Automatic Speech Recognition System with Dynamic Time Warping and Mel-Frequency Cepstral Coefficients

Kateryna Yalova, Mykhailo Babenko and Kseniia Yashyna

Dniprovsky State Technical University, Dniprobydivska str.2, Kamyanske, 51918, Ukraine

Abstract

The approach to speech recognition presented in this paper is used to create a system for automatic recognition of user commands for a graphical editor. The automatic speech recognition system is used as a recognition module in a plug-in for a graphics editor. The proposed automatic speech recognition system has a limited dictionary size, is speaker-dependent, and is used to recognize separate speech given by the user in the form of short speech commands. The list of commands contains 20 commands in Ukrainian language, the name of which corresponds to the name of the pictograms in the graphical editor. The user's voice command is used as an input, which is processed, recognized and presented as a command for the graphical editor. The user must use a microphone to transmit a voice command. The system allows processing user commands in real time. Isolated command words are used as commands. The stages of voice command recognition are next: analysis of an analog signal, its transformation into a digital signal, formation of a filter bank, comparison of the processed command with a template. To analyze the sound wave, it is proposed to use the Fourier transform. The Hamming function is used to reduce spectrum blurring. For feature extraction from voice commands, it is proposed to use the Mel-frequency cepstral coefficients algorithm. Matching voice commands with a template is carried out using the Dynamic time warping method. The use of the Mel-frequency cepstral coefficients and Dynamic time warping algorithm is justified by the fact that the vocabulary is limited and the commands are short. The accuracy of command recognition was evaluated for various speakers. The average recognition accuracy is 93%.

Keywords

Automated speech recognition, dynamic time warping (DTW), (MFCC).

1. Introduction

Speaking is the most natural form of human communication, and therefore the implementation of an interface based on the analysis of speech information is a promising direction for the development of intelligent control systems. One of the current unsolved problems in information and measurement systems is the construction of systems for automatic recognition of speech signals that are invariant to the speaker [1]. Its solution would make it possible to expand the range of users of such systems and significantly increase the efficiency of information exchange in man-machine systems. The task of language analysis includes a wide range of tasks. Traditionally, they are divided into three subclasses: identification, classification, and diagnostic tasks. Identification tasks include the tasks of verification and identification of announcers. The tasks of classification include the task of recognizing key words, recognizing fused speech, and the task of semantic language analysis. Diagnostic tasks include the task of determining the psychophysical state of the announcer.

The creation of language interfaces can be used in systems of various purposes: voice control for people with disabilities, answering machines, automatic call processing, implementation of "smart

COLINS-2023: 7th International Conference on Computational Linguistics and Intelligent Systems, April 20–21, 2023, Lviv, Ukraine
EMAIL: yalovakateryna@gmail.com (K. Yalova); mvbab130973@gmail.com (M. Babenko); yashinaksenia85@gmail.com (K. Yashyna)
ORCID: 0000-0002-2687-5863 (K. Yalova); 0000-0003-1013-9383 (M. Babenko); 0000-0002-8817-8609 (K. Yashyna)

home" commands. However, despite the rapidly increasing computing power, the creation of speech recognition systems remains an extremely difficult problem. This is due to both its interdisciplinary nature (it is necessary to have knowledge of linguistics, digital signal processing, acoustics, pattern recognition, etc.), and the high computational complexity of the developed algorithms [2]. The latter imposes significant limitations on automatic speech recognition systems – on the volume of the processed dictionary, the speed of receiving an answer, and its accuracy. The task of optimizing the quality and level of recognition in automatic speech recognition systems is a relevant scientific and practical task, the solution of which allows improving the quality of the developed SILK (speech, image, language, knowledge) interfaces of software systems and applications.

The purpose of the article is to present the results of applying the Dynamic warping method (DTW), fast Fourier transform (FFT), Mel-frequency cepstral coefficients (MFCC), to solving the problem of developing a system for automatic speech signal recognition. The authors implemented the following tasks in the course of achieving the goal:

- the peculiarities of the application of the DTW algorithm for speech signal recognition are analyzed. A fast Fourier transform was applied to the analysis of the input signal and MFCC were used to construct the input vector of features;
- a model of an automatic speech recognition system with characteristics has been designed: the developed system is command-based, dependent on the speaker with the type of structural unit – the speaker's command phrase, which sets a command for working with the text within the text editor;
- a speech recognition software module for text editors has been developed in the form of a plug-in, which allows a user to evaluate the quality of the proposed solutions and establish the recognition error.

1.1. Related works

During the recognition of the incoming speech signal, various methods are used: hidden Markov model (HMM), Decision Trees, Linear Predictive Codes, DTW, methods based on the use of neural networks of various architectures.

