UDC 004.93

Vladyslav Kholiev

Kharkiv National University of Radio Electronics, Kharkiv, Ukraine

CONCEPTUAL MODEL OF THE TECHNOLOGY FOR CALCULATING THE SIMILARITY THRESHOLD OF TWO AUDIO SEQUENCES

Abstract. *The paper is focused on the pressing* ***problem*** *of speaker verification by means of voice time series comparison. The* ***aim*** *of this paper is to determine the orders of mel-frequency cepstral coefficients that most accurately describe the difference between an authentic voice and an artificially generated copy for their further use as input to a neural network model in a resource-limited environment. To achieve this goal, the following* ***tasks*** *were accomplished: a conceptual model of the technology for determining the similarity threshold of two audio series was developed; the orders of fine-frequency cepstral coefficients with the most characteristic differences between the recording and the generated voice were determined on the basis of neural network analysis; an experimental study of the dependence of the execution time and computational load on the created feature vector when assessing the degree of similarity of two time series was conducted; and the optimal similarity threshold was determined on the basis of the chosen dataset. The developed model of the technology for determining the similarity threshold was tested on a dataset that is a combination of the DEEP-VOICE dataset and our own dataset. The demonstrated result of applying the developed technology showed an increase of 43% when using the specified MFCCs compared to using all of them. Based on experimental studies, the DTW acceptance threshold was set at 0.37.*

Key words: *machine learning; mfcc, dtw, feature extraction, speaker recognition; classification, voice cloning, siamese neural networks.*

Introduction

In recent years, the development of smart technologies, in particular the field of generative artificial intelligence (AI), has been gaining speed. Together with the development of computing power, this has ensured high availability and, consequently, the prevalence of AI-based services.

These services are used in many areas, such as natural text processing, audio-visual promotional materials generation, smart prompts when writing program code, personalized chatbots for consulting and support services, as well as text-to-speech (TTS) and voice cloning (Figure 1).

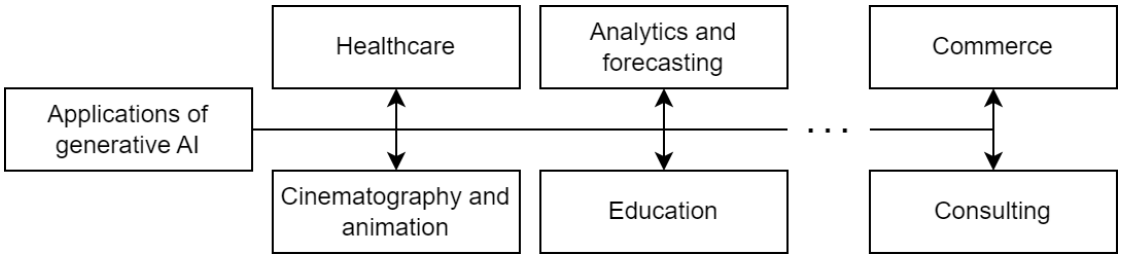


Figure 1 – Applications of generative artificial intelligence.

In particular, TTS with the use of cloned voice, or real-time cloning, has such applications as:

- personalized voice assistants for personal and commercial use in many areas;
- cloning the voice of people with speech impediments or injuries of the speech apparatus in medicine;
- reducing the cost of re-shooting takes, voice acting for animated characters, as well as digital doubles of live actors in cinema and animation;

- personalization and improvement of learning conditions for people with visual impairments and special learning needs [1];

At the same time, cases of impersonation of users and their various characteristics (voice, face, etc.) have significantly increased, and their severity is also growing. Static photos and voice are faked especially often to fabricate events involving or authenticate themselves as victims of fraud.

Voice falsification is the most accessible, and therefore the most widespread form of identity fraud.

In system of knowledge exchange of young scientists presented in [2] (Figure 2), user voice data plays a key role, as the system has the functionality of audio and video conferencing, and also authenticates users using voice.

