УДК 004.8

АУДІО-АУТЕНТИФІКАЦІЯ КОРИСТУВАЧА ЗА ГОЛОСОМ

Д. Зарецька, М. Баранов, С. Іванов

Львівський національний університет імені Івана Франка,
вул. Університеська, 1, Львів, 79000,
e-mail: dana.zaretska@lnu.edu.ua,
mykola.baranov@lnu.edu.ua, serhii.ivanov@lnu.edu.ua

Розв'язування задач ідентифікації та аутентифікації людини за голосом опираються на створення методів, що базуються на знаходженні відмінностей між мовцями. Ці методи широко застосовують для створення систем безпеки та методів діаризації (виділення кількох мовців з єдиного аудіопотоку). Системи розпізнавання мовців використовують голос людини як ключовий інструмент для перевірки особистості.

Разом з методами аудіоаналізу часто використовують моделі на підставі штучних нейронних мереж. Моделі, отримуючи інформацію з аудіо-сигналів, навчаються їх класифікувати, розпізнавати та по-різному з ними взаємодіяти, залежно від сформульованої задачі. Розпізнавання мовця може бути контекстно-залежним, або ж контекстно-незалежним. Від вигляду сформульованої задачі залежить специфіка алгоритму та особливості його розробки. У роботі з контекстно-залежним розпізнаванням від користувача вимагається повторення однієї фрази, тоді як контекстно-незалежне розпізнавання не потребує повторення фіксованого тексту, є зручнішим для користувача, але складнішим у розробці. Подано алгоритм контекстно-незалежного розпізнавання, що складається з двох основних частин. Першою є опрацювання та фільтрація вхідного аудіопотоку, а другою – тренування Siamese neural network. Основними математичними підходами, які використовували для розробки алгоритму, були швидке перетворення Фур'є як частина поняття спектра, мел-спектрограма, якою візуалізується розподіл частот, та triplet loss як важлива складова Siamese neural network. Дані для тренування моделі подавали у вигляді мел-спектрограм та були згенеровані з вхідного аудіопотоку в першій частині алгоритму. Запропонована модель була натренована на датасеті Speaker Recognition Audio Dataset fff61aed-e з точністю розпізнавання 97,9%.

Ключові слова: аудіоаналіз, нейронні мережі.

1. Задача аудіо-аутентифікації мовця

Застосування методів **аудіоаналізу** є вирішенням багатьох задач, – детекція сигналів, подій, розпізнавання музики чи мови, сентиментальний аналіз, генерація звуків. Методи аудіоаналізу також віднайшли своє місце в галузі біології (біоакустика, океанографія).

Дослідження, які стосувалися задачі розпізнавання голосу, вперше були проведені у 1937 році Франчесою Макгі. В 1942 в лабораторіїї Белла відкрили поняття спектрограми, застосування якої є дуже широким у задачах аудіоаналізу. Перші спроби створення системи розпівання мовця, використанням кореляції двох цифрових спектрограм були проведені ще у 1960 Прузнаським, також у лабораторії Белла. У 1970 компанія Texas Instrument systems створила першу автоматизовану систему верифікації мовця, використовуючи цифрові фільтри для спектрального аналізу. 1980 та 1990 характеризуються значним прогресом в цій ділянці, багато сучасних систем розпізнавання базуються на методах, що були винайдені в той проміжок

часу (такими є прихована марковська модель, кепструм, N-грама). Також у 1980 відбулися вдосконалені спроби застосування штучних нейронних мереж для задач розпізнавання сигналів. Перше представлення нейронних мереж відбулось ще у 1950, проте через труднощі практичної імплементації своєї популярності в цьому домені вони набули значно пізніше [1].

Задачі розпізнавання мовця поділяються на дві групи:

– контекстно-залежні (text-dependent methods);

– контекстно-незалежні (text-independent methods).

Контекстно-залежні характеризуються тим, що текст або ж фраза, яку промовляє людина, однакова для всіх мовців. Розв'язування таких задач дещо простіше. Контекстно-незалежні методи не передбачають начитку однакового тексту та мають працювати незалежно від того, що каже мовець [2].

