МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ В ПРИРОДНИЧИХ НАУКАХ ТА ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

MATHEMATICAL MODELING IN NATURAL SCIENCES AND INFORMATION TECHNOLOGIES

DOI: 10.31319/2519-8106.2(51)2024.317425
UDC 004.934:004.932.72

Yalova Kateryna, Candidate of Technical Sciences, Associate Professor, Head of the Department of the Systems software
Ялова К.М., кандидат технічних наук, доцент, кафедра програмного забезпечення систем
ORCID: 0000-0002-2687-5863
e-mail: yalovakateryna@gmail.com

Babenko Mykhailo, Candidate of Technical Sciences, Associate Professor, Associate Professor of the Department of the Systems software
Бабенко М.В., кандидат технічних наук, доцент, кафедра програмного забезпечення систем,
ORCID: 0000-0003-1013-9383
e-mail: mvbab@ukr.net

Sheliuh Kostiantyn, PhD student, Department of the Systems software
Шелюг Костянтин, здобувач третього (доктора філософії) рівня вищої освіти,
кафедра програмного забезпечення систем
email: kostia902@ukr.net

Dniprovsky State Technical University, Kamianske
Дніпровський державний технічний університет, м. Кам'янське

AUDIO SIGNAL PRE-PROCESSING WITHIN SPEECH RECOGNITION TASK

ПОПЕРЕДНЯ ОБРОБКА АУДІО СИГНАЛУ В ЗАДАЧІ РОЗПІЗНАВАННЯ МОВЛЕННЯ

The article presents a generalized description of the speech recognition task, consisting of the following stages: resampling, framing and windowing, feature extraction, vocal tract length normalization, and denoising. The importance of implementing pre-processing is emphasized, as the final result and overall quality of recognition depend significantly on the efficiency and effectiveness of this stage. It is proposed to use the Fast Fourier Transform to represent the input audio signal, and the Hamming window is applied to create audio signal segments for subsequent feature extraction using Mel-Frequency Cepstral Coefficients. The use of the Dynamic Time Warping algorithm for vocal tract length normalization and Recurrent Neural Network for denoising is described. The results of an experiment on pre-processing the audio signals of voice commands for controlling mobile phone applications on the Android operating system are presented.

Keywords*: speech recognition, audio signal pre-processing, Fast Fourier Transform, Mel-Frequency Cepstral Coefficients, Dynamic Time Warping algorithm, Recurrent Neural Network.*

Мова є найбільш природною формою людського спілкування, тому реалізація інтерфейсу, який базується на аналізі мовленнєвої інформації є перспективним напрямком розвитку інтелектуальних систем управління. Система автоматичного розпізнавання мовлення – це інформаційна система, що перетворює вхідний мовленнєвий сигнал на розпізнане повідомлення. Процес розпізнавання мовлення є складним і ресурсоємним завданням через високу варіативність промови, яка залежить від віку, статі та фізіологічних характеристик мовця. У статті представлено узагальнений опис задачі розпізнавання мовлення, що складається з етапів: передискретизація, кадрування та застосування вікон, виділення ознак, нормалізація довжини голосового тракту та шумопригнічення. Попередня обробка мовленнєвого сигналу є першим і ключовим етапом у процесі автоматичного розпізнавання мови, оскільки якість вхідного сигналу суттєво впливає на якість розпізнавання і кінцевий результат цього процесу. Попередня обробка мови складається з очищення вхідного сигналу від зовнішніх і небажаних шумів, виявлення мовленнєвої активності та нормалізації довжини голосового тракту. Метою попередньої обробки мовленнєвого сигналу є підвищення обчислювальної ефективності систем розпізнавання мови та систем керування із природньомовним інтерфейсом.

У статті запропоновано використання швидкого перетворення Фур'є для описування вхідного аудіо сигналу; вікна Hamming для створення сегментів аудіосигналу з подальшим визначенням ознак засобами Mel-Frequency Cepstral Coefficients. Описано використання алгоритму динамічного трансформування часової шкали для нормалізації довжини голосового тракту та рекурентної нейронної мережі для шумопригнічення. Наведено результати експерименту щодо попередньої обробки аудіо сигналу голосових команд для керування застосунками мобільного телефону з оперативною системою Android.

