

ВПЛИВ ТРИВАЛОСТІ АУДІОСИГНАЛІВ НА ТОЧНІСТЬ ІДЕНТИФІКАЦІЇ ГОЛОСІВ МОВЦІВ

Володимир Поворозник, аспірант, Ігор Микитин, д.т.н., проф.

Національний університет «Львівська політехніка», Львів, Україна,

e-mail: volodymyr.b.povoroznyk@lpnu.ua, igor.p.mykytyn@lpnu.ua

Анотація

У статті проведено дослідження здатності системи, яка використовує голосові ембедінги, ідентифікувати голоси мовців. Для цього використовується набір аудіозаписів голосів п'яти осіб, на основі якого сформовано аудіо-фрагменти різної тривалості – від 5 до 600 секунд. Ембедінги ruannote-audio отримано за допомогою нейронної мережі, після чого обчислено коефіцієнти подібності між ембедінгами фрагментів одного голосу (внутрішньо-особова подібність) та між ембедінгами фрагментів різних голосів (міжособова відмінність). Досліджено, як тривалість аудіо-фрагментів впливає на зону захисту під час розподілу осіб на «свій/чужий». Проведені дослідження показали, що існує певне значення тривалості аудіо-фрагментів, за якого отримано відносно широку зону захисту, що в свою чергу підвищує ймовірність точної ідентифікації особи за голосом. Отримані результати можуть використовуватися у майбутніх дослідженнях у сфері біомекоефіцієнтної верифікації.

Ключові слова

Голосова біометрія, голосовий ембедінг, нейронні мережі, голосова ідентифікація, коефіцієнт подібності.

1. Вступ

Сучасні системи голосової біометрії все частіше застосовуються у сфері безпеки, автентифікації та голосових інтерфейсах. Їхня ефективність значною мірою залежить від здатності точно розрізняти голоси різних мовців – навіть за умов обмеженої тривалості аудіо-фрагментів. Базовим елементом таких систем є ембедінги голосу – векторні представлення голосових характеристик, які дозволяють обчислювати подібність або відмінність між голосами.

Існує значна кількість досліджень, що демонструють можливості застосування ембедінгів (наприклад, x-vector, ECAPA-TDNN, ruannote-audio) у задачах ідентифікації мовця [1][2].

У цій статті описано результати дослідження залежності ширини зони захисту від тривалості фрагментів аудіо записів голосів мовців під час ідентифікації. Проведено порівняльний аналіз голосів п'яти мовців із використанням ембедінгів та розрахунком коефіцієнтів подібності таких як: відстані Брея-Кертіса, Канберри, Чебишева а також косинусна, Мангетенська, Евклідова та кореляційна відстані. Аналіз охоплює як внутрішньо-особову подібність, так і міжособову відмінність. Завданням даної роботи є визначення мінімальної довжини аудіо-фрагменту для надійної ідентифікації осіб за голосом, що має важливе значення під час розробки біометричних систем.

2. Недоліки

Серед основних обмежень сучасних систем ідентифікації мовця за голосом є достатньо висока ймовірність розпізнавання злоумисника як користувача (2-10% в залежності від методу перетворення аудіо в ембедінги) [3]. Крім того немає публічно доступних наборів даних з високоякісними аудіозаписами голосів різних тембрів (бас, тенор, сопрано тощо), які озвучують один і той самий текст достатньо великої тривалості, що є необхідним для коректного порівняння різних систем захисту з голосовою ідентифікацією. Також відсутній універсальний критерій вибору метрики для розрахунку коефіцієнтів подібності.

3. Мета

Метою цієї роботи є дослідження впливу тривалості аудіо-фрагментів на ширину зони захисту, збільшення якої дозволяє покращити характеристики систем ідентифікації мовця за голосом.

4. Опис та параметри дослідження

Для проведення дослідження було створено кілька аудіозаписів, які ділилися на фрагменти однакової тривалості. Для кожного фрагмента обчислювався ембедінг за допомогою нейронної мережі ruannote-audio. Після цього здійснювалось порівняння ембедінгів для фрагментів аудіозаписів одного й того самого мовця (внутрішньо-особова подібність) та фрагментів аудіозаписів різних мовців (міжособова відмінність). У кожному з випадків порівняння проводилося для всіх ембедінгів фрагментів першого аудіозапису з усіма ембедінгами фрагментів другого аудіозапису (Рис. 1).

