DOI 10.31891/2307-5732-2023-321-3-115-119
УДК 621

ДУМИН АНДРІЙ
Національний університет «Львівська політехніка»
ORCID ID: 0000-0003-2111-2899
e-mail: andrii.r.dumyn@lpnu.ua

СИСТЕМА АВТОМАТИЗОВАНОГО ОЗВУЧУВАННЯ З ЕЛЕМЕНТАМИ ШТУЧНОГО ІНТЕЛЕКТУ

Автоматичне озвучування текстів вже давно не є новинкою серед користувачів. Проте при автоматизованому озвучуванні художніх текстів або при автоматизованому переозвучуванні аудіо з інших мов втрачається емоційна складова. Емоційне перетворення голосу з урахуванням статі мовця, особливостей мовлення тощо має на меті зберегти мовний зміст та ідентичність мовця. У цій статі запропоновано архітектуру системи автоматизованого переозвучування аудіо та відео з вбудованими класифікаторами визначення тональності тексу, емоційного забарвлення мовця та модуля опрацювання метаданих мовця для збереження його ідентичності. Розроблена архітектура стане основою подальших досліджень за даною тематикою.

Ключові слова: ASR, автоматичне розпізнавання мовлення, розпізнавання емоцій, перетворення тексту в мовлення, перетворення мовлення в текст, аналіз голосу.

DUMYN ANDRII
Lviv Polytechnic National University

THE AUTOMATED VOICING SYSTEM WITH ELEMENTS OF ARTIFICIAL INTELLIGENCE

Automatic voicing of texts has not been a novelty among users for a long time. However, the emotional component is lost during the automated dubbing of artistic texts or audio from other languages. The emotional transformation of the voice, considering the gender of the speaker, features of speech, etc., aims to preserve the linguistic meaning and identity of the speaker. This work provides an overview of the latest research in the field of automated voicing, obtaining metadata from audio, and proposes a general architecture for an automated voicing system with elements of artificial intelligence, such as a classifier for determining the emotional coloring of speech, models for determining gender and speech features. The obtained work results will form the basis of further research in developing a group of classifiers for determining the emotional coloring of speech, gender, age, and features of human speech. Based on the proposed architecture, the corresponding system's design and development are planned. The proposed system will significantly simplify the consumption of foreign language content for users from different countries, regardless of the level of proficiency in one or another language. For this reason, automated translation and voiceover systems are widespread. However, the speaker's emotional component and other features need to be recovered during automated dubbing of texts or audio from other languages. That is why the automated voicing of texts or dubbing of audio or video will be relevant, taking into account the gender of the speaker, his age, emotional coloring and other features of speech. Such a system will simplify the process of adapting audio and video content to the users of one or another country. It will help make a large part of exciting content available to users. In education, this system can be used as an auxiliary tool when viewing lectures or parts of lectures from foreign lecturers, significantly expanding students' access to educational materials.

Keywords: Automatic Speech Recognition, emotional recognition, voice, Text-to-Speech, Speech-to-Text, voice analysis.

Постановка проблеми

Щоденно кількість аудіо та відео контенту в мережі Інтернет збільшується. За даними компанії Cisco [1] у 2022 році 82% глобального інтернет-трафіку належало перегляду потокового відео або завантаженню відео. Цьому сприяє зростаюча популярність стрімінгових платформ, таких як YouTube, Netflix, Amazon Prime Video та інших. Проте часто користувачі стикаються з проблемою, коли знайдений аудіо чи відео контент на цікаву їм тематику подається незнайомою мовою. За даними statista.com [2] найбільш поширеними мовами у світі є English, Chinese(Mandarin), Hindi, Spanish. Тільки на ютубі 33% відео є англійською мовою і відповідно 67% іншими мовами [3]. До прикладу, за звітом EF English Proficiency Index за 2022 рік [4], жителі більше 80 країн в середньому володіють англійською рівнем Moderate proficiency і нижче, що спричиняє складнощі у споживанні актуального для них контенту англійською мовою. Зважаючи на це, можна припустити, що запропонована система значно спростить споживання іншомовного контенту для користувачів з різних країн не залежно від рівня володіння тою чи іншою мовою. З цієї причини популярністю користуються системи автоматизованого перекладу та озвучення. Проте при автоматизованому озвучуванні текстів або при автоматизованому переозвучуванні аудіо з інших мов втрачається емоційна складова та інші особливості мовця. Саме тому актуальною буде система автоматизованого озвучування текстів або переозвучування аудіо чи відео, що враховуватиме й стать мовця, його вік, емоційне забарвлення та інші особливості мовлення. Така система дозволить спростити процеси адаптації аудіо та відео контенту під користувачів тієї чи іншої країни, допоможе зробити велику частину цікавого контенту доступним для користувачів. В сфері освіти дана система може використовуватись як допоміжний інструмент при перегляді лекцій або частин лекцій від іноземних лекторів, що значно розширить доступ студентам до навчальних матеріалів.