***Ключові слова**: розпізнавання мовлення, попередня обробка аудіо сигналу, швидке перетворення Фур'є, Mel-Frequency Cepstral Coefficients, алгоритм динамічного трансформування часової шкали, рекурентна нейронна мережа.*

Problem's Formulation

The most effective means of human-machine interaction are those that are implemented in a natural way: through visual images and speech [1]. The task of speech recognition is solved, in particular, to create a SILK (speech-image-language-knowledge) interface for controlling the activities of various devices, for example: voice control of various devices, systems, programs; answering machines; automatic call processing; voice user authentication, etc. Speech recognition is the process of transforming a speech signal into digital information [2]. An automatic speech recognition system is a system that converts an input speech signal into a recognised message. In this case, the message can be presented both in the form of the text of the message and immediately converted into a form convenient for its further processing in order to generate an appropriate system response. Automatic speech recognition systems are classified according to the following features [3]:

- vocabulary size (a limited set of words or a large vocabulary);
- dependence on the speaker (speaker-dependent or speaker-independent);
- type of speech (fused, split);
- purpose (dictation systems, command systems);
- recognition algorithm used;
- type of structural unit (phrases, words, and phonemes, etc.).

The creation of SILK interfaces is used to develop the idea of an automated human environment, where human voice commands are perceived as voice commands for electronic devices that surround a person in everyday life. The most successful examples of automatic speech recognition systems are virtual assistant applications built into the functionality of mobile devices, such as SIRI.

Despite the rapidly growing computing power and the rapid development of artificial intelligence methods, the creation of speech recognition systems remains an extremely challenging problem. This is due to both its interdisciplinary nature (it requires knowledge of linguistics, digital signal processing, acoustics, pattern recognition, etc.) and the high computational complexity of the developed algorithms [4]. The latter imposes significant limitations on automatic speech recognition systems —

on the size of the processed vocabulary, the speed of response and its accuracy. The quality of speech recognition is a crucial criterion for the quality of voice-controlled systems.

Analysis of recent research and publications

The first attempts to create automatic speech recognition systems were made in the 1950s and 1960s. The starting point of the initial research was the fundamental concepts of acoustic phonetics. However, in 1959, the starting point of the research changed fundamentally. At University College in the UK, D. Fry and D.B. Dean created a recogniser for four vowels and nine consonants based on statistical information to take into account the valid phoneme sequences in English. Also, among the successful developments of the late 60s, the research of R. Reddy at Carnegie Mellon University in the field of continuous speech recognition based on dynamic phoneme tracking should be mentioned. Since the late 70s, dynamic programming in a variety of variants, including the Viterbi algorithm, has become the dominant method of automatic speech recognition. In the 70s, the intensive development of pattern recognition ideas and dynamic programming methods continued, and a series of experiments aimed at developing speaker-independent speech recognition systems began. AT&T Bell Labs used a wide range of complex classification algorithms to determine the number of patterns required to represent all variants of different words to a wide range of users. Research in the field of speech recognition in the 1980s is characterized by a shift in methodology from the direct pattern recognition paradigm to the formal concept of statistical modelling using Hidden Markov Models, which became the most widely used method in virtually every laboratory around the world. During this period, the idea of using neural networks in speech systems was also reconsidered. Modern approaches to solving the speech recognition task now involve the application of neural networks.

Three main approaches to speech recognition have subsequently emerged [5]:

1. Acoustic signal processing. Probabilistic models, based on Markov chains and Monte Carlo integration methods, are used to build acoustic signal processing technology. These models have led to the development of an algorithm for processing acoustic signals reflected from obstacles.

2. Template-based methods. In template-matching approaches, the input message is compared with a set of pre-recorded words to find the best match, which is a fairly effective method for finding accurate word patterns. However, this approach has a significant drawback: speech variations can only be modelled by using a large number of templates for each word, which becomes impractical over time.

3. Neural networks. The neural network approach is the most modern and effective method for solving the speech recognition task [6]. The choice of a specific neural network architecture depends on the particular speech recognition system being developed. The most commonly used types of neural networks that have proven successful in speech recognition are: Recurrent Neural Networks (RNN), Convolutional Neural Networks, Transformers, and hybrid models, where different types of neural networks are combined into a single architecture.

Regardless of the approach or method chosen for solving the speech recognition task, the goal of research in this area is to achieve accurate recognition and processing of voice commands. Despite the vast amount of scientific work, researchers, and companies working on optimizing recognition algorithms, this task remains relevant and requires further development.

Formulation of the study purpose

Pre-processing of speech signals is the first and most crucial step in the process of automatic speech recognition, as the quality of the input signal significantly impacts the recognition quality and the final outcome of the process. Speech pre-processing involves cleaning the speech signal from surrounding and unwanted noise, detecting speech activity, and normalizing the vocal tract length. The goal of speech signal pre-processing is to make speech recognition systems computationally more efficient.

