Voice Recognition Management System

Olga E. Baklanova and Adil B. Ryspayev
D. Serikbayev East Kazakhstan State Technical University, Ust-Kamenogorsk, the Republic of Kazakhstan
E-mail: o_e_baklanova@mail.ru, adilko.vip@mail.ru

Abstract – **The report discusses the algorithm of the recognition system for isolated words, the main stages of recognition, the mechanism of a digital system for processing an audio signal, cepstral coefficients as well as the implemented software project - a computer model of a system for recognizing words in Visual Studio using built-in functions.**

Keywords – **recognition, digital processing system, discrete Fourier transform, Kotelnikov theorem, Hamming window**

I. INTRODUCTION

The relevance of this topic for many years remains the most fruitful area of research, both in mathematics and in the field of programming. The task of speech recognition remains relevant today.

Since the advent of the first computers (electronic computer), one of the most important issues in the development of computer technology has been the process of human-machine interaction. For a long time it was available only to narrow specialists - technologists "communicated" with the machine through an intermediary - a programmer. This situation lasted until the appearance of the dialog interface, when the user was able to personally enter the command addressed to the machine from the keyboard and receive a meaningful answer. The appearance of a graphical interface, in which there was no need for a person to know any commands, led to the widespread dissemination of personal computers.

However, man has always sought a more universal and natural way of interacting with computers. The creation of natural means for a person to communicate with a computer is currently the most important task of modern science, while the voice input of information is carried out in the most convenient way for the user.

Speech recognition problems are studied by: Carnegie Mellon University (USA) [1], University of Illinois (USA), Oregon Institute of Science and Technology (USA), Computer Center of the Russian Academy of Sciences (Yu. I. Zhuravlev, V. Ya. Chuchupal), Institute for Information Transmission Problems of the Russian Academy of Sciences (V. N. Sorokin), Institute of Mathematics SB RAS and Novosibirsk State University (N. G. Zagoruyko and V. M. Velichko), Moscow State University named after M.V. Lomonosov (O. F. Krivnova), Moscow State Technical University N. E. Bauman (Yu. N. Zhigulevtsev), Moscow Power Engineering Institute (A. I. Evseev), Moscow State Linguistic University (R. K. Potapova), Moscow Technical University of Communication and Informatics (Yu. N. Prokhorov), St. Petersburg State University (V. I. Galunov), St. Petersburg Institute of Informatics and Automation of the Russian Academy of Sciences. [2]

Such companies as IBM, Philips, Dragon Systems, Cognitive Technologies, Istrasoft, Sacrament and others conduct research in this area, which indicates its relevance.

Analysis of existing speech recognition systems [3] (Dragon NaturallySpeaking, Sakrament ASR Engine, Voice-Mode, SIRIUS, etc.) showed that developers create either large specialized speaker-independent systems or complexes for developing speech applications that give high recognition accuracy with a small dictionary of commands; or user applications that enable voice control of the computer and voice input of text. This class of systems needs training for a specific speaker and, in this regard, does not provide high quality recognition.

In most systems, Fourier transform is used to create models of speech units, only a few support the Russian language, since additional studies of the linguistic component of the process of recognizing Russian-language speech are needed.[4]

The digital audio signal processing system involves representing the analog speech signal in digital form. As a result of analog-to-digital conversion (ADC), a continuous signal is translated into a series of discrete time samples, each of which is a number. This number characterizes the signal at a point with a certain accuracy. The accuracy of the representation depends on the width of the range of numbers obtained, and, therefore, on the resolution of the ADC.

II. MATERIALS AND METHODS

The frequency of digitization is also of no small importance. The human ear perceives sound in the frequency range from 16 Hz to 22 kHz. In fact, the boundaries of audible frequencies depend on a particular person - his age, gender and health.

The possibility of transmitting a continuous signal by its discrete samples was justified by V. A. Kotelnikov in 1933 [5]

The human voice corresponds to a frequency range of 300-4000 Hz, therefore, in order to restore the voice signal without loss, it is necessary to use a sampling frequency greater than 8 kHz. The optimal value of the frequency in the conditions of the problem under consideration is 12 kHz.

According to Kotelnikov's theorem, if a continuous signal has a spectrum limited by frequency, then it can be completely and unambiguously reconstructed from its discrete samples taken at time intervals.

$$T = \frac{1}{2 \cdot F_{max}}, \quad (1)$$

With a frequency of $F_d \geq 2 \cdot F_{max}$, where

F_d is the sampling rate;

F_{max} is the maximum frequency of the signal spectrum.

In other words, the signal sampling frequency (ADC sampling frequency) should be at least 2 times the maximum signal frequency that we want to measure.

In order to calculate the signal spectrum from its discrete samples, the discrete Fourier transform (DFT) is used.