Аналіз останніх джерел

Наукова спільнота активно працює над вирішенням проблем аналізу голосу, отримання метаданих з нього, тощо. Зокрема автори [5] займаються побудовою моделі нейронної мережі для визначення статі мовця за голосом. Автори експериментально показують, що комбінація наборів функцій MFCC і Mel

забезпечує точність визначення статі 94,32%. У дослідженні [6] автори використовують структуру мереж Deeper Long Short Term Memory (LSTM) для прогнозування статі на основі набору аудіо даних. Для дослідження автори використовують 10 найбільш ефективних атрибутів даних, запропонований підхід дозволив успішно передбачити стать мовця з точністю 98,4%. Автори [7] представили метод використання супергаусивності мовних сигналів як ознаки для класифікації здорового та дизартричного мовлення. Експериментальні результати показали, що використання супергаусивності мовних сигналів дає значно вищу точність класифікації, ніж найсучасніші характеристики, такі як основна частота, тремтіння, мерехтіння, співвідношення гармонік до шуму або кепстральні коефіцієнти частоти Mel.

Щодо досліджень емоційності мовлення автори роботи [8] наводять короткий огляд найбільш актуальних розробок обчислювальної обробки емоцій у голосі. Автори пропонують потенційне використання розглянутих технологій, зокрема для інтерпретації в психологічних дослідженнях і використання в додатках цифрової охорони здоров'я та цифрової психології. У статті [9] автори пропонують нову структуру, засновану на варіаційному автоматичному кодуванні генеративної змагальної мережі Вассерштейна (VAW-GAN), яка використовує попередньо навчену модель розпізнавання мовних емоцій (SER) для передачі емоційного стилю під час навчання та відтворення. Таким чином мережа може передавати як видимий, так і невидимий емоційний стиль у нове висловлювання. Основною метою роботи [10] є покращення швидкості розпізнавання мовних емоцій з використанням різних алгоритмів виділення ознак. У роботі акцентується увага на попередній обробці отриманих звукових зразків, де шум із мовних зразків видаляється за допомогою фільтрів. Ці алгоритми виділення ознак підтверджені для універсальних емоцій, включаючи гнів, щастя, сум і нейтральність. У роботі [11] автори побудували моделі ієрархічної класифікації, щоб дослідити важливість ідентифікації віку та статі перед емоційною ідентифікацією. Автори порівняли продуктивність чотирьох різних моделей і представили зв'язок між віком/статтю та точністю розпізнавання емоцій. Результати дослідження показали, що використання окремої моделі емоцій для кожної статі та вікової категорії дає вищу точність порівняно з використанням одного класифікатора для всіх даних.

Метою роботи є формулювання вимог щодо побудови системи автоматизованого озвучування текстів та проєктування базової архітектури розроблюваної системи.

Аналіз існуючих аналогів

На ринку представлено багато сервісів від провідних компаній у цій галузі, що дозволяють в певній мірі адаптувати аудіо та відео контент під свої потреби.

У таблиці 1 наведено огляд найпопулярніших доступних сервісів для перетворення аудіо в текст.

Таблиця 1

Огляд найпопулярніших доступних сервісів для speech to text

Сервіс	Характеристики	Кількість мов
Amazon transcribe	Забезпечує транскрибування чисел і пунктуації, налаштування розпізнавання кількох динаміків.	37
IBM watson speech to text	Модель, що працює у реальному часі, забезпечує визначення ключових слів, мітки доповідачів, мітки часу по мовленню, альтернативи слів	14 + діалекти
Google cloud speech to text	Пакетні моделі та моделі роботи в реальному часі, визначають шумостійкість, фільтр для неправильних слів	120 (не всі мови підтримують всі функції)
Azure Cognitive Services speech to text	Модель у реальному часі, налаштування, форматування, нормалізація тексту, сценарії мовлення	96 + діалекти

Підчас дослідження вдалось знайти лише ряд готових до використання моделей для визначення емоційного забарвлення мовлення, проте такі моделі потребують додаткової інтеграції в якусь систему через відсутність користувацького інтерфейсу.