The purpose of this work is to present the results of implementing speech signal pre-processing mechanisms for further recognition.

Presenting main materials

Sound travels through the environment as a longitudinal wave at a speed that depends on the density of the medium. The simplest way to represent sound is through a sinusoidal graph. The shape of a sound wave is determined by three factors: amplitude, frequency, and phase. Amplitude refers to the displacement of the sinusoidal graphs above and below the time axis ($y = 0$), which corresponds to the energy of the loaded sound wave. Frequency is the number of sinusoid cycles per second. A cycle

of oscillation starts at the midline, reaches a maximum and a minimum, and then returns to the midline. Phase measures the position relative to the start of the sinusoidal curve. Although phase cannot be heard by humans, it can be determined in relation to the position between two signals.

The process of speech recognition is complex and cumbersome due to the high variability of speech, which depends on the speaker's age, gender, and physiology [2]. The generalized algorithm for speech recognition is divided into the following main steps (Fig. 1 [7]):

– receiving the input audio signal;
– pre-processing;
– feature extraction;
– acoustic and language modelling;
– outputting the recognition results.
–

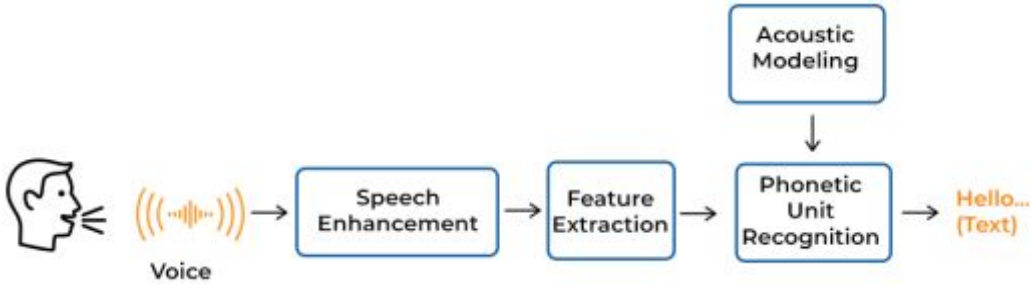

Fig. 1. Main steps of speech recognition process

As the speech recognition system for the presented research, a voice control system for a mobile phone on the Android operating system is being developed. The characteristics of the system under development are as follows:

1. The size of the dictionary used depends on the number of speakers, voice commands, and the number of templates per command. In this work, it is proposed to use 5 speakers, 10 commands, and 5 templates for each command.

2. The system is speaker-dependent, as recognition will be performed in real-time.

3. The system operates with command-based input using discrete speech (with pauses between individual structural units).

4. The structural unit is a phrase that corresponds to a voice command.

Tabl. 1 shows the list of voice commands that were recorded for the experiments on denoising and preparation of the audio signal for further recognition.

Table 1. Experiment’s voice command

Command number	Command text
1	Android, run a calculator
2	Android, run a calculator
3	Android, run a notepad
4	Android, run a notepad
5	Android run a compass
6	Android, run a compass
7	Android, run a scanner
8	Android, run a scanner
9	Android, turn up the sound
10	Android, mute

Tasks and steps in audio pre-processing are used to improve the efficiency and accuracy of recognition process ina whole. The main of them are as follows [8]:

1. Resampling — is a process of changing the sampling rate of an audio signal. The sampling rate is the number of samples of audio carried per second, measured in Hertz (Hz).

2. Framing and windowing — is a mechanism of dividing the continuous audio signal into frames and applying a window function to each frame. It is used for capturing the time-varying nature of speech.

3. Extract Features — is a process of transforming raw audio data into a set of representative, high-level features that can be used by machine learning models speech recognition effectively.

4. Vocal tract length normalization — is a technique used to compensate for the differences in vocal tract length across different speakers.

5. Denoising — is a process of removing or reducing background noise from an audio signal, ensuring that the signal is easier to process for the speech recognition.

There are various resampling methods, for example: linear interpolation, sinc function interpolation, polyphase filtering, and resampling with a windowed sinc filter [9]. Fourier's theorem is used to analyse sound waves and perform resampling, which states that any complex periodic oscillation can be represented as the sum of simple harmonic oscillations. As a result, a set of amplitudes, phases, and frequencies for each sinusoidal component of the wave has been obtained:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi}{N}kn}, \quad (1)$$

where N— is the number of signal values, K— is the number of frequencies, x_n— represents the signal values at specific points in time, X_k— are the complex amplitudes of the sinusoidal signals that constitute the original signal, $k=0,\ldots,K-1$ — is the frequency index, and $n=0,\ldots,N-1$ — represents the discrete time points at which the signal was measured.

