ISSN 2522-9818 (print)
ISSN 2524-2296 (online) *Innovative technologies and scientific solutions for industries. 2023. No. 4 (26)*

UDC 004.91 DOI: https://doi.org/10.30837/ITSSI.2023.26.016

O. BARKOVSKA, A. HAVRASHENKO

ANALYSIS OF THE INFLUENCE OF SELECTED AUDIO PRE-PROCESSING STAGES ON ACCURACY OF SPEAKER LANGUAGE RECOGNITION

The **subject matter** of the study is the analysis of the influence of pre-processing stages of the audio on the accuracy of speaker language regognition. The importance of audio pre-processing has grown significantly in recent years due to its key role in a variety of applications such as data reduction, filtering, and denoising. Taking into account the growing demand for accuracy and efficiency of audio information classification methods, evaluation and comparison of different audio pre-processing methods becomes important part of determining optimal solutions. The **goal** of this study is to select the best sequence of stages of pre-processing audio data for use in further training of a neural network for various ways of converting signals into features, namely, spectrograms and mel-cepstral characteristic coefficients. In order to achieve the goal, the following **tasks** were solved: analysis of ways of transforming the signal into certain characteristics and analysis of mathematical models for performing an analysis of the audio series by selected characteristics were carried out. After that, a generalized model of real-time translation of the speaker's speech was developed and the experiment was planned depending on the selected stages of pre-processing of the audio. To conclude, the neural network was trained and tested for the planned experiments. The following **methods** were used: mel-cepstral characteristic coefficients, spectrogram, time mask, frequency mask, normalization. The following **results** were obtained: depending on the selected stages of pre-processing of voice information and various ways of converting the signal into certain features, it is possible to achieve speech recognition accuracy up to 93%. The practical significance of this work is to increase the accuracy of further transcription of audio information and translation of the formed text into the chosen language, including artificial laguages. **Conclusions**: In the course of the work, the best sequence of stages of pre-processing audio data was selected for use in further training of the neural network for different ways to convert signals into features. Mel-cepstral characteristic coefficients are better suited for solving our problem. Since the neural network strongly depends on its structure, the results may change with the increase in the volume of input data and the number of languages. But at this stage, it was decided to use only mel-cepstral characteristic coefficients with normalization.

Keywords: mel-cepstral characteristic coefficients; spectrogram; time mask; frequency mask; normalization; neural network; voice; audio series; speech.

Introduction

Voice is the result of the vibration of the vocal folds (vocal cords) in the human larynx. These vibrations generate sound, which is then modified by various parts of the mouth and nasal cavities, as well as the lips and tongue. The physical basis of voice includes aspects such as frequency, amplitude, waveform, and other parameters [1, 2]. Here are some of the main criteria of the voice and how they are represented in a spectral diagram.

Frequency determines the pitch of the voice and is measured in hertz (Hz). Figure 1 shows the frequencies increasing along the vertical axis.

Amplitude defines the loudness of a sound and is measured in decibels (dB). The legend on the right (Fig. 1) shows that color intensity increases with density.

Dynamics indicates variations in the volume of the speech, such as pauses and quiet zones. Pauses in speech are displayed with minimal dynamics. In the figure, we can see a pause in the first 0.5 seconds, and from 4.5 to 6 seconds.

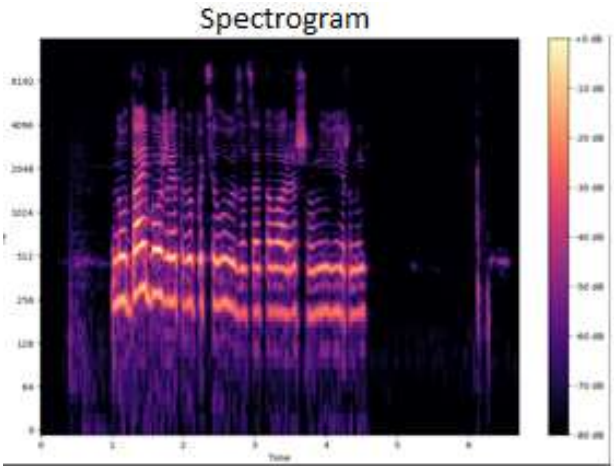

Fig. 1. Example of a spectrogram

Timbre defines the unique characteristics of a voice that distinguish it from other voices. Different timbres are represented by different shapes on the spectrogram.