Сучасні алгоритми голосової біометрії часто базуються на використанні мел-частотних кепстральних коефіцієнтів, динамічної трансформації часової шкали, прихованої марковської моделі. Алгоритм, який використали ми, бере за основу використання зображень мел-спектрограм як голосових ознак і нейронну модель-класифікатор для розпізнавання.

2. Набір даних, який використовували

Для дослідження використано набір даних Speaker Recognition Audio Dataset fff61aed-e, представлений на платформі для змагань з аналітики та передбачувального моделювання Kaggle у 2018. Дані складаються з 50-ти аудіодоріжок, кожна з яких містить запис голосу однієї людини тривалістю 60 хвилин. Текст, який промовляється, є унікальним для кожного мовця, тому для таких даних задача полягатиме у **контекстно-незалежному** розпізнаванні. Згодом дві аудіодоріжки були видалені з набору через наявність спотворювальних шумів та інших голосів у записах.

3. Опрацювання даних

Для тренування нейронної мережі дані були опрацьовані. Над кожним аудіозаписом відбувся процес видалення тиші. Вирізалися ті часові проміжки, в яких максимальна гучність не перевищувала 15 дБ.

Отримані дані було поділено на рамки тривалістю 4 секунди кожна. Тренувальні дані містили 8-15 рамок на кожного мовця. У валідаційних даних таких рамок було 5-8. Тестові дані містили по три рамки на кожну людину. Дані повністю унікальні для тренувального, валідаційного та тестового набору даних, не повторюються в одному чи різних наборах.

Для опису перетворень, які застосовували над даними, варто визначити поняття *спектра*. Аудіосигнал складається з кількох одночастотних звукових хвиль. Перетворення Фур'є дає змогу розкласти сигнал на окремі частоти та амплітуди цих частот. Іншими словами, перетворює сигнал з часового виміру в частотний. Результат такого перетворення називається **спектром**. На практиці для цього використовують швидке перетворення Фур'є, яке є швидкою версією дискретного перетворення Фур'є [4].

Формула дискретного перевторення Фур'є подана нижче

$$A_k = \sum_{n=0}^{N-1} e^{-i\frac{2\pi}{N}kn} a_n, \tag{1}$$

де a_n – сигнал, з якого потрібно знайти спектр. Дані кожної рамки були відображені у вигляді *мел-спектрограми*.

Звичайна **спектрограма** є відображенням спектра частот у часі, демонструє залежність між силою спектруму та часом. Нею можна відобразити зміну гучності сигналу у часі. Вісь Y представляє частоти, а вісь X – зміну в часі. Нижче відображено пари **спектрограм** для двох мовців.

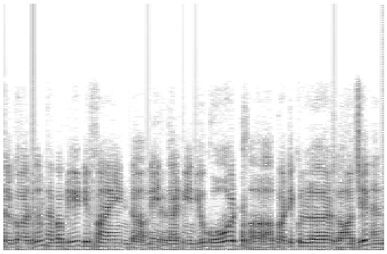

Рис. 1. Спектрограма 1 мовця А

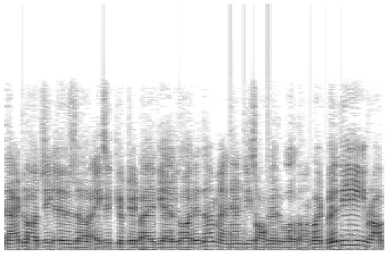

Рис. 2. Спектрограма 2 мовця А

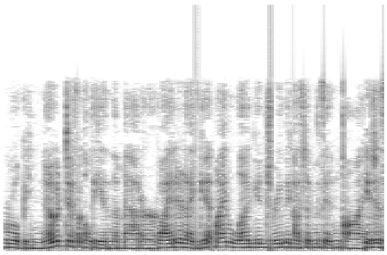

Рис. 3. Спектрограма 1 мовця Б

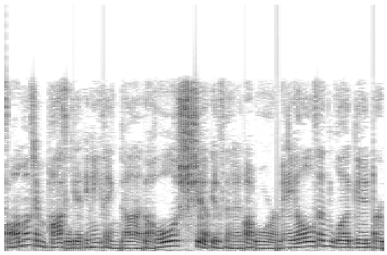

Рис. 4. Спектрограма 2 мовця Б

Для порівняння варто відобразити мел-спектрограми. **Мел-спектрограма** утворена за допомогою певного логарифмування значень по осі Y. Зазвичай перетворення лінійної шкали у мел-шкалу відбувається за такою формулою:

$$H(f) = 1127 \cdot \ln\left(1 + \frac{f}{700}\right). \tag{2}$$

Нижче зображено пари **мел-спектрограм**, згенерованих для різних мовців.

Саме таке відображення звукових даних використане для тренування моделі. Причина цього те, що логарифмічна спектрограма [3] досить репрезентативне відображення для кожного мовця. Чітко проглядаються патерни, за якими можна

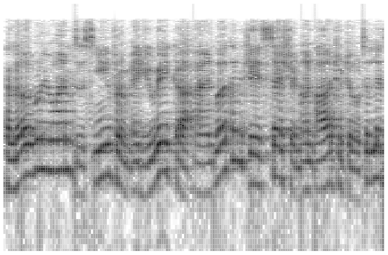

Рис. 5. Мел-спектрограма 1 мовця А

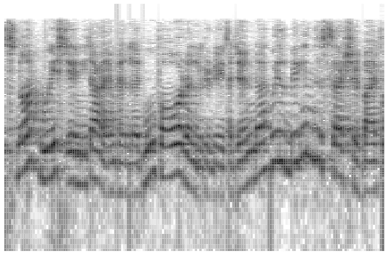

Рис. 6. Мел-спектрограма 2 мовця А

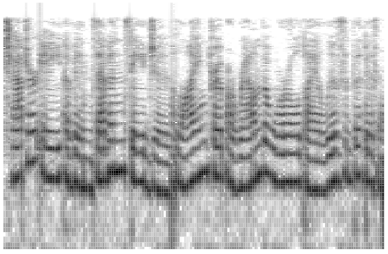

Рис. 7. Мел-спектрограма 1 мовця Б

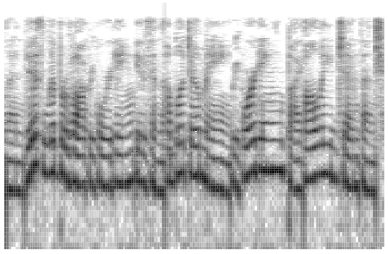

Рис. 8. Мел-спектрограма 2 мовця Б

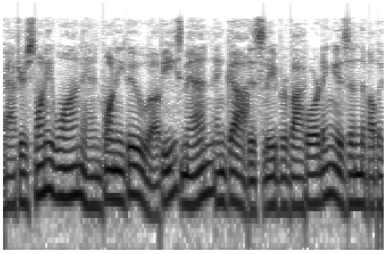

Рис. 9. Мел-спектрограма 1 мовця В

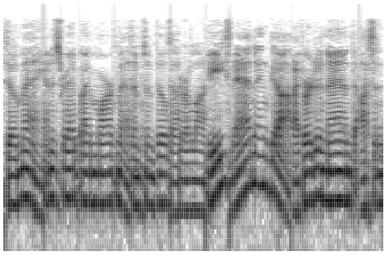

Рис. 10. Мел-спектрограма 2 мовця В

помітити схожість двох зразків голосу однієї людини та відмінність між різними мовцями. Такі відмінності не помітні у разі використання звичайної спектрограми.

4. Запропонована архітектура нейронної мережі

Як класифікаційну модель використано **сіамську нейронну мережу** [5]. Сіамська нейронна мережа особлива тим, що на вхід приймає позитивні та негативні пари. Позитивними парами у випадку цієї задачі будуть дві мел-спектрограми,

згенеровані з голосів однієї людини, а негативними – ті, що згенеровані з голосів двох різних мовців. Отже, модель вчиться помічати подібності між двома звуковими відображеннями одного мовця та знаходити відмінності між різними. Нейронна мережа тренувалася на 10 епохах і має таку архітектуру.