The classic approach for digital sound processing is the conversion of the amplitude-time dependence into the frequency form. It allows you to adequately describe in frequency coordinates all types of signals; frequency-truncated Fourier components describe the data more plausibly than any other power series. The individual components are sinusoids and are not distorted during transmission through linear systems, which allows them to be used as good test signals.

The direct Fourier transform has the form:

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

where X_k are the spectral samples, x_n are the samples of the discrete signal, N is the conversion length.

Human speech is a sequence of sounds. Sound, in turn, is a superposition (superposition) of sound vibrations (waves) of various frequencies. The wave, as we know from physics, is characterized by two attributes - amplitude and frequency. In order to save the audio signal on a digital medium, it is necessary to divide it into many gaps and take some "average" value on each of them.

The process of solving the problem of recognizing isolated words can be divided into four main stages:

- signal input from the external environment into the system;
- finding the endpoints of a word;
- extraction of signal characteristics;
- determination of the recognition result.

Separating a word from a continuous stream of incoming information is a difficult task due to the characteristics of the voice, the environment and the equipment with which the sound signal is recorded.

Sound signals, propagating in real conditions, passing through various real mediums of sound transmission, change their shape beyond recognition. In the Nature of sound waves, there is only one wonderful sound wave, in which something remains always unchanged, regardless of which mediums of sound transmission it passes through and from which many surfaces this wave is not reflected (adding up with all its many reflected waves). [6]

Definition of word endpoints

The main unit of processing a digitized signal is a frame - an array of samples corresponding to a specific time period.

Frames are a more suitable unit of data analysis than specific signal values, since it is much more convenient to analyze waves at a certain interval than at specific points. The location of the overlapping frames allows smoothing the results of the analysis of frames, turning the idea of frames into a certain "window" moving along the original function (signal values).

It was experimentally established that the optimal frame length should correspond to a gap of 10ms, "overlap" - 50%. Considering that the average word length (at least in my experiments) is 500ms - this step will give us approximately 500 / (10 * 0.5) = 100 frames per word. [7]

To isolate a word from a continuous stream of information in real time, a simple but at the same time quite effective method for determining the Rabiner-Sambur endpoints based on the calculation of the frame energy and the frequency of zero-crossing can be used. This method requires less computation due to the lack of additional signal conversion from the time domain to the frequency domain.

In this case, the frame energy is understood as the normalized sum of the absolute values of the amplitudes of the discrete samples of the signal (1).

$$E = \frac{1}{K} \sum_{n=1}^{N} A_n, \qquad (3)$$

where K is the normalization coefficient, N is the frame length.

To calculate the energy value, other calculation methods can be used, for example, finding the Euclidean norm. Since the minimum memory unit for signal processors is 2 bytes, as well as the fact that the resolution of the audio codec used to digitize the audio signal is 16 bits, arrays containing double-byte elements are preferred as structures for storing information. The normalization of the final value is necessary in order to avoid overloading the discharge grid.

The normalization coefficient is selected from the following considerations: since the resolution of the codec is 16 bits, and the amplitude value can be either positive or negative, the maximum value possible in absolute value, which can be stored in a two-byte signed type, is 216-1 = 32768 , and the maximum sum of the absolute values of the amplitudes is 32768 * 512. Based on the foregoing, as well as the fact that the energy value is stored in a double-byte signed type, the normalization coefficient is chosen equal to the length of the frame.

The zero-crossing frequency is defined as the number of times the original signal changes sign and its value is above the noise threshold. This value does not need to be normalized, since the maximum value of the parameter is N-1.

III. SOLATION OF CHARACTERISTICS OF SPEECH INFORMATION. CHALK-CEPSTRAL COEFFICIENTS

One of the stages of the speech recognition process after extracting a word from the input data stream is the extraction of parameters (characteristics), where different techniques are used, for example, the method of searching for cepstral coefficients. The main task at this stage is to isolate certain parameters of the signal, such that the number of these parameters should be minimal in order to speed up the comparison with the sets of characteristics from the library, and at the same time, these parameters should be such that they can be used fairly accurately compute a specific word. [9]

Mel-cepstral coefficients are a peculiar representation of the energy of the spectrum of an audio signal.

Human hearing aids have the property of frequency masking - a situation where normally audible sound is covered by another loud sound with a close frequency. This characteristic is dependent on the signal frequency and varies from 100 Hz for low frequencies to 4000 Hz for high frequencies.

The range of audible frequencies can be divided into a certain number of critical bands (division into 24 critical bands is accepted), indicating a drop in ear sensitivity for the highest frequencies. Critical bands are considered another parameter of sound, similar to its frequency. But, unlike the frequency, completely independent of the organs of hearing, the critical bands are calculated in accordance with auditory perception. As a result, they form some measures of frequency perception, for which the units of measurement - bark and chalk.