Intonation determines the melody and rhythm of speech. Changes in intonation can be detected in the shape and location of intense sectors on the spectrogram.

Сучасний стан наукових досліджень та технологій в промисловості. 2023. № 4 (26)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

Each language can have its own peculiarities in the spectral composition of the voice, due to the phonetic features and nature of the sounds that make up it. Voice spectral diagrams can be used to analyze these differences and to study various language features.

The relevance of this study lies in the fact that in today's world, where globalization and international communication are becoming a necessity, speaker recognition systems with subsequent translation into the chosen language are becoming extremely important. These technologies are revolutionizing the way we communicate and interact in different cultural and linguistic environments. Thanks to this, speaker recognition systems can become an integral part of our digital lives, helping to make the world more connected and diverse. The practical significance of this work is to increase the accuracy of further transcription of audio information and translation of the generated text into the chosen language.

Materials and methods

A generalized traditional voice signal processing pipeline is shown in Figure 2.

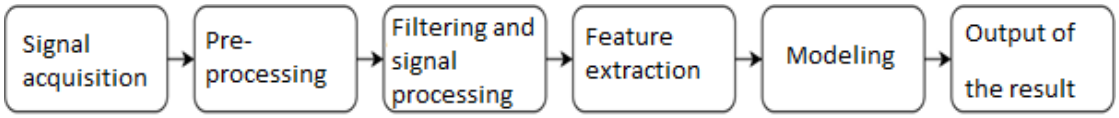

Fig. 2. Traditional voice signal processing pipeline

Each of these stages has certain tasks, the quality of which will affect the next step. The purpose and tasks of each stage are described below.

– signal acquisition is recording or selecting an audio file to be processed;

– pre-processing involves converting the signal from analog to digital format for further processing;

– filtering and signal processing usually involves the use of various filters to extract useful information, such as reducing background noise or enhancing harsh sounds;

– feature extraction is necessary to reduce the amount of input data, which speeds up the processing time of incoming audio files without losing processing accuracy;

– modeling in the context of voice signal processing means creating mathematical models or algorithms that can learn and recognize certain properties of voice signals;

– the result is displayed in a form suitable for the user, with the possibility of further processing of the received data.

In this paper, the research is focused on the feature extraction and modeling stages, so let's consider them in more detail. Among the existing possible ways to convert a signal into certain features, we can distinguish:

– mel-cepstral characteristic coefficients

– linear coding with prediction;

– spectrogram.

The shallow mel-cepstral characteristic coefficients are one of the frequently used features used in a variety of applications, especially in voice signal processing, such as speaker recognition, voice recognition, and gender identification. The mel-cepstral characteristic coefficients can be calculated by performing five sequential processes (Figure 3), namely, framing the signal, calculating the power spectrum, applying a bank of chalk filters to the obtained power spectra, calculating the logarithms of all filter banks, and finally applying the discrete cosine transform [3, 4, 5].

Fig. 3. Scheme for determining fine-frequency capstral coefficients

Speech signal analysis aims to find more informative, compact, and relevant knowledge than the raw speech signal data itself. Vocal tract features (also called segmental, spectral, or system features) [6] are one of the well-known representations of speech analysis. Among the most well-known applications that use the mel-cepstral characteristic coefficients of the speech signal are emotion recognition in speech, speech and dialect recognition, and speech recognition, as shown in the following subsections.

Linear Predictive Coding (LPC) is a method of analyzing and modeling an audio signal. Linear prediction analysis can provide a set of speech signal model parameters that accurately represent the spectral

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2023. No. 4 (26)

amplitude of the speech signal, taking into account a variety of articulatory and phonetic features. LPC can mimic the human articulatory system very well and, therefore, has some advantages in extracting speech feature parameters [7, 8].

The obtained LPC coefficients can serve as an acoustic signature of the voice, which allows them to be used for speaker recognition and other tasks in the field of speech processing.