A frequency or phase point, combined with the amplitude, is called the spectrum. In the conducted research, Fast Fourier Transform (FFT) was used to accelerate the process of sound wave processing. FFT works with complex numbers and transformation sizes that are powers of two. For a signal length of $N=2m$, the FFT can be computed as:

$$X[k] = \sum_{n=0}^{N-1} x[n] \cdot W_N^{kn}, \quad (2)$$

where $W_N = e^{-j\frac{2\pi}{N}}$ is the N-th root of unity.

The purpose of applying FFT is to obtain a modified sampling rate by processing the frequency representation of the input signal [10]. The next step after resampling is framing and windowing to create audio signal segments for subsequent feature extraction. Framing is used to increase the effectiveness of speech recognition, based on the assumption that the audio signal is stationary with unchanging characteristics. The goal of framing is to represent the audio signal as a set of short overlapping frames [11]. This process can cause spectral leakage. To prevent this and smooth the edges, windowing is applied. The simplest window is the rectangular window: a constant value of 1 that does not alter the signal. It is equivalent to the absence of a weighting window. One of the windows that is used in audio pre-processing for speech recognition — is the Hamming window [12]. The Hamming window creates a weighted emphasis to the center of the frame. It can be found with a formula:

$$w(n) = 0.54 - 0,46 \cdot \cos\left(\frac{2\pi n}{N-1}\right), \quad (3)$$

where $n=(0,1,\ldots,N\text{-}1)$, N — is the number of samples in the frame.

The use of the Hamming window reduces the level of spectral leakage by approximately 40 dB relative to the main peak. However, the spectral description still contains a lot of redundant information that is unnecessary for automatic speech recognition, that is why after framing and windowing, feature extraction is performed. Mel-Frequency Cepstral Coefficients (MFCCs) are the most commonly used feature extraction techniques in audio pre-processing for speech recognition [13]. The frequency spectrum is passed through a set of Mel filters, which are triangular bandpass filters. The Mel scale can be found using a formula (4), where frequency f is converted to Mel scale:

$$M(f) = 1127 \cdot \ln\left(1 + \frac{f}{700}\right), \tag{4}$$

Afterward, it is necessary to summarize the energy in each of these segments to get data of how much energy exists in different frequency regions. This is called a filter bank:

$$H_m(k) = \begin{cases} 0, k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)}, f(m-1) \le k \le f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)}, f(m) \le k \le f(m+1) \\ 0, k > f(m+1) \end{cases} \tag{5}$$

where *m*— is a number of MFCC, *k*—is a current frequency.

The first filter is very narrow and indicates how much energy exists near 0 Hz. As the frequency increases, the filters become wider because human hearing is less sensitive to higher frequencies. After calculating the energy in the filter bank, the logarithm of the energy values needs to be computed, as humans do not perceive loudness on a linear scale. This operation makes coefficients more aligned with human sound perception. The final step is to calculate the Discrete Cosine Transform of the logarithmic energies from the filter bank:

$$X_k = \sum_{n=0}^{N-1} x_n \cos\left(\frac{\pi}{N}\left(n + \frac{1}{2}\right)k\right). \tag{6}$$

The next step is vocal tract length normalization. Pronunciation of the same word typically has different durations depending on the speaker. Even if the word is pronounced with the same duration, the duration of individual parts of the word may vary. Therefore, to obtain a measure of similarity between two speech signals in the form of a vector, time alignment — vocal tract length normalization — must be performed. One of the effective method for time alignment is the Dynamic Time Warping (DTW) algorithm. DTW is an algorithm used to measure similarity between two time-dependent sequences (see Fig. 2) [14]. Let's assume that the first sequence is $X=(x_1,x_2,...,x_N)$with N time steps, and the second is $Y=(y_1,y_2,...,y_M)$ with M time steps. DTW method operates on segments of sequences, meaning that feature analysis involves processing feature vectors at regular intervals. Since the feature vector can have a large number of segments, there is a need for a method to calculate the local distance between points of signal X and template Y in an n-dimensional space (Euclidean distance). The distance between X and Y at the steps i and j can be calculated as:

$$d(x_i, y_j) = \sqrt{\sum_{k=1}^{K} \left(x_i^k - y_j^k\right)^2}, \tag{7}$$

where *K*— is a number of *X* and *Y* features in our case the number of MFCC coefficients.