The earliest attempts to create systems for automatic speech recognition began in the 1950s, when the first speaker-dependent system that recognized numbers was developed [3], and resonances of vowel sounds in words were used as characteristics of the input signal. In the 70s, the DTW and the method of linear predictive coding (Linear Predictive Coding – LPC) were discovered. Despite the rapid development of the neural network approach to solving the problem of automatic speech recognition, DTW remains a popular method.

Such scientists as K. Chakraborty, A. Talele, S.R. Suralkar, A. C. Wani, M. Mahajan, A. Katuri, A. G. Siva and many others devoted their works to the use of DTW in the task of automatic speech recognition. Methods for using DTW and MFCC in an automatic speech recognition system for certain native languages were presented in their works by M.S. Nguyen, L. Muda Awad, H. Omar, Y. Farghaly. In the works of X. Sun, Y. Miyanaga, B. Sai, it is proposed to use multireferences DTW to reduce the calculation cost. Scientists S. Joshi, S. Nagar, A. Ismail, S. Abdlerazek proposed to apply the DTW algorithm within language-dependent and speaker-dependent speech signal recognition systems. In scientific papers A. S. Haq, C. Setianingsih, R. Martinek, J. Vanus, J. Nedoma, M. Fridrich, J. Frnda, T. Desot the effectiveness of using DTW and MFCC in the task of creating automatic speech recognition systems for home automation and creating systems for interacting with smart home devices is substantiated. In the works of E. Principi, V. Prasad, M. A. Anusuya, K. Sharma, S.D. Dhingra, B.J. Mohan carried out the development of automatic speech recognition systems that transform an acoustic signal into text. The versatility of DTW and MFCC application areas testifies to the effectiveness of their use. The speech recognition problem solution has a number of different applications from voice control of digital devices, to determining the owner of the voice and even recognizing the species of birds by their sounds.

Despite the significant number of scientific works devoted to the problem of speech signal recognition, the tasks of improving the quality of speech recognition and developing new approaches to the implementation of automatic speech signal recognition systems remain relevant scientific and practical tasks [4].

2. Proposed methodology

The input data for the proposed system are the user's voice commands entered through the microphone. Since the system must process voice signals in real time, it is proposed to use FFT to analyze the speech signal. To change the size of the spectrum, removed after the FFT stop, it is recommended to use the Hamming window.

In this paper, it is proposed to use MFCCs, a method that allows extracting features that an acoustic signal received from a speaker has. It was introduced by Davies and Mermelstein in the 1980s and has been relevant ever since, supplanting linear prediction coefficients (LPCs) and linear predictive cepstral coefficients (LPCCs), which were previously the main features for automatic speech recognition, especially with HMMs classifiers. The disadvantage of this algorithm is a significant dependence on the correctness of the process of converting an analog signal into a digital one. Also, the recognition process will be negatively affected by extraneous noise and speech defects of the speaker. After building the filter bank, you can proceed to compare the processed commands with templates.

DTW method is an automatic speech recognition method based on pattern matching. DTW allows to find the difference between two 2-time series of voice commands that have different durations. Although the accuracy of speech recognition using DTW is lower than that of methods using neural networks, it is still popular and is used for speech recognition systems with a limited vocabulary. The expediency of using DTW when developing a plug-in for a graphic editor as an effective method for recognizing speech commands is justified by the characteristics of the automatic speech recognition system, namely: the limited size of the dictionary, recognition of separate speech given by the user in the form of short speech commands.

The adequacy of the proposed methods was evaluated by determining the level of recognition accuracy for different speakers and voice commands.

3. Results

Speech recognition is the process of transforming a speech signal into digital information [5]. The automatic voice recognition system is an information system that converts an incoming speech signal into a recognized message. At the same time, the message can be presented both in the form of the text of this message, and immediately transformed into a form convenient for its further processing [6].

Most speech recognition systems (Automatic Speech Recognition – ASR) consist of an analog signal analysis and processing process and a recognition process. When analyzing an analog signal from speech, properties are extracted that are used further in the recognition process to determine what was said. Automatic voice recognition systems are classified according to the following characteristics: dictionary size (limited set of words or a large dictionary), dependence on the announcer (announcer-dependent or announcer-independent), language type (fused, separate), purpose (dictation systems, command systems), recognition algorithm that is used, the type of structural unit (phrases, words, phonemes, etc.), principles of selection of structural units [7].

The general scheme of the speech signal recognition process is next: receiving the acoustic signal that comes from the user's microphone, digitizing the sound signal, obtaining the characteristics of the signal and comparing the feature vector of the input signal with templates.