One of the main advantages of LPC is its ability to reduce the data dimensionality in a well-controlled manner, which makes it effective for transmitting and storing speech information. Another important property of LPC is the ability to reconstruct a signal based on a limited number of parameters, which makes it effective in resource-constrained systems such as real-time speech recognition systems or mobile applications.

A spectrogram is a visual representation of the time-frequency analysis of sound. It allows you to analyze changes in sound intensity as a function of frequency and time. A spectrogram is usually created in one of two ways: approximated as a set of filters derived from a series of bandpass filters (this was the only way before the advent of modern digital signal processing methods), or calculated from a time signal using the windowed or fast Fourier transform. These two methods actually produce different quadratic frequency-time distributions, but are equivalent under certain conditions.

Table 1. *Comparison of feature extraction methods*

	Mel-cepstral characteristic coefficients	Linear prediction coding	Spectrogram
Description	Mel-cepstral characteristic coefficients are coefficients obtained from the taken cepstral coefficients of the sound spectrum calculated on the basis of the chalk scale (chalk scale).	LPC models the voice signal as a linear combination of previous samples.	A spectrogram visualizes a time series as a function of time and frequency. It shows which frequencies are present in the signal at a particular moment in time.
Advantages	Effective for speech recognition, takes into account the peculiarities of auditory perception.	It reproduces formants (resonant peaks) in speech well, which is important for speech recognition.	It shows the time and frequency characteristics of a signal well, and helps to separate different sound sources.
Disadvantages	Requires a lot of calculations to obtain the cepstral coefficients and may be limited in higher frequencies.	It can be sensitive to noise, and its effectiveness decreases with non-stationary signals.	It is not always effective for speech recognition due to the lack of distinct speech characteristics.

Spectrograms are widely used in audio analysis, music production, speech processing, medical research, and many other fields to visualize information about the time and frequency character of sound signals [9, 10].

The result of the comparative analysis of feature extraction methods is shown in Table 1.

In further work, we will consider in detail the fine-frequency cepstral coefficients and spectrograms, because these methods generate compact representations, which facilitates their processing and analysis. In addition, they help to extract important acoustic features, such as formants in speech.

A variety of mathematical models are used for voice recognition, among which the most common are neural networks (deep and shallow), support vector method (SVM), random forest, decision trees, and hidden Markov models.

The result of the comparative analysis of mathematical models for audio data analysis is shown in Table 2.

Table 2. *Comparative analysis of mathematical models for audio data analysis*

Model	Advantages	Disadvantages
Neural networks	Ability to learn complex dependencies, great flexibility in working with different types of data.	High computational resource requirements, need for a large amount of data for training.
Support vector method	Effective in high-dimensional spaces, works well with small data sets.	May be sensitive to noise, limited in cases where data is not linearly resolved.
Decision tree	Easy to interpret, does not require significant preliminary data analysis.	May be prone to overfitting, does not always work well with complex dependencies in data.
Random forest	Reduces the tendency to overlearn, works well with large amounts of data.	Can be computationally expensive, loss of interpretability compared to a separate decision tree.
Hidden Markov models	Good for modeling time series and event sequences.	Assumes independence between states, limited in modeling complex dependencies [15].

Сучасний стан наукових досліджень та технологій в промисловості. 2023. № 4 (26)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

These mathematical models are used in various voice technologies, including speech recognition, speaker identification, emotion analysis, and others. The choice of a particular model may depend on the specific task, data availability, processing volume, and other factors. In this work, modeling is performed using neural networks because they can learn complex relationships and have great flexibility in working with different types of data.

Training a neural network is impossible without a prepared training and test set. The dataset was selected based on the results of the analysis of existing audio corpora.

For the test data, we chose the corpus described in [16]. The Common Voice corpus is a multilingual collection of transcribed speech intended for research and development of speech technologies. Common Voice is designed for the purposes of automatic speech recognition, but can be useful in other areas (e.g., language detection). To achieve scale and sustainability, Common Voice uses crowdsourcing for both data collection and data validation. The latest release includes 29 languages, and as of November 2019, a total of 38 languages are being collected. More than 50,000 people have participated so far, resulting in 2,500 hours of audio. To the best of our knowledge, this is the largest publicly available audio corpus for speech recognition in terms of both hours and languages.