When considering the functionality of the system in each mode, it is important to understand what the inputs and outputs of each mode are:

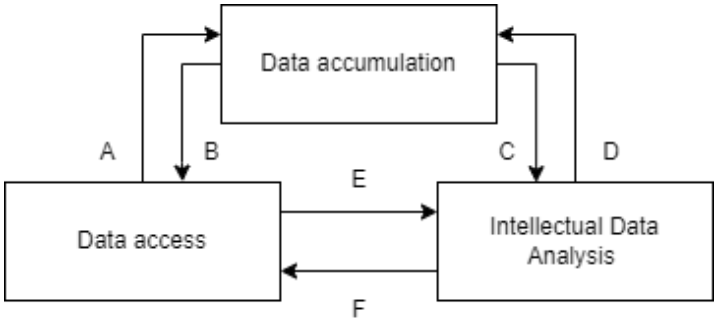


Figure 2 – Functional diagram of the proposed system of knowledge exchange of young scientists.

- marker A: data is accumulated in the form of user voice data for further training, as well as in the form of electronic documents of scientific papers for vectorization and further clustering of papers similar in topic.
- marker B: the repository subsystem receives various requests: for access to scientific papers in various forms, for user data and metadata (for example, for the subsystem that provides social rooms), as well as requests for verification and identification of users.
- marker C: the relevant modules receive voice data for training, or classification and research documents for further vectorization and clustering.
- marker D: as a result of the analysis mode, neural network models are trained on the basis of voice data and their weights are saved to the repository, and based on the uploaded research documents, their vectorized representation is formed and saved, and the documents themselves are assigned to a group with similar topics.
- marker E: if there is a need to identify or verify users, requests are sent with the relevant data (user ID(s) and/or conference IDs, etc.);
- marker F: in response to the queries, either individual neural networks or a pseudo-ensemble module is deployed based on the stored weights of the trained networks.

Thus, improving the recognition of cloned generated voice is a high priority to reliably authenticate users and maintain their security and the integrity of their data.

In general, several approaches are used to solve the problem of comparing data in the context of binary comparison (“equals” or “not equals”) (Figure 3), such as: neural network (the decision is made using a neural network model), algorithmic or traditional (the result is obtained as a result of the algorithm of logical and mathematical comparison of part or all data), and mixed (the previous two approaches are combined in varying percentages).

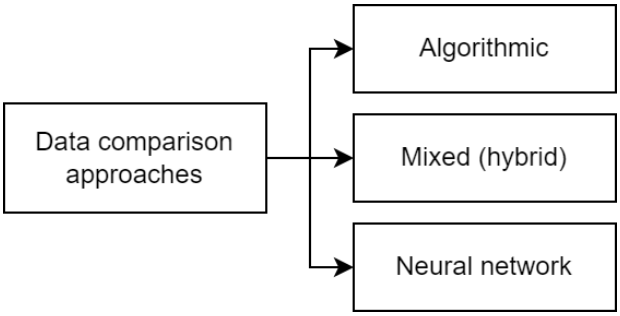


Figure 3 – Generalized classification of approaches to data comparison.

The solution proposed in this paper is a hybrid one, i.e., it combines traditional and neural network methods.

Standard audio sequences classification pipelines usually include the following steps:

- Speech segment detection;
- Pre-processing (silence removal, noise filtering, etc.);
- Feature extraction;
- Audio sequence analysis based on the obtained features and assignment to one of the classes;
- Optional post-processing.

The feature extraction stage is extremely important, as it has the greatest impact on the accuracy of audio classification. The most common methods include:

- spectrograms;
- mel-frequency cepstral coefficients (MFCC);
- constant-Q transform (CQT);
- continuous wavelet transform (CWT);
- and others.

The analysis of [1,3-5] showed that MFCCs are less sensitive to background noise and amplitude variations than similar methods, and also show high efficiency in speech recognition. These advantages ensure the widespread use of mel-frequency cepstral coefficients in various practical areas of life (Figure 4).