3.1. Speech signal analysis

To analyze a sound wave, let’s use Fourier’s theorem, which states that any complex periodic oscillation can be decomposed into the sum of simple harmonic oscillations. As a result, we will get a set of amplitudes, phases and frequencies for each sinusoidal wave component:

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

where N – the number of signal values, K – the number of frequencies, x_n – the value of the signal at certain points in time, X_k – complex amplitudes of sinusoidal signals that make up the initial signal, k=0,…K-1 – frequency index, n=0,…,N-1 – discrete time points at which the signal was measured.

The frequency or phase point taken together with the amplitude is called the spectrum. To restore a discrete signal from the spectrum, we use the inverse Fourier transform:

$$x_n = \frac{1}{K}\sum_{k=0}^{K-1} X_k e^{\frac{2\pi}{K}kn}. \quad (2)$$

To monitor changes in the spectrum of a signal over time, you can use a spectrogram - a visualization of changes in the spectrum over the entire sound segment. For its construction, a windowed Fourier transform is used – the spectrum is calculated from successive windows of the signal, each of which overlaps a part of the previous window. To significantly speed up the spectrogram construction process, the FFT algorithm was used, which works with complex numbers and transformation sizes representing powers of two [8]. If the frequency of the tone coincides with one of the frequencies of the FFT grid, then the spectrum will look “perfect”: a single sharp peak will indicate the frequency and amplitude of the tone. If the frequency of the tone does not coincide with any of the frequencies of the grid, then the FFT will “collect” the tone from the frequencies available in the grid, combined with different weights. It is worth taking into account that the FFT decomposes the signal not according to the frequencies that are actually present in the signal, but according to a fixed uniform frequency grid. Such blurring is usually undesirable, as it can mask weaker sounds at nearby frequencies.

To reduce the effect of spectrum blurring, the signal before FFT calculation is multiplied by weight windows – functions falling to the edges of the interval [8]. They reduce the blurring of the spectrum due to some deterioration of the frequency resolution. In this work, it is proposed to use the Hamming window to reduce the blurring of the spectrum obtained after applying the FFT. The formula by which the Hamming function can be determined can be presented in the form [9]:

$$w(n) = 0.54 - 0.46 * cos\left(\frac{2\pi n}{N-1}\right), \quad (3)$$

where n – total amount of samples in a single frame.

Applying a Hamming window reduces the level of spectral blurring by about 40 dB relative to the main peak. The input signal was divided into intervals of 20-40 ms, since the size of such an interval is sufficient to obtain a reliable spectral estimate. To compensate for peak broadening when applying weight windows, longer FFT windows can be used: for example, 8192 counts instead of 4096.

3.2. Building a bank of filters

The sound signal is constantly changing, so for simplicity, let’s assume that the sound signal hardly changes over a short period of time. That is why the paper suggests dividing the signal into intervals of 20-40 ms. If the frame is much shorter, we don’t have enough samples to get a reliable spectral estimate, if longer, the signal changes too much over the entire frame. Spectral estimation determines which frequencies are present in the frame.

Spectral estimation still contains a lot of information that is not needed for automatic speech recognition. In particular, it is not possible to distinguish between two closely spaced frequencies. This effect becomes more pronounced with increasing frequency. To construct a vector of characteristics, it is advisable to use the MFCC algorithm, which will allow dividing the spectrum into sections that will be represented by frequency projections in the corresponding range on the Mel scale [10]:

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

where f – the frequency that is projected onto the Mel scale.

The obtained values need to be converted back to the frequency form:

$$h(m) = 700 * \left(exp\left(\frac{m}{1127}\right) - 1\right), \quad (5)$$

where m – frequency projection on the Mel scale

$$f(i) = Floor\left((n+1) * \frac{h(i)}{R}\right), \tag{6}$$

where n –FFT window size; R – signal sampling rate.

The equation is used to form the filter bank:

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)}, & f(m-1) \le k \le f(m)) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)}, & f(m) \le k \le f(m+1) \\ 0, & k > f(m+1) \end{cases}, \tag{7}$$

where m – amount of MFCC; k – current frequency.

The Mel filter bank contains triangular-shaped overlapping filters [11] as shown in Figure 1.

