ТЕХНИЧЕСКИЕ НАУКИ
И ТЕХНОЛОГИИ

АҚПАРАТТЫҚ ЖҮЙЕЛЕР
ИНФОРМАЦИОННЫЕ СИСТЕМЫ
INFORMATION SYSTEMS

DOI 10.51885/1561-4212_2022_4_179
МРНТИ 44.29.01

D. Muratuly[1], N.F. Denissova[1], Yu.V. Krak[2].
[1]D. Serikbayev East Kazakhstan Technical University, Ust-Kamenogorsk, Kazakhstan
*E-mail: Muratulydidar@gmail.com**
E-mail: NDenisova@edu.ektu.kz
[2]Taras Shevchenko National University of Kyiv, Kyiv, Ukraine
E-mail: Yuri.krak@gmail.com

SPEECH DETECTION AND RECOGNITION FOR USE IN ONLINE PROCTORING SYSTEMS: A REVIEW AND RESEARCH OF TECHNOLOGIES

ОНЛАЙН ПРОКТОРИНГ ЖҮЙЕЛЕРІНДЕ ҚОЛДАНУ ҮШІН СӨЙЛЕУ ЖӘНЕ ТАНУ: ТЕХНОЛОГИЯЛАРДЫ ШОЛУ ЖӘНЕ ЗЕРТТЕУ

ОБНАРУЖЕНИЕ И РАСПОЗНАВАНИЕ РЕЧИ ДЛЯ ИСПОЛЬЗОВАНИЯ В СИСТЕМАХ ОНЛАЙН-ПРОКТОРИНГА: ОБЗОР И ИССЛЕДОВАНИЕ ТЕХНОЛОГИЙ

Abstract. *Speech detection and recognition is a classification task that determines if there is a voice in a particular audio segment. This process is an important pre-processing step that can be used to improve the performance of other tasks such as automatic real-time voice detection during an exam. This article provides an overview of methods, libraries for speech detection and recognition. Speech recognition research spans many subject areas such as computer technology, artificial intelligence, digital signal processing, pattern recognition, acoustics, linguistics, and cognitive science. The aim of the study is to develop an online proctoring module that records and transcribes audio streams during the exam.*
Keywords: *distance learning, exam session, speech detection.*

Аңдатпа. *Сөйлеуді анықтау және тану – белгілі бір дыбыс сегментінде дауыстың бар-жоғын анықтайтын жіктеу тапсырмасы. Бұл процесс емтихан кезінде нақты уақытта автоматты түрде дауысты анықтау сияқты басқа тапсырмалардың өнімділігін жақсарту үшін пайдалануға болатын маңызды алдын ала өңдеу қадамы болып табылады. Бұл мақалада сөйлеуді анықтау және тануға арналған әдістерге, кітапханаларға шолу жасалады. Сөйлеуді тану зерттеулері компьютерлік технологиялар, жасанды интеллект, цифрлық сигналдарды өңдеу, үлгіні тану, акустика, лингвистика және когнитивтік ғылым сияқты көптеген пәндік салаларды қамтиды. Зерттеудің мақсаты - емтихан кезінде аудио ағындарды жазатын және транскрипциялайтын онлайн-прокторинг модулін әзірлеу.*
Түйін сөздер: *қашықтықтан оқыту, емтихан сессиясы, сөйлеуді анықтау.*

Аннотация. *Обнаружение и распознавание речи – задача классификации, которая определяет, наличие голоса в определенном аудиосегменте. Этот процесс является важным этапом предварительной обработки, который можно использовать для повышения производительности других задач, таких как автоматическое обнаружение голоса в режиме реального времени во время экзамена. В этой статье представлен обзор методов, библиотек для обнаружения и распознавания речи. Исследования распознавания речи охватывают многие предметные области, такие как компьютерные технологии, искусственный интеллект, цифровая обработка сигналов, распознавание образов, акустика, лингвистика и когнитивные науки. Целью исследования является разработка модуля онлайн-прокторинга, который записывает и*

расшифровывает аудиопотоки во время экзамена.