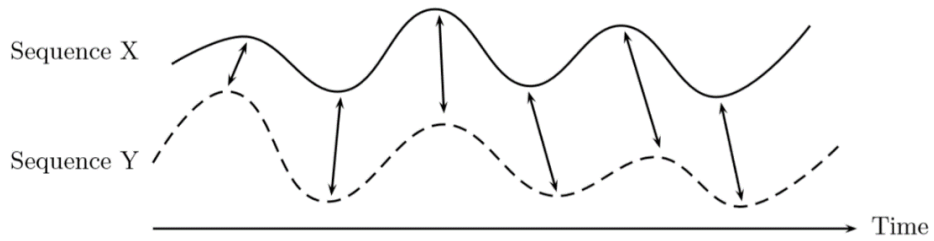

Fig. 2. Difference in the tract length

Thus, the input signal is compared with all templates. The result of the comparison will be the template for which the minimum divergence between the input signal and the template was found, which is the sum of the local distances between segments of the signal and the template.

Audio pre-processing is ended with the denoising process. Common denoising methods used in audio pre-processing are: spectral subtraction, Wiener filtering, Kalman filtering, non-local means denoising, deep-learning-based denoising, wavelet transform-based denoising. Also there are some applications that are used to reduce unwanted noise in the input audio, for example: Audacity, Adobe Audition, iZotope RX, Waves NS1, Krips, NVIDIA RTX Voice, Acon digital deNoise, etc. Fig. 3 shows the spectrogram of the voice command «Android, open calculator», where the audio recording of the template was made outdoors with background noise.

The use of NVIDIA RTX Voice improved the quality of the input audio signal and filtered out background noise (see Fig. 4). By comparing the spectrograms shown in Fig. 3 and 4, it can be observed that the input message has undergone significant changes, and the level of background noise has been reduced.

