DEVELOPING A SYSTEM FOR ASSESSING THE CORRECT PRONUNCIATION OF WORDS IN UZBEK LANGUAGE

A. Kakhkharov[1], M. Abdullayeva[1], U. Khasanov[2], N. Tolibova[3], N. Elbobo kizi[4]
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, associate professor[1]
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, assistant of professor[2]
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, master's degree student[3]
Shakhrisabz Branch of Tashkent Institute of Chemical Technology, assistant of professor[4]
https://doi.org/10.5281/zenodo.14930150

Abstract. *This work presents the stages and results of implementing a system for assessing the correct pronunciation of words in Uzbek language. Filtering, segmentation, spectral representation, feature extraction, and evaluation based on the DTW algorithm were performed on the recorded speech signal. During the testing of the results, the pronunciation of the same word was 77.50%, and word recognition was 100%. Based on the results, users had the opportunity to work on themselves and improve their pronunciation accuracy.*

Keywords: *speech signal, filtering, framing, segmentation, pronunciation, spectral, algorithm, feature extraction, DTW, analysis, histogram, recognition.*

Introduction. The general range of the speech signal is in the frequency range of 50 Hz - 1500 Hz, and the processing algorithms operate in the range of 50 Hz - 3400 kHz. The classification of speech sounds depends on the voice range and the gender of the person speaking. Despite the fact that there are two main measurement criteria, various, non-repeating speech sounds arise during the measurement process. Because of the classification of speech sounds and the generalization of the process of speech sound formation, it is possible to determine the criteria of the speech signal.

A speech signal is a signal formed because of complex vibrations. Secondly, the composition of one speech signal does not correspond to the composition of another speech signal. Thirdly, the formant frequency of a speech signal is in a different frequency range for individuals. Taking into account these criteria, speech signal processing is carried out based on complex algorithms.

Initial processing stage. In the initial processing of speech signals, the speech signal is converted from analog to digital representation. Analog-to-digital converters perform this task. In this case, discrete quantization and encoding are performed. In this case, the speech signal is often converted to a .wav file format, the sampling frequency is quantized in the range of 8 kHz - 19.2 kHz, and the recording channel is selected as mono. In real-time processing of speech signals, the speech signal is quantized at a sampling frequency of 22 kHz.

When choosing a speech signal, these parameters are considered standard. This is because the .wav file format of the speech signal is an uncompressed representation of the speech signal, and the sampling frequency of 22 kHz is the standard sampling frequency chosen by scientists

when processing speech signals. After the speech signal is represented in digital form, it is processed using the following initial processing algorithms [1].

Normalization. The digital representation of the speech signal is brought to a standard form in various intervals. Such normalization intervals are [0:1], [-1:1], [-3:3], [-128:128], [-256:256], [-512:512] for speech signals. These intervals ensure that the values of the speech signal fall into one standard interval. The interval [-1:1] was chosen when processing children's speech signals.

Filtering. Speech signal filtering is used to clean the speech signal of noise and eliminate external influences. In this case, finite impulse characteristic filters and infinite impulse characteristic filtering algorithms can be used [2]. After the filtering process, silence states in the speech signal are eliminated.

Segmentation. Segmentation is the process of dividing the values of the speech signal into frames of a specified size. Currently, in modern speech signal processing systems, speech signal frames contain 64, 128, 256, 512, 1024, 2048 speech signal coefficients. In speech signal processing, one of these segmentation intervals is selected depending on the problem. As a result, the speech signal is divided into frames [3].

Windowing. The windowing process of a segmented speech signal is to approximate the initial and final values of the detected frame to zero. The main reason for this is to prevent the spectral values from tending to infinity in parametric processing. There are types of windowing algorithms such as rectangular window, Hanna window, Hamming window, Blackman window, Kaiser Window. Hamming window is one of the most commonly used window types for speech signals by scientists [4].

Parametric processing stage. At the parametric processing stage, the acoustic, spectral, formant frequency, and cepstral parameters of the speech signal frames are determined [5]. In general, the selection of a particular parameter is selected depending on the problem. Each parameter has its own mathematical algorithm. In this case, the completeness of the parameters is checked.