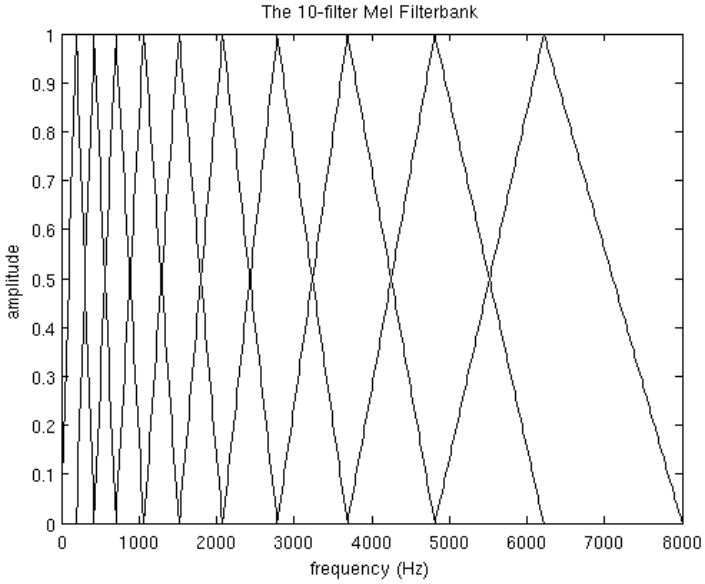

Figure 1: View of the filter bank

After calculating the energy of the filter bank, it is necessary to calculate their logarithm, since humans do not hear loudness on a linear scale. This operation makes Mel coefficients more similar to human perception of sound. The last step is to calculate the Discrete Cosine Transform (DCT) of the energy of the logarithms of the filter bank:

$$X_k = \sum_{n=0}^{N-1} x_n * cos\left[\frac{\pi}{N}\left(n + \frac{1}{2}\right)k\right]. \tag{8}$$

In the process of language recognition, the most difficult thing is to carry out the procedure of comparing two language elements, which are also characterized by a length in time, therefore there are quite a lot of such procedures and methods.

3.3. Dynamic time warping application

DTW allows to find the minimum distance between two sequences or time series depending on certain values, such as time scales, which is effectively used in automatic speech recognition systems [12].

Let's assume there are two numerical sequences P=(p_1, p_2, ..., p_n) and Q=(q_1, q_2, ..., q_m), schematically the alignment of the sequences in time can be represented as shown in Figure 2.

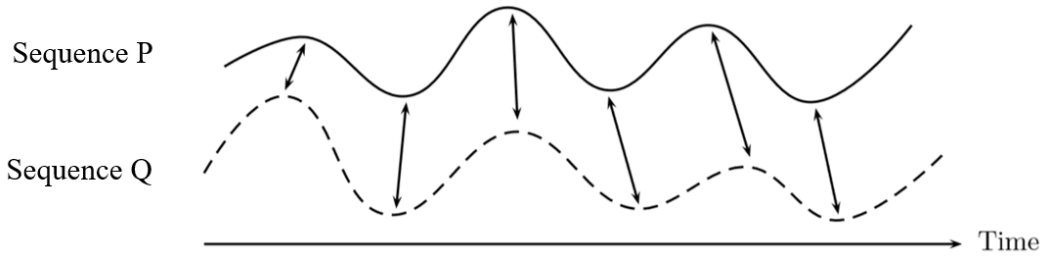

Figure 2: Scheme of time alignment of two sequences

To calculate local deviations between elements of two sequences, you can calculate the absolute deviation of the values of two elements (Euclidean distance) [13,14]. As a result, the matrix of deviations d will be obtained:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}, \tag{9}$$

Next, it is necessary to calculate the matrix of minimum distances between the sequences. Its elements are according to the following formula:

$$md_{ij} = d_{ij} + min(md_{i-1,j-1}, md_{i-1,j}, md_{i,j-1}), \tag{10}$$

where md_{ij} – minimum distance between the i-th and the j-th elements of sequences P and Q.

After that, we find the minimum path in the obtained matrix, which can be followed from the element md_{nm} to md_{00} by following the next rules [15-16]:

1. The path is laid only forward – indices i and j are never increased.
2. Indices are only decremented by one per iteration.
3. The path starts in the lower right corner and ends in the upper left corner of the matrix.

Based on the obtained path, we estimate the global deformation:

$$GC = \frac{1}{k}\sum_{i=1}^{k} w_i, \tag{11}$$

where w_i – elements of the minimum deformation path; k – the number of elements of the deformation path.

4. Implementation

To evaluate the quality of the proposed solutions regarding the speech signal recognition process, a command-based, announcer-dependent automatic speech signal recognition system was designed that converts the announcer's input speech signal into a text formatting command within the text editor. C# language is selected as the programming language. For software implementation, an object-oriented approach to the analysis of the data domain and the construction of the architecture of the internal classes of the system was used. Testing of the program code was carried out in manual mode.

