VOL. 14, NO. 10, MAY 2019 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

# FOUNDATION OF A MATHEMATICAL METHOD FOR ANALYSIS OF VOICE COMMANDS

Tymchenko S. E.[1], Tymchenko E. M.[1], Vlasov S. F.[2,6], Vlasov V. S.[3], Kovalenko V. L.[4,7] and Kotok V. A.[5,7]
[1]Department of Higher Mathematics, National Technical University Dnipro Polytechnic, Dmytra Yavornytskoho Ave., Dnipro, Ukraine
[2]Department of Underground Mining, National Technical University Dnipro Polytechnic, Dmytra Yavornytskoho Ave., Dnipro, Ukraine
[3]Department of Software Engineering, National Technical University Dnipro Polytechnic, Dmytra Yavornytskoho Ave, Dnipro, Ukraine
[4]Department of Analytical Chemistry and Food Additives and Cosmetics, Ukrainian State University of Chemical Technology, Gagarin Ave., Dnipro, Ukraine
[5]Department of Processes, Apparatus and General Chemical Technology, Ukrainian State University of Chemical Technology, Gagarin Ave., Dnipro, Ukraine
[6]Department of Building Production Technology, Vyatka State University, Moskovskaya St., Kirov, Russian Federation
[7]Department of Technologies of Inorganic Substances and Electrochemical Manufacturing, Vyatka State University, Moskovskaya St., Kirov, Russian Federation
E-Mail: Valeriy_e-ch@ukr.net

**ABSTRACT**

Development of solutions for problems of automatic speech recognition and understanding is getting more and more scientific and practical value. The aim of the work is to evaluate the effectiveness of using mathematical method for recognition of speech sounds. State of the problem was analyzed, requirements to developed algorithm were reviewed and method for solving the set problem have been chosen. A hypothesis has been proposing that each sound has its own specific spectra and the possibility of realizing recognition of individual speech sounds regardless of speakers have been studied. Analysis of the experimental data was conducted. Spectra of recorded sounds were compiled and algorithms based on discrete Fourier transform were chosen for their processing. To evaluate the quality of speech recognition (probability of correct recognition), the program was used to process all 180 recorded sounds (6 for 30 voices) and data on a number of correct and false recognitions for each sound with all four methods was compiled. The obtained data were used to develop and programmatically implement an algorithm for recognition of vowel sounds. The program realizes few different algorithms for sound recognition and outputs result for each of them and total recognition result. The novelty of technical solution is in that developed system is able to recognize sounds with a sufficient degree of probability without the need for pre-processing. The practical value of the project is in the realization of vowel sound recognition as a fragment of a new approach to the development of a system for voice command recognition for Slavic language family. Result for algorithm revealed that for proposed method, the probability of correct recognition is higher than 84%, which allows concluding that proposed method can be used for development of fundamentally new approach in recognition of voice command for multi-user systems with voice control.

**Keywords:** phonetics, digital signal processing, voice control, speech sounds, spectral analysis, discrete Fourier transform.

## 1. INTRODUCTION

### 1.1 Problem statement and its relation to scientific and practical problems

Voice control is a way of interacting with a device using speech, which is an important trend these days. The main problem with integration of voice control into various fields of human life is insufficient accuracy of speech recognition and interpretation process [1]. Many specialists across the world have devoted their work to the creation of speech recognition methods that are insensitive to various distortions (external noise, speech variation, syntax deviations etc.). There are two main types of speech recognition methods. First is speaker-dependant i.e. a user has to teach the system to recognize their voice before it can function [2]. In other words, phrases with partial errors are discarded on word chain recognition level. This limitation cannot be eliminated using an existing approach (exhaustive iteration, for instance), as that would significantly complicate the recognition model. As a result, such systems are used individually. The second type is speaker-independent speech recognition i.e. system is capable of recognizing any speech regardless of the speaker. Speaker-dependent speech recognition systems are designed for a single user [3-5]. Systems of the second type are (voice-independent) are developed for any user of a given type (English language, Russian language etc.). These are the most complex and expensive systems, with lower accuracy that systems of first type [6-7].

### 1.2 Literature analysis

First systems for automatic speech recognition were only capable of recognizing a limited number of words and require preliminary tuning for the user. One of most known systems for recognition of Russian speech is system "Речь", developed in the early 80's of past century under the supervision of T. K. Vintsiuk. The system is based on the concept of sequential processing of speech information based on dynamic programming and time interpretation of speech as a result if non-linear compression and stretching [8-10]. Another direction in speech recognition was founded by V. N. Trunin-Donskoy. Here, the main attention is paid to acoustic

features of where and how the speech is formed (time, frequency, amplitude) for making a decision on each step of the speech process. The main difference of this approach from one uses in by T.K. Vinsiuk and colleagues, is primary basis on mathematical method.

Among most known developments are speech recognition-synthesis devices MARS-1 and MARS-2 based on formant analysis and synthesis. In the middle of 90's of XX century speaker-independent systems with a dictionary of 1000 words have been developed, which had recognition accuracy of 87-99%, regardless of vocabulary. These systems were based on the principle of hierarchic recognition, and processing procedure was based on dynamic programming. Another rather effective approach was based on the computation of minimal matching using gradient decline method, which was later used in the development of speech recognition device DIS-332. The device was designed for recognition of over 200 commands with 96-98% accuracy and was based on K580IK80 microprocessor [11-13].

In recent years intellectual systems, designed for telecommunication and various information services, are being developed. More people prefer using new intellectual technologies that simplify access information and save time. As such, the primary characteristics speech-recognition systems became speaker-independency, which eliminates the need for preliminary tuning for a specific user and all them to immediately dialog with the system [14-15].