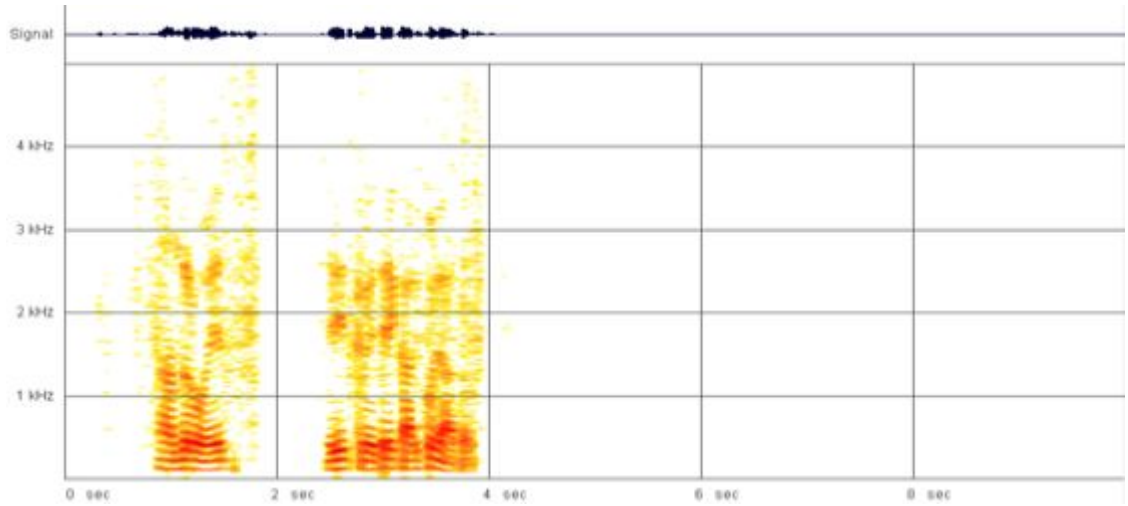

Fig. 3. Voice command'sspectrogram before denoising

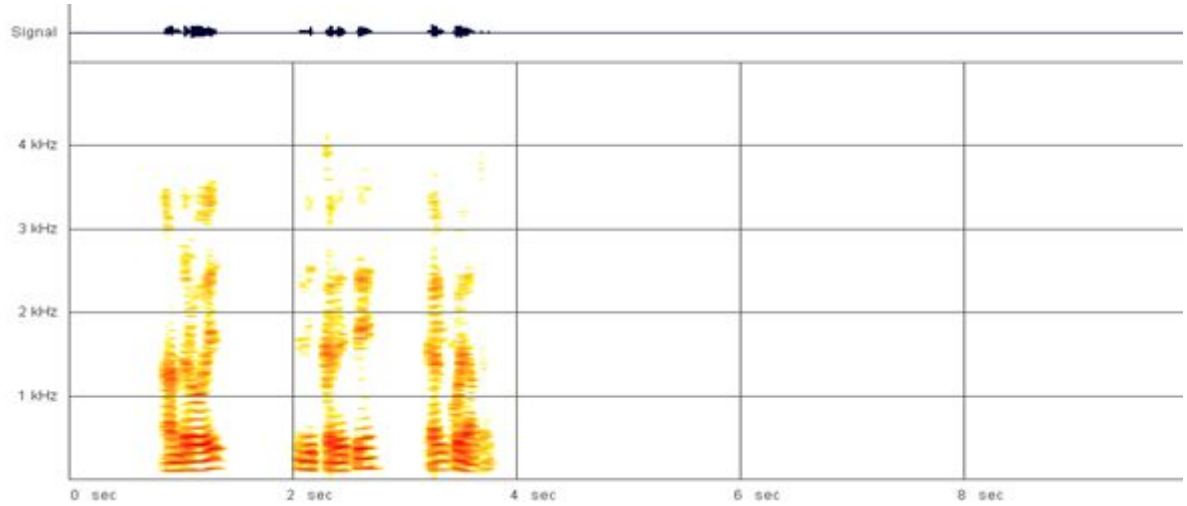

Fig. 4. Voice command's spectrogram after RTX Voice application

To improve filtering, it was decided to implement filtration using neural networks (NN). The architecture used to build the NN was Recurrent Neural Networks (RNN). The effectiveness of applying RNN for audio signal preprocessing is due to their ability to retain information over time [15]. With the segmented input message and the corresponding bank of MFCCs, the sequence of feature vectors needs to be passed to the NN. The NN uses hidden layers to process each vector. At each time

step t, the RNN takes an input feature vector $x_t$and computes a new hidden state h_{t-1}, which is influenced by both the current input and the previous hidden state h_{t-1}. This allows the RNN to remember important information from previous inputs:

$$h_t = \sigma(W_h x_t + U_h h_{t-1} + b_h), \quad (8)$$

where h_t — is the hidden state at time step t, W_h and U_h are weight matrices, $x_t$is the current input MFCC features, b_h is the bias term, σ — is an activation function.

The network was trained using backpropagation through time.

Fig. 5 shows the spectrogram of the audio signal after applying the NN for removing unwanted noise.