Another significant feature of Common Voice is its constant updating and expansion with fresh and representative data to improve and refine speech recognition models.

The reasons for choosing this dataset are shown in Table 3. The main reason for choosing this particular corpus was the large number of different languages and the large number of speakers in each set. In addition, this dataset is public under the CC0 license. Its use will allow us to simplify the model because there is no noise and will help train the model on different accents of different languages.

Table 3. *Advantages of the chosen corpus*

Selected corpus for training an artificial voice command analyzer	Advantages	Consequences of the advantages
Common Voice	A large number of languages	Possibility of scaling
	Absence of noise	No need to eliminate noise that could reduce accuracy
	Different audio authors	No false dependencies are created due to person/gender
	Public license	Can be used in this project.

To build the neural network, we used the TensorFlow library, which can help a researcher train neural networks to detect and decipher patterns and correlations, similar to the learning and understanding used by humans[17].

The goal is to analyze the impact of audio data preprocessing methods on the accuracy of speaker speech detection for use in further neural network training for different ways of converting signals into features, namely spectrograms and fine-frequency cepstral coefficients.

To achieve this goal, the following tasks should be solved

– analysis of methods for converting signals into features;

– review of mathematical models for performing audio order transformation by selected features;

– development of a generalized model for real-time translation of a speaker's speech;

– planning the experiment depending on the selected stages of audio pre-processing;

– raining and testing of the neural network for the planned experiments;

– comparative analysis of the training results.

Solution of the stated problem. Discussion of the results

In this paper, we propose a generalized model of real-time translation of a speaker's speech into one of the selected languages in Figure 4. The model consists of stages that sequentially determine the speaker's language, convert the speech to text, and translate the prepared text into one of the selected languages, including artificial languages [18, 19, 20].

These mathematical models are used in various voice technologies, including speech recognition, speaker identification, emotion analysis, and others. The choice of a particular model may depend on the specific task, data availability, processing volume, and other factors. In this work, modeling is performed using neural networks because they can learn complex relationships and have great flexibility in working with different types of data.

This study examines the impact of preprocessing stages on speech recognition accuracy.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2023. No. 4 (26)

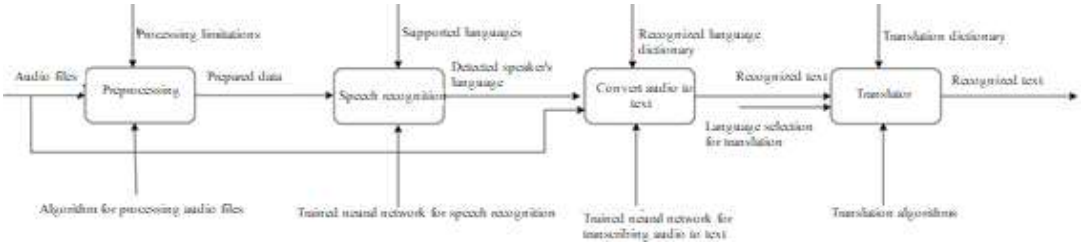

Fig. 4. Generalized model for real-time translation of a speaker's speech

The experiments were planned depending on the selected audio preprocessing stages – normalization, frequency mask and time mask for two main ways of transforming the signal into certain features – for fine-frequency cepstral coefficients and for spectrograms.

The time-domain severity transform coefficients and the spectrogram represent different ways of representing an audio signal in the form of feature vectors, and each of these methods can determine different aspects of the audio signal.

For each case, a separate neural network was created, and, taking into account the peculiarities of the results of each type of processing, the structure of the network was changed - the number of network layers, cutoff coefficients, number of epochs, etc.

A total of 8 experiments were generated for two different ways of converting the signal into features – for low-frequency capstral coefficients and for spectrograms.

The neural network was trained on datasets containing the same number of audio tracks for different languages:

1. Ukrainian;
2. English;
3. French;
4. German.

The system accepts randomly selected files from the dataset as input. The dataset was divided into training and test samples in the following ratio: 80% of the input data (out of 1000 audio files) was included in the training set, 20% is the test data on which we will test our neural network (200 test files).