Ключевые слова: *дистанционное обучение, экзаменационная сессия, обнаружение речи.*

Introduction. The elevated speech system plays an important role in distance learning during the exam. The proctoring system module for speech development divides speech into different sound waveforms, analyzes each sound waveform using different algorithms.

Voice activity detection is a technique that detects the presence or absence of human speech. Voice activity detection systems must distinguish speech from noise and silence [1].

A good speech detection system should provide an excellent balance between low computation/latency and decent modern quality [2].

Advances in speech signal processing techniques have made it possible to accurately detect the presence of speech in an incoming signal, a problem in the industry in a variety of noise environments. The separation of the speech segment from the non-speech segment in the audio signal is achieved using voice activity detectors.

Materials and methods of research.

Basic VAD works on the principle of extracting measured characteristics from the incoming audio signal, which is divided into frames of 5-40 ms.

These extracted features from the audio signal are then compared to a threshold limit, which is typically estimated from noise-only periods of the input signal, and a VAD solution is computed.

If the characteristic of the input frame exceeds the assumed threshold value, a VAD decision (VAD=1) is computed, which declares that speech is present. Otherwise, a VAD decision is calculated (VAD = 0), which declares the absence of speech in the input frame [3].

Speech signals are sound signals defined as pressure fluctuations propagating through the air. These pressure changes can be described as waves and are accordingly often referred to as sound waves [4].

In this context, we are primarily interested in the analysis and processing of such signals in digital systems.

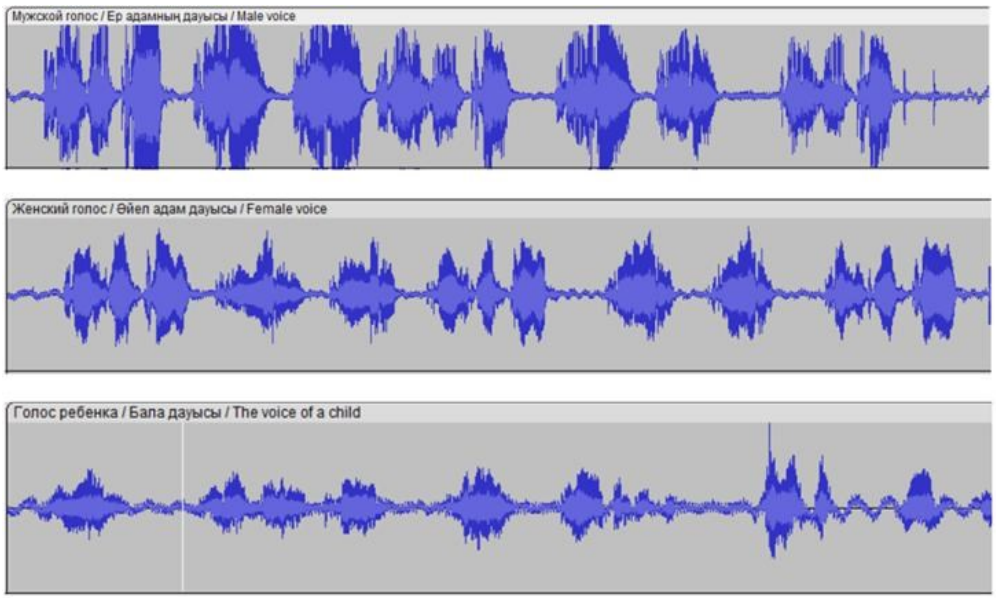

Figure 1. Illustrative representation of waveforms

Recent advances in audiovisual speech recognition (AVSR) have established the importance of incorporating visual components into the speech recognition process to improve reliability [5].