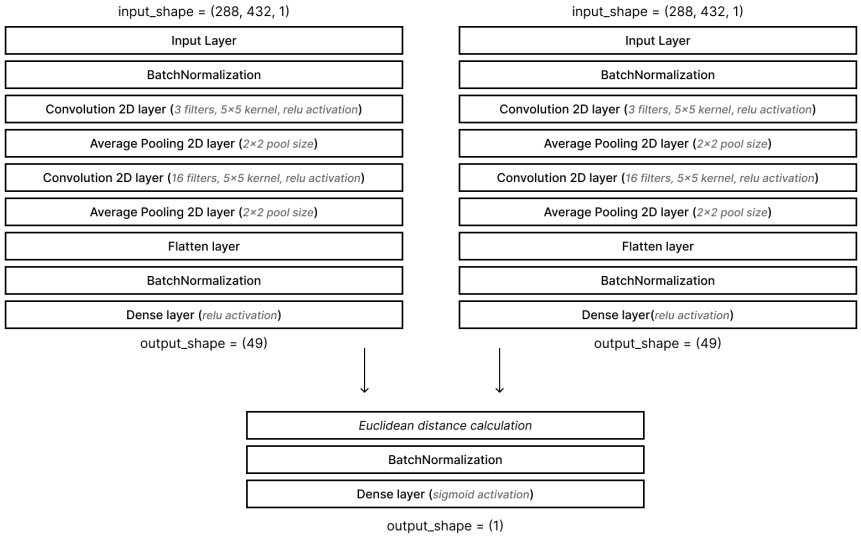

Рис. 11. Архітектура нейронної моделі

Опис шарів мережі:

- ***Input layer*** – вхідні дані(зображення мел-спектрограми розміром 432x288).
- ***Batch normalization layer*** – нормалізація вихідних з минулого шару даних.
- ***Convolutional 2D layer*** – згортковий шар нейронної мережі.
- ***Average Pooling 2D layer*** – усереднене агрегування.
- ***Flatten layer*** – перетворення матриці у вектор.
- ***Dense layer*** – шар, всі нейрони якого з'єднані з кожним нейроном минулого шару.

Приклад позитивної та негативної пари, які використовували для тренування моделі.

Як функцію-оптимізатор для моделі обрали *Adam*. Функцією втрат є *Triplet Loss*, хід роботи якої описується так [6].

1. З даних вибирається одне зображення(*якір*), а також позитивний і негативний зразок до вибраного якору(позитивним буде мел-спектрограма того ж мовця, негативним – іншого).
2. Вибрані зображення подають у вигляді векторів. Обчислюється **евклідова відстань** між якорем та позитивним і негативним зразком.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}. \tag{3}$$

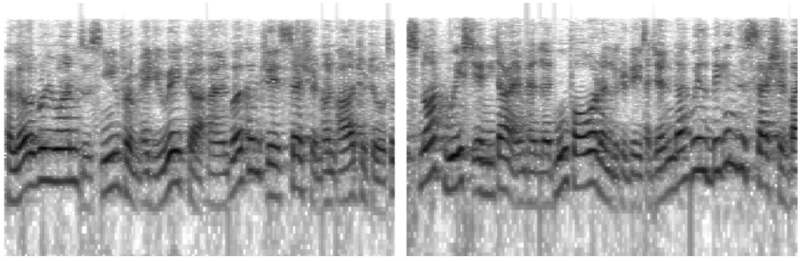

Рис. 12. Позитивна пара

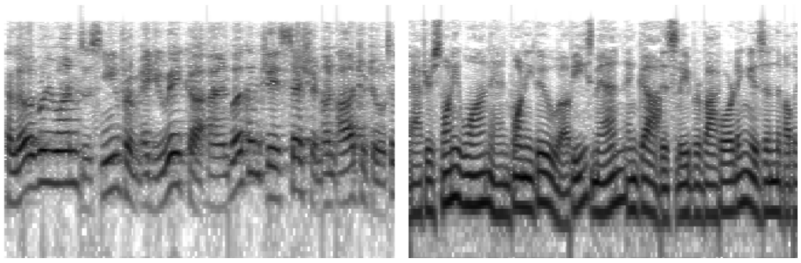

Рис. 13. Негативна пара

3. Обчислюється функція втрат(тут **triplet loss**) як різниця між відстанями між якорем і позитивним зразком й між якорем і негативним зразком.