Intellectual processing stage. The intellectual processing stage is an important stage. At this stage, the grouped set of parameters is processed based on decision-making algorithms. The group of algorithms of this stage includes dynamic programming algorithms, Markov chains, and neural network algorithms. The results obtained from the intellectual processing stage are summarized. That is, the results of the processed frames are analyzed and a conclusion is drawn about the state of the speech signal [6].

After dividing the speech signals into parameters, it is necessary to make a decision about the speech signal. Intellectual processing algorithms [7, 8] perform this task.

Speech signal processing methods. In world experience, the following intelligent algorithms are widely used in speech signal processing today:

DTW (dynamic time warping) algorithm;

Markov chains;

neural networks.

DTW (Dynamic Time Warping) algorithm is an optimal algorithm for finding a match between time sequences. This algorithm was originally used in speech recognition, that is, to compare two different pronunciations of the same speech signal. Later, it was also used in other fields. An algorithm calculates the difference between speech signals by measuring the distance

between two signals (also called Euclidean distance). The DTW algorithm is an algorithm that measures the distance between lines by ignoring local and global shifts between them.

Markov chain. The application of Markov chains in the process of digital processing of speech signals is used in the processing of languages of many countries. The application of Markov chains to the process of speech processing requires a very strong knowledge of the basics of graph theory. During the processing of speech based on Markov chains, it is necessary to set a certain boundary. The boundary can be the speech-processing domain or the speech processing process at a certain time.

The correct choice of boundary conditions allows you to build a model of a complete Markov chain. A Markov chain is used in the process of determining and evaluating the next state of a speech signal, having studied several states. Based on Markov chains, it is possible to create a model necessary for determining and evaluating the unknown state of speech after a known state in the process of digital processing of speech signals. The more states of a point are studied in Markov chains, the more complex the speech process is analyzed and its internal structure is studied [9].

Artificial neural network. Human biology plays a large role in the development of artificial neural networks. Researchers are described by organizing mental activity using terms appropriate to the configuration and algorithm of the existing network. It is based on knowledge of the functioning of the human brain.

Each synapse is characterized by its synaptic connection or its weight. From a physical point of view, this is equivalent to the electrical conductivity of a transducer. The current state of the neuron is determined by the approximate sum of the values at its input.

A number of Uzbek scientists have also conducted research on the processing of Uzbek speech signals. The practical results of these studies were analyzed in the processing algorithms.

Kamoliddin Shukurov (and others)’s scientific work is about adaptive methods for filtering noise affecting speech commands in real operating conditions. The method of adaptive NLMS filters is proposed for speech command recognition. The recognition accuracy of speech command has increased due to the use of the NLMS method. Digital signal filters are used to eliminate interference in speech recognition systems. However, the effectiveness of the use of classical filtering algorithms for speech signals with a complex appearance is very low. In such cases, the use of adaptive filters can meet the system requirement. Flexible filters use Least Mean Square (LMS), Normalized LMS (NLMS) algorithms to eliminate noise. These algorithms differ from each other in Global Signal-to-noise ratio (GSNR). If the GSNR is good, this algorithm is more useful for the system. Using these algorithms, the accuracy of the speech recognition system was found. Thus, they conclude that adaptive filtering algorithms are more efficient than classical filtering algorithms [10].

In voice control and speech recognition systems, it is important to eliminate external noise and parasitic signals. In this case, an adaptive filtering approach based on the NLMS (Normalized Least Mean Square) algorithm is used. At the beginning of the process, a 5-second noisy signal is recorded through the microphone, and filter coefficients are determined based on this signal. Then, speech commands coming through the microphone are filtered using these coefficients. Due to the variability of the amplitude and frequency of external noise and parasitic signals in different environments and conditions, the filter coefficients require constant adaptation. As a result, filter

adaptation is required before each system start-up, which can affect the usability of the system; therefore, adaptive algorithms in real time can increase the efficiency of the system [10, 11].