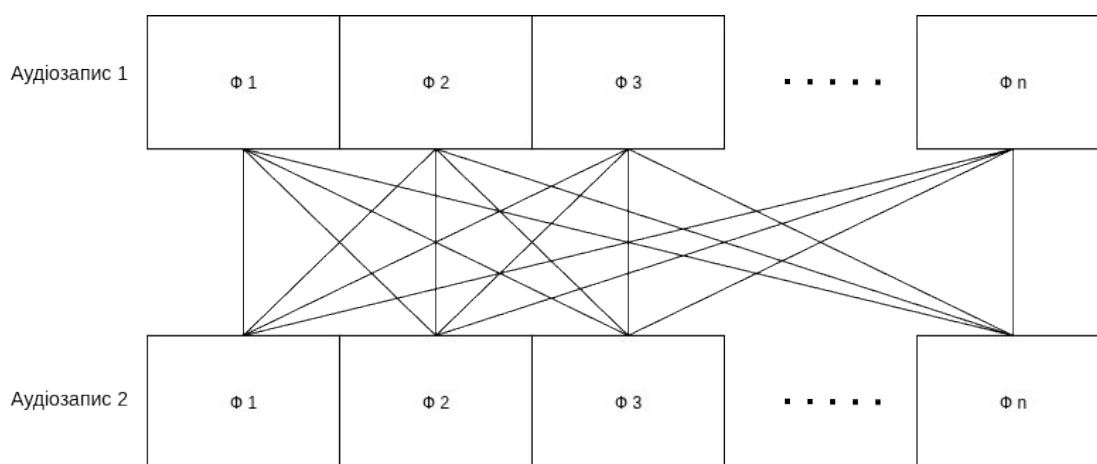


Рис. 1. Схематичне зображення способу порівняння ембедінгів фрагментів аудіозаписів: Φ_i – i -ий фрагмент аудіозапису, де $i = 1, 2, \dots, n$

Для визначення коефіцієнтів подібності двох ембедінгів застосовувались різні відомі метрики [4], зокрема:

- косинусна відстань,
- евклідова відстань,
- манхеттенська відстань,
- відстань Канберри,
- відстань Брея-Кертиса,
- відстань Чебишева,
- кореляційна відстань.

Обчислені значення коефіцієнтів подібності використовувалися для визначення ширини зони захисту (Рис. 2), яка однозначно пов'язана із точністю та стабільністю ідентифікації «свій/чужий», а також отримання залежності ширини зони захисту від тривалості фрагментів аудіозапису.

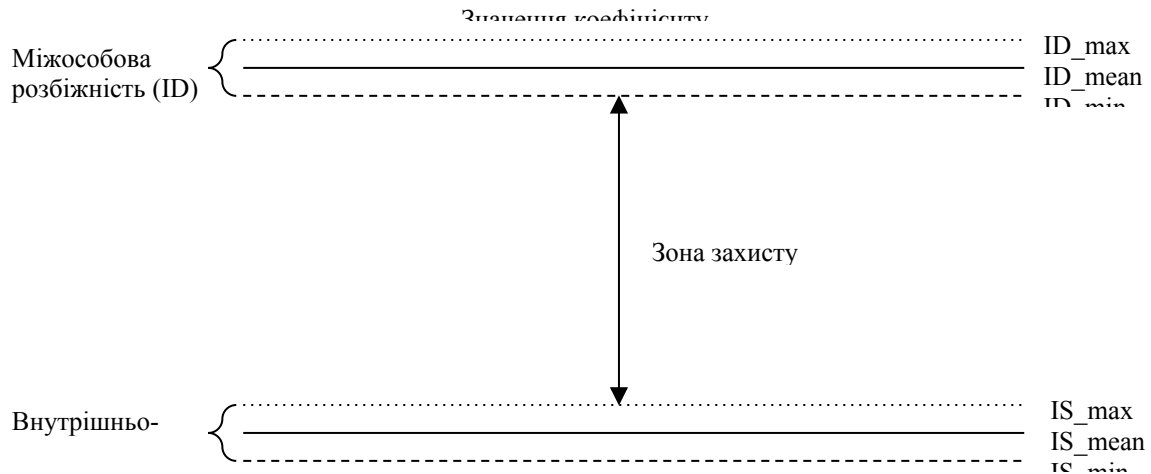


Рис. 2. Графічне зображення зони захисту

Ширина зони захисту ΔZ обчислюється за формулою:

$$\Delta Z = ID_{min} - IS_{max} ,$$

де ID_{min} – мінімальне значення коефіцієнтів подібності матриці невідповідності для міжособової розбіжності, IS_{max} – максимальне значення коефіцієнтів подібності матриці невідповідності для внутрішньо-особової подібності.

Крім того проводився порівняльний аналіз коефіцієнтів подібності для визначення оптимальної метрики, яка б забезпечувала найбільшу ширину зони захисту.

4.1. Формування набору даних та отримання ембедінгів

Набір даних для проведення досліджень складається з 5 аудіозаписів, кожен з яких озвучує однаковий текст англійською мовою різними голосами. У вибірці присутні 2 жіночі та 3 чоловічі голоси. Ці аудіозаписи згенеровано нейромережею, Google TTS Chirp-3, яка дозволяє перетворити текст в аудіозапис [5]. Такий спосіб створення аудіозаписів зумовлений тим, що немає в наявності довготривалих аудіозаписів однакового тексту різними мовцями.

Всі аудіозаписи створені з такими параметрами:

- Формат: WAV.
- Частота дискретизації: 24 кГц.
- Бітова швидкість: 384 кбіт/сек.
- Канали: 1.
- Тривалість: 60 хв.

Для отримання ембедінгів фрагментів аудіозаписів обрано нейронну мережу ruannote-audio [6], яка перетворює аудіосигнал у ембедінг із 512 числових значень. Попередньо проведений аналіз показав, що ця нейронна мережа забезпечує більш точне розпізнавання мовців у довгих або змішаних аудіозаписах [7]. Проводились дослідження для різної тривалості фрагментів аудіозаписів, а саме: 5, 6, 8, 10, 12, 15, 20, 24, 30, 40, 60, 144, 150, 180, 200, 225, 240, 300, 360, 400, 450 та 600 секунд. Тривалість фрагментів вибиралась такою, щоб аудіозапис ділився на n рівних частин без остачі, наприклад, 720 фрагментів по 5 секунд, 600 фрагментів по 6 секунд тощо.

4.2. Опис метрик

Метрика “Косинусна відстань” – вимірює кут між двома векторами у багатовимірному просторі, незалежно від їхньої довжини. Часто використовується в задачах розпізнавання різноманітних об’єктів, де важливий напрям, а не модуль вектора. Дана метрика розраховується за формулою:

$$d_{\cos}(A, B) = 1 - \frac{\sum_{i=1}^n a_i \cdot b_i}{\sqrt{\sum_{i=1}^n a_i^2} \cdot \sqrt{\sum_{i=1}^n b_i^2}},$$

де A, B – вектори, a_i, b_i – координати вектора A, B відповідно.

Метрика “Евклідова відстань” є стандартною функцією у n -вимірному евклідовому просторі, значення якої обчислюється як квадратний корінь із суми квадратів різниць відповідних координат двох векторів:

$$d_{euc}(A, B) = \sqrt{\sum_{i=1}^n (a_i - b_i)^2}.$$

Метрика “Манхеттенська відстань” визначається як сума абсолютних значень різниць між координатами векторів:

$$d_{manh}(A, B) = \sum_{i=1}^n |a_i - b_i|.$$

Ця метрика залежить від обертання системи координат, але не залежить від дзеркального відбиття відносно осі координат або паралельного перенесення [8].

Метрика “Відстань Канберри” є зваженою нормалізованою формою манхеттенської відстані, що обчислюється за формулою:

$$d_{canb}(A, B) = \sum_{i=1}^n \frac{|a_i - b_i|}{|a_i| + |b_i|}.$$

Метрика характеризується підвищеною чутливістю до малих значень координат, оскільки кожен доданок нормалізується відносно абсолютної суми компонент [9].