Visual features have great potential to improve the accuracy of existing speech recognition methods and are becoming increasingly indispensable in modeling speech recognizers. The combination of machine learning methods combined with visual features can provide promising solutions to the problem of understanding speech in a noisy environment [6].

The performance of a pattern recognition method is usually determined by the ability to extract useful features from the available data in order to efficiently characterize and distinguish patterns [7].

The method of extracting features from speech signals generates spectrograms, which are time-frequency representations of the original signal.

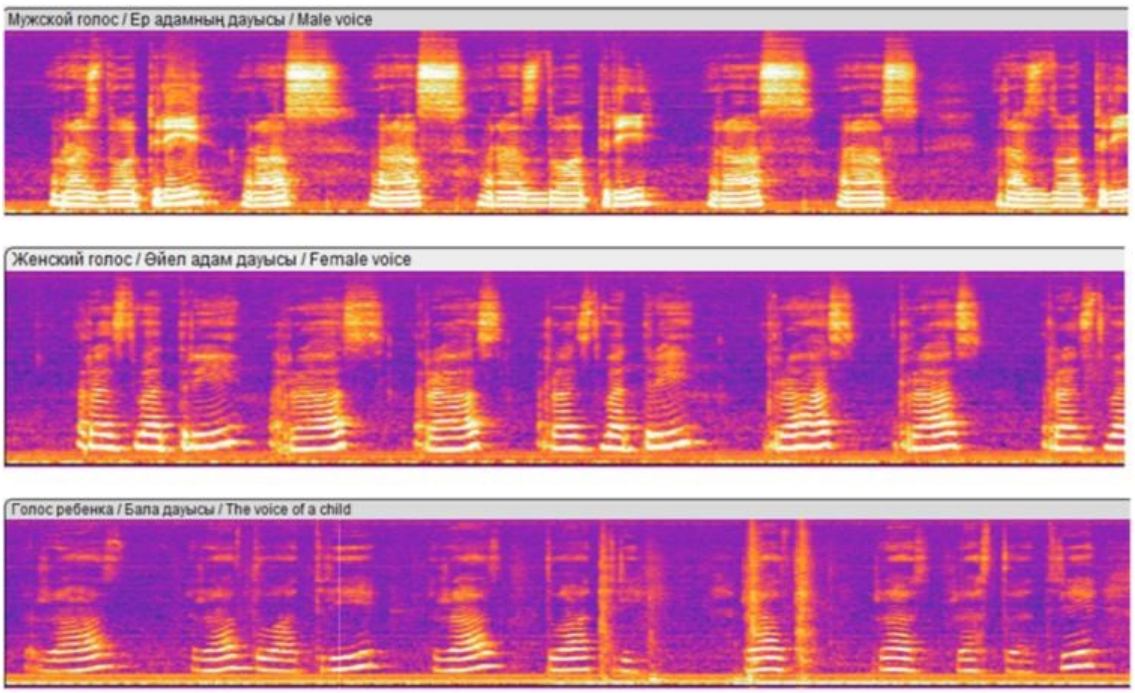

Figure 2. An illustrative representation of an audio stream as a spectrogram

Speaker recognition, also known as voice print recognition, is an important branch of speech signal processing.

The spectrogram shows that the female voice has more high-frequency components than the male voice. Some people may see more high-frequency components in a male voice - this is also correct, depending on which frequency range the attention is directed to.

Since frequencies above 8000 Hz are usually not very significant for speech detection, the 0-8000 Hz range is usually considered. Many researchers have used the spectrogram as an acoustic feature in combination with the artificial neural network method for speaker recognition [8].

Common speech recognition techniques include hidden markov models (HMMs), Gaussian mixed models (GMMs), vector quantization, dynamic time transformation, support vector machines (SVMs), and artificial neural networks.

For more than two decades, the Gaussian Mixed Model - Universal Background Model (GMM-UBM) has become a widely used paradigm in speaker recognition systems due to its good performance in speaker recognition.