The neural network itself consists of several levels and is shown in Figure 5.

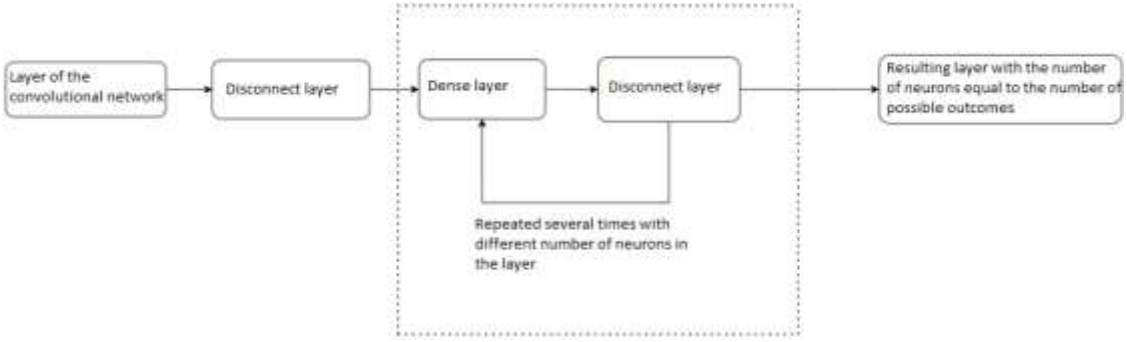

Fig. 5. Generalized structure of the neural network

The first layer uses a Convolutional Neural Network in the TensorFlow library. After this layer, we will have levels of convolutional representation, which can then be aligned and passed to the fully connected layer for further analysis and classification.

Next are several Dense layers with different numbers of neurons on them. With each subsequent layer, the number of neurons decreases. This is due to the dimensionality reduction technique. Reducing the number of neurons helps to control the number of parameters in the model, which is important to avoid overfitting, especially when data is limited.

In addition, to reduce overfitting, each Dense layer is followed by a random disconnect layer. This is one of the regularization techniques that helps to avoid overfitting the model. In each training epoch, a certain portion of the layer weights is randomly selected and set to zero at that stage. This helps to generalize the model and increase its resistance to overfitting.

The last layer is the result layer, in which the number of neurons is equal to the number of classes of possible outcomes.

An assessment of the impact of audio pre-processing methods on the accuracy (probability of

Сучасний стан наукових досліджень та технологій в промисловості. 2023. № 4 (26)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

correct recognition) of speaker speech recognition in accordance with the experiments conducted using fine-frequency cepstral coefficients is shown in Table 4.

Table 4. *Experimental results using mel-cepstral characteristic coefficients*

Normalization	Time mask	Frequency mask	Accuracy of speaker language detection, %
–	–	–	90
+	–	–	93
–	+	–	24
+	+	–	24
–	–	+	24
+	–	+	25
–	+	+	22
+	+	+	23

As can be seen from the results, the time mask and frequency mask remove useful information and cannot be used with mel-cepstral characteristic coefficients.

An assessment of the impact of audio preprocessing methods on the accuracy (probability of correct recognition) of speaker speech recognition in accordance with the experiments conducted using spectrograms is given in Table 5.

Table 5. *Experimental results when using a spectrogram*

Normalization	Time mask	Frequency mask	Accuracy of speaker language detection, %
–	–	–	61
+	–	–	65
–	+	–	63
+	+	–	67
–	–	+	63
+	–	+	68
–	+	+	65
+	+	+	73

Based on the results of the spectrogram, we can conclude that masking helps to increase accuracy, but this method is still worse than mel-mean characteristic coefficients in this task, since mel-mean characteristic coefficients also take into account the features of the audio signal that are important for speech or sound recognition, and include high-frequency and low-frequency coefficients, differentiation, and other characteristics. A spectrogram, on the other hand, simply represents the distribution of signal energy at different frequencies over time. It may include less information about specific aspects of the audio signal that are important for certain tasks, such as speech recognition.

