ВІННИЦЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ

Факультет комп’ютерних систем та автоматики

Кафедра комп’ютерних систем управління

Спеціальність 151 Автоматизація та комп'ютерно-інтегровані технології

Освітньо-професійна програма Інформаційні системи та інтернет речей

ЗАТВЕРДЖУЮ

Завідувач кафедри АІІТ

Квєтний Р.Н.

«___» _____ 2019року

МАГІСТЕРСЬКА КВАЛІФІКАЦІЙНА РОБОТА

Розробка і дослідження методів і моделей дикторонезалежної ідентифікації фонотипів в інтелектуальних системах автоматизації

08.01.МКР.012. 00.000 ТЗ

Студент групи ЗАКІТ-18мСтецюк В.В.

Керівник: к.т.н., проф.Биков М.М.

Рецензент:к.т.н., доцент Паламарчук Є.А.

Вінниця ВНТУ – 2019

VINNYTSIA NATIONAL TECHNICAL UNIVERSITY

Faculty of Computer Systems and Automation

Computer System Control Department

Specialty 151 Automation and computer-integrated technologies

Educational and professional program Information Systems and Internet of Things

APPROVED

Head of the AIIT department

D.Sc., prof. R.N. Kvyetnyy

«___» _____ 2019year

MASTER'S QUALIFICATION WORK

Development and research of methods and models of speaker independent identification of phonotypes in intelligent automation systems

08.01.MKP.012. 00.000 ТЗ

Student of the group 3AKIT-18mStetsiukV.V.

Supervisor: Ph.D., prof.Bykov M.M.

Reviewer: Ph.D., docentPalamarchuk Y.A.

VinnytsiaVNTU – 2019

ВІННИЦЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
Факультет комп'ютерних систем та автоматики
Кафедра комп'ютерних систем управління
Освітньо-кваліфікаційний рівень магістр
Спеціальність 151 Автоматизація та комп'ютерно-інтегровані
Освітньо-професійна програма Інформаційні системи та інтернет речей

ЗАТВЕРДЖУЮ
Завідувач кафедри АІІТ
Квєтний Р.Н.
«___» ____09_____ 2019року
Протокол № 1 засідання кафедри
АІІТ від 2.09.2019р.

З А В Д А Н Н Я

НА МАГІСТЕРСЬКУ КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТУ

Стецюку Вадиму Валерійовичу

(прізвище, ім'я, по батькові)

1. Тема магістерської кваліфікаційної роботи«Розробка і дослідження методів і моделей дикторонезалежної ідентифікації фонотипів в інтелектуальних системах автоматизації»
керівник магістерської кваліфікаційної роботи БиковМикола Максимович, к. т. н., професор_____
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)
затверджені наказом вищого навчального закладу від "02" 10 2019 року № 254
2. Строк подання студентом магістерської кваліфікаційної роботи 10.12.2019 року
3. Вихідні дані до магістерської кваліфікаційної роботи Склад користувацької системи — необмежений; динамічний діапазон – 60 дБ; відношення сигналу/шум – 40 дБ; частотний діапазон мовного сигналу – [100..4200] Гц; тип мови – розділені слова, достовірність розпізнавання мови – 95%, середовища моделювання – Matlab, VisualStudio.
4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити)вступ, огляд та аналіз методів автоматизованого розпізнавання та ознакового опису мовних образів, розробка дикторонезалежних методів ідентифікації мовних образів на моделі фаззі-логічного опису та моделі квазі-частотного детектора, ідентифікація параметрів частотнодетектуючих і частотносегментуючих функцій за допомогою нейронної мережі, розробка алгоритмів та програмного забезпечення ідентифікації мовних образів, машинний експеримент.
5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень)методи виділення ознак (плакат) – 2 аркуші, моделі фазі логічного та квазичастотного методів ідентифікації фонотипів (плакат) – 1 аркуш, схеми програм обробки

сигналу та виділення ознак (креслення) – 4 аркуші, схема взаємодії програмних модулів (креслення) - 1 аркуш, результати машинного експерименту (плакат) – 3 аркушів.