In recent years, the application of deep learning technology in the field of speech recognition has greatly improved both recognition speed and reliability, and the results obtained with deep learning neural networks continue to encourage the use of neural networks for speech recognition [9].

Currently, speech feature extraction methods commonly used in speaker recognition systems include cepstral linear prediction coefficients.

Different types of electromagnetic waves have different frequency ranges. For example, radio, microwave, infrared, visible light, ultraviolet and X-rays. Acoustic sound waves that humans can hear typically have a frequency between 20 hertz and 20 kilohertz [10].

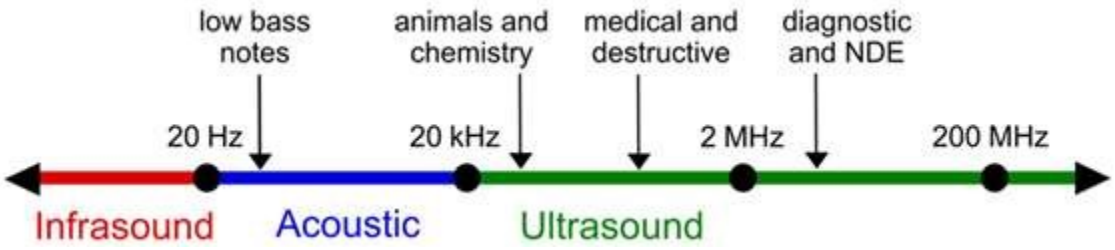

Figure 3. Types of electromagnetic waves and their frequency ranges

The intensity corresponds to the signal strength. The signal is a function F of time t with period t in uppercase. The signal strength can be calculated by taking the integral of the function.

This means that if the power is less than p zero, then the person cannot hear the signal. Then the signal intensity can be defined as the power rhythm divided by zero p, with the unit of intensity being the decibel, not the power [11].

$$P = \frac{1}{T}\int_0^T \left(f(t)\right)^2 dt \tag{1}$$

Intensity is also determined by sound pressure. The numerator is the RMS value of the sound pressure. Under the denominator is the standard reference sound pressure equal to 20 microparticles.

When intensity is determined by pressure, the factor is 20, and when determined by force, the factor is ten.

$$L_{dB} = 20log_{10}\left(\frac{Prms}{Pref}\right) \tag{2}$$

In this case: *Prms* is the root mean square value of the sound pressure measure, *Pref* is the standard reference sound pressure [12].

Speech activity detection refers to the task of determining whether a signal contains speech or not. Thus, this is a binary solution. A related problem is to determine the probability that the input signal contains speech or not, called the Speech Presence Probability (SPP).

Table 1. Human perception of intensity

Noise source	Decibel
The sound, at the moment of firing from a gun «unsuppressed» (without a silencer), medium caliber, near the muzzle of the barrel.	150
Sandblaster, jackhammer at a distance of less than one meter	120
The sound of a freight railway car seven meters away	90
Loud conversations of people, at a distance of less than one meter	70
The sound is characterized by loud conversations at a distance of one meter	60
Normal everyday speech, calm conversation of people. This threshold is the daily norm for residential premises.	40
Quite a distinct whisper, also the sound is comparable to the ticking of a wall clock.	30
The sound is comparable to the distant rustle of leaves.	10

The SPP is then usually expressed as a probability in the range of 0 to 1. The probability of the presence of speech is usually an intermediate step in the detection of speech activity, so that the classification of speech activity is obtained by thresholding the output of the speech presence probability estimator [13].

Figure 4. Types of electromagnetic waves and their frequency ranges

Although the subjective assessment of auditory perception depends on subjective factors and has problems such as low accuracy and stability, it cannot be denied that the subjective assessment of auditory perception is the only reference standard for testing and evaluating the effectiveness of objective voice test parameters and voice function [14].

Results and discussion. In many speech signal processing applications, voice activity detection (VAD) plays an important role in separating the audio stream into time slots containing speech activity and time slots in which there is no speech [15].