A comparison of the results presented in the tables is shown in Figure 6.

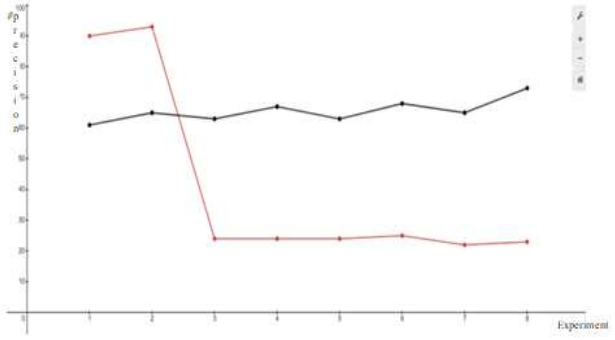

Fig. 6. The resulting graph of speaker language detection accuracy for each experiment. Red – mel-cepstral characteristic coefficients, black – spectrogram

Conclusions

This paper analyzes the impact of audio data preprocessing methods on the accuracy of speaker language detection for use in further neural network training for different methods of converting signals into features, namely spectrograms and mel-cepstral characteristic coefficients. Depending on the chosen methods of pre-processing voice information and different ways of converting the signal into features, we managed to achieve 93% accuracy in determining the speaker's language. The practical significance of this work is to increase the accuracy of further transcription of audio information and translation of the generated text into the chosen language, including artificial languages.

Based on the results of the study, a neural network model using mel-cepstral characteristic coefficients with normalization at the stage of input data preprocessing was chosen for further experiments on speaker speech recognition. This decision is based on several important factors that confirm the advantages of mel-cepstral characteristic coefficients in this context:

1. mel-cepstral characteristic coefficients allow to reduce the amount of input data. This allows to increase the speed of both processing and training of the neural network.

2. They effectively recognize speech and take into account the peculiarities of auditory perception, which interacts well with neural networks.

3. Experiments have shown that mel-cepstral characteristic coefficients have high classification accuracy for different sizes of training samples. This demonstrates their ability to provide information for effective analysis and generalization of patterns in the data.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) *Innovative technologies and scientific solutions for industries. 2023. No. 4 (26)*

4. mel-cepstral characteristic coefficients showed a stable result regardless of the input languages of the training samples. This may be an important aspect for further research, as it allows scaling up the study for more classes of input data.

All the justifications confirm that further speaker recognition and translation into the chosen language, including artificial languages, will use mel-cepstral characteristic coefficients as a method of extracting features of the input audio sequence, since the accuracy of speaker recognition using spectrograms, as shown experimentally in this paper, is 22% worse. In addition, it was concluded that the use of time or frequency masks with mel-cepstral characteristic coefficients reduces the recognition probability to the level of guessing. Therefore, these feature extraction methods should be abandoned.

So, taking into account all the data and analysis results, we can conclude that mel-cepstral characteristic coefficients with preliminary data normalization is a reasonable pickup line for speaker speech detection. Its high accuracy and reduced input data size make this method the best choice for further research and implementation in practical applications.

References

1. Zhang, Z.(2016), "Mechanics of human voice production and control". *The journal of the acoustical society of america* Vol.140.4. P. 2614–2635. DOI: https://doi.org/10.1121/1.4964509

2. Garellek, M.(2022), "Theoretical achievements of phonetics in the 21st century: Phonetics of voice quality". *Journal of Phonetics* Vol.94(24). DOI: https://doi.org/10.1016/j.wocn.2022.101155

3. Abdul, Z. K., Al-Talabani A. K.(2022), "Mel Frequency Cepstral Coefficient and its Applications: A Review", *IEEE Access*, Vol. 10, P. 122136–122158. DOI: https://doi.org/10.1109/ACCESS.2022.3223444

4. Ayvaz, U.(2022), "Automatic speaker recognition using mel-frequency cepstral coefficients through machine learning". *CMC-Computers Materials & Continua*. Vol.71(3), P. 5511–5521. DOI: https://doi.org/10.32604/cmc.2022.023278