Presently, many developments on speaker-independent recognition of Russian speech have emerged, which are based on statistic language model utilizing various speech units as a basis (word-form, lemma, morpheme etc.). Development of faster computers leads to the use of a statistical method that is based on a computationally complex hidden Markov model (HMM), which created new possibilities for speech recognition [16-18].

### 1.3 Outline of previously unsolved part of the general problem

It is known, that speech recognition technologies have found their use in various fields. However, there are many unsolved problems, and many ideas require additional development. For instance, programs operating with isolated words have reached high accuracy in command systems - for most spread modern application the accuracy of speech recognition is 95-99% on average and is mainly affected by the noise level. At the same time, the problem of recognizing of continuous speech with sufficient accuracy is yet to be solved, while for cases of the limited dictionary, such systems do exist (VoxReports based on ViaVoice, Verbmobil) and show high accuracy. Presently, many works are devoted to the problem of recognizing continuous speech (ICS RAN, «Istra-soft», IBM http://masters.donntu.edu.ua/2012/iii/akopyan/library/article1.htm - lib), as this type of voice interaction is considered as most promising [5-6].

The most important stage of speech processing during recognition is the isolation of information features that characterize speech signal. The is a number of mathematical method for analysis of speech spectrum. Fourier transform, which is known from digital signal processing theory, is most used. This mathematical apparatus has proven itself in this field; with many known methods for signal processing being based on Fourier transform [7].

However, when comparing the results of modern systems with those of early stage of speech recognition science, it can be said that there was not much advancement over past decades. This makes some specialists doubt the possibility of implementing speech interfaces in near future. Others think that the problem is almost solved. The majority expert agrees that development of speech recognition requires some time.

### 1.4 Problem statement

The present paper proposes a hypothesis that each sound has a unique spectrum, regardless of voice timbre or individual pronunciation nuances. Thus systems build on recognition of individual sounds and not whole words stop being individual and become mutely-user, which significantly broadens their application range. The aim of the work is to study the possibility of recognizing individual sounds, regardless of the speaker, based on their spectral analysis. Successful recognition of individual sounds directly influences the recognition accuracy of voice commands. One should consider that recognition of sounds and breakdown of recording into separate sounds can be complicated by various noises. As such, this requires the development of an algorithm for recognition of specific features of each sound spectra for successful identification from noise background.

The developed algorithm should meet the following requirements:

- Must not require preliminary training i.e. dictation of sounds;
- The possibility of correct recognition must not be below 80%;
- The algorithm must not require preliminary noise remove when recognizing sounds;
- Signal processing time should be minimal for program execution on embedded platforms with limited system resources.

To achieve the set aim it is necessary to:

- Analyze the spectra of vowel sounds recorded from different voices;
- Develop an algorithm, based on obtained data, for recognition of vowel sounds;
- Programmatically implement the developed algorithm and analyze the possibility of its practical application.

## 2. EXPERIMENTAL

Vocal sound - a sounds produced by speech organs of humans with aim of language communication. Speech organs include throat, an oral cavity with tongue,

lungs, nasal cavity, lips and teeth. Science about speech sounds is called phonetics. In general, speech sounds are split into noises and tones: tones are produced as a result of oscillating vocal cords; noises are produced as a result of non-periodic oscillation of air leaving lungs. The tone is usually vowels; almost all voiceless consonants are considered noises. Voiced consonants are formed from merging of noises and tones.

Spectra of speech sound can be split into tone (periodic) and noise (non-periodic) components. Tone sounds are formed with help of vocal cords, noise - by obstacles in the oral cavity. By the presence of these components, speech sounds can be classified as:

- Vowels - tone;
- Voiceless consonants - noise;
- Sonorant consonants - tone with a low amount of noise;
- Voiced consonants - noise with the inclusion of tone.

In the Russian language there are 10 vowels: А, Е, Ё, И, О, У, Ы, Э, Ю, Я and 6 vowel sounds: [А], [Э], [И], [О], [У], [Ы]. Some source does not differentiate between sounds [Ы] and [И]. For instance, the Ю is translated by sounds [ЙУ], or by softening of preceding consonant and sound [У].

To achieve the set aim - mathematical analysis of speech sounds - experimental studied have been conducted. For experimental data, six vowel sounds were recorded: [А], [Э], [И], [О], [У], [Ы]. The recording was conducted using laptop's built-in microphone, under real conditions, without special equipment, and with different environmental noises. Each sound was recorded by 15 different male and 15 female voice, 30 voices total. Recordings were cut into fragments of about 100 ms with about equal signal amplitude over the duration of the fragment, and were saved into the uncompressed format - PCM WAV, single channel (mono), 22050 Hz sampling rate, 8-bit depth. All recording was normalized. For normalization, file amplitude was scaled so the maximum signal amplitude was equal to a possible maximum value determined by bit depth of audio file. Normalization coefficient:

$$K_{norm} = \frac{2^{n-1} - 1}{A_{max}},$$

where

n - bit depth of audio file,

$A_{max}$ - maximum absolute value of signal amplitude.

All signal values were multiplied by the normalization coefficient.

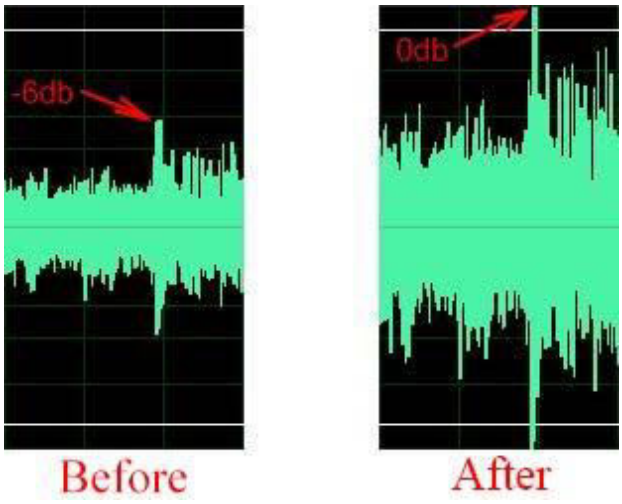



**Figure-1.** Oscillogram fragment before and after normalization.

Normalization is important for the further spectral analysis of sounds as coefficient values (amplitudes) depend on the amplitude of the analyzed signal. Only their ratio is independent of general amplitude.