The detection of speech activity is usually associated with a binary decision about the presence of speech for each frame of a noisy signal.

Approaches that localize speech fragments in the time and frequency domain, such as speech presence probability estimation or ideal binary mask estimation, can be considered as extensions of speech activity detection [16].

The Short Time Fourier Transform (STFT) and Modified Discrete Cosine Transform (MDCT) are selected for audio processing and their built patterns are passed to the CNN [17].

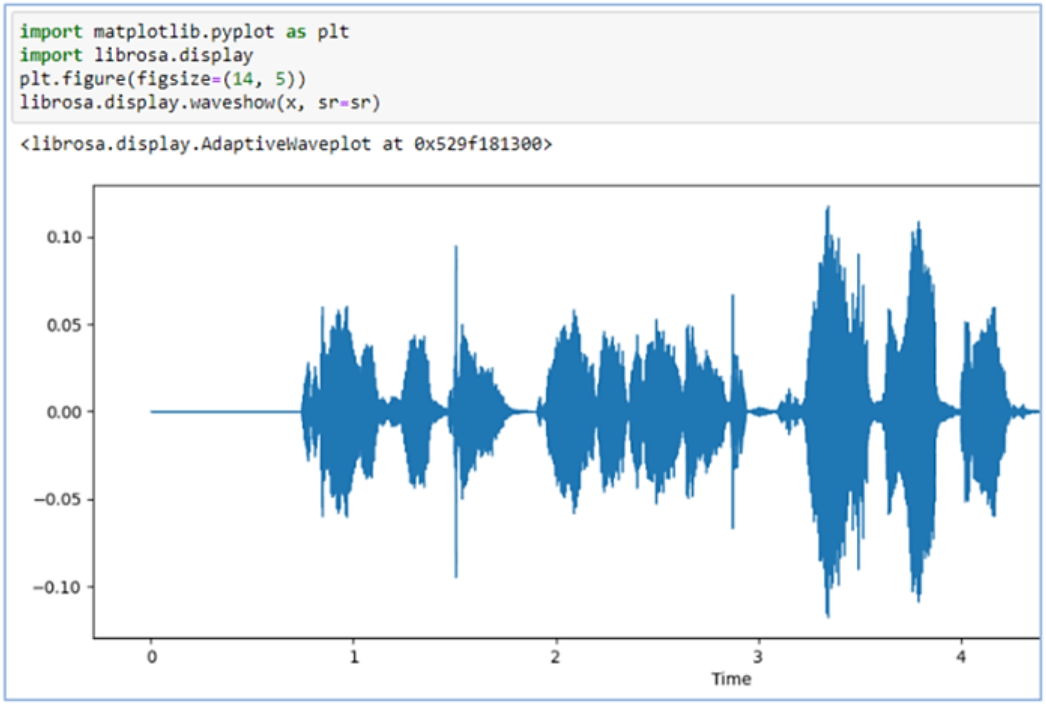

Figure 5. Representing an audio stream as waveforms

A spectrogram is an image showing the dependence of the spectral power density of a signal on time. Spectrograms are used for speech identification, animal sound analysis, various areas of music, radio and sonar, speech processing, seismology, and other areas.

Spectrograms are speech spectrum maps originally developed during World War II to detect submarines and decipher enemy codes, but later came to be used in the field of linguistics.

The vertical and horizontal axes of the gray scale spectrogram represent frequency and time, respectively, while the scale value of each pixel reflects the energy density of the signal at the corresponding time and at the corresponding frequency [18].

Spectrograms can display changes in fundamental frequency, pitch period, and formant intensity in a speaker's utterance over time in a two-dimensional view. They are usually used to represent the long-term frequency characteristics of a speech signal, but may not reflect the detailed characteristics of pronunciation [19].