Метрика “Відстань Брея-Кертіса” є симетричною і нормалізованою та визначається за функцією:

$$d_{bc}(A, B) = \frac{\sum_{i=1}^n |a_i - b_i|}{\sum_{i=1}^n (a_i + b_i)}.$$

Значення функції лежить у межах від 0 до 1, де 0 вказує на повну ідентичність, а 1 – на повну розбіжність двох векторів.

Метрика “Відстань Чебишева” визначається як максимальна абсолютна різниця між відповідними координатами векторів:

$$d_{cheb}(A, B) = \max(|a_i - b_i|).$$

Ця метрика найчастіше використовується в задачах, де важливими є максимальні відхилення за координатами вектора.

Метрика “Кореляційна відстань” визначається із використанням коефіцієнту кореляції Пірсона за формулою:

$$d_{corr}(A, B) = 1 - \frac{\sum_{i=1}^n (a_i - \underline{a})(b_i - \underline{b})}{\sqrt{\sum_{i=1}^n (a_i - \underline{a})^2 \cdot \sum_{i=1}^n (b_i - \underline{b})^2}},$$

де $\underline{a}$ та $\underline{b}$ – середні значення векторів.

Ця метрика може відображати лише лінійну кореляцію змінних, і не враховує інші типи взаємозв'язків.

5. Дослідження залежності похибки ідентифікації клонованого голосу від тривалості аудіосигналу

Враховуючи те, що коефіцієнти подібності пораховані за деякими метриками можуть мати значення за межами діапазону від 0 до 1, то для визначення оптимальної метрики, яка б забезпечувала найширшу зону захисту, необхідно пронормувати всі результати досліджень. Нормування проводимо до діапазону від 0 до 1. Потрібно наголосити, що також нормуються і коефіцієнти, які вже знаходяться в цьому діапазоні. Це необхідно для коректного порівняння, оскільки коефіцієнти майже ніколи не приймають максимальне значення рівне 1.

Нормування проводилось за таким алгоритмом:

1. Визначалися матриці невідповідності коефіцієнтів подібності.
2. Визначалось максимальне значення коефіцієнтів подібності.
3. Розраховувався коефіцієнт нормалізації за формулою $k = 1/\max$.
4. Всі коефіцієнти подібності матриці множилися на коефіцієнт нормалізації.

На Рис. 3 представлено матриці невідповідності, які отримано із використанням метрики “Відстань Канберри”, одна з яких є ненормованою (Рис. 3,а), інша – нормованою (Рис. 3,б).