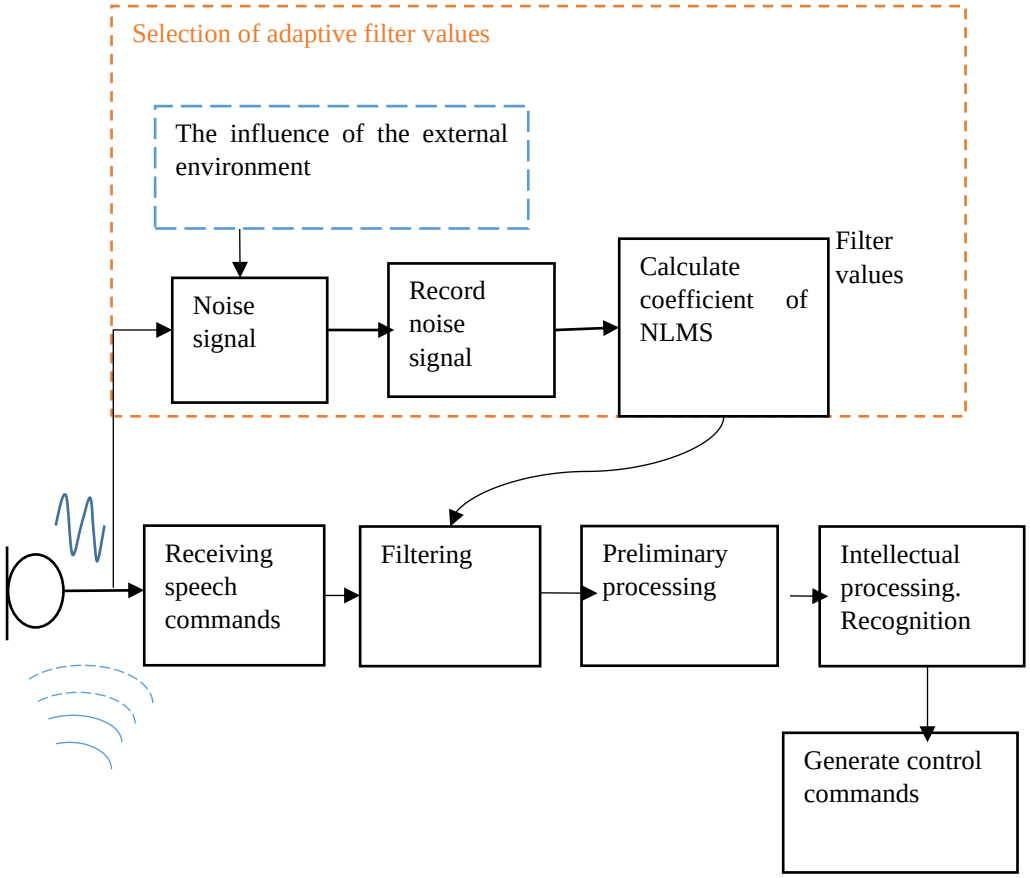

Figure 1. Algorithm for performing adaptive filtering of speech signals in voice control systems

In this study, methods for adaptive filtering of speech signals are proposed, showing that adaptive filter algorithms meet the requirements of speech recognition tasks in various technical environments. The accuracy of the system was evaluated using the proposed method, and speech signals filtered through adaptive filters achieved a recognition accuracy of 94-96%. This result demonstrates the effectiveness of adaptive filtering in improving the performance of speech recognition systems in noisy conditions [10].

Speech signals are a complex dynamic process, and their processing requires the recognition of random characters. Intelligent algorithms for processing these signals require large computational resources and power. The open source CMU Sphinx environment is an effective tool for implementing speech recognition systems and algorithms in various languages, and a special language model has been created for Uzbek speech commands. Developed in the Python programming language, this program runs on platforms such as Windows, Linux, and Raspberry Pi3 and was able to recognize isolated Uzbek speech commands with 80-95% accuracy [12, 13].