Зокрема хорошими продуктами для подальшої інтеграції є аналізатор мовних емоцій від Мітеша Путрана [12]. Побудована науковцем модель [13] машинного навчання може виявляти емоції під час постійної розмови один з одним.

Також варто звернути увагу на Demfier multimodal speech emotion recognition [14]. У цьому застосунку створено легкі багатомодальні моделі машинного навчання та виконано їх порівняння із важчими та менш інтерпретованими аналогами глибокого навчання. Автори зазначають, що полегшені моделі можна порівняти з базовими лініями глибокого навчання, а в деяких випадках навіть перевершити їх, досягаючи кращої продуктивності.

Ще одним відомим продуктом є застосунок для розпізнавання мовних емоцій від Xuanji Pe [15]. Даний продукт використовує згорткові рекурентні нейронні мережі TensorFlow для розпізнавання мовних емоцій (SER) у базі даних IEMOCAP [16]. Для покращення проставлення міток емоційності продукт використовує три стратегії об'єднання, щоб створити функції рівня висловлювання для SER.

У таблиці 2 міститься огляд найбільш популярних доступних сервісів автоматичного озвучування текстів.

В результаті огляду не було знайдено жодної системи, яка б поєднувала в собі функції одночасного перекладу тексту з аудіо та автоматичного озвучування з урахуванням емоційного забарвлення та інших характеристик мовця. Це, у свою чергу, підтверджує актуальність даної тематики.

Таблиця 2

Огляд найпопулярніших доступних сервісів для text to speech

Сервіс	Характеристики	Кіл-ть мов
Amazon polly	Модель, що працює у реальному часі, вимова, гучність, висота звуку, швидкість тощо	25 + діалекти
IBM watson text to speech	Забезпечує налаштування вимови, визначає власні слова, виразність, час звучання слів	20 + діалекти
Google cloud text to speech	Визначає вимову, SSML, стать мовця	45 + діалекти
Azure Cognitive Services text to speech	Визначає вимову, гучність, висоту голосу	96 + діалекти

Виклад основного матеріалу

Аналізуючи вже існуючі продукти та останні наукові дослідження, а також запити потенційних користувачів, можна сформулювати перелік наступних вимог щодо розроблюваної системи:

1) підтримка функції перетворення аудіо в текст;
2) підтримка функції визначення тональності аудіо;
3) підтримка функції розпізнавання статі та інших особливостей людини (вік, дизарктричне мовлення, особливості вимови, акцент, тощо) на основі голосу;
4) підтримка функції автоматизованого озвучування текстів з урахуванням особливостей мовця;
5) зручний та зрозумілий користувацький інтерфейс;
6) можливість задавання власних налаштувань щодо адаптації;
7) швидка робота системи.

Для тренування розроблених моделей було вирішено використовувати наступні набори даних:

- Корпус LibriSpeech — колекція прочитаного англійського мовлення, створеного з аудіокниг. Цей корпус є частиною проекту LibriVox, і містить 1000 годин мовлення з частотою дискретизації 16 кГц [17].
- Корпус Common Voice — це багатомовна колекція транскрибованого мовлення. Цей корпус розроблено компанією Mozilla та містить 2 500 годин аудіо 38-а різними мовами[18], зокрема українською.
- Корпус Acted-Emotional-Speech-Dynamic-Database — це загальнодоступний набір даних розпізнавання мовних емоцій для дослідницьких цілей [19, 20].
- Корпус The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) — динамічний мультимодальний набір виразів обличчя та голосу північноамериканською англійською мовою [21]. Даний набір містить 7356 записів від 24 професійних акторів, які озвучують лексично відповідні твердження з нейтральним північноамериканським акцентом.
- Корпус emotional speech database (ESD) [22] — база даних, що складається з 350 паралельних висловлювань та 29 годин мовних даних, що охоплює 5 класів емоцій (нейтральні, щастя, гнів, смуток і здивування).

На рисунку 1 зображено базову архітектуру системи автоматизованого озвучування.

Загалом розроблювана система повинна складатись із ряду модулів, які можна налаштовувати та розширювати, наприклад, для підтримки різних мов чи покращення їх роботи. У ролі вхідних даних виступає аудіо файл у довільному форматі. Користувачу повинна надаватись можливість задавати певні вхідні параметри для системи, як-от мова перекладу, зберігання проміжних кроків та голос або голоси озвучення.