The barks scale (Figure 1) is associated with critical hearing bands, and since the width of these bands is uneven, increases with increasing frequency of sound vibrations, the scale is uneven. The direct and inverse relationships between the tone frequency in Hz and the pitch in barks are calculated by formulas (4) and (5), respectively:

$$b = 13\,\mathrm{atan}(0.00076 \cdot f) + 3.5\mathrm{atan}(\tfrac{f}{7500})^2 \quad (4)$$

$$f = \frac{52548}{b^2 - 52.56b + 690.39} \quad (5)$$

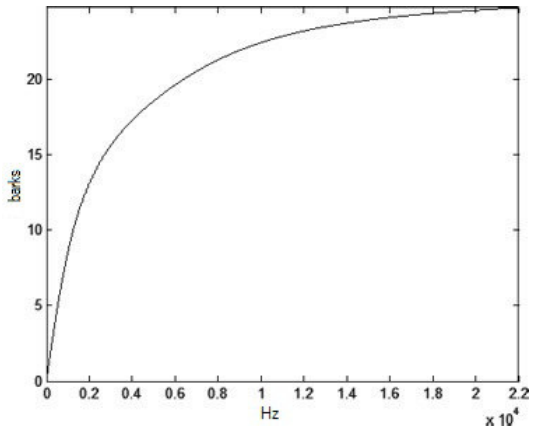

Figure 1 - Audible frequency range: barks scale.

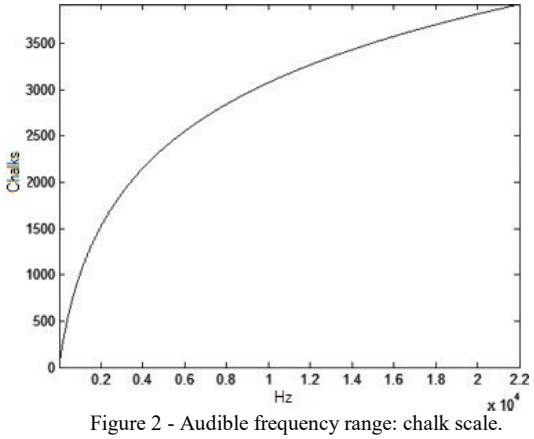

Figure 2 - Audible frequency range: chalk scale.

The chalk scale is uneven (Figure 2). The chalk scale is based on the statistical processing of a large amount of data on the subjective perception of the pitch of sound tones. According to the results of studies, we can say that the pitch of the sound is mainly related to the frequency of oscillations, however, it depends on the sound volume level and its timbre. The direct and inverse relationships between the tone frequency in Hz and the pitch in chalk are calculated by formulas (5) and (6), respectively:

$$m = 1127.01048 \cdot \ln\left(1 + \frac{f}{700}\right) \quad (5)$$

$$f = 700(e^{\frac{m}{1127.01048}} - 1) \quad (6)$$

The frequency range of a person's voice is limited and ranges from 300-4000 Hz. Therefore, by modeling a band-pass filter, one should discard the frequency components that are located outside this range and, therefore, do not carry a semantic load.

The studied signal is divided into frames based on the Welch periodogram method - the signal sample vector is divided into overlapping segments (using 50% overlap), after which each frame is multiplied by the weight function and the discrete Fourier transform is calculated for it (2). The Hamming window (7), which is shown in Figure 3, is used as the weighting function. Using the weighting function allows weakening the spreading of the spectrum at the junction of frames.

$$x_n = 0.53836 - 0.46164cos\frac{2\pi n}{N-1} \quad (7)$$

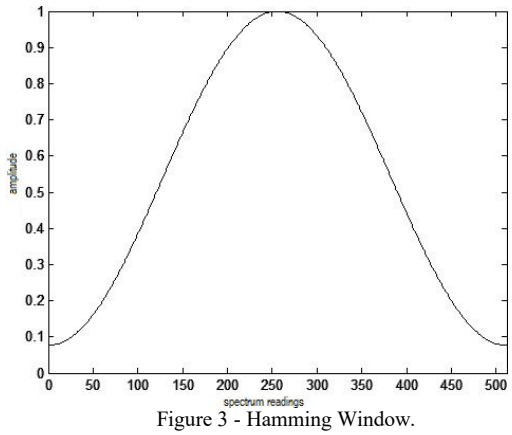

Figure 3 - Hamming Window.

Research results in the field of psychophysical perception say that the most significant information is in the actual frequency spectrum, therefore, after the Fourier transform algorithm for further analysis, the actual signal spectrum is determined, where phase information can be omitted.

Results and Discussion

Speech processing begins with an assessment of the quality of the speech signal. At this stage, the level of interference and distortion is determined. The evaluation result is sent to the acoustic adaptation module, which controls the module for calculating the speech parameters necessary for recognition.