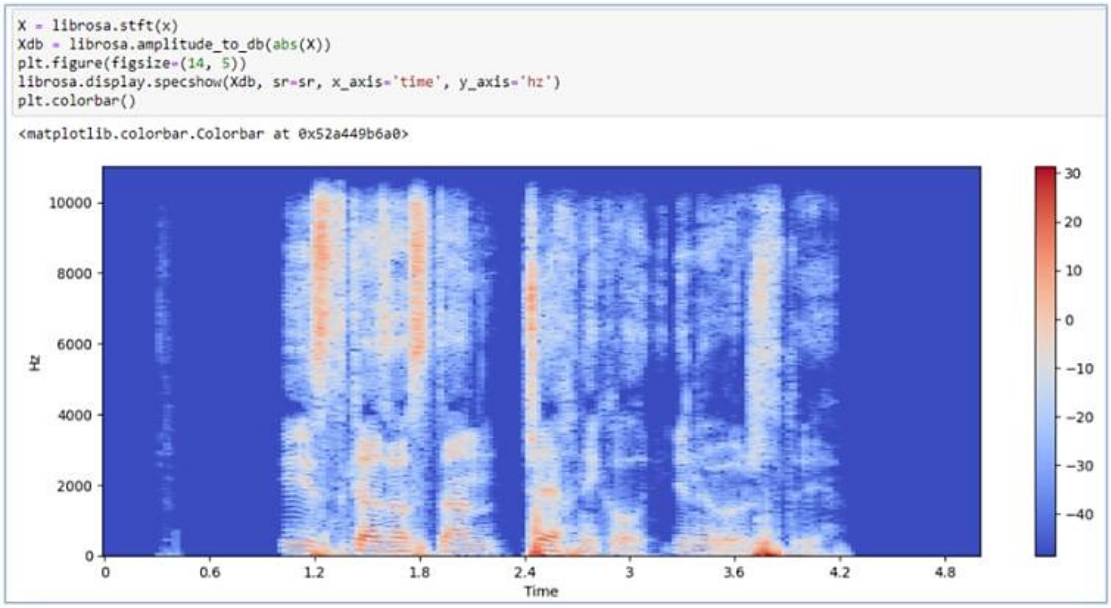

Figure 6. Representing an audio stream as a spectrogram

From a physiological point of view, given the short-term stationary characteristics of human pronunciation, long utterances can be divided into several shorter speech segments, each of which is represented as a single frame. Then each segment of a short-term speech signal can be considered as a short-term stationary signal [20].

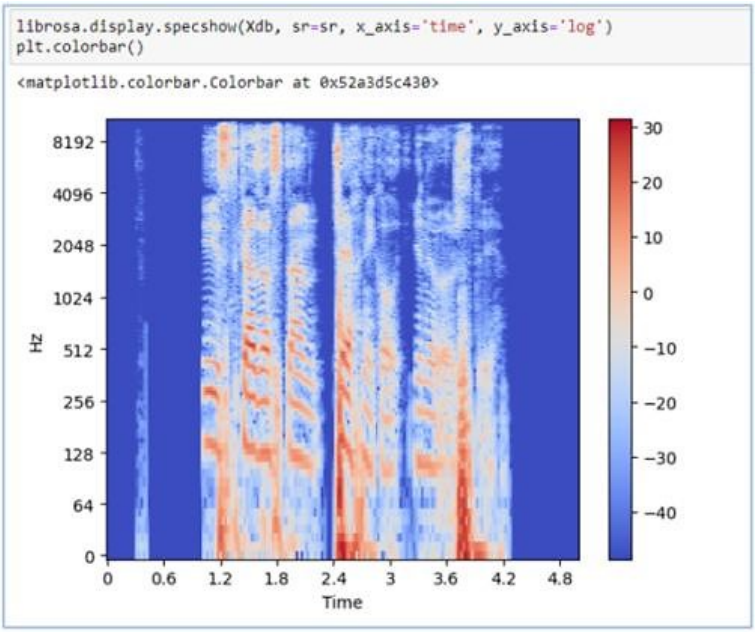

Figure 7. Short-term stationary signal of the spectrogram

```
type(audio)
speech_recognition.AudioData

r.recognize_google(audio)
'East Kazakhstan Technical University'
```

Figure 8. Code fragment for converting an audio stream to a text file

The amplitude-frequency spectrum of sound represents data on the relative intensity of the frequency components of sound. This data can be obtained using filter banks or a frequency tunable filter. The more intense the frequency contained in the sound spectrum, the stronger the "response" of the corresponding filter will be. Since the spectral analysis uses electrical vibrations.