6. Консультанти розділів магістерської кваліфікаційної роботи

Розділ змістової частини роботи	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
1	к.т.н, професор кафедри КСУ Биков М.М.		
2	к.т.н, професор кафедри КСУ Биков М.М.		
3	к.т.н, професор кафедри КСУ Биков М.М.		
4	к.т.н, професор кафедри КСУ Биков М.М.		

Календарний план

№ з/п	Назва етапів роботи	Строк виконання етапів роботи	Примітка
1	Вступ, огляд існуючих методів розпізнавання та опису мовних образів та обґрунтування вибраних методів. Постановка задач дослідження	02.10.2019р.	
2	Розробка і дослідження методів дикторонезалежної ідентифікації мовних образів на основі фаззі-логічної моделі та моделі квазі-частотного детектора	20.10.2019 р.	
3	Ідентифікація параметрів частотнодетектуючих і частотносегментуючих функцій за допомогою нейронної мережі.	02.11.2019 р.	
4	Розробка алгоритмів та програмного забезпечення ідентифікації мовних образів, машинний експеримент.	15.11.2019 р.	
5	Оформлення пояснювальної записки, графічного матеріалу і презентації	25.11.2019 р.	
6	Апробація результатів дослідження. Публікації	10.12.2019 р.	
7	Захист МКР	12.12.2019 р.	

Дата видачі завдання “02”092019року

Студент _____ Стецюк В.В.

Керівник магістерської кваліфікаційної роботи _____ БиковМ.М.

VINNYTSIA NATIONAL TECHNICAL UNIVERSITY
Faculty of Computer Systems and Automation
Computer System Control Department
Educational qualification level: Master's degree
Specialty 151 Automation and computer-integrated technologies
Educational and professional program Information Systems and Internet of Things

APPROVED
Head of the AIIT department
D.Sc., prof. R.N. Kvyetnyy
«___» _____ 2019year
Protocol №1of meeting of
AIITdepartment from2.09.2019
y.

T A S K

FOR THE MASTER'S QUALIFICATION WORK OF THE STUDENT

Stetsiuk Vadym V.

(Last name,first name, middle name)

1. Title of master's qualification work«Development and research of methods and models of speaker-independent identification of phonotypes in intelligent automation systems»
Head of the master's qualification workBykov Mykola M., Ph.D.,professor
(Last name, first name, middle name, degree, academic status)
approved by order of higher education institution from "02" 10 2019year №254
2. Deadline for submission of master's qualification work 10.12.2019 year
3. Baseline data for master's qualification workThe composition of the user system — unlimited; dynamic range – 60 db; signal ratio /noise – 40 db; frequency range of the speech signal – [100..4200] Hz; type of speech– separated words, authenticity of speech recognition– 95%, modeling environment – Matlab, VisualStudio.
4. Contents of the explanatory note (list of questions to be worked out)introduction, review and analysis of methods of automated recognition and feature description of speechpatterns, development of speaker-independent methods of identification of speech patterns on the model of fuzzy-logical description and model of quasi-frequency detector, identification of parameters of frequency-detecting and frequency-segmenting functions using a neural network, development of algorithms and software for speech identification, machine experiment.
5. List of graphic material (with exact indication of the drawings)methods of selection of features (poster) - 2 sheets, models of fuzzy logical and quasi-frequency phonotype identification methods (poster) - 1 sheet, schemes of signal processing and feature selection (drawing) - 4 sheets, scheme of interaction of program modules (drawing) - 1 sheet, machine experiment results (poster) - 3 sheets.

6. Sections consultants of master's qualification work

Content section of the work	Last name, initials and academic status	Signature, date	
		Provided by	Received by
1	Ph.D., professor pf CSC Department Bykov M.M.		
2	Ph.D., professor pf CSC Department Bykov M.M.		
3	Ph.D., professor pf CSC Department Bykov M.M.		
4	Ph.D., professor pf CSC Department Bykov M.M.		

Schedule

№	The name of the stages of work	Deadline	Notes
1	Introduction, overview of existing methods of recognition and description of speechpatterns and justification of the selected methods. Formulation of research objectives	02.10.2019y.	
2	Development and research of methods of speaker-independent identification of speech patterns based on the fuzzy-logic model and the model of quasi-frequency detector	20.10.2019 y.	
3	Identification of parameters of frequency-detecting and frequency-segmenting functions using a neural network	02.11.2019 y.	
4	Development of algorithms and software for speech patterns identification, machine experiment.	15.11.2019 y.	
5	Explanatory note, graphic material and presentation	25.11.2019 y.	
6	Testing the results of the study. Publications	10.12.2019 y.	
7	Master’s work defense	12.12.2019 y.	