## 3. RESULTS AND DISCUSSIONS

### 3.1 Obtaining acoustic spectra using DFT

For sound analysis, a program was written that analysis spectral analysis of data based on discrete Fourier transform. The transform result is the output as a graph. Harmonics were plotted onto (number of harmonics is labeled above, below - their frequency values), and their ordinates were plotted onto ordinate. Figure-2 shows programs main window.

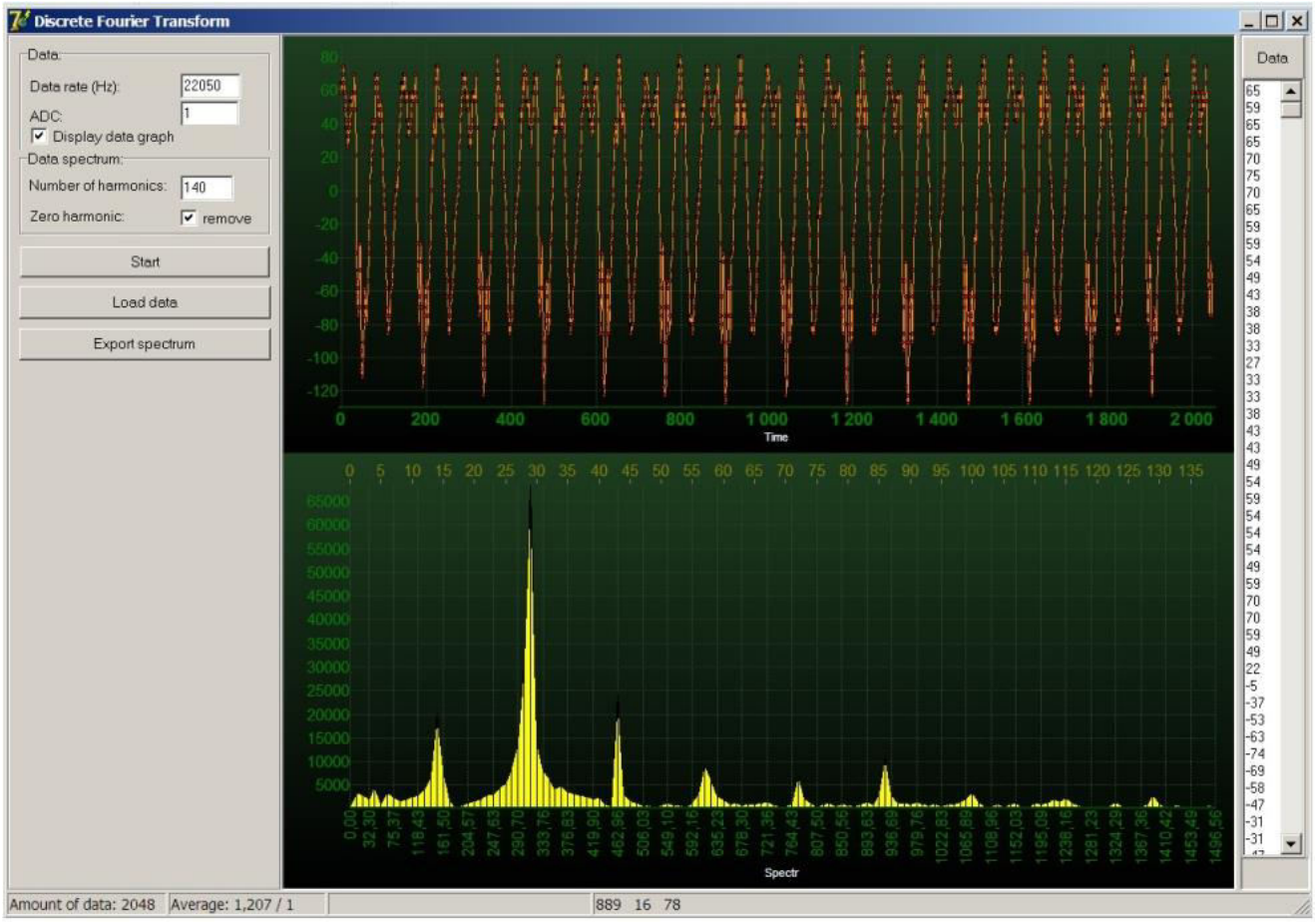



**Figure-2.** The main window of the spectral analysis program.

Fourier transform - is an operation that compares functions of different variables. This new function describes coefficients (amplitudes) upon decomposition of an initial function into base components - harmonic oscillations with different frequencies. Discrete Fourier transform - transform of finite number (complex) sequences, which is like in continuous case, transforms convolution into point wise multiplication. It is used in digital signal processing and other cases when it fast convolution transformation is required, for instance, multiplication of large numbers. A generalized formula for direct Fourier transform is written as:

$$S(f) = \int_{-\infty}^{\infty} e^{-i2\pi ft} x(t)dt$$

For frequency analysis, the correlation dependency between signal represented in time and complex exponent with set frequency is calculated. The complex exponent is decomposed into real and imaginary part according to Euler's formula:

$$e^{-i2\pi ft} = \cos(2\pi \cdot f \cdot t) - i\sin(2\pi \cdot f \cdot t)$$

The formula for direct Fourier transform uses integration in time from negative to positive infinity. This is, of course, a mathematical abstraction. In the real world, we can integrate from a given moment in time, which we can label as 0, to moment in time T. The direct Fourier transform formula then transforms into:

$$S(f) = \int_{0}^{T} e^{-i2\pi ft} x(t)dt$$

As a result, properties of Fourier transform change. Instead of a continuous function, the signal spectra becomes a discrete sequence of numbers. Now the minimal frequency and step of frequency values of spectra become:

$$f_{\min} = \Delta f = \frac{1}{T} \; ; \quad \Omega = \frac{2\pi}{T}$$

Fourier transform for digital signal samples is called a discrete Fourier transform and is written as follows:

$$X(k) = \sum_{n=0}^{N-1} e^{-i\frac{2\pi}{T}kn} x(n) =$$
$$\sum_{n=0}^{N-1} [\cos(\frac{2\pi}{T}kn) - i\sin(\frac{2\pi}{T}kn)]x(n)$$

where

$N$ - the number of digital samples;

$k$ - a number of signal harmonic.

The frequency of harmonic is equal to $\frac{k}{T_{meas}}$,

where $T_{meas}$ - sampling period.

Data for analysis can be loaded into the program in two ways. The first way is to directly paste data table column into the respective field on the right of the program's main window. The second way is to load the WAV file. WAV file has a bit depth of 8 bits and analyzed data must be at the start of the file. When reading WAV file, first 46 bytes are ignored as they are used to store file header. Only first 2048 values are analyzed.

Field «Data rate» is used for input of data rate. Field ADC - multiplier. Because spectral composition is independent of its amplitude and for better performance only integer part of each value is processed. When processing real number it is necessary to set the multiplier to remove decimal part of the values. To obtain correct amplitudes of spectral components, the values are divided by the value of ADC before output. Next field is used for input of number (first N) of harmonics which would be selected.

During experiments, it was found that for vowels there is no need to select more than 140 harmonics. During studied various recordings of speech sounds were subjected to spectral analysis with a number of process harmonics up to 512 (Figure-3).

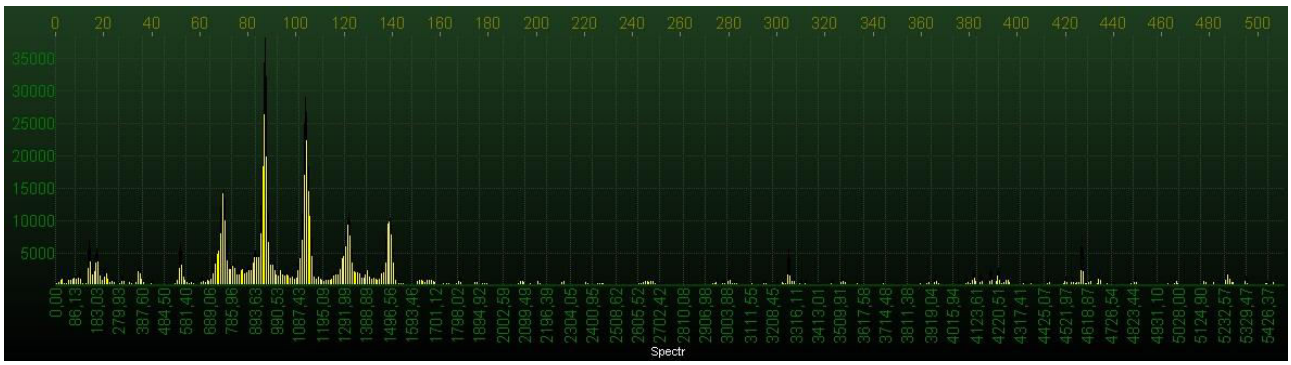



**Figure-3.** Spectrogram of Russian sound [A] with a number of harmonics equal to 512.

Visual analysis of obtained spectrogram revealed that harmonics above 140 have no valuable information while processing time significantly increases with increasing number of processed harmonics. This is in agreement with common practice in voice transfer over low bandwidth wired and wireless communication channels, in which frequencies below 300 Hz and above 3 kHz are filtered out, which does not cause loss of information. The frequency of 3 kHz approximately corresponds to 280$^{th}$ harmonic, but such high frequencies appear only in whistling or hissing sounds, which are not subject of the present study.