Першим етапом в системі є попереднє опрацювання аудіо, приведення до потрібного аудіо кодеку та частоти дискретизації, визначення кількості голосів – модуль попередньої обробки (1). Це модуль також відповідатиме за розбиття аудіо на структурні одиниці за звучанням окремого голосу. Також на цьому етапі здійснюватиметься побудова матриці звучання певного голосу та визначення міток часу його звучання.

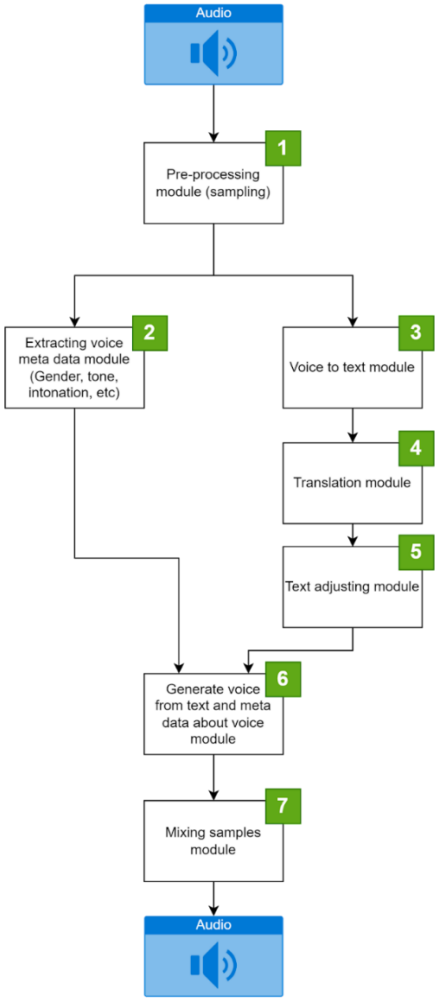

Рис. 1. Запропонована архітектура розроблюваної системи.

Далі система на основі попередньо підготовлених даних повинна визначати емоційне забарвлення фраз, стать, вік (дитина, дорослий, людина похилого віку) та інші особливості мовлення (акцент, дизартричне мовлення тощо) – модуль (2). Для цього необхідно розробити групу відповідних класифікаторів, результати роботи яких взаємодоповнюватимуть одні одних. Модуль (3) відповідає за перетворення підготовлених на першому етапі даних у текст. На цьому етапі здійснюватиметься транскрибування аудіо чи відео та буде складено матрицю тривалості звучання фраз з аудіо чи відео.

Для забезпечення перекладу тексту на обрану користувачем мову використовуватиметься модуль (4). Також на цьому етапі буде складено матрицю тривалості потенційного звучання перекладених фраз. Важливим є модуль (5), що відповідатиме за коригування та виправлення можливих помилок в тексті, а також адаптуватиме переклад до вікна часу в оригіналі (що буде корисно для відео). В результаті коригування буде виконуватись оновлення матриці тривалості потенційного звучання, а також при потребі враховуватиметься зміна перекладених фраз.

На базі підготовленого тексту, який отримано на виході з модуля (5) та комплексу метаданих з модуля (2) буде виконуватись перетворення тексту в голос із врахування матриць тривалості звучання фраз та потенційного звучання перекладених фраз – модуль (6). Для забезпечення такого функціоналу буде розроблено комплекс моделей для автоматичного генерування голосу з урахуванням емоційної складової, віку, статі тощо.

Завершальним модулем системи є модуль (7), що забезпечує об’єднання усіх аудіозаписів в один, при потребі можлива функція певного вирівнювання звукової доріжки. Також даний модуль буде додавати аудіо доріжку до відеоряду (при переозвучуванні відео).

Висновки

В даній роботі наведено огляд останніх досліджень в галузі автоматизованого озвучення, отримання метаданих з аудіо та запропоновано загальну архітектуру для системи автоматизованого озвучування з елементами штучного інтелекту, такими як класифікатор визначення емоційного забарвлення мовлення, моделі для визначення статі та особливостей мовлення. Отримані результати роботи ляжуть в основу подальших досліджень при розробленні групи класифікаторів для визначення емоційного забарвлення мовлення, визначення статі, віку, особливостей мовлення людини. На базі запропонованої архітектури планується проєктування та розроблення відповідної системи.

References

1. Service Provider Network and Technology Services. Cisco. URL: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html (date of access: 17.01.2023).

2. Statista Search Department (2023, Mar 9th) The most spoken languages worldwide 2022 [Infographic]. Statista. URL: https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/(date of access: 9.03.2023).

