity matrix for Dataset #1

Таблиця Б2. Матриця косинусної подібності Dataset # 3
Table B2. Cosine similarity matrix for Dataset #3

У таблицях Б1 та Б2 наведено матриці косинусної подібності між d-vector ембеддингами коротких голосових команд для Dataset #1 (чоловічий голос) та Dataset #3 (жіночий голос) відповідно. Наведені значення дозволяють кількісно оцінити внутрішньокласову варіативність мовних сигналів та стабільність формування голосових представлень. Порівняльний аналіз демонструє високий рівень внутрішньокласової подібності в обох випадках, переважно в діапазоні 0.75-0.90, що свідчить про здатність моделі формувати узгоджені голосові ембеддинги для коротких мовних фрагментів. Для жіночого голосу (Dataset #3) спостерігається дещо вища середня подібність і менша дисперсія значень, що може бути пов’язано з більш стабільними спектральними характеристиками. Разом із тим, наявність окремих комбінацій команд із пониженими значеннями подібності (≈0.65-0.75) відображає вплив фонетичної структури та артикуляційної варіабельності, що необхідно враховувати під час вибору порогів прийняття рішень у системах голосової автентифікації.

Додаток В

Таблиця В1. Фізичні характеристики dataset #5
Table C1. Acoustic characteristics of Dataset #5

	Dataset #1. Розмір файла (байти)	Dataset #3. Розмір файла (байти)	Dataset #5 Розмір файла (байти)	Середній фактор збільшення розміру файла	Dataset #1. Довжина аудіозапису (сек)	Dataset #3. Довжина аудіозапису (сек)	Dataset #5 Довжина аудіозапису (сек)	Середній фактор збільшення довжини аудіозапису
s	222802	286162	606802			49	3	
	257362	238162	589522			2	3	
a	288082	353362	698962			84	3	
	272722	353362	579922			84	3	
	247762	228562	616402			1	3	
	284242	291922	581842			52	3	
	293842	255442	606802			3	3	
	280402	263122	514642			37	8	
	272722	265042	554962			38	3	
	268882	247762	549202			0	86	
	272722	230482	505042			2	2	
	257362	272722	524242			42	2	
	305362	284242	560722			48	2	
	230482	282322	566482			47	2	
	195922	220882	533842			15	2	
	226642	228562	514642			9	2	
watch	236242	280402	549202			46	2	

Trusov Mykhaylo *PhD student, V. N. Karazin Kharkiv National University, 4 Svobody Sq., Kharkiv, 61022, Ukraine*

Turuta Oleksiy *Associate Professor of the Department of Program Engineering, Kharkiv National University of Radioelectronics, 14 Nauky Ave., Kharkiv 61166, Ukraine*

Uzlov Dmitro *Associate Professor of the Department of Theoretical and Applied Informatics, V. N. Karazin Kharkiv National University, 4 Svobody Sq., Kharkiv, 61022, Ukraine*

Analysis of the effectiveness of the Resemblyzer library for short-command voice authentication

Relevance. Voice interaction is widely used in Internet of Things systems and autonomous embedded devices. However, its practical deployment is constrained by security and privacy requirements as well as the limited computational resources of edge platforms. This creates a demand for fully local voice authentication solutions capable of operating without reliance on cloud services. **Goal.** The objective of this study is to evaluate the capabilities of the open-source Python library Resemblyzer for implementing autonomous user voice authentication based on short voice commands under conditions of no access to cloud computing and limited hardware resources. **Research methods.** The study was conducted using several audio datasets with varying duration, quality, and file size. Voice embeddings generated by the Resemblyzer library were used for feature representation. Quantitative similarity assessment between recordings was performed using the cosine similarity metric in scenarios involving comparisons of recordings from the same speaker and from different speakers.
Results. The results demonstrate that reliable voice authentication is achieved for audio recordings with a duration of at least 2.63 seconds and a file size of no less than 495 kB. Short fragments with durations of 1-1.5 seconds were found to be insufficiently informative for stable speaker discrimination, particularly when compared against a high-quality reference recording. A clear dependence of authentication performance on the amount of acoustic information contained in the voice signal was identified.
Conclusions. The obtained results confirm the aplicability of Resemblyzer for the development of fully autonomous real-time voice biometric authentication systems. Practical requirements for the minimum duration and informational richness of voice commands are formulated, which may be interpreted as technical constraints on the entropy of voice passwords in secure IoT applications.
Keywords: *voice control, user authentication, Resemblyzer, short voice commands, Internet of Things (IoT), resource-constrained devices, voice verification, voice embedding, cosine similarity.*

Надійшла у першій редакції 09.10.2025; в останній 12.11.2025; опубліковано 30.12.2025.
The first version has been received on 09.10.2025; the final version on 12.11.2025; published 30.12.2025.