Assigned“02”092019year

Student _____ Stetsiuk V.V.

Head of master's qualification work _____ __BykovM.M.

Анотація

Магістерська кваліфікаційна робота присвячена розробці та дослідженню методів і моделей дикторонезалежної ідентифікації фонотипів в неперервному мовному сигналі в інтелектуальних системах автоматизації.

Розроблено математичні моделі, алгоритм тапрограмне забезпечення дляаналізумовних сигналів, а саме: фільтрації,знаходження енергії сигналу, кореляцій енергій, виділення інформативних ознак комбінованим фаззі-логічим методом.

Розроблено метод, алгоритми та програмне забезпечення для дикторонезалежного опису мовних образів,що базується на основі моделі "квазічастотного модулятора".Розроблено метод для сегментування неперервного сигналу, що базується на основі моделі "квазічастотного" сегментатора.

Результати роботи можуть бути застосовані у системах розпізнавання мови для вдосконалення методу автоматичного виділення фонотипів.

Annotation

Master's qualification work is devoted to the development and research of methods and models of speaker independent identification of phonotypes in continuous speech signal in intelligent automation systems.

Mathematical models, algorithm and software for analysis of speech signals have been developed, namely: filtering, finding of signal energy, correlations of energies, isolation of informative features by the combined fuzzy-logic method.

A method, algorithms and software for speaker independent description of speech patterns based on the "quasi-frequency modulator" model have been developed. A method for segmenting a continuous signal based on the "quasi-frequency" segmenter model has been developed.

The results of the work can be applied in speech recognition systems to improve the method of automatic selection of phonotypes.

CONTENT

INTRODUCTION

Actuality of topic. For many decades, and especially in recent years, due to the development of technology and systems for automatic recognition and synthesis of speech, the acoustic characteristics of speech signals are studied, and attempts are made to establish a connection between the acoustic parameters and the phonetic features of speech signals, namely attempts to understand how the brain, by receiving information about the nature of changes in sound pressure over time, receives information about the content of speech.

Recognition is the task of converting input information, which considers some of the parameters and features of recognizable patterns to output, which is a conclusion to which class this pattern belongs. Therefore, given that cybernetics is the science of general laws of information transformation in complex systems, pattern recognition is one of the branches of this science.

Requirements for complex mechanization and automation of production, creation of robots, within wide limits of solving problems of technical and medical diagnostics, economic and sociological phenomena and processes, determination of the most probable directions of their transformation and determined considerable efforts of scientific thought aimed at solving theoretical and applied problems of recognition problem.

One of the most important and poorly formalized speechrecognition problems today is the task of automating the process of extracting informative parameters to describe speech patterns. It is especially relevant when parameters are sought in the form of signal energies in different frequency bands, since it is necessary to analyze a very large number of variants of the frequency range of the speech signal into subbands. Therefore, the problem of software development for parametric identification of sounds in the speech recognition modules of these systems, which is solved in this paper, is quite relevant.

Relationship with scientific programs, plans, topics. The work was performed according to the plan of the departmental research work №52K4 “Decision-making methods based on the principles of computational intelligence”.

The purpose of this work is to increase the efficiency of the process of parametric identification of sounds in the module of recognition of sounds of automation systems by developing methods and algorithms for automation of the process of distinguishing features and dictation-independent description of speech patterns.

To achieve this purpose it is necessary to solve the following tasks:

– development of a combined correlation-logical method for automation of the process of selection of informative features of speech patterns;

– development of algorithms and programs for automation of selection of informative features on the basis of the offered method;

– development of a method of announcer-independent description of speech patterns on the basis of the “quasi-frequency modulator” model and a method of segmentation of a continuous signal based on the model of “quasi-frequency” segmenter;

– identification of the parameters of the “quasi frequency” segment of the torus on the neural network model;

– development of algorithms and programs for implementation of "quasi-frequency" description and segmentation of the speech signal.

The object of study is the speechsignal.

The subject of the study are models, methods and algorithms for automatic selection of features of speech sounds and their speaker independent description.