The developed system is equipped with functions: formation of a dictionary of announcer commands, training of the system for a specific announcer, and execution of recognized commands. The automatic speech recognition system is used as a software module for the implementation of a graphic editor plug-in, the use of which allows you to use the functionality of the graphic editor through voice control commands. The developed voice recognition system has a limited command dictionary size, aimed at recognizing 20 isolated command words. The developed application has the following functional requirements:

- the possibility of training the system for a specific announcer;
- recognition of received voice commands;
- conversion of voice commands to commands of a graphic editor

The scheme of application of DTW and MFCC, used during the development of the system of automatic recognition of voice commands, is presented in Figure 3.

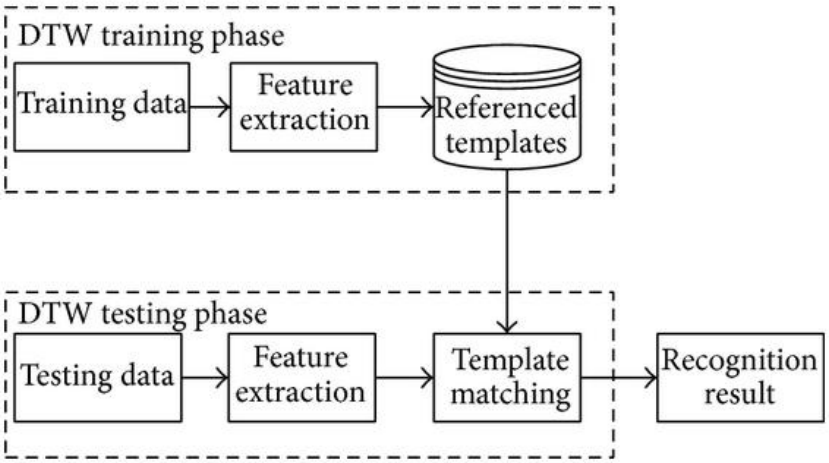

Figure 3: Application scheme of DTW and MFCC

The use of an object-oriented approach made it possible to present the program code in the form of a set of classes available for repeated use. The architecture of the developed application is presented in the form of a class diagram in Figure 4.