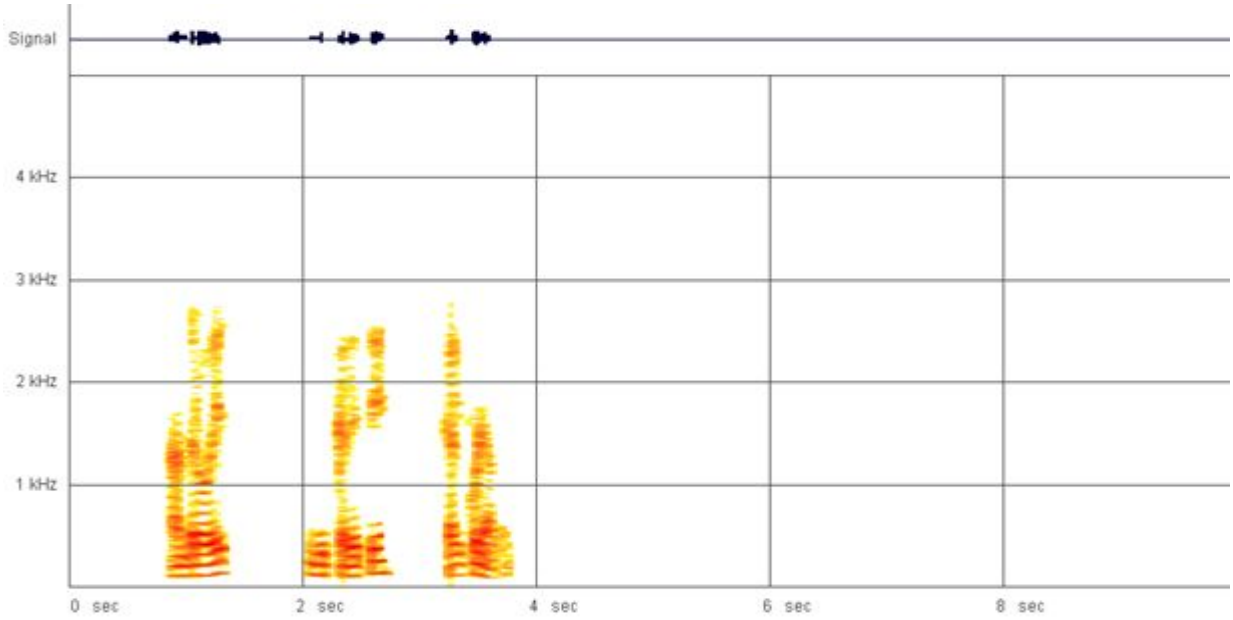

Fig. 5. Voice command's spectrogram after neural network application

By comparing the data in Fig. 3, 4, and 5, we can conclude that the proposed approach for audio signal preprocessing in the speech recognition task is effective and holds promise for further use.

Conclusions

The task of speech recognition is a relevant scientific and practical problem, falling within the fields of pattern recognition and natural language processing. Audio signal pre-processing in speech recognition is a crucial stage that significantly influences the automated recognition system's accuracy and efficiency. It involves preparing the audio signal for analysis by removing noise, normalizing, segmenting, and transforming the signal into a form suitable for input into the model for further analysis. Proper pre-processing allows the extraction of the most important features for speech recognition, reduces computational complexity, and improves noise robustness. This paper outlines an approach to the stages of audio signal pre-processing, employing Fast Fourier Transform to describe the input audio signal, the Hamming Window for creating audio signal segments, followed by feature extraction using Mel-Frequency Cepstral Coefficients. The use of the DTW algorithm is an effective means of implementing vocal tract length normalization, while the application of RNN demonstrates efficiency in minimizing unwanted noise levels.

References

[1] Pahwa, R.,Tanwar, H., & Sharma, S. (2020). Speech recognition system: a review. International Journal of Future Generation Communication and Networking, 13, 2547—2559.

[2] O`Shaughnessy, D. (2024). Trends and developments in automatic speech recognition research. Computer speech and language, 83, 1—15. doi: 10.1016/j.csl.2023.101538.
[3] Al-Fraihat, D.,Sharrab, Y., Alzyoud, F., Qahmash, A., & Maaita, A. (2024). Speech recognition utilizing deep learning: a systematic review of the latest developments. Human-centric Computing and Information Sciences, 15. doi: 10.22967/HCIS.2024.14.015.
[4] Zhang, L., & Sun, X. (2021). Study on speech recognition method of artificial intelligence deep learning. Journal of Physics: Conference Series,1754. doi: 10.1088/1742-6596/1754/1/012183.
[5] Barkovska, O., Havrashenko, A. (2023). Analysis of the influence of selected audio preprocessing stages on accuracy of speaker language recognition. Innovative Technologies and Scientific Solutions for Industries, 4 (26), 16—23. doi: https://doi.org/10.30837/ITSSI.2023.26.016.
[6] Keerio, A., Mitra, B., Birch, P., Young, R. & Chatwin, C. (2008). On preprocessing of speech signals. World Academy of Science, Engineering and Technology, 47, 317—323.
[7] Top 10 Speech Recognition Software and Platforms in 2022. Retrieved from: https://www.spiceworks.com/tech/artificial-intelligence/articles/speech-recognition-software/.
[8] Rajaratnam, K., Shah, K., Kalita, J. (2018). Isolated and ensemble audio preprocessing methods for detecting adversarial examples against automatic speech recognition. *Proceedings of the 30th Conference on Computational Linguistics and Speech Processing* (ROCLING'18), pp. 16—30, Hsinchu, Taiwan.
[9] Labied, M., Belangour, A., Banane, M., & Erraissi, A. (2022). An overview of automatic speech recognition preprocessing techniques. *Proceedings of the International Conference on Decision Aid Sciences and Applications* (DASA'22), pp. 804—809, Chiangrai, Thailand.
[10] Lee, S.-J., &Kwon, H.-Y.(2020). A preprocessing strategy for denoising of speech data based on speech segment detection. Applicated science, 10(20), 7385. doi: 10.3390/app10207385.
[11] Raj, V. A., & Dhas, M. D. K. (2022). Analysis of audio signal using various transforms forenhanced audio processing. International Journal of Health Sciences, 6(S2), 12989—13001.doi:https://doi.org/10.53730/ijhs.v6nS2.8890.
[12] Vreca, J., Pilipovic, R., &Biasizzo, A. (2024). Hardware-software co-design of an audio feature extraction pipeline for machine learning applications. Electronics, 13(5), 875. doi:https://doi.org/10.3390/electronics13050875.
[13] Durairaj, P. & Sriuppili, S. (2021). Speech processing: MFCC based feature extraction techniques- an investigation. Journal of Physics: Conference Series, 1717. doi:10.1088/1742-6596/1717/1/012009.
[14] Shaohua, J., & Zheng, C. (2023). Application of dynamic time warping optimization algorithm in speech recognition of machine translation. Heliyon, 9(11), 1—10. doi: 10.1016/j.heliyon.2023.e21625.
[15] Boyko, N., & Hrynyshyn, A. (2021). Using recurrent neural network to noise absorption from audio files. *Proceedings of the 2nd International Workshop on Computational & Information Technologies for Risk-Informed Systems* (CITRisk'2021), pp.1—19, Kherson, Ukraine.