5. Shalbbya, A. (2020), "Mel frequency cepstral coefficient: a review". *ICIDSSD*, P. 1–10. DOI: https://doi.org/10.4108/eai.27-2-2020.2303173

6. Ramakrishnan, S. (2012), "Recognition of emotion from speech: A review". *Speech Enhancement, Modeling and Recognition Algorithms and Applications. Rijeka, Croatia: InTech*, P. 121–136. DOI: https://doi.org/10.5772/39246

7. Wang, L.(2022), "A Machine Learning Assessment System for Spoken English Based on Linear Predictive Coding". *Mobile Information Systems*, Vol. 2022 (5). P. 1–12. DOI: https://doi.org/10.1155/2022/6131572

8. Darling, D., Hinduja, J.(2022), "Feature Extraction in Speech Recognition using Linear Predictive Coding: An Overview". *i-Manager's Journal on Digital Signal Processing* Vol. 10.2. 16 p. DOI: https://doi.org/10.26634/jdp.10.2.19289

9. Lonce, W. (2017), "Audio spectrogram representations for processing with convolutional neural networks". *arXiv preprint arXiv*. P. 37–41. DOI: https://doi.org/10.48550/arXiv.1706.09559

10. Gong, Y., Chung, Y., Glass, J. (2021), "Audio spectrogram transformer". arXiv preprint arXiv:Version 3. available at: https://arxiv.org/abs/2104.01778

11. Qiuqiang, K. (2020), "Large-scale pretrained audio neural networks for audio pattern recognition". *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28 (2020). P. 2880–2894. DOI: https://doi.org/10.48550/arXiv.1912.10211

12. Sandhya, P.(2020), "An Analysis of the Impact of Spectral Contrast Feature in Speech Emotion Recognition". *Third International Conference on Advances in Electronics, Computers and Communications (ICAECC). IEEE*, Vol. 9. No. 2. P. 87–95. DOI: https://doi.org/10.3991/ijes.v9i2.22983

13. Charbuty, B., Abdulazeez, A. (2021), "Classification based on decision tree algorithm for machine learning". *Journal of Applied Science and Technology Trends*, Vol. 2.01 (2021). P. 20–28. DOI: https://doi.org/10.38094/jastt20165

14. Breiman, L. (2001), "Random forests". *Machine learning* Vol. 45 (2001). P. 5–32. DOI: http://dx.doi.org/10.1023/A:1010933404324

15. Deshmukh, A.(2020), "Comparison of hidden markov model and recurrent neural network in automatic speech recognition", *European Journal of Engineering and Technology Research*, Vol. 5.8 (2020). P. 958–965. DOI: https://doi.org/10.1051/itmconf/20235401016

16. Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., Weber, G. (2020), "Common Voice: A Massively-Multilingual Speech Corpus". *Proceedings of the Twelfth Language Resources and Evaluation Conference*, P. 4218–4222. DOI: https://doi.org/10.48550/arXiv.1912.06670

17. Goldsborough, P. (2016), "A tour of tensorflow". arXiv preprint arXiv:1610.01178 (2016). available at: https://arxiv.org/abs/1610.01178

18. Havrashenko, A., Barkovska, A. (2023), "Analysis of word search algorithms in the dictionaries of machine translation systems for artificial languages". *Computer systems and information technologies.* No 2. P. 17–24. DOI: https://doi.org/10.31891/csit-2023-2-2

Сучасний стан наукових досліджень та технологій в промисловості. 2023. № 4 (26)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

19. Havrashenko, A., Barkovska, O.(2023), "Analysis of text augmentation algorithms in artificial language machine translation systems". *Advanced Information Systems.* No.7(1). P. 47–53. DOI: https://doi.org/10.20998/2522-9052.2023.1.08

20. Barkovska, O, Havrashenko, A., Kholiev, V., Sevostianova, O.(2021), "Automatic text translation system for artificial languages", *Computer systems and information technologies*. No. 3. P. 21–30. DOI: https://doi.org/10.31891/CSIT-2021-5-3

Received 30.11.2023

Відомості про авторів / About the Authors

Барковська Олеся Юріївна – кандидат технічних наук, доцент, Харківський національний університет радіоелектроніки, доцент кафедри електронних обчислювальних машин, Харків, Україна; e-mail: olesia.barkovska@nure.ua; ORCID ID: https://orcid.org/0000-0001-7496-4353