The Speech - to-Text API allows developers to convert sound to text in over 125 languages and variants by applying powerful neural network models in an easy-to-use API.

Thus, speech signals are used as a basis and converted from analog to binary format. And it goes through three main stages.

First, the path shape encoder, which takes the actual waveform and creates a series of ones and zeros as its representation.

Secondly, a hybrid encoder that uses both wave and parametric principles in this production.

Third, a parametric encoder that attempts to determine certain characteristics of speech, such as pitch and amplitude, that were mentioned earlier.

Today, there are quite a few services and programs that perform various tasks with speech recognition (voice control, voice typing, etc.). Ideally, all these systems should help and simplify the performance of the tasks assigned to them.

All modern speech recognition systems are based on statistical methods that make it possible to use the powerful apparatus of mathematical statistics and probability theory, which, in turn, significantly improves the quality of recognition.

Conclusions. The article examines the development of modern society under the influence of globalization processes, which lead to the emergence of new requirements for the subject of any field of activity, including higher education.

The pandemic has contributed to the active use of distance learning technologies. In the future, the experience of their implementation should become the basis for developing innovative teaching methods, improving the efficiency of the educational process and improving information and resource support.

During the pandemic, after intensive research, the speech recognition system has carved its niche and can be seen in many areas of life. The accuracy of speech recognition systems remains one of the most important research challenges.

Speech recognition is a complex task. In this article, we have tried to provide an overview of how much this technology has advanced over the previous years. The performance of a speech recognition system mainly depends on the quality of the signal pre-processing step.

Most of the research done so far explains the fact that speech is a very subjective phenomenon. Common problems are speaker variation, background noise, and continuous speech. Perhaps the most obvious source of performance degradation in speech recognition is noise.

Summing up the results of the work done, I would like to note that the tasks set have been completed:

1) An idea about the field of digital processing of audio signals was obtained, the features of speech activity were studied, as well as machine learning methods for its detection.

2) A review of existing speech activity detection algorithms is described, their advantages and disadvantages are considered.

3) A review of the tools with which you can develop a voice activity detector is made.

4) Modules for data sampling and model training in the Python programming language have been developed and implemented.

It should also be noted that the approach used in this study can be applied directly to speech detection and recognition during an exam using scalable online proctoring tools.