   Описується так:

$$L = max(d(a,p) - d(a,n) + margin, 0). \tag{4}$$

   Для програмної реалізації використовували мову Python, бібліотеку librosa для роботи з аудіосигналом, numpy для опрацювання векторних даних, tensorflow та keras для роботи з нейронними мережами.

5. Результати та метрики

Для того, щоб описувати метрики для оцінки роботи моделі, варто визначити кілька понять, що будуть використовуватися під час розрахунку самих метрик. Такими є **TP**, **TN**, **FP** та **FN**.

Опис та інтерпретація нижчезазначених метрик стосується саме задач бінарної класифікації. Нехай існує дві групи об'єктів: ті, які належать до позитивного класу, та ті, що належать до негативного. В контексті цієї задачі об'єктом позитивного класу вважатимемо ту пару мел-спектрограм, де обидві згенеровані з голосу тієї самої людини. Тоді об'єктом негативного класу буде пара, де мел-спектрограми згенеровані з голосів, що належать різним людям.

- *TP (True Positive)* – кількість об'єктів позитивного класу, які були коректно класифіковані як ті, що належать до позитивного класу.
- *TN (True Negative)* – кількість об'єктів негативного класу, які були коректно класифіковані як ті, що належать до негативного класу.

- *FP (False Positive)* – кількість об'єктів негативного класу, які були некоректно класифіковані як ті, що належать до позитивного класу.
- *FN (False Negative)* – кількість об'єктів позитивного класу, які були некоректно класифіковані як ті, що належать до негативного класу.

Користуючись вищезазначеними поняттями, опишемо набір метрик, які використовують для оцінки роботи поданої класифікаційної моделі.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{8}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FT)(TP + FP)(TN + FP)(TN + FN)}}. \tag{9}$$

Значення кожної з вищеописаних метрик для задачі авторизації користувача такі:

Результати експериментів

Метрика	Accuracy	Recall	Precision	F1	MCC
Значення	0,9792	1,0	0,96	0,9796	0,9592

Оцінку моделі було проведено на тестувальній вибірці, тобто на тих даних, які не були використані для тренування моделі.

Як видно з таблиці, натренована модель здатна розрізняти голоси різних людей і коректно ідентифіковувати кілька зразків як ті, що належать одній людині.

Нижче зображено графіки того, як змінюється *loss*-функція та *Accuracy* на кожній епосі. Суцільною лінією позначено таку зміну на тренувальних даних, пунктиром – на валідаційних (рис. 14 та 15).

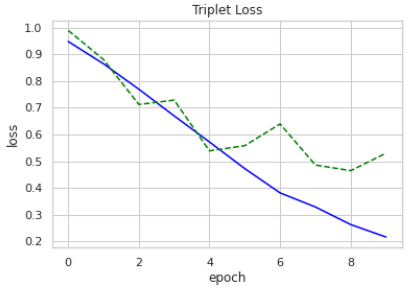

Рис. 14. Зміна *triplet loss* в залежності від кількості епох

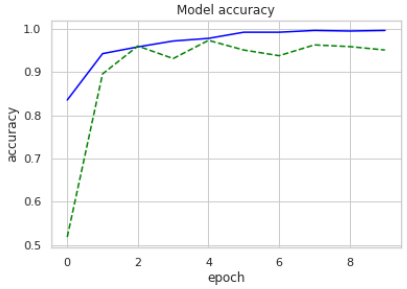

Рис. 15. Зміна *Accuracy* в залежності від кількості епох

Наведений алгоритм аудіо-аутентифікації має певні обмеження:

1) Наявність шумів в аудіодоріжці спотворює процес видалення тиші, рамки можуть бути подані некоректно, що ускладнить процес правильної авторизації. Способом вирішення цієї проблеми може бути застосування певних фільтрів перед етапом видалення тиші;
2) Наявність кількох голосів в аудіозаписі може непередбачувано впливати на роботу класифікатора.