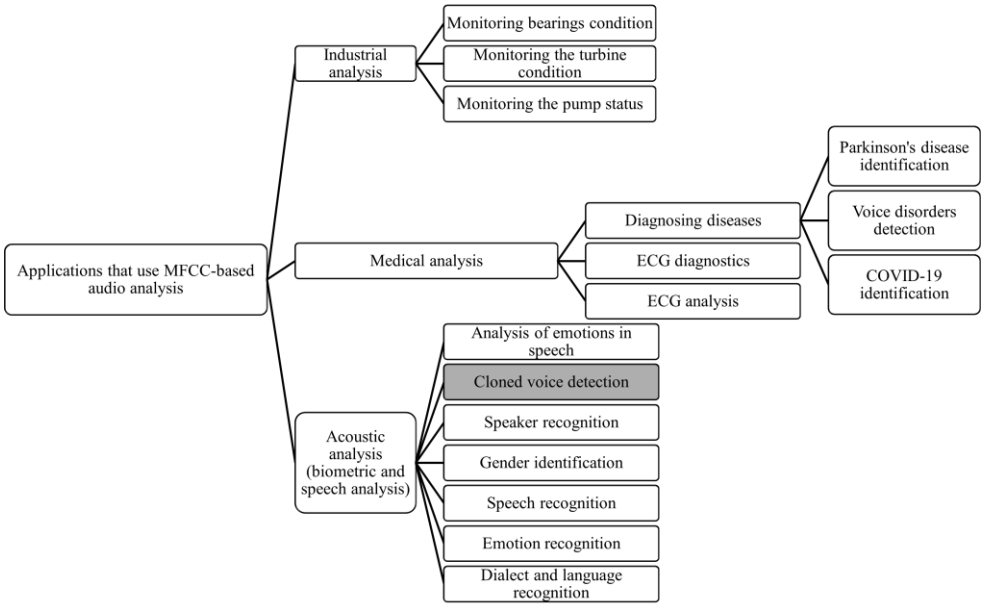


Figure 4 – Practical applications based on MFCC.

Therefore, further research is focused on the use and study of mel-frequency capstral coefficients to solve the problem.

The identified features are the basis for further analysis of the audio sequence or comparison of two (or more) audio sequences. The result of comparing two audio sequences based on the extracted features is a proximity score, which can be determined using metrics such as:

- Jaccard's coefficient;
- Euclidean distance;
- Hamming distance;
- Pearson correlation coefficient;
- Signal to noise ratio (SNR/PSNR);
- Dynamic Time Warping (DTW).

Among them, DTW stands out for its relatively low computational complexity (O(n) with low order of n), as well as adaptability to work with time series [6-7].

The aim of this work is to determine the orders of the mel-frequency cepstral coefficients that most accurately

describe the difference between an authentic voice and an artificially generated copy for their further use as input to a neural network model under limited resources. To achieve this goal, the following tasks must be performed:

- to develop a conceptual model of the technology for calculating the threshold of similarity between two audio orders;
- based on the neural network analysis, determine the orders of the mel-frequency cepstral coefficients with the most characteristic differences between the recording and the generated voice;
- to conduct an experimental study of the dependence of the execution time and computing unit load on the created vector of characteristic features when assessing the degree of similarity of two time series;
- determine the optimal threshold of similarity based on the use of the DTW algorithm in the context of the selected dataset;
- analyze the obtained results.

To accomplish these tasks, a method of analyzing voice information based on a hybrid approach of neural network and algorithmic analysis was proposed.

Results an discussion

In this paper, a hybrid technology for analyzing voice information is proposed based on a combination of neural network analysis of mel-frequency cepstral coefficients and their comparison using the dynamic time warping (DTW) algorithm. The conceptual model of the technology for determining the threshold of similarity between two audio sequences is shown in Figure 5.