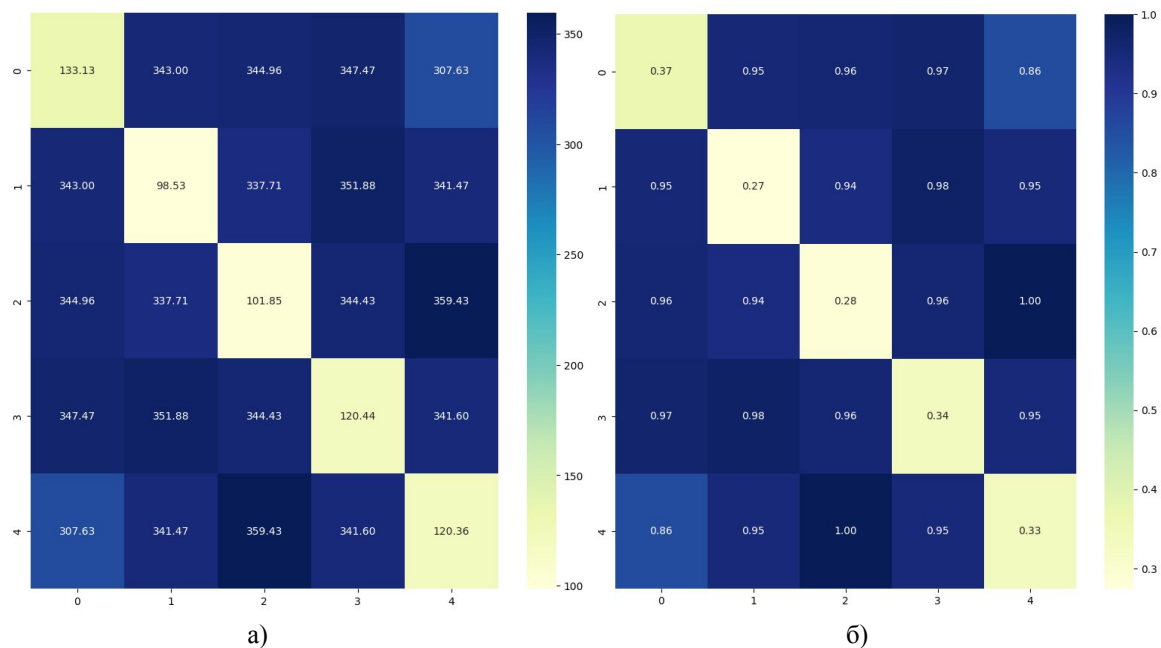


Рис. 3. Матриці невідповідності коефіцієнтів подібності, які отримано із використанням метрики “Відстань Канберри”: а – до нормування, б – після нормування.

На Рис. 4 зображено пронормовані матриці невідповідності коефіцієнтів подібності, які отримано із використанням метрики “Косинусна відстань” для двох значень тривалості фрагментів аудіозаписів: 5 с (Рис. 4,а) та 400 с (Рис. 4,б).

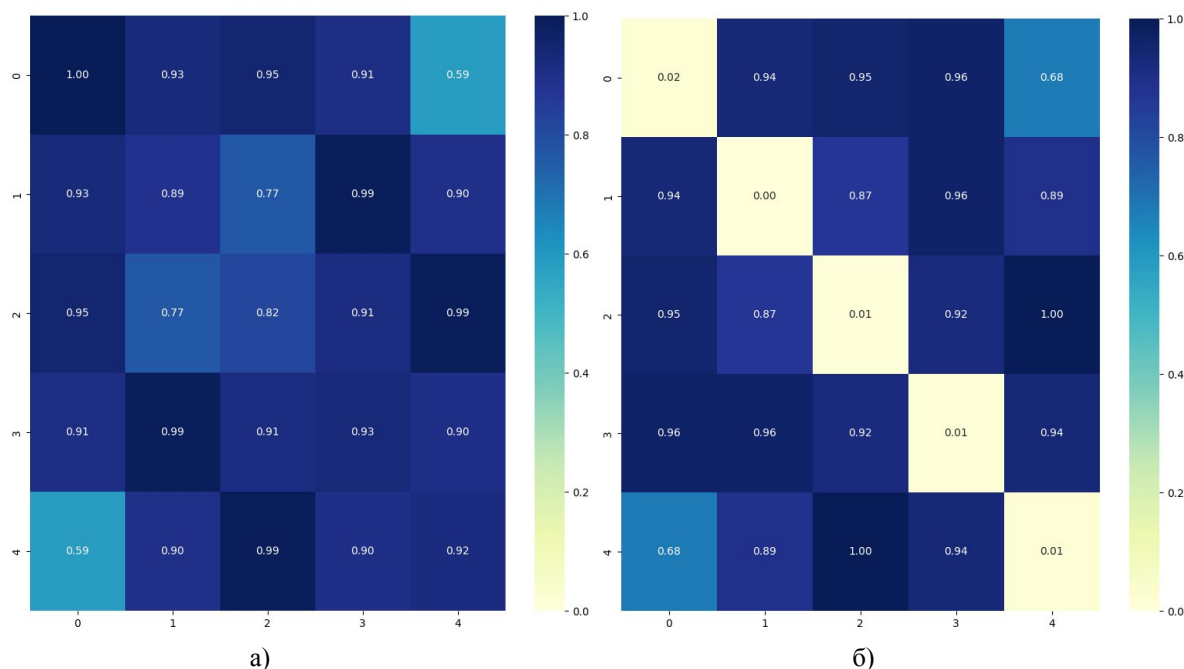


Рис. 4. Матриці невідповідності коефіцієнтів подібності, які отримано із використанням метрики “Косинусна відстань”: а – тривалість фрагментів аудіозаписів 5 с, б – 400 с.