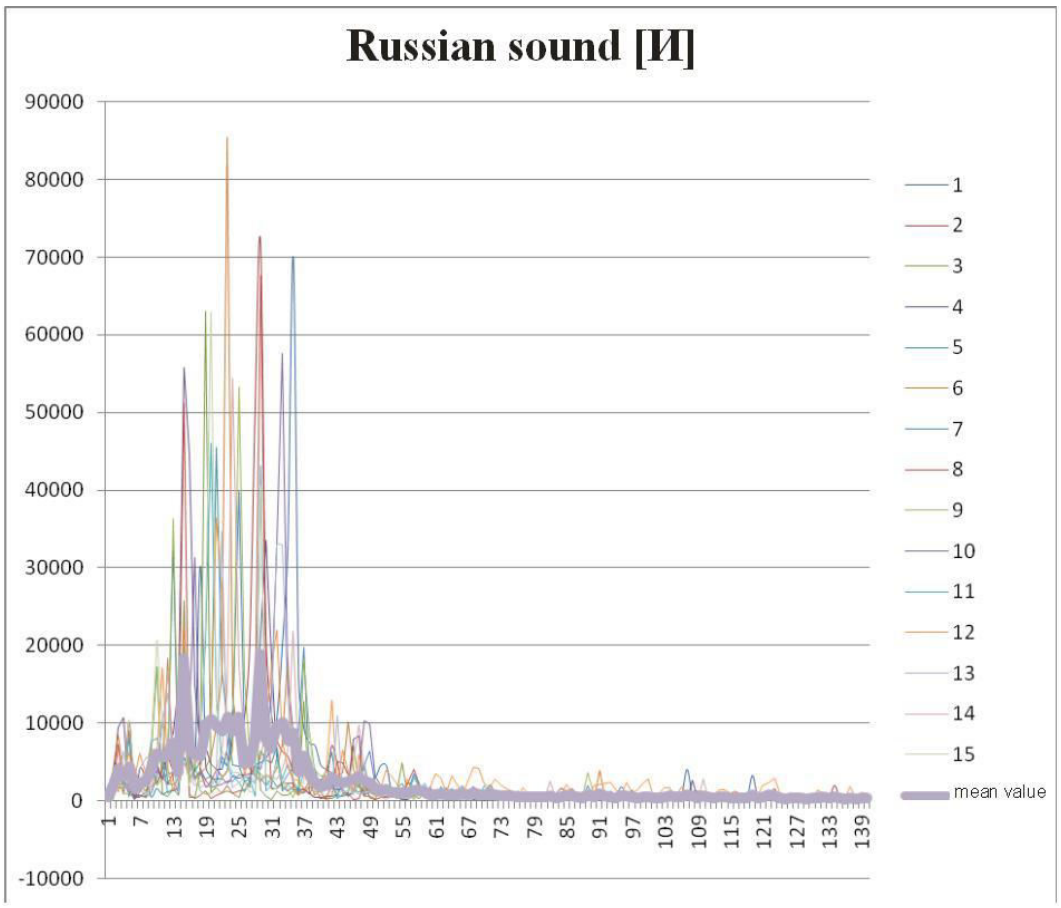


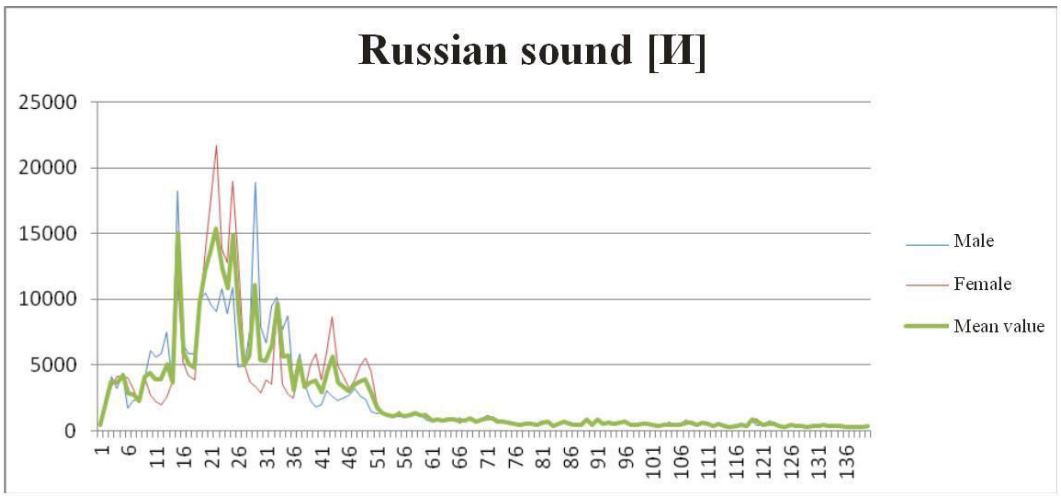

**Figure-4.** Spectrograms of Russian sound [И] for 15 male voices

The program can also export analysis result into a text file composed of three columns - harmonic number, frequency and amplitude values. This is necessary for the further analysis of signal spectra using MS Excel. The text file is output in the same folder from which initial data file was loaded from and in case of manual input - into program's folder.

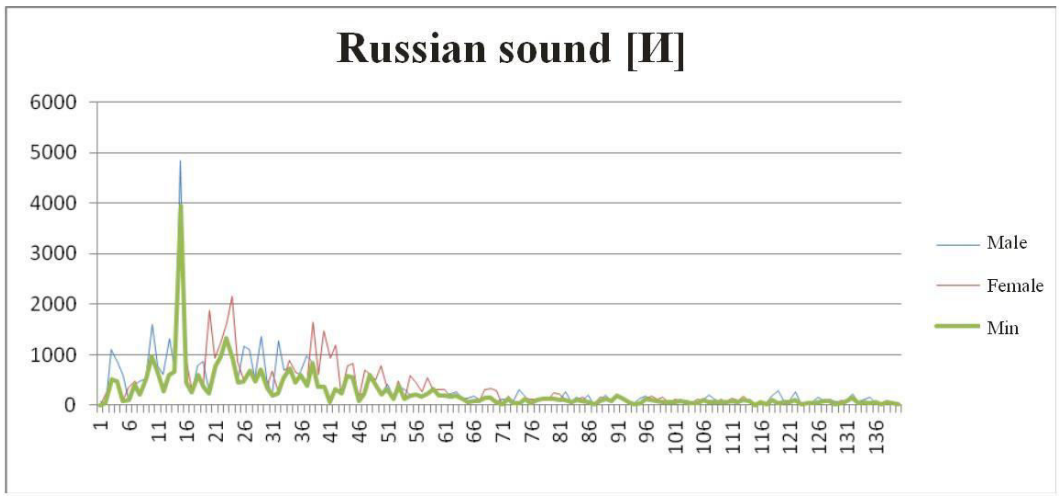
For each of 30 voices, spectra for each of 6 sounds were compiled into MS Excel table with separation into a male and female voice. A graph was plotted for each sound. Graphs show spectrograms of this sound for all recorded 15 male and 15 female voice separately. Digital version of is appended onto a disk, a sample of spectrogram for sound [И] for 15 male voices is shown in Figure-4.