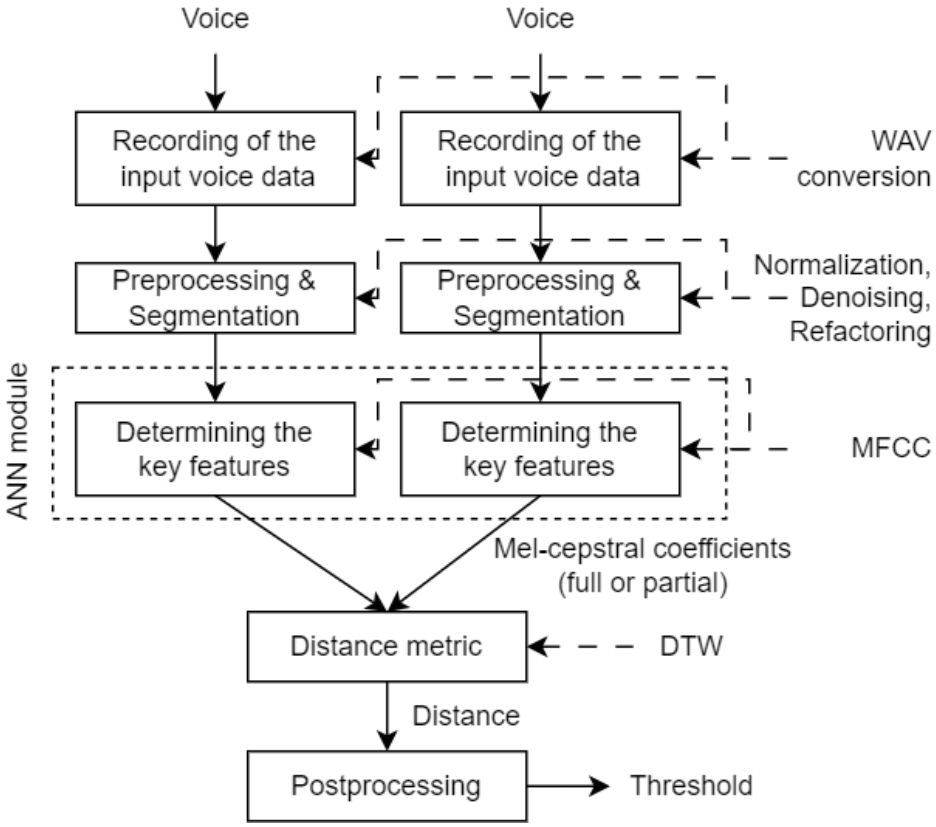


Figure 5 – The conceptual model of the technology for determining the threshold of similarity between two audio sequences.

In this work, a combination of own dataset, formed from own audio recordings and cloned audio sequences, and DEEP-VOICE dataset, presented in [8] and developed for a related topic, is used.

DEEP-VOICE consists of recordings from public speeches of famous individuals (Table 1), as well as cloned audio sequences generated using the open text-to-speech framework “Retrieval-based-Voice-Conversion-WebUI” [9].

The cloned audio sequences repeat the texts of other recordings, but are spoken in the voice of each individual. In total, the dataset includes 64 audio sequences - 8 real recordings and 56 generated audio files. The length of the recordings is limited to 10 minutes. The recordings have varying degrees of recording quality and background noise levels to represent real-world conditions and to ensure the

diversity of the dataset. There are recordings of both men and women, but the dataset is not balanced.

Table 1.

The data collected for training, validation, and testing for the experiments in [6] (sorted alphabetically by last name). Audio fragments truncated to ten minutes.

Individual	**Source**	**Length (MM:SS)**
Joe Biden	Victory Speech	10:00
Ryan Gosling	Golden Globes Speech	1:33
Elon Musk	Commencement Speech	10:00
Barack Obama	Victory Speech	10:00
Margot Robbie	BAFTAs Speech	1:19
Linus Sebastian	Stepping Down Monologue	9:30
Taylor Swift	Women in Music Speech	10:00
Donald Trump	Victory Speech	10:00
Total		***62:22***

Own dataset, combined with the one mentioned above, has a similar structure, namely, it consists of recordings of the same texts together with generated audio sequences of similar texts. In total, there are 16 audio files in this subset.

The resulting dataset was split into training and test samples in the ratio of 80/20 [10] and the k-fold cross validation method was applied.

One of the key features of the proposed technology is the stage of identification and selection of key features after their extraction. Since the features are represented by mel-frequency cepstral coefficients, this means that in this context, the identification of key features is the identification of individual orders of coefficients that best describe the features of the audio signal that distinguish the authentic voice from the generated copy, for their further use in a certain proximity measure using one of the metrics. Narrowing the number of features used will reduce the time for comparing two audio sequences, as well as reduce the computational load of the technology, which is an advantage when computer resources are limited.

Key features are identified using so-called Siamese neural networks and DTW is used as a proximity measure.

Siamese neural networks are a specialized architecture that usually consists of two parallel identical neural networks that have the same weights in order to evaluate or compare the similarity between two input objects [11] (Figure 6).