Result and discussion. My master's degree research project involves assessing the correct pronunciation of words in Uzbek language. There are many words in Uzbek language and they are considered unique. It was decided to implement the research project based on a limited vocabulary. The words «boshla», «oldinga», «orqaga», «ortga», «pastga», «tepaga», «o’ngga», «chapga» va «yakunla».

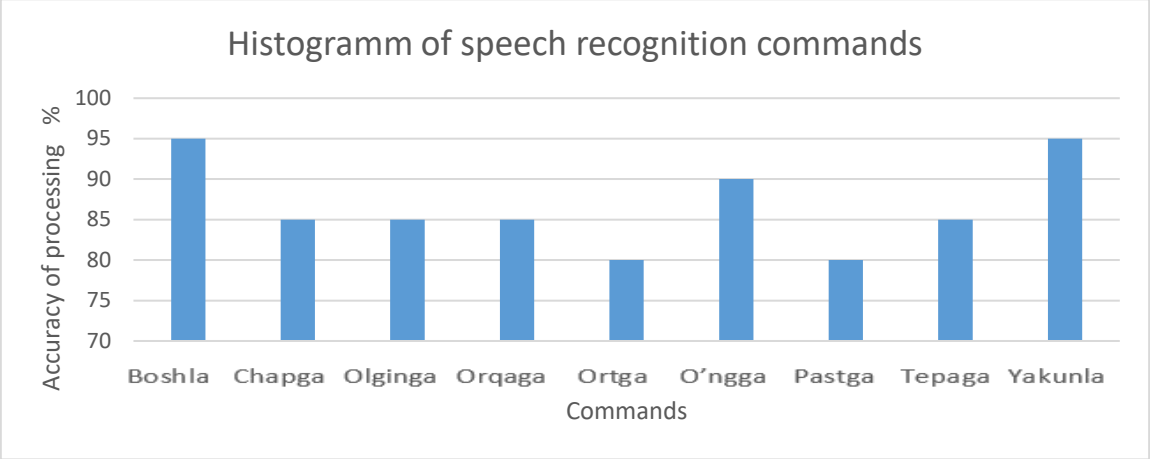

Figure 2. Accuracy of processing Uzbek speech commands.

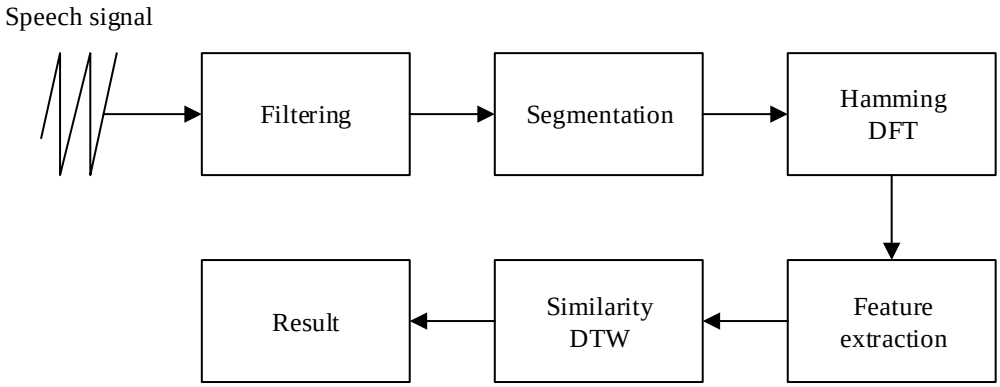

Figure 3. Stages of implementing a scientific work

The main stages are as follows;
recording the speech signal;
filtering the speech signal;
segmenting the speech signal;
transferring the speech signal to the spectral domain;
extracting the features of the speech signal;
calculating the pronunciation and word similarity using the DTW algorithm;
two different values are output in the output section; the first is the pronunciation, that is, the result of pronouncing the word in Uzbek, and the second is the result of pronouncing the word correctly.

