

УДК 004.934:[004.85+519.6]

Д-р. ф., старший викладач Жук І. С., асистент Громова В.В.,
студент Романецький М. С.

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

МЕТОД ДЛЯ ВИДІЛЕННЯ ГОЛОСУ ЦІЛЬОВОГО МОВЦЯ ЗА АУДИОЗАПИСАМИ

Abstract

Ivan Zhuk, PhD, Senior Lecturer; assistant Gromova V.V.,
Romanetskyi Mykyta, student.

Voice isolation application for identifying suspect from audio recordings

The paper presents the development of an application for isolating a suspect's voice from multispeaker audio recordings. The approach combines short-time Fourier transform (STFT), speech segmentation, embedding extraction and speaker clustering to identify and mute non-target speakers. The system achieves clear separation of individual voices in realistic recordings with ambient noise.

Вступ

Обробка мовлення є одним із ключових напрямів сучасних досліджень у галузі штучного інтелекту, цифрової обробки сигналів та криміналістики. Аудіозаписи часто містять накладання голосів, сторонні шуми або ефекти відлуння, що ускладнює розпізнавання окремих мовців.

У сучасних системах аналізу мовлення активно впроваджуються методи глибинного навчання, які показують високу ефективність у завданнях ідентифікації мовців і розділення джерел. Зокрема, роботи Wang і Chen (2018) [1] показали, що використання нейронних архітектур типу Deep Clustering і Deep Attractor Network дозволяє значно підвищити якість відокремлення голосів у шумових умовах. Сучасні моделі, такі як SepFormer (Subakan et al., 2021) [2] і Conv-TasNet (Luo & Mesgarani, 2019) [3], забезпечують якість розділення, близьку до людського сприйняття.

Водночас, більшість сучасних досліджень зосереджені на відокремленні джерел у контрольованих умовах. Як зазначає Xiao-Lei Zhang [4], розроблення стійких до шуму систем ідентифікації мовців та моделей, здатних працювати у складних акустичних середовищах, є одним із ключових напрямів розвитку сучасної аудіо-аналітики. Це підкреслює актуальність досліджень, спрямованих на створення програмних засобів для

автоматичного виділення голосу цільового мовця в умовах реальних записів.

Постановка задачі

У сучасних умовах швидкого розвитку систем розпізнавання мовлення та аудіо-аналізу зростає потреба у створенні технологій, здатних працювати записами, що містять декількох мовців, накладання голосів та фонові шуми.

Метою дослідження є розробка методу, що забезпечує ідентифікацію цільового мовця за аудіозаписом і приглушення голосів інших учасників розмови. Об'єктом дослідження є процес цифрової обробки мовленнєвих сигналів, а предметом – математичні методи та алгоритми виділення голосу окремого мовця. Очікуваним результатом є програмний застосунок, здатний автоматично визначати часові межі мовлення цільового мовця, ізолювати його голос і зменшувати гучність або видаляти фрагменти, що належать іншим мовцям.

Методи та алгоритми

Запропонована система складається з послідовних етапів обробки. На першому етапі сигнал $x(t)$ розкладається на часово-частотне представлення за допомогою короткочасного перетворення Фур'є [5]:

$$X(\tau, \omega) = \int_{-\infty}^{\infty} x(t) \omega(t - \tau) e^{-j\omega t} dt \quad (1)$$

де $\omega(t)$ – віконна функція, що обмежує ділянку сигналу в межах короткого інтервалу часу. Для дискретного сигналу це перетворення має вигляд:

$$X[m, k] = \sum_{n=0}^{N-1} x[n + mH] \omega[n] e^{-j2\pi kn/N} \quad (2)$$

Де N – довжина вікна, H – крок зсуву, m – індекс вікна, k – частотна компонента.

Отримане представлення $X[m, k]$ дозволяє виявляти частотні особливості мовлення кожного мовця у часі та використовується як основа для подальших етапів аналізу.

На наступному кроці виконується сегментація мовлення [6] з метою виділення ділянок, що містять мовну активність. Кожен сегмент сигналу описується набором ознак, серед яких енергія:

$$E_i = \frac{1}{N} \sum_{n=0}^{N-1} s_i^2[n] \quad (3)$$

та спектральна відстань:

$$D_i = \sum_f ||S_i(f)| - |S_{i-1}(f)|| \quad (4)$$

Що допомагають виявити зміни станів мовлення у часі.

Після сегментації з кожного фрагмента екстрагуються ембедінги [7] мовця, які формуються глибинною нейронною мережею ESCAPA-TDNN [8]. Вектор ознак $e \in R^D$ є компактним описом голосу:

$$\tilde{e} = \frac{e}{\|e\|_2} \quad (5)$$

Це дозволяє обчислювати косинусну подібність [9] між різними сегментами:

$$\cos(\theta_{ij}) = \tilde{e}_i^T \tilde{e}_j \quad (6)$$

На основі цих значень ембедінги кластеризуються методом агломеративного об'єднання [10].

Результатом кластеризації є групи сегментів, кожна з яких відповідає окремому мовцю:

$$\hat{S}(t, f) = M(t, f) \cdot S(t, f) \quad (7)$$

$$M(t, f) \in [0; 1] \quad (8)$$

де маска $M(t, f)$ маска приглушення, що визначає ступінь зниження амплітуди для нецільових мовців.