### 3.2 Processing of vowel speech sound spectra

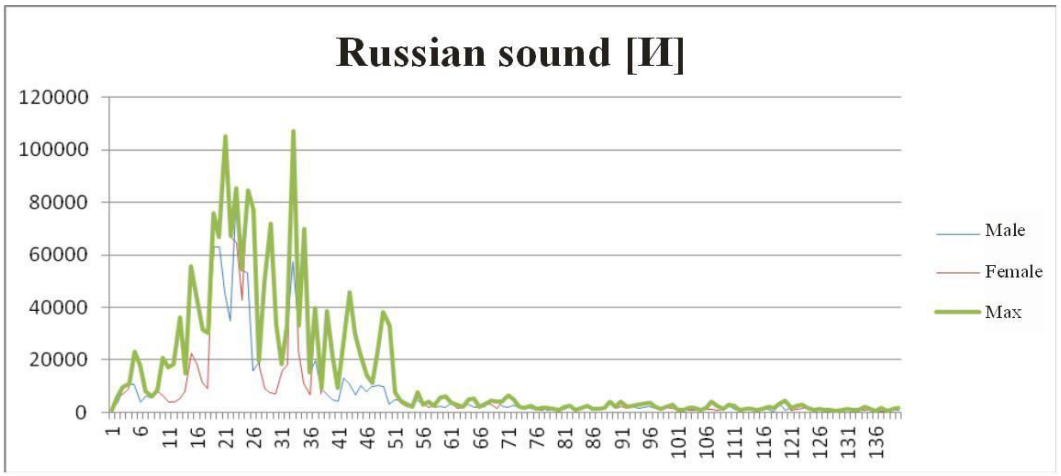
For processing of sounds recorded over the course of the experiment, for possible methods were considered. First three methods are similar between each other and assume plotting of graphs with mean, maximum and minimal values of spectrogram separately for male and female voices and a whole set of data. Examples of these spectrograms for Russian sound [И] are shown in Figure-5. Then, each recording is subjected to spectral analysis and counting of the sum of absolute difference values between the amplitude of each harmonic of a given recording and respective value of maximum, minimum and mean amplitude of this harmonics for each sound.




a)




b)




c)

**Figure-5.** Graphs of amplitude values: a - mean values, b - minimum values, c - maximum values.

Each of these three methods assumes that sound for which value of this sum (sum of absolute amplitude errors) is lowest and corresponds to studied sound.

The fourth method is based on assumption that spectrograms of a single sound but spelled by different people with different voice timbre would be similar but shifted relative to each other by a difference of the mean frequency of voices. Figure-6 shows spectra of Russian sound [И], recorded in different voices.

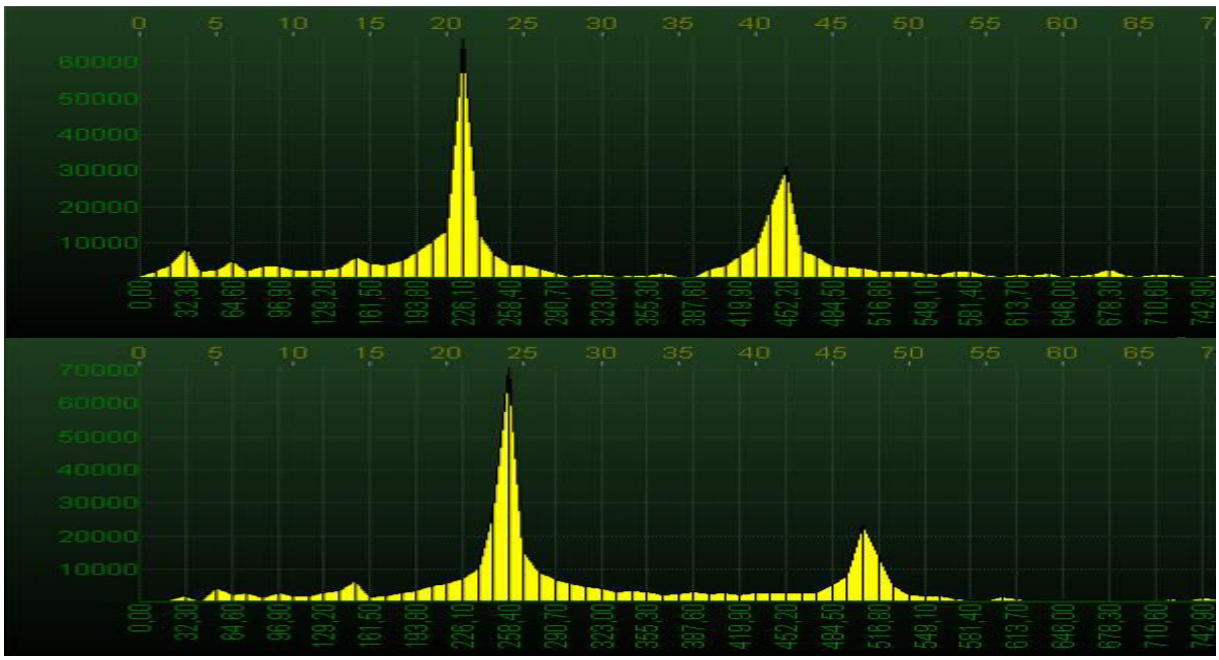


**Figure-6.** Spectra of Russian sound [И] of two different voices.

Maximum amplitude values for each sounds spelled with each voice were derived from experimental data. Spectra were compared based on maximum amplitude values. Among compared spectra, fragments with an equal maximum length of 80 harmonics were selected. Analogs comparison was conducted for male and female voices for each sound, and samples with a length of 64 harmonics were selected. For studied sound, a spectral analysis was conducted. The spectrogram is compared with maximum value with each of fragment obtained during analysis of experimental data. Sound recognition as in first three methods is based on the lowest sum of absolute amplitude error, with the exception that not all 140 harmonic but only studied fragment was compared i.e. $80^{th}$ or $64^{th}$ harmonics.