За головною діагоналлю матриці відкладалися отримані максимальні значення коефіцієнтів подібності для фрагментів аудіозаписів одного й того самого мовця. Решта елементів матриці відображали мінімальні значення коефіцієнтів подібності для фрагментів аудіозаписів різних мовців. За тривалості фрагментів аудіозаписів 400 с (Рис. 4,б) забезпечується достатньо широка ширина зони захисту, яка відображається чіткою границею між результатами, отриманими для внутрішньо-особової подібності та міжособової відмінності. За тривалості фрагментів аудіозаписів 5 с (Рис. 4,а) зони захисту практично немає – деякі мінімальні значення коефіцієнтів подібності для міжособової відмінності є більшими за максимальні значення коефіцієнтів подібності для внутрішньо-особової подібності.

В Табл. 1 представлено результати дослідження коефіцієнтів подібності ID_min, IS_max та ширини зони захисту ΔZ для деяких значень тривалості фрагментів аудіозаписів, а саме 5, 20, 100, 200, 300 та 400 секунд, для семи досліджуваних метрик.

Табл 1. Результати дослідження коефіцієнтів подібності ID_min, IS_max та ширини зони захисту ΔZ

	5			20			100		
Метрики	ID_min	IS_max	ΔZ	ID_min	IS_max	ΔZ	ID_min	IS_max	ΔZ
Косинусна	0,59	1	-0,41	0,57	0,54	0,03	0,55	0,16	0,39
Евклідова	0,77	1	-0,23	0,75	0,73	0,02	0,8	0,42	0,38
Манхетенська	0,76	1	-0,24	0,74	0,72	0,02	0,78	0,42	0,36
Канберри	0,81	1	-0,19	0,81	0,85	-0,04	0,85	0,62	0,23
Брея-Кертіса	0,68	1	-0,32	0,64	0,64	0	0,69	0,33	0,36
Чебишева	0,46	1	-0,54	0,74	0,98	-0,24	0,75	0,5	0,25
Кореляційна	0,59	1	-0,41	0,56	0,54	0,02	0,63	0,18	0,45

	200			300			400		
Метрики	ID_min	IS_max	ΔZ	ID_min	IS_max	ΔZ	ID_min	IS_max	ΔZ
Косинусна	0,57	0,09	0,48	0,60	0,06	0,54	0,68	0,04	0,64
Евклідова	0,81	0,32	0,49	0,68	0,07	0,61	0,83	0,19	0,64
Манхеттенська	0,79	0,31	0,48	0,79	0,25	0,54	0,8	0,19	0,61
Канберри	0,85	0,52	0,33	0,84	0,44	0,40	0,86	0,37	0,49
Брея-Кертіса	0,69	0,24	0,45	0,69	0,19	0,5	0,65	0,13	0,52
Чебишева	0,71	0,33	0,38	0,72	0,26	0,46	0,7	0,19	0,51
Кореляційна	0,65	0,1	0,55	0,68	0,07	0,61	0,68	0,04	0,64

На Рис. 5 зображено залежності максимальних значень коефіцієнтів подібності для внутрішньо-особової подібності (крива IS_max) та мінімальних значень коефіцієнтів подібності для міжособової відмінності (крива ID_min), які отримано із використанням метрики “Косинусна відстань”, від тривалості фрагментів аудіозаписів.