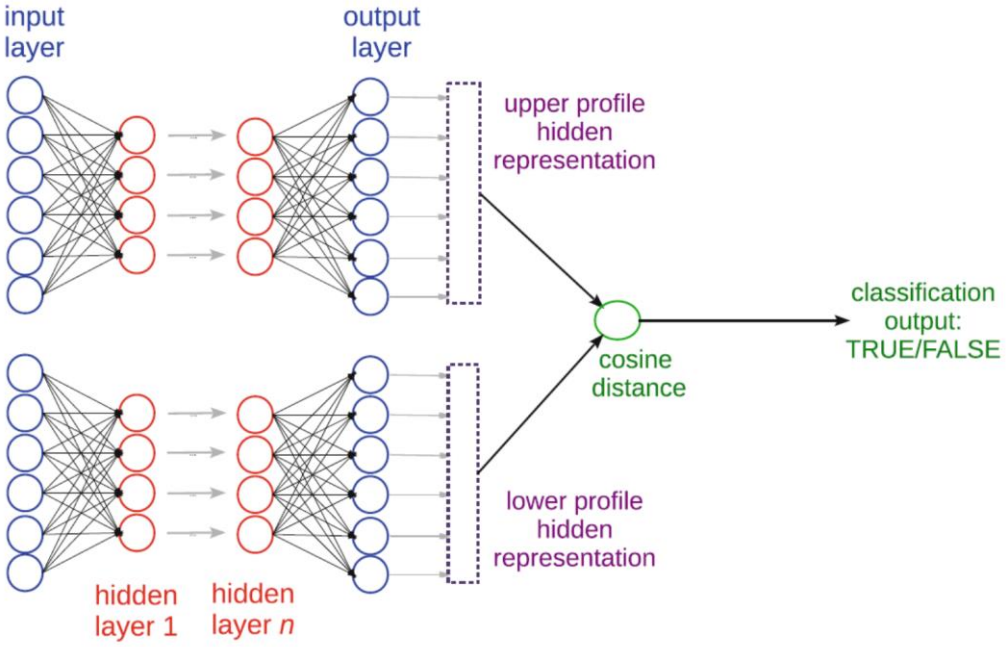


Figure 6 – Generalized architecture of the Siamese neural network.

Such a neural architecture is able to learn from a limited data set by generalizing information about feature vectors represented as MFCCs. In this work, one of the models in the pair will be trained on authentic voices, while the other will be trained on cloned copies. As a result, both models will provide generalized mel-frequency cepstral coefficients for authentic and cloned voices, respectively.

Usually, for such architectures, the cosine similarity coefficient is used as a measure of proximity, but due to the peculiarities of the input and output data, i.e. MFCC, it is appropriate to use DTW. The returned values are indices corresponding to the orders of the mel-frequency cepstral coefficients.

The coefficients used for further determination of proximity were those with a DTW distance of at least 0.37.

The results of the analysis are shown in Table 2. The average length of an audio sequence is 600 seconds.

Table 2. – *The dependence of execution time and computing unit load on the created vector of characteristic features when assessing the degree of similarity of two time series*

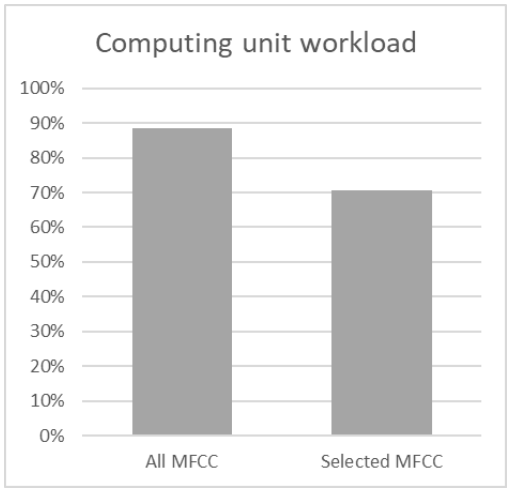
Experiment number	**A degree of proximity (DTW metric)**	**Average DTW value**	**Proximity calculation time, m:s**	**Computing unit load, %**
Comparison of audio sequences using all MFCC orders for identical voices				
1	0.35	0,33	6:57	88%
2	0.33		7:04	89%
3	0.36		6:43	89%
4	0.34		7:15	88%
5	0.37		5:04	89%
6	0.27		4:35	88%
7	0.30		7:01	87%
Comparison of audio sequences using all MFCC orders for different voices				
8	0.42	0,55	7:54	88%
9	0.38		8:44	89%
10	0.63		7:33	88%
11	0.59		9:15	88%
12	0.48		6:54	89%
13	0.71		6:15	89%
14	0.66		9:20	88%
Comparison of audio sequences using partial (predefined) MFCC orders for identical voices				
15	0.32	0.33	3:18	77%
16	0.34		2:15	65%
17	0.35		2:22	71%
18	0.38		2:42	68%
19	0.29		3:16	74%
20	0.28		2:55	69%
21	0.35		3:15	70%
Comparison of audio sequences using partial (predefined) MFCC orders for different voices				
22	0.55	0.54	2:45	76%
23	0.41		3:51	77%
24	0.68		3:22	68%
25	0.46		3:24	69%
26	0.54		2:59	70%
27	0.61		3:55	64%
28	0.53		2:51	71%