Research methods. In the course of the study applied: information theory; theory of digital signal processing; recognition theory; theory of algorithms; computer simulation for analysis and validation of theoretical findings.

Scientific novelty of the obtained results.

1. For the first time, a method of distinguishing features of speech patternsis proposed, based on the use of the alpha level of the correlated energy of the speech signal and the implicit matrix of spectral components, which allows automating the process of selection of informative features of the sound type.

2. For the first time, an announcer-independent description of speech patternshas been proposed based on a "quasi-frequency modulator" model, which allows to increase the accuracy of identification of sound types and to reduce the amount of descriptive data by an order of magnitude.

3. The model of the neural network sets the values of the weight coefficients of the parameters of the frequency-segmenting function, which makes it possible to automate the process of segmentation of the speech signal into speech patterns.

The practical value of the work is in the development of software that automates the process of extracting the features of speech sounds necessary for their parametric identification, and forms a speaker independent description of speechpatterns for their segmentation and recognition in a continuous speech signal.

The validity of the theoretical provisions of the master's qualification work is confirmed by the rigorous formulation of problems, the correct application of mathematical methods in the proof of scientific provisions, the strict derivation of analytical relations, the comparison of the results obtained by the methods developed in the work with the known, and the convergence of the results of mathematical results during the implementation of the developed software and hardware.

Publications. On the subject of the study published work in the repository of VNTU.

1 VARIANT ANALYSIS AND TECHNICAL AND ECONOMIC REASONING OF THE SELECTED METHODS OF THE PROBLEM

1.1 Analysis of methods of selection of features of speech patterns

1.1.1 Essence of the technical problem that arose at the present stage of development of science

Automatic speechrecognition systems refer to information and measurement systems. Their peculiarity is that the acoustic speech signal to be analyzed contains a considerable amount of unnecessary information. Shortening the information by simply processing the signal, such as limiting the frequency range or compressing the dynamic range does not allow you to get the results you want. Therefore, in order to effectively solve the problem of recognition of speech patterns, it is necessary to move from the n-dimensional of observation $Y= \{y_1, . . . , y_n\}$ acoustic signal to the m-dimensional feature $X=\{x_1, . . ., x_m\}$, containing only relevant information for the classifier. The requirement to fulfill the condition $m << n$ is understandable. Choosing a set of information features is a key recognition issue, as no classification procedure can compensate for the inadequacy of the feature system. The problem of choosing features must be related to the overall performance of the system.

One of the most important and poorly formalized speechrecognition problems today is the task of automating the process of extracting informative parameters to describe speech patterns. It is especially relevant when parameters are sought in the form of signal energies in different frequency bands, since it is necessary to analyze a very large number of variants of splitting the frequency range of the speech signal into sub bands [1,2,3,4].

The main problems in the development of the system of recognition of speech patterns are related to the considerable variation of the parameters of the speech signal, which is explained by the different parameters and form of the speech tract of the speakers, defects in pronunciation and other features of the speaker's speech.

The main problems that arise with automatic speechrecognition are the following [5,6,7]:

– the word can be spoken at any speed. Because of this, the duration of the spoken word does not match the duration of the corresponding standard;

– incorrect word boundaries;

– background noise can override portions of the speech signal, which necessitates a combination of speech recognition systems and interference reduction systems;

– speaker-independence of many speech recognition systems;

– when using large-volume dictionaries, there is a problem with the cost of training and storing models (standards) for recognition.

Existing recognition systems have already achieved sufficiently high authentication accuracy. The main trends for developing new systems are:

– increasing the speed of the recognition process that can be achieved by using a neural network approach;

– increasing the vocabulary of words to be recognized, which requires a new method of pattern encoding;

– optimization of the vector of speech signal features used in recognition, that is, creating a new model of signal description;

– the transition from single word recognition to fusion speechrecognition that can be achieved by using formal grammar theory.

Therefore, a speech signal is a complex signal that is hierarchically organized so that simpler single-level patterns are uniquely matched into more sophisticated next-level patterns. Thus, according to some complex code, a multilevel hierarchical system with a fault-tolerant structure that carries information is being built [8].