Список використаної літератури

1. Pahwa R., Tanwar H., Sharma S. Speech recognition system: a review. *International Journal of Future Generation Communication and Networking*. 2020. Vol. 13. P. 2547–2559.
2. O`Shaughnessy D. Trends and developments in automatic speech recognition research. *Computer speech and language*. 2024. Vol. 83. P. 1–15. DOI: 10.1016/j.csl.2023.101538.
3. Al-Fraihat D., Sharrab Y., Alzyoud F., Qahmash A., Maaita A. Speech recognitionutilizing deep learning: a systematic review of the latest developments. *Human-centric Computing and Information Sciences*. 2024. Vol.15. DOI: 10.22967/HCIS.2024.14.015.
4. Zhang L., Sun X. Study on speech recognition method of artificial intelligence deep learning. *Journal of Physics: Conference Series.* 2021. Vol. 1754. DOI: 10.1088/1742-6596/1754/1/012183.

5. Barkovska O., Havrashenko A. Analysis of the influence of selected audio pre-processing stages on accuracy of speaker language recognition. *Innovative Technologies and Scientific Solutions for Industries*. 2023.No. 4, I. 26, P. 16–23. DOI: https://doi.org/10.30837/ITSSI.2023.26.016.
6. Keerio A., Mitra B., Birch P., Young R., Chatwin C. On preprocessing of speech signals. *World Academy of Science, Engineering and Technology*. 2008.Vol. 47. P. 317—323.
7. Top 10 Speech Recognition Software and Platforms in 2022. URL: https://www.spiceworks.com/tech/artificial-intelligence/articles/speech-recognition-software/ (дата звернення: 07.06.2024).
8. Rajaratnam K., Shah K., Kalita J. Isolated and ensemble audio preprocessing methods for detecting adversarial examples against automatic speech recognition. *Proceedings of the 30th Conference on Computational Linguistics and Speech Processing*, Hsinchu, Taiwan, 2018. pp. 16–30.
9. Labied M., Belangour A., Banane M., Erraissi A. An overview of automatic speech recognition preprocessing techniques. *Proceedings of the International Conference on Decision Aid Sciences and Applications*, Chiangrai, Thailand, 2022. pp. 804–809.
10. Lee S.-J., Kwon H.-Y. A preprocessing strategy for denoising of speech data based on speech segment detection. *Applicated science*. 2020. Vol. 10, I. 20.P. 7385. DOI: 10.3390/app10207385.
11. Raj V. A., Dhas M. D. K. Analysis of audio signal using various transforms for enhanced audio processing. *International Journal of Health Sciences*.2022. Vol. 6, I. 2. P. 12989–13001. DOI:https://doi.org/10.53730/ijhs.v6nS2.8890.
12. Vreca J., Pilipovic R., Biasizzo A. Hardware-software co-design of an audio feature extraction pipeline for machine learning applications. *Electronics*. 2024. Vol. 13, I. 5. P. 875. DOI:https://doi.org/10.3390/electronics13050875.
13. Durairaj P., Sriuppili S. Speech processing: MFCC based feature extraction techniques —an investigation. *Journal of Physics: Conference Series*. 2021. 1717. DOI:10.1088/1742-6596/1717/1/012009.
14. Shaohua J., Zheng C. Application of dynamic time warping optimization algorithm in speech recognition of machine translation. *Heliyon*. 2023. Vol. 9, I. 11. P. 1–10. DOI: 10.1016/j.heliyon.2023.e21625.
15. Boyko N., Hrynyshyn A. Using recurrent neural network to noise absorption from audio files. *Proceedings of the 2nd International Workshop on Computational & Information Technologies for Risk-Informed Systems* (CITRisk'2021), Kherson, Ukraine, 2021. pp. 1–19.

Надійшла до редколегії 03.06.2024