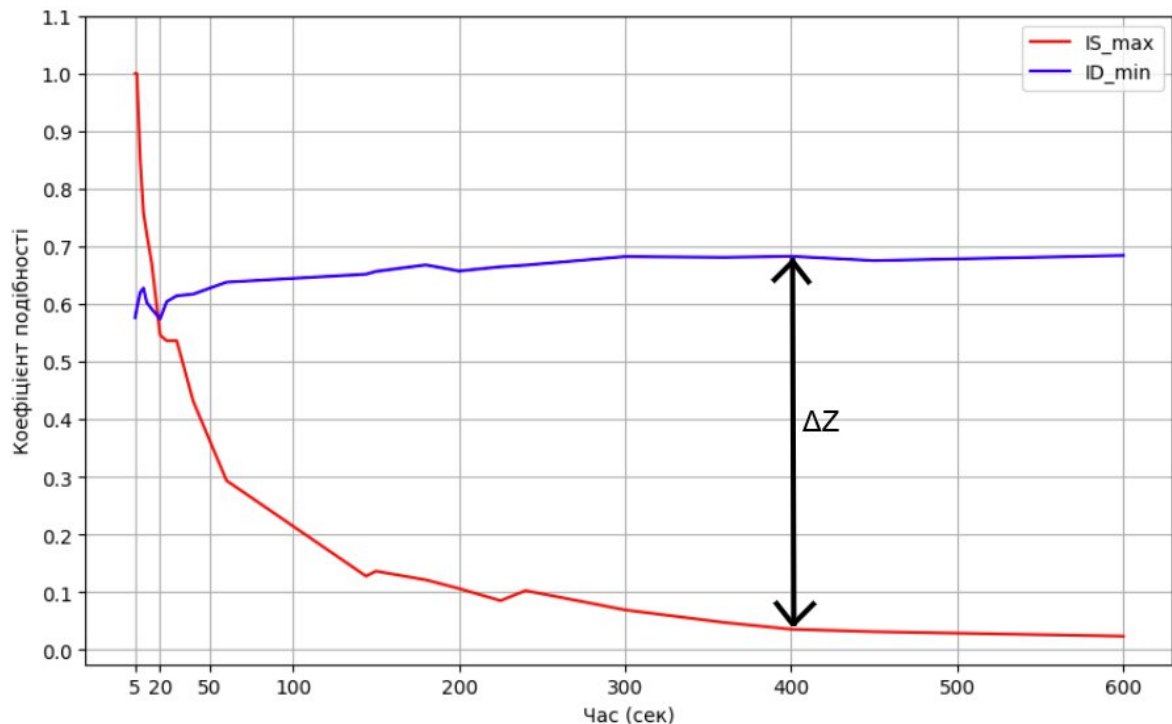


Рис. 5. Залежність ширини зони захисту від тривалості фрагментів аудіозаписів

Як видно з Рис. 5, для значення тривалості фрагментів аудіозаписів 5 секунд, спостерігається від'ємне значення зони захисту (Табл. 1), що свідчить про низьку точність ідентифікації мовця за голосом. Проте збільшення тривалості фрагментів (Рис. 5) призводить до появи та розширення зони захисту, що забезпечує вищу надійність розпізнавання «свій/чужий».

Для діапазону значень тривалості фрагментів аудіозаписів від 5 до 20 секунд отримано від'ємні або близькі до 0 значення зони захисту, що свідчить про її відсутність. Для значень тривалості фрагментів більших за 20 секунд маємо чітко виражену зону захисту (Рис. 5). Так, наприклад, для значення 100 секунд ширина зони захисту для метрики “Косинусна відстань” становить 0,39, а для 400 секунд – 0,64 (Табл. 1).

За результатами порівняння різних метрик (Табл. 1) для тривалості фрагментів аудіозаписів 400 секунд найширшу зону захисту (більше 0,61) забезпечують 4 метрики: “Косинусна відстань”, “Кореляційна відстань”, “Евклідова відстань” та “Манхеттенська відстань”. Проте зі зменшенням тривалості найкращі результати

забезпечує метрика “Кореляційна відстань”. Так для значення 100 секунд ширина зони захисту становить 0,45, а для інших метрик ширина не перевищує 0,39.