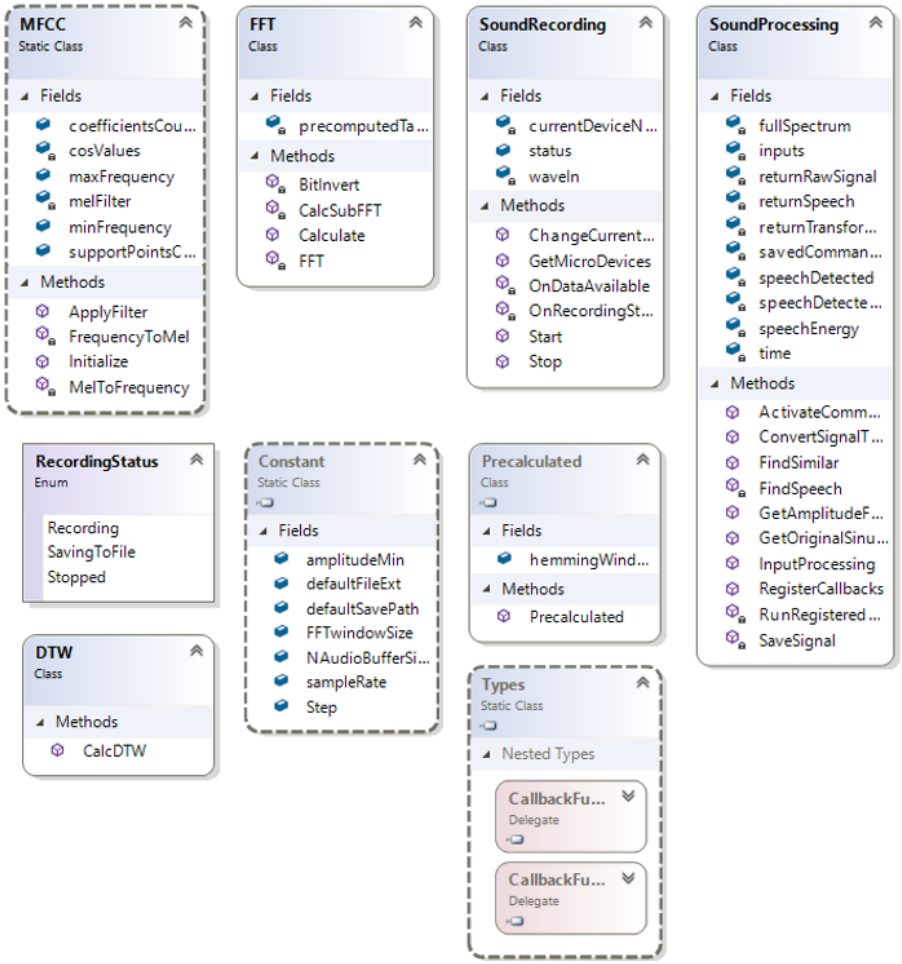

Figure 4: Architecture of software application classes

The presented class diagram was generated automatically within the software environment VisualStudio.net during the software implementation. The architecture of the class does not contain inheritance relations. To increase the software flexibility, the composition relations were implemented for interaction between classes. A composition relationship is a relationship in which a data consumer class has a data provider class type field. This type of relation is not displayed in the class diagram generated in the VisualStudio.net.

The SoundRecording class is responsible for receiving the digitized sound signal that comes from the user's microphone. The OnDataAvailable method is called every 100ms, and passes the digitized signal to the InputProcessing method of the SoundProcessing class. This class manages all stages of recognition – obtaining the spectrum of the signal window by Fourier transformation (FFT class), calculating low-frequency cepstral coefficients (MFCC class), comparing the input signal feature vector with templates (DTW class). Auxiliary classes are also used to store constant values (Class Constant), common to the project, and a class to store sequentially calculated values (Class Precalculated), which significantly speeds up calculations, thanks to the calculation of values only once – when the program is started.

The result of the developed program code execution is a new item displaying – the Addins menu item in the main toolbar of the graphic editor. After clicking the Addins button, the Start button is displayed, which activates the command recognition mode and Preferences, which starts the application settings mode. This part of the software is implemented as a desktop application within Windows Application Form in C# programming language. The functionality of the developed plugin is divided into two parts:

- system training with the ability to create and save a command template
- command execution mode.

By pressing the Start button and starting the voice command recognition mode, you can speak commands and evaluate the result of their execution. The recognition mode stops functioning when the Stop button is pressed.

In the program settings mode, you can change the voice signal transmission device, add the voice equivalent for the command to the dictionary, analyze the sound wave and its spectrum. Figure 5 shows the application settings window.

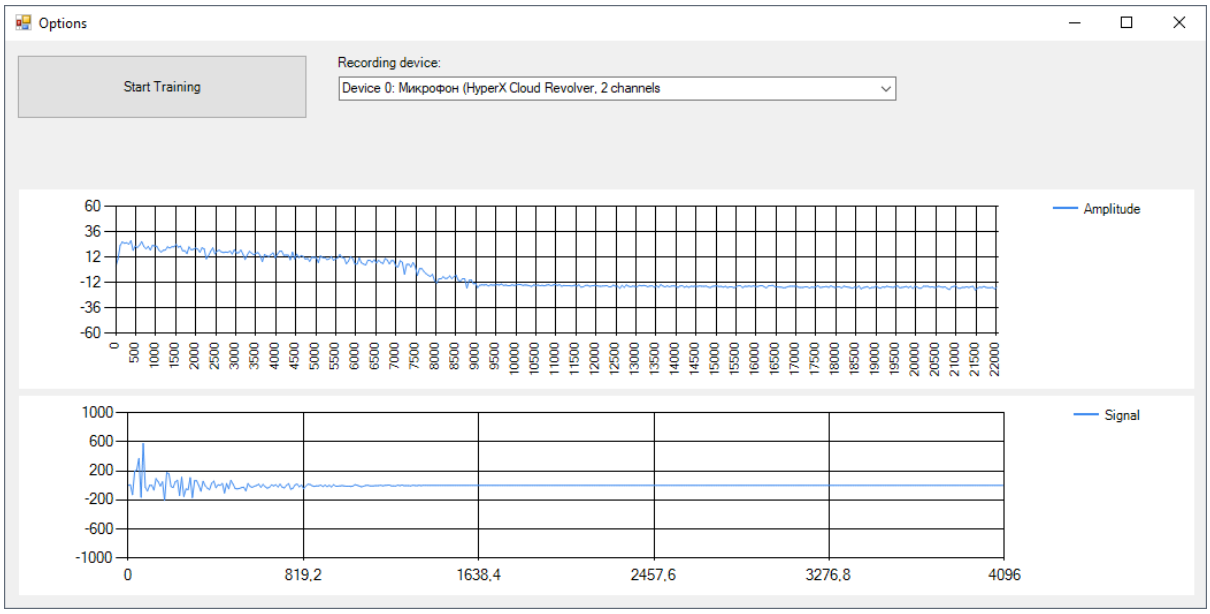

Figure 5: Screen form of software application settings

When the Start Training button is pressed, the user needs to say the command, after which the training results window will open, where you can see which command the spoken signal is similar to. If necessary, it can be added to the command dictionary with the "Save as template" button.