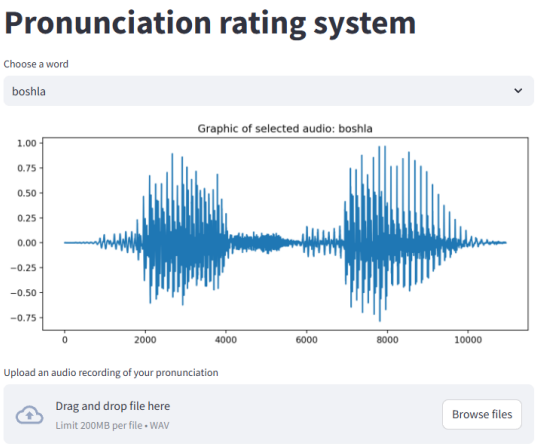

Figure 4. System home page.

In the section "Choose a word", the users select the word they want to pronounce, and the pronounced word is uploaded to the "Browse files" section.

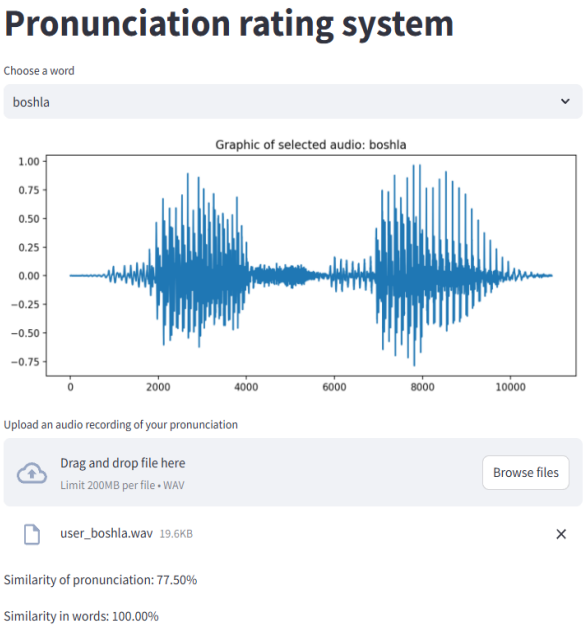

Figure 5. Results obtained.

When analyzing the results, the user's pronunciation is 77.50%, and the word accuracy is 100%. Word accuracy is analyzed by dividing the file sent by this user into Uzbek sounds, and each sound is compared against a "dataset". If an audio file of another word is sent to this system, selecting the starting word, the following result is obtained:

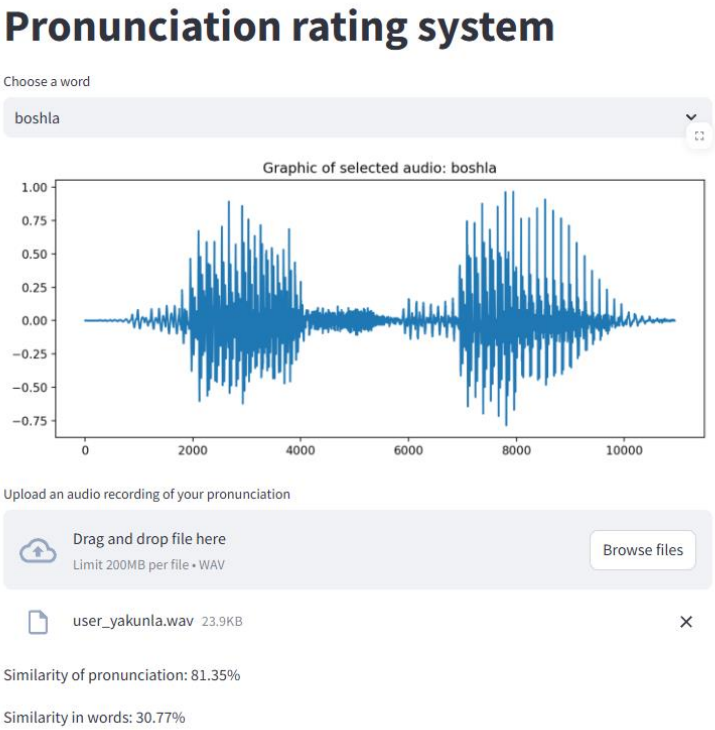

Figure 6. Analysis of results.