1.1.2 Approaches to solve a technical problem

An analysis of existing recognition work shows that the interpretation of the term "sign" is often quite arbitrary, resulting in some cases it is confused with the

concept of "the way the sign is presented." As a result, different ways of representing the same property resulting from different transformations are used as separate features. The definition of the pattern attribute has caused a great deal of controversy, so there is still no clear definition in the theory of pattern recognition. The reason, obviously, is that the notion of a sign is inextricably linked to the notion of the plural, nor can it be initially defined because of the more general concepts. Therefore, it is possible to give only its description, thus defining the characteristic features of the trait.

Developing a truly effective and speaker-independent speech recognition system requires two main tasks: the first is the selection of a reduced set of informative speech signal parameters that is used to construct the feature space; the second is the choice of rational presentation (description) of the traits obtained [9,10].

Very often, automatic speech recognition systems use spectral transform parameters, which are obtained in the form of signal intensities in a number of frequency channels, to construct a character description. The use of these parameters is based on a linear model of speech production.

The analysis of the existing methods of recognition of speech signals and systems, the work of which is based on the use of these methods, revealed the main problems and disadvantages inherent in the methods of recognition, as well as problems arising in the operation of these systems [11].

The vast majority of AWP systems implement word recognition, although there are some attempts at post-phonemic analysis [12, 13]. When sequentially recognized, it is necessary to increase the number of recognizable words twice, to increase the memory capacity to store word samples twice, and to increase the memory and processor speed four times, which is usually faced with known limitations, both in terms of memory and performance (see Table 1.1).

Table 1.1 - Difficulty classes of solved SARM tasks

Purpose of the system	Dictionary volume	Speechrestrictions	Limitations on grammatical structure	Announcer	Noise properties
1. Isolated word recognition	10 – 60	team	–	trained	noise of cars
2. Recognizing phrases from isolated words	50 – 200	speechfrom a limited area	high (grammar grid)	–“”–	quiet room
3. Recognition of merging phrases	50 – 200	–“”–	–“”–	–“”–	–“”–
4. Understanding merged phrases	50 – 200	close to natural	medium	–“”–	–“”–
5. Understanding of merged speech	not limited	natural	low	arbitrary	real environment

In order to give satisfactory results, features for describing speech patterns should be distinguished by transformations based on the principles of acoustic signal formation in the vocal tract. Ways of presenting features should take into account mechanisms of speech auditory system that are invariant to the interference.

The means of presentation of the features must satisfy the following requirements [14,15,16]:

1) to report features of the invariance property of interference of various kinds; 2) not to diminish its information content; 3) the cost of separating the trait must be within the allowable range. In part, these requirements can be fulfilled if some principles of operation of the auditory system are used to describe a speech signal, for example, to produce a frequency analysis of a signal in the bands corresponding to critical hearing bands, to distribute energies in different frequency bands in the form of a logarithm of the ratio of these energies, to use the effect disguise. Smoothing

real-time data with high-speed algorithms and devices is also effective to increase the invariance of obstacle features.

But the existing ways of solving the problem of identification and segmentation of the speech signal are complex and imperfect, so the task of this thesis is to create a method and develop a software and device for its implementation, which would allow to effectively and simply identify informative features for the selection of syllables in the speech signal.

1.1.3 Rationale and choice of baseline

The basic analogue is the method of selecting informative features with the teacher. The purpose of this analogue is to reduce more data to some of the most informative parameters that can be used to calculate similarities. One of the known methods of obtaining the desired parameters is the use of Fourier analysis, but it requires more time. For convenience, the preprocessor program should generate the desired parameters close to real time. It would also be desirable that it was logically simple so that it could replace an equivalent specialized device, then the whole algorithm would be considered satisfactory.

The purpose of the primary segmentation procedure is to group together similar complex minimum segments. The acoustic similarity of the two segments can be determined as follows: Let each minimum be represented by n-parameters defined by an n-dimensional vector in space. Sometimes determine the closeness of two segments in terms of Euclidean distance between two points. However, such simply defined metrics are unsatisfactory in speechstudies.

The purpose of secondary segmentation is to correct possible primary segmentation errors by taking into account cross-segmental parameter changes. Each time a segment is created, the total change for each parameter is remembered. If these changes exceed a certain limit, the segment is divided into smaller segments. The boundary calculation is based on the same assumptions as the primary segmentation and depends on the type of parameter, the parameter value and the length of the segment.