Гаврашенко Антон Олегович – Харківський національний університет радіоелектроніки, аспірант кафедри електронних обчислювальних машин, Харків, Україна; e-mail: anton.havrashenko@nure.ua; ORCID ID: https://orcid.org/0000-0002-8802-0529

Barkovska Olesia – PhD (Engineering Sciences), Associate Professor, Kharkiv National University of Radio Electronics, Associate Professor at the Department of Electronic Computers, Kharkiv, Ukraine.

Havrashenko Anton – Kharkiv National University of Radio Electronics, PhD student at the Department of Electronic Computers, Kharkiv, Ukraine.

НЕЙРОМЕРЕЖНА МОДЕЛЬ У ЗАДАЧАХ ОБРОБЛЕННЯ ТА АНАЛІЗУ АУДІОФАЙЛІВ

Предметом дослідження є аналіз впливу етапів попереднього оброблення аудіоряду на точність визначення мови спікера. Значущість такого оброблення помітно зросла в останні роки завдяки її ключовій ролі в різноманітних застосуваннях, зокрема: зменшення обсягу інформації, фільтрація та шумопригнічення. Унаслідок збільшення попиту на рішення задач класифікації аудіоінформації оцінювання та порівняння різних методів оброблення аудіоряду стають важливими для визначення точності та ефективності отриманого рішення. **Мета роботи** – аналіз впливу методів попереднього оброблення аудіоінформації на точність визначення мови спікера для використання в подальшому навчанні нейромережі для різних способів перетворення сигналів в ознаки, а саме спектрограми та мелчастотні кепстральні коефіцієнти. Для досягнення поставленої мети були визначені такі **завдання**: проаналізувати способи перетворення сигналу в ознаки та аналіз математичних моделей для виконання аналізу аудіоряду за обраними ознаками; розробити узагальнену модель перекладу мови спікера в реальному часі та спланувати експеримент залежно від обраних етапів попереднього оброблення аудіоряду; змоделювати експеримент способом навчання та тестування згорткової нейроної мережі. Використані такі **методи**: мелчастотний кепстральний аналіз, спректральний аналіз, математичні методи штучного інтелекту. **Досягнуті результати**: залежно від обраних методів попереднього оброблення голосової інформаціїї та різних способів перетворення сигналу в ознаки вдалося досягти 93% точності визначення мови спікера. Практичною значущістю цієї роботи є збільшення точності подальшого транскрибування аудіоінформації та перекладу сформованого тексту обраною мовою, зокрема штучними мовами. **Висновки**. У процесі роботи було обрано найкращу послідовність етапів попереднього оброблення аудіоінформації з метою використання в подальшому навчанні нейромережі для різних способів перетворення сигналів в ознаки. Для вирішення окресленої задачі краще підходять мелчастотні кепстральні коефіцієнти. Оскільки точність нейромережі залежить від її структури, то зі збільшенням обсягів вхідної інформації та кількості мов результати можуть змінюватися. Але на певному етапі було прийнято рішення використовувати лише мелчастотні кепстральні коефіцієнти з нормалізацією на етапі попереднього підготовлення вхідної інформації.

Ключові слова: мелчастотні кепстральні коефіцієнти; спректрограма; часова маска; частотна маска; нормалізація; нейромережа; голос; аудіоряд; мова.

Бібліографічні описи / Bibliographic descriptions

Барковська О. Ю., Гаврашенко А. О. Нейромережна модель у задачах оброблення та аналізу аудіофайлів. *Сучасний стан наукових досліджень та технологій в промисловості*. 2023. № 4 (26). С. 16–23. DOI: https://doi.org/10.30837/ITSSI.2023.26.016

Barkovska, O., Havrashenko, A. (2023), "Analysis of the influence of selected audio pre-processing stages on accuracy of speaker language recognition", *Innovative Technologies and Scientific Solutions for Industries*, No. 4 (26), P. 16–23. DOI: https://doi.org/10.30837/ITSSI.2023.26.016