3. Pew Research Center (2019 July 25th) Popular YouTube channels produced a vast amount of content, much of it in languages other than English. Washington, D.C. URL: https://www.pewresearch.org/internet/2019/07/25/popular-youtube-channels-produced-a-vast-amount-of-content-much-of-it-in-languages-other-than-english/ (date of access: 8.03.2023)

4. EF English Proficiency Index (2022). URL: https://www.ef.com/wwen/epi/ (date of access: 08.03.2023).

5. Chachadi, K., Nirmala, S. R. 2022. Voice-based gender recognition using neural network. In Information and Communication Technology for Competitive Strategies (ICTCS 2020) (pp. 741-749). Springer, Singapore. DOI=https://doi.org/10.1007/978-981-16-0739-4_70.

6. Ertam, F. 2019. An effective gender recognition approach using voice data via deeper LSTM networks. Applied Acoustics, 156, 351-358. DOI=https://doi.org/10.1016/j.apacoust.2019.07.033.

7. Kodrasi, H. Bourlard. 2019. "Super-gaussianity of Speech Spectral Coefficients as a Potential Biomarker for Dysarthric Speech Detection," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 6400-6404, DOI=10.1109/ICASSP.2019.8683107.

8. Schuller, D. M., & Schuller, B. W. (2021). A Review on Five Recent and Near-Future Developments in Computational Processing of Emotion in the Human Voice. Emotion Review, 13(1), 44–50. DOI=https://doi.org/10.1177/1754073919898526.

9. Zhou, K., Sisman, B., Liu, R., Li H. 2021. "Seen and Unseen Emotional Style Transfer for Voice Conversion with A New Emotional Speech Dataset," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 920-924, DOI= 10.1109/ICASSP39728.2021.9413391.

10. Koduru, A., Valiveti, H.B., Budati, A.K. 2020. Feature extraction algorithms to improve the speech emotion recognition rate. Int J Speech Technol 23, 45–55 (2020). https://doi.org/10.1007/s10772-020-09672-4.

11. Shaqra, F. A., Duwairi, R., Al-Ayyoub, M. 2019. Recognizing Emotion from Speech Based on Age and Gender Using Hierarchical Models. Procedia Computer Science, 151, 37-44. DOI=https://doi.org/10.1016/j.procs.2019.04.009.

12. Ringeval, F. 2011. Ancrages et modèles dynamiques de la prosodie : application à la reconnaissance des émotions actées et spontanées. (Speech anchor and dynamic models of prosody : application to acted and spontaneous emotion recognition).

13. GitHub - MiteshPuthran/Speech-Emotion-Analyzer: The neural network model is capable of detecting five different male/female emotions from audio speeches. (Deep Learning, NLP, Python). GitHub. URL: https://github.com/MiteshPuthran/Speech-Emotion-Analyzer (date of access: 10.01.2023).

14. Sahu, G. 2019. Multimodal Speech Emotion Recognition and Ambiguity Resolution. ArXiv. DOI=https://doi.org/10.48550/arXiv.1904.06022

15. Mingyi Chen, Xuanji He, Jing Yang, Han Zhang. 2018. "3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition", IEEE Signal Processing Letters, vol. 25, no. 10, pp. 1440-1444, 2018.

16. Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower, E., … Narayanan, S. "IEMOCAP: Interactive emotional dyadic motion capture database," Journal of Language Resources and Evaluation, vol. 42, no. 4, pp. 335-359, 2008.

17. Panayotov, V., Chen, G., Povey, D., Khudanpur, S. 2015. "Librispeech: An ASR corpus based on public domain audio books," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 5206-5210, DOI=10.1109/ICASSP.2015.7178964.

18. Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Weber, G. 2019. Common voice: A massively-multilingual speech corpus. arXiv preprint arXiv:1912.06670. DOI= https://doi.org/10.48550/arXiv.1912.06670

19. Vryzas, N., Kotsakis, R., Liatsou, A., Dimoulas, C. A., Kalliris, G. 2018. Speech emotion recognition for performance interaction. Journal of the Audio Engineering Society, 66(6), 457-467.

20. Vryzas, N., Matsiola, M., Kotsakis, R., Dimoulas, C., Kalliris, G. 2018. Subjective Evaluation of a Speech Emotion Recognition Interaction Framework. In Proceedings of the Audio Mostly 2018 on Sound in Immersion and Emotion (p. 34). ACM.

21. Livingstone, S. R., & Russo, F. A. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLOS ONE, 13(5), e0196391. DOI=https://doi.org/10.1371/journal.pone.0196391

22. Zhou, K., Sisman, B., Liu, R., Li, H. 2021. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 920-924). IEEE.