Висновок

На основі проведених досліджень встановлено, що зі збільшенням тривалості фрагментів аудіозаписів спостерігається суттєве зростання ширини зони захисту, що дозволяє забезпечити вищу точність та стабільність ідентифікації «свій/чужий». Існує певне значення тривалості фрагментів (400 секунд), збільшення якого не призводить до суттєвого зростання ширини зони захисту. Для значення тривалості фрагментів аудіозаписів 5 секунд, яке переважно використовується у задачах ідентифікації мовців, отримано від’ємне значення зони захисту, що свідчить про низьку точність ідентифікації мовця за голосом. Чітко виражену зону захисту отримано для значень тривалості фрагментів аудіозаписів більших за 20 секунд. За результатами порівняння різних метрик для відносно довготривалих фрагментів аудіозаписів, найширшу зону захисту забезпечують 4 метрики: “Косинусна відстань”, “Кореляційна відстань”, “Евклідова відстань” та “Манхеттенська відстань”. Проте зі зменшенням тривалості найкращі результати забезпечує метрика “Кореляційна відстань”.

В подальшому планується застосування отриманих результатів для відсіювання синтезованих голосів під час ідентифікації мовця.

Подяка

Автори вдячні співробітникам кафедри інформаційно-вимірювальних технологій НУ «Львівська політехніка» за допомогу в оформленні статті до публікації.

Конфлікт інтересів

Автори заявляють про відсутність будь-якого фінансового або іншого можливого конфлікту, який стосується цієї роботи.

Список літератури

- [1] Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. *Interspeech 2020*. [Online]. Доступ: https://www.isca-speech.org/archive/Interspeech_2020/pdfs/1137.pdf
- [2] Bredin, H., Laurent, A., & Gillies, A. (2020). Pyannote.audio: neural building blocks for speaker diarization. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)*. [Online]. Доступ: <https://ieeexplore.ieee.org/document/9053198>
- [3] Х. Руда, Д. Сабодашко, Г. Микитин, М. Швед, С. Бордуляк, Н. Коршун (2024). “Особливості створення та розповсюдження фішингових веб-ресурсів,” *Cybersecurity: Education, Science, Technique*, vol. 2, no. 12, pp. 80–88, 2021. [Online]. Доступ: <https://csecurity.kubg.edu.ua/index.php/journal/article/view/645/508>
- [4] A. Levy, B. Riva Shalom, and M. Chalamish, “A guide to similarity measures and their data science applications,” *Journal of Big Data*, vol. 12, Art. no. 188, 2025. [Online]. Доступ: <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-025-01227-1>
- [5] Google Cloud, “Chirp 3 HD: High-quality, low-latency text-to-speech model,” *Google Cloud Text-to-Speech*, 2024. [Online]. Доступ: <https://cloud.google.com/text-to-speech/docs/chirp3-hd>
- [6] B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification,” *arXiv preprint arXiv:1911.01255*, 2019. [Online]. Доступ: <https://arxiv.org/abs/1911.01255>
- [7] H. Bredin, A. Laurent, and Y. Zhong, “End-to-end domain-adversarial voice activity detection,” *Proc. Interspeech 2021*, pp. 4658–4662, 2021. [Online]. Доступ: https://www.isca-archive.org/interspeech_2021/bredin21_interspeech.pdf
- [8] E. F. Krause, *Taxicab Geometry: An Adventure in Non-Euclidean Geometry*. New York, NY, USA: Dover Publications, 1986.
- [9] T. Souravlas, I. Roumeliotis, C. Roumeliotis, and C. Zissis, “Time series similarity measures and deep learning: State-of-the-art review,” *arXiv preprint arXiv:2412.20574*, 2024. [Online]. Доступ: <https://arxiv.org/abs/2412.20574>

Дані про авторів

Поворозник Володимир Богданович
Аспірант
Кафедра «Інформаційно-вимірювальні технології»
Національний університет «Львівська політехніка»
вул. С. Бандери, 12, м. Львів, Україна, 79013
E-mail: volodymyr.b.povoroznyk@lpnu.ua
Контактний тел.: +38 098 22 72 971

Микитин Ігор Петрович
Доктор технічних наук, професор
Кафедра «Інформаційно-вимірювальні технології»
Національний університет «Львівська політехніка»
вул. С. Бандери, 12, м. Львів, Україна, 79013
E-mail: ihor.p.mykytyn@lpnu.ua
Контактний тел.: +38 097 66 31 411
Кількість статей у загальнодержавних базах даних – 80
Кількість статей у міжнародних базах даних – 6
Номер ORCID: <https://orcid.org/0000-0002-7157-3725>