The start window of the program is shown in Figure 4. The process of recognizing isolated words begins after a user clicks on the picture (Figure 4). After clicking, the working form opens, where (by opening the form) the microphone of the computer is turned on. The user speaks certain words from the software "dictionary", which are statically registered in the program. Speech information is fed to the PC microphone, if the degree of similarity of words spoken by the user with the dictionary is more than 50 percent, the corresponding result will appear on the screen.

In the signal, sections containing speech are highlighted, and speech parameters are evaluated. Phonetic probabilistic characteristics are distinguished for syntactic and semantic analysis. (Evaluation of information about parts of speech, word form, and statistical relationships between words.)

Figure 4 - The start window of the program " Word Recognition".

Further, the speech parameters enter the main unit of the recognition system - the decoder. This is a component

that compares the input speech stream with the information stored in acoustic and language models, and determines the most likely sequence of words, which is the end result of recognition. The work of the program "Word Recognition" is shown in Fig. 5-6.

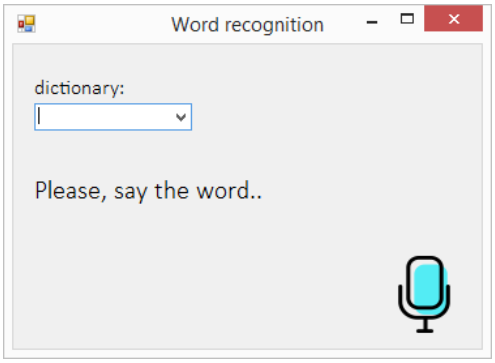

Figure 5 - The window of the program “Word Recognition”.

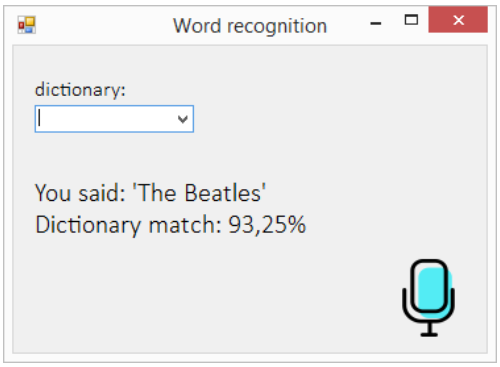

Figure 6 - The result of the program.

In the course of the work, a mathematical model for the automatic recognition of individual words was developed, and an algorithm for a program for recognizing isolated words was also built.

Conclusion

The result of the work is a computer implementation of the algorithm for recognizing isolated words using Visual Studio 2018. In the work, a program was developed in the high-level programming language C # .net, which implements the described algorithm for modeling speech recognition.

The results obtained showed the possibility of using the distinguished parameters of speech signals for speech recognition.

References

[1] Allen, L. Scripting for Dragon NaturallySpeaking 9/L. Allen. – San Francisco: SoftNet Systems Inc, 2006. – 130 p.

[2] Iconin, SJ. Automatic Speech Recognition System SPIRIT ASR Engine / CU. Iconin, D.V. Saran // Digital signal processing. - 2003. - No. 4. - From 5–13.

[3] Gale, T. IVOICE launches iVoice speech software developers kit v.3.0 / T. Gale. - Boynton Beach: Worldwide Videotex, 2003 .-- 7 p.

[4] Ronzhin, A.L. Russian Speech Recognition System SIRIUS / A.L. Ronzhin, A.A. Karpov, I.V. Lee St. Petersburg Institute of Informatics and Automation RAS. - SPb., 2005. - 112 p.

[5] Dyakonov V.P., Abramenkova I.V. MATLAB. Signal and image processing. St. Petersburg: Peter, 2002. 608 p.

[6] Sergienko A. B. Digital signal processing. - 2nd. - St. Petersburg: Peter, 2006. - P.751.

[7] Eificher, Emmanuel S., Jervis, Barry W. Digital Signal Processing:practical approach, 2nd edition .: Per. from English - M.: Publishing House “Villamo”, 2004. - 992 p.: Ill. - Paral. tit. English Sergienko

[8] Methods of automatic speech recognition / Ed. W. Lee - M .: Mir, 1983 .-- 716 p.

[9] Kristalinsky R.E., Kristalinsky V.R. Fourier and Laplace transforms in computer mathematics systems. M .: Hot line, Telecom, 2006.216 s.

[10] Voinarovsky, M. Psychologist [Electronic resource] / M. Voinarovsky // Programming. Fast Fourier Transform. - 2002.

[11] Eificher, Emmanuel S., Jervis, Barry W. Digital Signal Processing:

[12] A Practical Approach, 2nd Edition .: Per. from English - M.: Publishing House

[13] “Villamo”, 2004. - 992 pp., Ill. - Paral. tit. English Sergienko