### 3.3 Experimental evaluation of proposed algorithms

All four proposed method for sound recognition were realized in the program previously created for spectral analysis of sounds. The left panel shows the separate result for each method. Additionally, each of these methods conducts separate comparison for male voices, separate comparison for female voices and separate comparison for combined data. Such a solution would allow evaluating how the qualities of sound recognition change if user's gender is considered by the program. Figure-7 shows the main window of the program after sound recognition is realized.

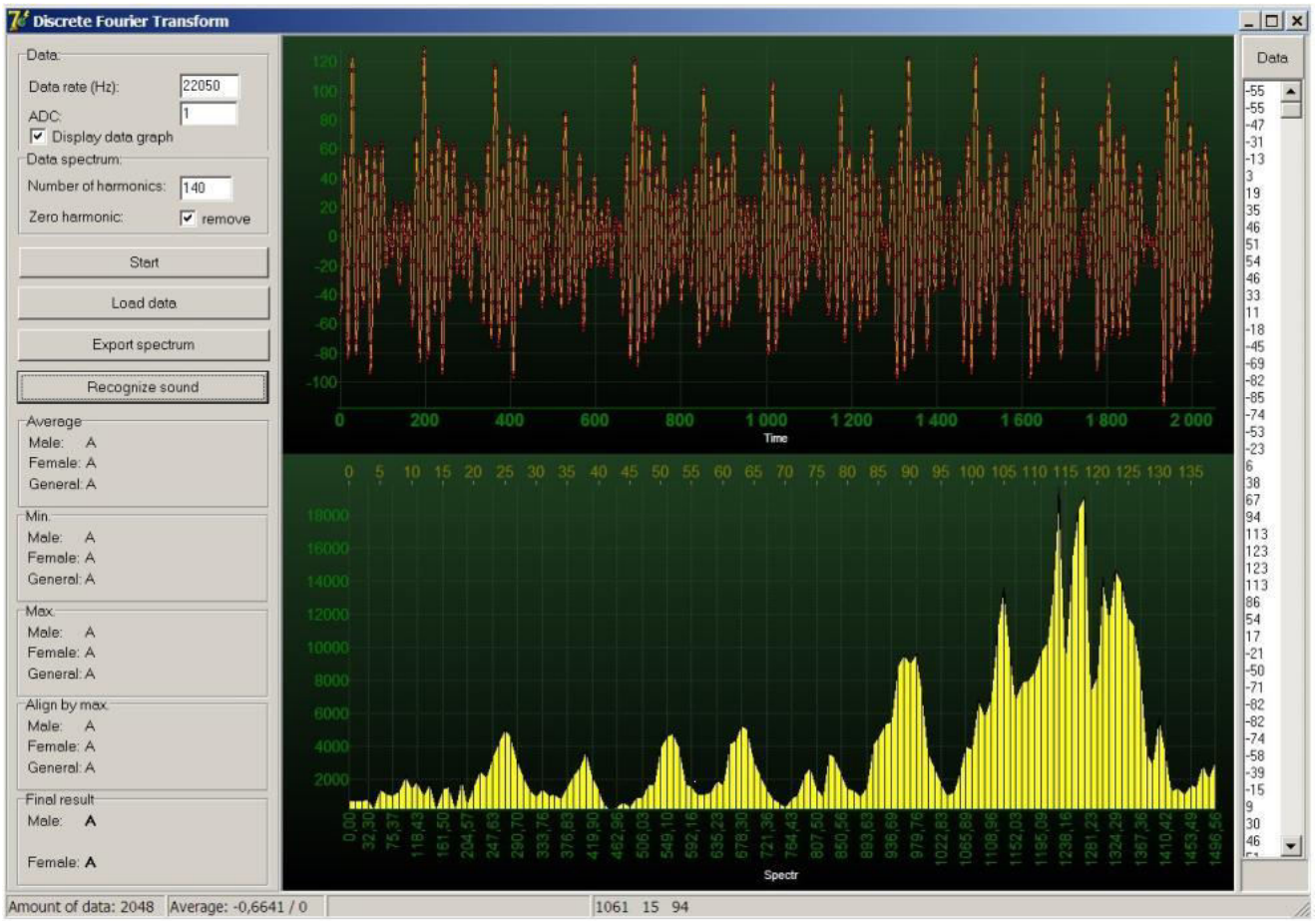



**Figure-7.** Main with of program for spectral analysis with realized sound recognition methods.

To evaluate the quality of sound recognition (a probability of correct recognition), all 180 previously recorded sounds (6 sounds for each of 30 voices) were processed by the program and data on a number of correct and incorrect recognition by each of four proposed methods was compiled. Static evaluation of results was conducted according to the literature [8].

For different sounds, different method produces the best results. For better probability, a parameter that indicated the user's gender must be present. For instance, for sound [A] the best result is produced by the first method. Notably, for male voices only data obtained from analysis of male voices must be used. At the same time, for female voice, a total data set is required It at least one method (with total data set or with a set for respective gender) produces the same result then it can be said that recognized sound is - [A].

Sound [O] is only recognized by the first method for data set corresponding to the user's gender. The probability of correct recognition is about 93% for male and 73% for female voices.

Russian sound [Э] is recognized by the second method for total data set regardless of gender. However, the second method sometimes produces false recognition of diphthongs similar [Э]. Their exclusion requires either confirmation with any other method (for total or corresponding data set) or exclusion of other sounds (i.e. sound [Э] is recognized last).