The system selected the word "boshla" and sent an audio file with the pronunciation of the word "yakunla". The system analyzed the file and found that the pronunciation was 81.35%, which is the pronunciation level in Uzbek language. The word similarity was 30.77%.

Conclusion. This work has developed a system for assessing the pronunciation of words in Uzbek language based on speech signal processing methods. Speech signal filtering and segmentation algorithms have been developed, which allow removing silences and cleaning the signal from noise caused by external influences and increasing the accuracy of processing. A system has been developed that processes sounds, syllables and words in Uzbek language with high accuracy.

Acknowledgement. In the future, it is planned to expand this work. The system is intended to develop a system based on data that includes not only a limited vocabulary, but also an expanded and general vocabulary.

REFERENCES

1. Ле Н. В. Предварительная обработка речевых сигналов для системы распознавания речи / Н. В. Ле, Д. П. Панченко // Молодой ученый. — 2011. — №5. Т.1. — С. 74-76.
2. Ермоленко Т.В., Федоров Е.Е. Методы подавления шумов в речевом сигнале /Речевые технологии, №3, Москва, 2009, с.3-13. Баронин С.П.
3. Vulic I., Smet W., Moens M.-F. Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora // Information Retrieval. — 2012. — Pp. 1–38.
4. Автокорреляционный метод выделения основного тона речи. Речевые технологии, № 2, Москва, 2008, с.3-12.
5. Шелепов В.Ю. Ниценко, А.В. Жук, Д.С. Азаренко. О распознавании фонем с помощью анализа речевого сигнала в частотной и временной областях. // Речевые технологии. – 2008.– №2. – С. 43-52.
6. Айфичер Э., Джервис Б. Цифровая обработка сигналов. Практический подход. 2-е издание. Вильямс, 2004. — 992 с.
7. Бекмуратов Т.Ф., Ботиров Ф.Б., Набиев М.М. Машинали ўқитиш ва чукур ўқитишнинг вазифалари ва киберхавфсизлик, “Муҳаммад ал- Хоразмий авлодлари” илмий амалий ва ахборот таҳлилий журнали, 3(9)/2019, с. 3-7. 33.
8. М.М. Мусаев. “Параллельные вычисления в цифровой обработке сигналов” // Вестник ТУИТ, Ташкент, 2013, №.3, стр.5-10.
9. Сорокин В.Н., Цыплихин А.И. Верификация диктора по спектрально-временным параметрам речевого сигнала.// Информационные процессы.2010, т.10, № 2, с.87-104.
10. K. Shukurov, U. Berdanov, U. Khasanov, S. Kholdorov and B. Turaev, «The role of adaptive filters in the recognition of speech commands,» 2021 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, 2021, pp. 1-4, doi: 10.1109/ICISCT52966.2021.9670084.
11. Xasanov Umidjon Komiljon o’g’li, Xoldorov Shohruhmirzo Imomali o’g’li, To’rayev Boburxon Shuxrat o’g’li, «NUTQ SIGNALLARI ORQALI SO’ZLOVCHINI TANIB OLISH», апр. 2023, doi: 10.5281/zenodo.7856104.
12. S. Kamoliddin Elbobo ugli, K. Shokhrukhmirzo Imomali ugli and K. Umidjon Komiljon ugli, «Uzbek speech commands recognition and implementation based on HMM,» 2020 IEEE 14th

International Conference on Application of Information and Communication Technologies (AICT), Tashkent, Uzbekistan, 2020, pp. 1-6, doi: 10.1109/AICT50176.2020.9368591.

13. Qaxxarov A’loxon Abrorovich, Xoldorov Shohruhmirzo Imomali o‘g‘li, Xasanov Umidjon Komiljon o‘g‘li. «SO‘ZLOVCHINI TANIB OLISHDA TANIB OLISH USULLARI TAHLILI». // EJTI. 2024. №6. URL: https://cyberleninka.ru/article/n/so-zlovchini-tanib-olishda-tanib-olish-usullari-tahlili.