6. Висновки

На підставі дослідження різних методів відображень аудіосигналу було запропоновано алгоритм авторизації користувача за голосом з використанням відображень мел-спектрограми та використанням сіамської нейронної мережі як моделі класифікатора. Проведений аналіз метрик підтверджує ефективність цього алгоритму ($Accuracy = 97.9\%$). В подальшому планується розширити такий алгоритм використанням фільтрів для видалення шумів з аудіосигналу.

Список використаної літератури

1. Sadaoki Furui 50 Years of Progress in Speech and Speaker Recognition Research / Furui Sadaoki // ECTI Transactions on Computer and Information Technology. – November 2005.
2. Mobiny A. Text-Independent Speaker Verification Using Long Short-Term Memory Networks / A. Mobiny, M. Najarian // Department of Electrical and Computer Engineering, University of Houston. – September 2018.
3. Pham Lam A Multi-spectrogram Deep Neural Network for Acoustic Scene Classification / Lam Pham, Ian McLoughlin, Huy Phan, Ramaswamy Palaniappan // Detection and Classification of Acoustic Scenes and Events. – 2019
4. Heckbert P. Fourier Transforms and the Fast Fourier Transform (FFT) Algorithm / P. Heckbert. – Feb. 1995.
5. Koch G. Siamese Neural Networks for One-shot Image Recognition [Електронний ресурс] / G. Koch, R. Zemel, R. Salakhutdinov // Department of Computer Science. – University of Toronto. – Режим доступу: https://www.cs.cmu.edu/ rsalakhu/papers/oneshot1.pdf
6. Baranwal Tanish Facial Re-Identification Using Siamese Nets [Електронний ресурс] // Tanish Baranwal. – 2020. – Режим доступу: https://towardsdatascience.com/facial-re-identification-using-siamese-nets-d36df39da7c0

Стаття: надійшла до редколегії 15.09.2022
доопрацьована 12.10.2022
прийнята до друку 19.10.2022

PERSON AUTHENTICATION BY VOICE

D. Zaretska, M. Baranov, S. Ivanov

Ivan Franko National University of Lviv,
Universytetska str., 1, Lviv, 79000, Ukraine,
e-mail: dana.zaretska@lnu.edu.ua,
mykola.baranov@lnu.edu.ua, serhii.ivanov@lnu.edu.ua

Creating methods that allow finding some distinctions between different speakers form the basis when solving tasks of identification and authentication of individuals by their voice. Those methods are used extensively in developing security systems and methods

of diarisation (partitioning an input audio stream into homogeneous segments according to the speaker identity). Voice recognition systems use the voice of a person as a crucial instrument for human identification. Different methods of audio analysis are often accompanied by neural network models. Those models learn the ways to classify, recognise and interact with audio signals. The audio analysis methods appeared to be the solution to many problems, such as signals or events detection, language and music recognition, sentimental analysis, and sound generation. Moreover, those methods found their place in the field of biology (bioacoustics, oceanography). Two types of speaker verification systems are text-dependent and text-independent. Text-dependent speaker verification requires the speaker to say exactly the given text or password, while text-independent speaker verification is more convenient because the person can speak freely to the system, but more difficult in creating. In this work, we present an algorithm for a text-independent speaker verification system. The algorithm consists of two important parts. The first part is processing and filtering the input audio stream in order to create training data, and the second is training the Siamese neural network model. The main mathematical methods used in the algorithm implementation were Fourier transform, as part of the spectrum concept, the mel-spectrogram, which shows the frequencies that make up the sound, and the triplet loss, as an important part of the Siamese network model. The data used for model training was a set of mel-spectrograms and it was produced directly from the audio stream in the first part of the algorithm. We trained our model on Speaker Recognition Audio Dataset fff61aed-e and achieved 97,9% accuracy.

Key words: audio analysis, neural networks.