The software application was tested in the following mode: 20 test commands, 10 announcers, 10 requests to pronounce each command, the sampling frequency of speech signals is 44.1 kHz, the

resolution of the speech sample is 16 bits, the number of channels is 2. The templates of the test commands coincide with the name of the commands. presented on the Microsoft Word quick access panel in the form of icon buttons. In order to establish the adequacy of the proposed solutions and estimate the speech signal recognition error, a transaction log was developed, which received a description of each speaker command recognition operation. Analysis of the transaction log data allowed for the average value of recognition accuracy for each command. Recognition accuracy values can be determined as [8,17]:

$$Accuracy = \frac{corectly\ detected\ commands}{number\ of\ comands\ in\ data\ set}. \quad (12)$$

Table 1 shows the average results of recognition level estimation for the described conditions of the experiment.

Table 1
The results of the conducted experiment

Command	Accuracy Rate
Align left	95%
Align right	93%
Justify	98%
Center	96%
Bold	93%
Strikethrough	90%
Italic	97%
Bullets	92%
Multilevel list	91%
Superscript	91%
Numbering	93%
Cancel input	98%
Display all signs	94%
Repeat the input	96%
Subscript	89%
Underline	91%
Page break	95%
Save	99%
Increase font size	94%
Decrease font size	94%

Based on the data presented in table 1, it is possible to determine the average level of command recognition accuracy. The average recognition error does not exceed 7%. It should be noted that the main factor of erroneous recognition is the diction of the user and the similarity of voice commands [18], for example, such as: “Superscript”, “Subscript”. The obtained result of recognition and its comparison with the recognition accuracy of systems using DTW and MFCC allows to draw conclusions regarding the adequacy and effectiveness of the proposed design solutions.

5. Conclusions

Automatic speech recognition allows computers, machines and digital devices to understand natural speech and perform appropriate actions. The generalized speech recognition algorithm includes the steps of:

1. Receiving an analog signal.
2. Transformation of the received signal into a digital one.

3. Spectogram formation using FFT and Hamming algorithm.
4. Definition of a filter bank when using MFCC.
5. DTW application to compare the received command and template.

There are various techniques, methodologies and algorithms that are used in each stage of the speech recognition process. Combining different approaches allows to achieve the desired level of recognition accuracy. The paper presents the results of the development of an automatic speech recognition system using DTW and MFCC. DTW, FFT and MFCC are well known methods that are commonly used in the speech recognition, but even systems implemented using these methods have different levels of recognition accuracy. Recognition accuracy varies significantly, for example, from 46% [19] to 96,4% [20] or even 99.5% for the dictionary with 8 isolated words [21] and largely depends on the implementation of the recognition system, on the volume of the dictionary, input signal transmission method, the language that is used for recognition. The proposed system is speaker-dependent, has a limited dictionary and is designed to recognize 20 short user commands specified by the user in Ukrainian using a microphone. The developed system is used as a program module of a plug-in for a graphic editor. The use of such a plug-in allows you to control the functionality of the graphic editor through the user's voice commands. The language of the software implementation is C #. The software has been developed using an object-oriented approach within the Windows Application Form as a desktop application to set option and record the results of the experiments and as a plug-in for the graphical redactor.

A feature of the developed system of automatic speech recognition is that the system is able to work with the user in real time, unlike other developments, where it is necessary to transmit an audio recording of a certain format to the input. At the same time, the system response time (recognition time) does not exceed 2 seconds on average.

The DTW speech command recognition algorithm is an effective method for accounting for temporal variations when comparing related time series in the task of automatic speech recognition and can be effectively applied in systems with a limited dictionary, as presented in the paper. Using (12), the accuracy of command recognition was estimated, the average value of accuracy for all commands and recognition variants was approximately 93%. Comparing the obtained level of accuracy with the results of the other authors is not indicative, since systems use different dictionaries, with different commands or isolated words, use different natural languages. However, the obtained level of recognition indicates a high efficiency of the proposed solutions and also demonstrates the promise of DTW, FFT and MFCC approaches combining for speech recognition in Ukrainian language. The methods, principles and algorithms used in the research, as well as the results of developed system testing, make it possible to state that the results of the research are valid and reproducible.

6. References

[1] C. Agarwal, P. Chakraborty, S. Barai, V. Goyal, Quantitative analysis of feature extraction techniques for isolated word recognition, Advances in computing and data sciences (2019) 618–627. doi: 10.1007/978-981-13-9942-8_58.

