Andrii Dumyn

Lviv Polytechnic National University, Lviv

ORCID: 0000-0003-2111-2899

**The AI-based automatic speech recognition system**

The amount of audio and video content on the Internet is increasing daily. However, users often need help finding audio or video content on a topic of interest presented in an unfamiliar language. The growing popularity of streaming platforms such as YouTube, Netflix, Amazon Prime Video, and others facilitates this. According to statista.com [1], the most common languages in the world are English, Chinese (Mandarin), Hindi, and Spanish. Only on YouTube, 33% of videos are in English and 67% in other languages [2]. For this reason, automated translation and voiceover systems are prevalent. However, the speaker's emotional component and other features must be recovered during the automated dubbing texts or audio from other languages. 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.

The scientific community is actively working on solving the problems of voice analysis, obtaining metadata from it. In particular, the authors of [3] are building a neural network model for determining the speaker's gender by voice. Concerning research on the emotionality of speech, the authors of [4] provide a brief overview of the most relevant developments in the computational processing of emotions in the voice. The main goal of the work [5] is to improve the speed of recognition of speech emotions using various feature extraction algorithms.

In general, the developed system should consist of several modules that can be customized and extended, for example, to support different languages or improve their operation.

The first stage in the system is pre-processing the audio. This module will be responsible for breaking the audio into structural units based on the sound of a single voice. The system should determine the emotional coloring of phrases, gender, age (child, adult, elderly), and other speech features (accent, hoarseness) based on previously prepared data. For this, developing a group of appropriate classifiers is required, the results of which will complement each other. The following module converts the data prepared at the first stage into text. At this stage, the audio or video will be transcribed, and a matrix of the duration of the phrase will be compiled. After that, a matrix of the duration of the potential sound of the translated phrases will be compiled. A set of models will be developed for automatic voice generation considering the emotional component, age, gender. The final module of the system ensures the unification of all audio recordings into one; if necessary, the function of a sure leveling of the soundtrack is possible. Also, this module will add an audio track to the video sequence (when dubbing the video).

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 interconnected system's design and development are planned.

**References**

1. 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).
2. 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)
3. 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.
4. 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.
5. 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.