For recognition of Russian sounds [И], [У] and [ЫI] results of few methods must be taken into account. Their spectra are similar which result into false recognition of these three sounds. Firstly, sound [У] need to be recognized as a probability if its correct recognition is lowest. Sound [У] is recognized by the first or second method for male voices and by first or fourth for female. The probability of correct recognition is about 73% for male and female voices

If recognition of sound [И] and [ЫI] is not required, then a first or third method for total data set can be used. However, it is difficult to differentiate between these sounds with sufficient probability. For female voices, the third method reliably differentiates between sound [И] and [ЫI]. In the present research, the method had no false recognition of sound [И] and [ЫI] for female voices. In practice, the probability is likely below 100%, but it should close to that value. While for male voices first and fourth method with initial data set for male voices point to sound [И], the recognized sound is indeed [И] with a probability of 80%, otherwise, the recognized sound - [ЫI] with a probability of about 93%.

In the case when the listed method has failed to individually recognize the sound, it is recognized by a maximum number of a method that returned the same

result for data set corresponding to the user's gender. In this case the probability of correct recognition is low, but in general, this approach increases the final probability of correct recognition.

### 3.4 Development and evaluation of a final algorithm for sound recognition

After conducted analysis of study results, a new algorithm, the result of which is final recognized sound, was added to the program. This algorithm was used for processing of experimental sound recording and the ratio between correct recognition and a total number of the recording was determined i.e. the final probability of correct recognition. To verify the obtained results, two new voices, one male, and one female that were not used for building etalon spectra were processed.

Final probability of correct recognition for each sound is listed in Table-1. Total result for male voices is 78 correct recognitions out of 90 or 87%. There was four error when recognizing sound [ЫI] and [И]. For female voices, the result is 74 correct recognitions out of 90 or 82%. There were no errors when recognizing sound [ЫI] and [И]. The final result when considering male and female voices is 152 correct recognitions out of 180 or slightly above 84%.

**Table-1.** Results of sound recognition.

| Sound | Male | Female | Final |
|---|---|---|---|
| А | 93% | 100% | 97% |
| И | 80% | 80% | 80% |
| О | 93% | 80% | 87% |
| У | 73% | 73% | 73% |
| ЫI | 93% | 80% | 87% |
| Э | 87% | 80% | 83% |

Processing of two addition voices supports this result. When processing of additional female sound, all six vowel sounds were recognized correctly. When analyzing an additional male voice there was a false recognition of Russian sounds [И].

## 4. CONCLUSIONS

The aim of the present work was to study the possibility of implementing recognition of separate speech sounds that is independent of the speaker, base on their spectral analysis. Analysis of currently existing voice control systems had been conducted. The necessity of a multi-user system with voice contras was reviewed. A hypothesis was proposed about the spectral shape being independent of individual features of the users.
Over the course of study, the following problems were solved:

- Problem domain have been studied;
- Spectra of vowel sounds recorded from different voice have been obtained using DFT;
- The obtained data were used to develop an algorithm for recognition of vowel sounds;
- The developed algorithm was realized programmatically;
- Result for algorithm revealed that for proposed method, the probability of correct recognition is higher than 84%, which allows to conclude that proposed method can be used for development of fundamentally new approach in recognition of voice command for multi-user systems with voice control.

## REFERENCES

[1] Lyons Richard 2006. Digital Signal Processing. 2nd edition. Binom-Press, Moscow.

[2] Kupriyanov M. C., Matyushkin B. D. 2000. Digital signal processing. 2nd edition. Polytechnic, St. Petersburg.

[3] Sergienko A. B. 2006. Digital Signal Processing. 2nd edition. BHV-Peterburg, St. Petersburg.

[4] Ed. Li W. 1983. Methods of automatic speech recognition. Mir, Moscow.

[5] Zinder L. R. 1979. General phonetics. Higher School, Moscow.

[6] Zlatoustova L. V., Potapova R. K., Trunin-Donskoy V.N. 1986. General and applied phonetics. MGU, Moscow.

[7] Potapova R. K. 1989. Speech control robot. Radio and communication, Moscow.

[8] Shcherbakov P., Klymenko D., Tymchenko S. 2017. Statistical research of shovel excavator performance during loading of rock mass of different crushing quality. Scientific Bulletin of National Mining University. 1: 49-54.

[9] Andrianov I, Olevskyi V., Olevska Yu. 2016. Analytic approximation of periodic Ateb functions via elementary functions in nonlinear dynamics. AIP Conference Proceedings. 1773:1.

[10] Drobakhin O. O., Olevskyi V. I., Olevskyi O. V. 2017. Study of eigenfrequencies with the help of Prony's method. AIP Conference Proceedings. 1895: 1.

[11] Kagadiy T. S., Shporta A. H. 2015. The asymptotic method in problems of the linear and nonlinear elasticity theory. Naukovyi Visnyk Natsionalnoho Hirnychoho Universytetu. 3: 76-81.

[12] Shashenko O. M., Sdvyzhkova O. O., Babets D. V. 2010. Method of argument group account in geomechanical calculation. 12th International Symposium on Environmental Issues and Waste Management in Energy and Mineral Production SWEMP 2010, May 24-26: 488-493.

[13] Stylianou Y. 2001 Apply the harmonic plus noise model in concatenative speech synthesis. IEEE Trans. on Speech and Audio Process. 9(1): 21-29.

[14] Zavarehei E., Vaseghi S., Yan Q. 2007. Noisy speech enhancement using harmonic-noise model and codebook-based post-processing. IEEE Trans. on Speech and Audio Process. 15(4): 1194-1203.

[15] Huang N. E., Zheng S., Steven R. L. 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society of London. 454: 903-995.

[16] Tychkov A. Yu. 2013. The software solutions of the problems of the biomedical information processing. Models, systems, networks in economics, technology, nature and society. 5: 114-116.

[17] Huang E. Huang, Samuel S. P. Shen. 2005. Hilbert-Huang Transform and its application. Interdisciplinary mathematical sciences. 5: 324.

[18] Goldenstein S., Gomes J. 1999. Time warping of audio signals Computer Graphics International, Proceedings. 52-57.