As a result of the neural network analysis, the coefficients that had the greatest difference when comparing the generalized coefficients obtained as a result of training the Siamese neural network for authentic and cloned voices were identified.

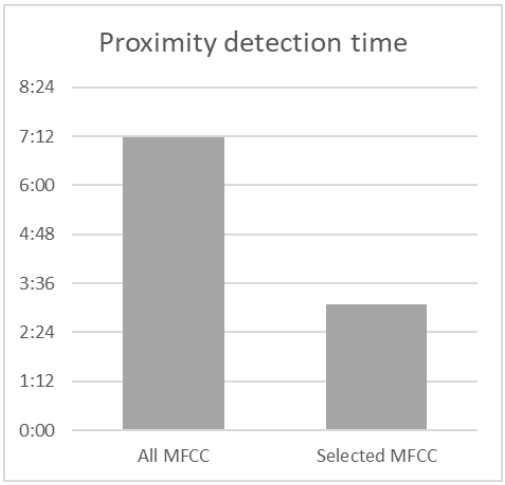
For further comparison and final proximity measure threshold determination, the DTW method was used again. It compares the audio sequences based on the selected coefficients. To make the final decision, it is necessary to determine the optimal similarity threshold

in the context of the selected dataset. The threshold was chosen according to the following principle: among the distance indicators for pairs of real voices, the average DTW of 0.33 was chosen. Meanwhile, the minimum value of 0.395 was taken for the distance indicators of different votes.

The resulting threshold is the arithmetic mean of the two indicators, which is 0.37. This threshold will be used in further research.



(а) computing unit load



(б) proximity detection time

Figure 7 – Graphs of the dependence of the proximity determination time and the computing unit load on the number and composition of mel-frequency capstral coefficients.

Table 2 and the graphs in Figure 7 show that at close values of the acceptance threshold, the technology of using partial MFCCs shows a significant increase in execution speed (43%) due to the reduction in computational complexity due to the absence of complications in the Siamese network architecture and the reduction of data to be processed. Further use of the selected MFCC coefficients is appropriate in the context of determining voice authenticity based on the metric of proximity to the cloned audio recording.

Conclusions

In this paper, a conceptual model of the technology for calculating the similarity threshold of two audio sequences was proposed. The technology is based on a mechanism for determining significant orders of mel-frequency cepstral coefficients to reduce the dimensionality of input data and, as a result, reduce the execution time and computing unit load, which is an advantage in conditions of limited resources. The dynamic time warping algorithm was chosen as a measure of proximity because of its relatively low computational complexity and adaptability to work with time series.

The obtained results show a significant increase in the speed of execution (43%) due to the reduction of computational complexity due to the absence of complications in the Siamese network architecture and the reduction of data for processing. Further use of the selected MFCC coefficients is appropriate in the context of determining voice authenticity based on the metric of proximity to the cloned record.

Further development includes the application of the obtained parameters and characteristics to the neural network model in order to develop a speaker verification module or improve its performance.