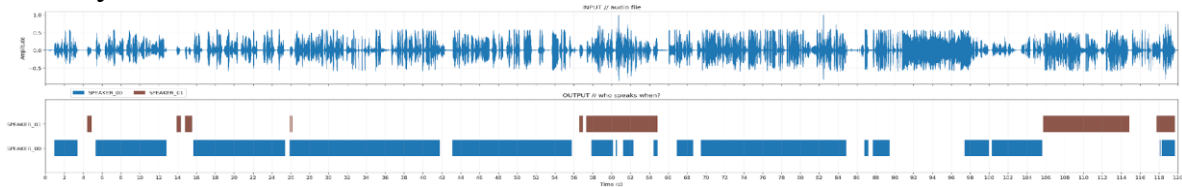


Рис. 1. Вхідний аудіо-сигнал (угорі) та часові інтервали активності мовців після обробки застосунком (унизу)

Цей підхід дає змогу відновити спектр цільового голосу з мінімальними спотвореннями, водночас ефективно приглушуючи небажані джерела.

Архітектура розробленого застосунку

Розроблений застосунок реалізує повний цикл обробки мовлення – від імпорту сирого аудіосигналу до отримання очищеного запису з голосом цільового мовця.

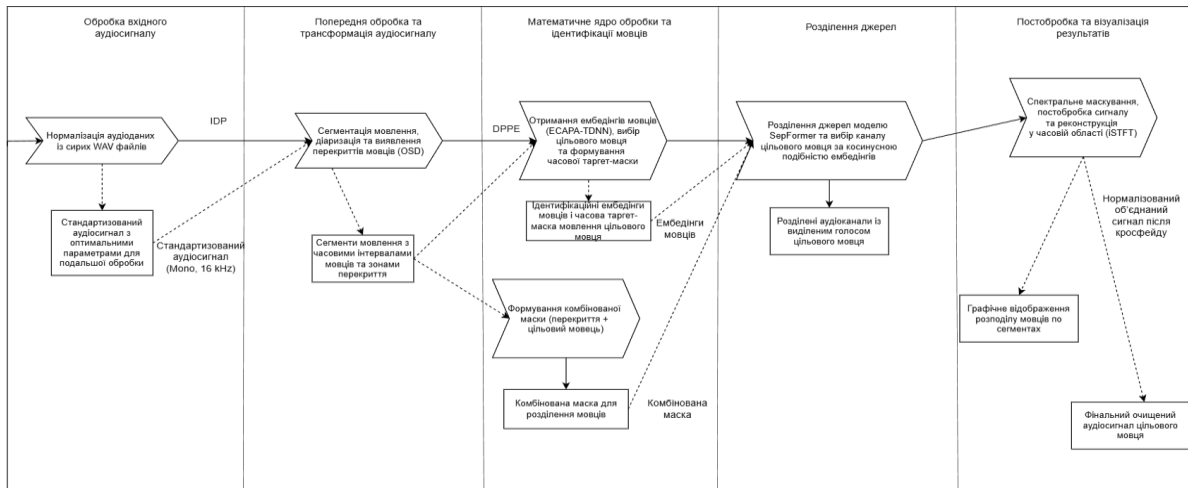


Рис. 2. Діаграма діяльності застосунку першого рівня

Діаграма діяльності першого рівня відображає загальну логіку функціонування застосунку.

Результати

Під час тестування на реальних аудіозаписах із кількома мовцями система успішно виділяла голос цільового мовця та приглушувала інші голоси. Після обробки фрагменти мовлення цільового мовця зберігали зрозумілість.

Висновки

Розроблений застосунок забезпечує автоматичне виділення мовлення цільового мовця з аудіозаписів із кількома мовцями. Метод поєднує аналітичні та нейронні підходи, використовуючи STFT, сегментацію, ембедінги та кластеризацію для ідентифікації цільового мовця та приглушення інших. Отримані результати підтверджують ефективність методу в умовах реального шуму та перекриття голосів.

Література

1. WANG, DeLiang; CHEN, Jitong. Supervised speech separation based on deep learning: An overview. *IEEE/ACM transactions on audio, speech, and language processing*, 2018, 26.10: 1702-1726.
2. SUBAKAN, Cem, et al. Attention is all you need in speech separation. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021. p. 21-25.
3. LUO, Yi; MESGARANI, Nima. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM*

- transactions on audio, speech, and language processing*, 2019, 27.8: 1256-1266.
4. BAI, Zhongxin; ZHANG, Xiao-Lei. Speaker recognition based on deep learning: An overview. *Neural Networks*, 2021, 140: 65-99.
 5. GRIFFIN, Daniel; LIM, Jae. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 1984, 32.2: 236-243.
 6. ORALBEKOVA, Dina, et al. Recent Methods and Algorithms in Speech Segmentation Tasks. In: *International Conference on Computational Collective Intelligence*. Cham: Springer Nature Switzerland, 2024. p. 273-283.
 7. DEHAK, Najim, et al. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 2010, 19.4: 788-798.
 8. DESPLANQUES, Brecht; THIENPOND, Jenthe; DEMUYNCK, Kris. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*, 2020.
 9. MUFLIKHAH, Lailil; BAHARUDIN, Baharum. Document clustering using concept space and cosine similarity measurement. In: *2009 International conference on computer technology and development*. IEEE, 2009. p. 58-62.
 10. MÜLLNER, Daniel. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*, 2011.