References

1. Jassim W.A., Harte N., Voice activity detection using neurograms // IEEE International Conference on Acoustics. – 2018. – PP. 5524-5528, https://doi.org/10.1109/ICASSP.2018.8461952
2. Sunil Kumar S.B., Rao K.S., Voice/non-voice detection using phase of zero frequency filtered speech signal// Speech Communication. – 2016. – PP.90-103, https://doi.org/10.1016/j.specom.2016.01.008
3. Liu F., Demosthenous A., A computation efficient voice activity detector for low signal-to-noise ratio in hearing aids// 2021 IEEE International Midwest Symposium on Circuits and Systems. – 2021. – PP.524-528, https://doi.org/10.1109/MWSCAS47672.2021.9531915
4. Çolak R., Akdeniz R., A novel voice activity detection for multi-channel noise reduction // IEEE Access. – 2021. – PP. 91017-91026, https://doi.org/10.1109/ACCESS.2021.3086364
5. Zhang X., Wang D., Boosting contextual information for deep neural network based voice activity detection // IEEE/ACM Trans Audio Speech Lang Process. – 2016. – Pp. 252-264, https://doi.org/10.1109/TASLP.2015.2505415
6. Shah V.H., Chandra M., Speech Recognition Using Spectrogram-Based Visual Features // Advances in Machine Learning and Computational Intelligence. – 2020. – Pp. 695-704, https://doi.org/10.1007/978-981-15-5243-4_66
7. Muratuly D., Denissova N., Krak Y., Apayev K., Biometric authentication of students to control the learning process in online education // Scientific Journal of Astana IT University. – 2022. – Pp. 22-32, https://doi.org/10.37943/LYFW8581
8. Nguyen B.T., Wakabayashi Y., Iwai K., Nishiura T., Analysis of derivative of instantaneous frequency and its application to voice activity detection // Applied Acoustics. – 2021. – Pp. 108-116, https://doi.org/10.1016/j.apacoust. 2021.108116
9. Makowski R., Hossa R., Voice activity detection with quasi-quadrature filters and GMM decomposition for speech and noise// Applied Acoustics. – 2020. – Pp. 107-114, https://doi.org/10.1016/j.apacoust.2020.107344
10. Bahdanau D., Chorowski J., Serdyuk D., End-to-end attention-based large vocabulary speech recognition // IEEE International Conference on Acoustics, https://doi.org/10.1109/ICASSP. 2016.7472618
11. Zazo R., Sainath TN., Simko G., Feature learning with raw-waveform CLDNNs for voice activity detection // Interspeech. – 2016. – PP.3668-3672, https://doi.org/doi:10.21437/Interspeech.2016-268.
12. Fan Z., Bai Z., Zhang X., Rahardja S., Chen J., AUC optimization for deep learning based voice activity detection// IEEE International Conference on Acoustics. – 2019. – Pp. 6760-6764, https://doi.org/doi:10.1109/ICASSP.2019.8682803
13. Hebbar R., Somandepalli K., Narayanan S., Robust speech activity detection in movie audio: data resources and experimental evaluation// IEEE International Conference on Acoustics. – 2019. – Pp. 4105-4109, https://doi.org/doi:10.1109/ICASSP.2019.8682532
14. Kim J., Hahn M., Voice activity detection using an adaptive context attention model // IEEE Signal Processing Letters. – 2018. – Pp. 1181-1185, https://doi.org/doi:10.1109/LSP.2018.2811740
15. Wilkinson N., Niesler T., A Hybrid CNN-BiLSTM Voice Activity Detector// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2021. – Pp. 6803-6807, https://doi.org/10.1109/ICASSP39728.2021.9415081
16. Tong S., Gu H., Yu K., A comparative study of robustness of deep learning approaches for VAD// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2016. – Pp. 5695-5699, https://doi.org/10.1109/ICASSP.2016.7472768

17. Mihalache S., Ivanov I.-A., Burileanu D., Deep neural networks for voice activity detection// 44th international conference on Telecommunications and Signal Processing (TSP). – 2021. – Pp. 191-194, https://doi.org/10.1109/TSP52935.2021.9522670
18. Abdullah S., Zamani M., Demosthenous A., A discrete wavelet transform-based voice activity detection and noise classification with sub-band selection //IEEE International Symposium on Circuits and Systems (ISCAS). – 2021. – Pp. 1-5, https://doi.org/10.1109/ISCAS51556.2021.9401647
19. Lee Y., Min J., Han D.K., Spectro-temporal attention-based voice activity detection// IEEE Signal Processing Letters. – 2020. – Pp. 131-135, https://doi.org/doi:10.1109/LSP.2019.2959917
20. Van Segbroeck M., Tsiartas A., Narayanan S.S., A robust frontend for VAD: exploiting contextual, discriminative and spectral cues of human voice// Proceedings of the annual conference of the International Speech Communication Association. – 2013. – P. 704-708, https://doi.org/10.21437/Interspeech.2013-198