[2] I. Sultana, N.K. Pannu, Automatic speech recognition system, International journal of advance research, ideas and innovations in technology 4 (2018) 277–279.

[3] D. Pandey, K. K. Singh. Implementation of DTW algorithm for voice recognition using VHDL, in: Proceedings of the International Conference on Inventive Systems and Control, ICISC '17, Coimbatore, 2017, pp. 1– 4. doi: 10.1109/ICISC.2017.8068638.

[4] M. Sood, S. Jain, Speech recognition employing MFCC and Dynamic time warping algorithm, Innovations in information and communication technologies (2021) 235–242. doi: 10.1007/978-3-030-66218-9_27.

[5] H. Pawar, N. Gaikwad, A. Kulkarni, A study of techniques and processes involved in speech recognition system, International journal of engineering and technology 7 (2020) 1905–1911.

[6] S. K. Ali, Z. M. Mahdi, Arabic voice system to help illiterate or blind for using computer, Journal of physics 1804 (2021) 1–11. doi:10.1088/1742-6596/1804/1/012137.

[7] L. Lerato, T. Niesler, Feature trajectory dynamic time warping for clustering of speech segments, Journal on Audio, Speech and Music 6 (2019) 1–9. doi: 10.1186/s13636-019-0149-9.
[8] H. F. C. Chuctaya, R. N. M, Mercado, J. J. G. Gaona, Isolated automatic speech recognition of quechua numbers using MFCC, DTW and KNN, International journal of advanced computer scince and application 9 (2018) 24–29. doi: 10.14569/IJACSA.2018.091003.
[9] S. Lokesh, M. R. Devi, Speech recognition system using enchanced mel grequency cepstral coefficient with windowing and framing method, Cluster Computing 22 (2019) 1669–1679. doi: 10.1007/s10586-017-1447-6.
[10] I. D. Jokić, S. S. Jokić, V.D. Delić, Z.H. Perić, One solution of extension of Mel-frequency cepstral coefficients feature vector for automatic speaker recognition, Information, technology and control 49 (2020) 224–236. doi: 10.5755/j01.itc.49.2.22258.
[11] A. Awad, H. Omar, Y. Ahmed, Y. Farghaly, Speech Recognition System Using MFCC and DTW 4 (2018).
[12] A. S. Haq, C. Setianingsih, M. Nasrun, M.A. Murti, Speech recognition emplemetation using MFCC and DTW algorithm for home automation Proceeding of the Electrical Engineering Computer Science and Informatics 7 (2020) 78-85. doi: 10.11591/eecsi.v7.2041.
[13] B. Kurniadhani, S. Hadiyoso, S. Aulia, R. Magdalena, FPGA-based implementation of speech recognition for robocar control using MFCC, Telekomnika 17(4) (2019) 1914–1922. doi: 10.12928/telkomnika.v17i4.12615.
[14] Y. Permanasari, E. H. Harahap, E. P. Ali, Speech recognition using Dynamic Time Warping (DTW), Journal of Physics 1366 (2019) 1–6 doi:10.1088/1742-6596/1366/1/012091.
[15] R. G. Kanke, R. M. Gaikwad, M. R. Baheti, Enhanced Marathi Speech Recognition Using Double Delta MFCC and DTW, International Journal of Digital Technologies 2(1) (2023) 49–58.
[16] I. Wibawa, I. Darmawan, Implementation of audio recognition using mel frequency cepstrum coefficient and dynamic time warping in wirama praharsini, Journal of Physics 1722(2021) 1–8. doi: 10.1088/1742-6596/1722/1/012014.
[17] B. Paul, R. Paul, S. Bera, S. Phadikar, Isolated Bangla spoken digit and word recognition using MFCC and DTW, Engineering mathematics and computing 1042 (2022) 235–246. doi: 10.1007/978-981-19-2300-5_16.
[18] R. Harshavardhini, P. Jahnavi, S.K. Zaiba Afrin, S. Harika, Y. Annapa, N. Naga Swathi, MFCC and DTW based speech recognition, International research journal of engineering and technology 5 (2018) 1937–1940.
[19] I. Khine, C. Su, Speech recognition system using MFCC and DTW, International Journal of Electrical, Electronics and Data Communication 6 (2018) 29–34.
[20] S. Riyaz, B. L. Bhavani, S. V. Kumar, Automatic Speaker Recognition System in Urdu using MFCC and HMM, International Journal of Recent Technology and Engineering 7 (2019) 109–113.
[21] N. Adnene, B. Sabri, B. Mohammed, Design and implementation of an automatic speech recognition based voice control system, EasyChair Preprint 5116 (2021) 1–7.