References

1. Sidhu, Manjit & Latib, Nur & Sidhu, Kirandeep. (2024). MFCC in audio signal processing for voice disorder: a review. Multimedia Tools and Applications. 1-21. 10.1007/s11042-024-19253-1.
2. Холєв В., Барковська О. COMPARATIVE ANALYSIS OF NEURAL NETWORK MODELS FOR THE PROBLEM OF SPEAKER RECOGNITION //СУЧАСНИЙ СТАН НАУКОВИХ ДОСЛІДЖЕНЬ ТА ТЕХНОЛОГІЙ В ПРОМИСЛОВОСТІ. – 2023. – №. 2 (24). – С. 172-178.
3. Zheng, Fang & Zhang, Guoliang & Song, Zhanjiang. (2001). Comparison of Different Implementations of MFCC.. J. Comput. Sci. Technol.. 16. 582-589. 10.1007/BF02943243.
4. Dave, Namrata. (2013). Feature extraction methods LPC, PLP and MFCC in speech recognition. International Journal For Advance Research in Engineering And Technology(ISSN 2320-6802). Volume 1.
5. Sharma, Garima & Umapathy, Kartikeyan & Krishnan, Sridhar. (2020). Trends in audio signal feature extraction methods. Applied Acoustics. 158. 107020. 10.1016/j.apacoust.2019.107020.

6. Abdullah Mueen, Eamonn J. Keogh: Extracting Optimal Performance from Dynamic Time Warping. KDD 2016: 2129-2130
7. Yurika Permanasari et al 2019 J. Phys.: Conf. Ser. 1366 012091
8. Bird, J. J., & Lotfi, A. (2023). Real-time Detection of AI-Generated Speech for DeepFake Voice Conversion. arXiv preprint arXiv:2308.12734.
9. Retrieval-based-Voice-Conversion-WebUI. (n.d.). github.com. https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI/
10. Kholiev, V., Barkovska, O. (2023), "Analysis of the of training and test data distribution for audio series classification", Information and control systems at railway transport, No. 1, P. 38-43. DOI: https://doi.org/10.18664/ikszt.v28i1.276343
11. Chicco, D. (2021). Siamese Neural Networks: An Overview. In: Cartwright, H. (eds) Artificial Neural Networks. Methods in Molecular Biology, vol 2190. Humana, New York, NY. https://doi.org/10.1007/978-1-0716-0826-5_3.

Received (Надійшла) 14.06.2024.

Концептуальна модель технології визначення порогу подібності двох аудіорядів

Vladyslav Kholiev, Olesia Barkovska

Анотація. Робота присвячена актуальній **проблемі** верифікації спікерів шляхом порівняння голосових часових рядів. **Метою** даної роботи є визначення порядків мелчастотних кепстральних коефіцієнтів, які найточніше описують різницю автентичного голосу від штучної згенерованої копії для подальшого їх використання у якості вхідних даних нейромережевої моделі в умовах обмежених ресурсів. Для досягнення цієї мети були виконані наступні **задачі**: розроблено концептуальну модель технології визначення порогу подібності двох аудіорядів, на основі нейромережевого аналізу визначено порядки мелчастотних кепстральних коефіцієнтів з найхарактернішими відмінностями запису від згенерованого голосу, проведено експериментальне дослідження залежності часу виконання та завантаженості обчислювача від створеного вектору характерних ознак при оцінюванні міри подібності двох часових рядів, а також визначено оптимальний поріг подібності на основі використання алгоритму DTW у контексті обраного датасету. Розроблена модель технології визначення порогу подібності була протестована на наборі даних, що являє собою комбінацію набору даних DEEP-VOICE та власного датасету. Продемонстрований результат застосування розробленої технології показав приріст у 43% при використанні визначених MFCC порівняно з використанням усіх. На основі експериментальних досліджень був встановлений поріг прийняття DTW у 0.37.

Ключові слова: machine learning; mfcc, dtw, feature extraction, speaker recognition; classification, voice cloning;.siamese neural networks

About the authors / Відомості про авторів

Холєв Владислав Олександрович – асистент кафедри “Електронно обчислювальних машин”, Національний університет радіоелектроніки «ХНУРЕ», Харків, Україна;

Vladyslav Kholiev – Professor Assistant at Electronic Computers Department, Kharkiv National University of Radio Electronics "NURE", Kharkiv, Ukraine.
e-mail: vladyslav.kholiev@nure.ua; ORCID ID: https://orcid.org/0000-0002-9148-1561.