This method requires a lot of calculations, calculations, assumptions, which leads to inaccuracy of choice of features, low reliability and very time-consuming. Compare the technical parameters of the new method with the analogue and present them in the form of table 1.2.

Table 1.2 - Comparison of the technical characteristics of the new method and analogue

Technical indicators	New method	Analog
Accuracy, %	99,8	75
Reliability, %	99	80
The amount of time spent, min	5	90

Therefore, the table shows that by all the technical characteristics of the newly developed method has much better characteristics, which indicates the feasibility of its further development and implementation.

1.1.4 Specification of technical requirements for the design object

The object of study in this system is a speech signal, the means of information is a personal computer, and the object of automation is the process of processing a speech signal in order to segment it into separate sounds and produce informative ones, that is, in the form of a sequence of sounds.

There are no standards and specifications for this new development. Such technology can be used in automatic speech recognition systems, which in turn can be used in various voice control systems. For example, in the management of work at a distance by commands.

Let's specify the input and output data of this software:

1) Characteristics of the input data:

Initial data that is loaded for signal processing into audio is audio files that must be in .wav format, with a sound file length limitation of 2000 (0.2 s) to 120,000 (12 s).

The audio file is uploaded to the signal processing subroutine, where it is filtered, its energy is calculated and correlation is made.

2) Characteristics of the source data:

The output of the whole system is informative features, that is, time intervals that correspond to specific sounds and correlations on them, the energies of which are informative for the data intervals of the audio signal.

Software Requirements:

1) Input files with the extension ".wav";

2) Availability of installed MatLab software.

The main problems with the development of the universal program are related to the considerable development of the parameters of the speech signal, which is explained by the different form and length of the speech tract of the speakers, defects in pronunciation, dialectal features of the speakers, etc. The developed system will reduce these shortcomings by automatically segmenting the speech signal into individual sound types and highlighting informative features for these sound types.

1.1.5 Stages of solving the problem of automation of selection of informative features

The main objective of the thesis is to develop a method for automating the selection of informative features of speech patterns.

To automate the process of finding informative parameters, a combined fuzzy-logic approach is proposed. His idea came from an analysis of a model of the human hearing system, supplemented by a neural network.

To implement such a process, a number of such sequential actions must be performed:

– frequency band formation: select their number, center frequency, bandwidth, filter type;

– filtering;

– energy calculation;

– correlation;

– finding α-levels for each correlation and forming "1" intervals;

– finding those "1" overlapping intervals;

– forming an implicit matrix for them;

– formation on the implicit matrix of ranges whose energies will be informative (by the operation of a conorm).

Therefore, one of the objectives of the study is to develop a method for automating the selection of informative features for known patterns (in this case, spectral components for given speech sounds). With the introduction of a new software product, the automatic speechrecognition system will be much faster to recognize due to the rapid selection of informative featuresof speech patterns.

1.2 Existing ways to solve software for voice recognition systems

1.2.1 Essence of the technical problem

One of the possible ways to solve the problem of the dictator's independence of the speechrecognition system is to choose the speechelements of recognition, which are characterized by the parameters of the speech signal, which are largely independent of the individual features of the speaker. Such parameters include component segments, phonetic classes (their presence and number, location and order of follow) [17,18,19]. The disadvantage of this approach is the limitation of its use only for dictionaries of small volume.

Another approach to the problem of speech recognition of an arbitrary speaker is based on the adaptation of the device of speech input to the speaker due to the transformation of standards in the implementation of speech. However, it has not yet been possible to achieve high accuracy in recognition.

More precision is given by the use of many standards for one and the same element of a recognizable speech, where each standard corresponds to a separate

speaker. Although this approach can compensate for the impact on recognition of most of the factors that determine individual differences in demand, however, the increase in hardware and computing costs for its implementation makes the high cost of speech recognition systems and thus makes their application limited.

For this reason, it is urgent to construct an effective feature description of speech signals that is invariant to the features of the speakers.

The closest analogue to the software being developed will be the voice-independent voice command software developed by Microsoft [20]. The given system allows to recognize words its dictionary makes 264 words, recognition accuracy 97-98%. The cost of the software of this system is 45 UAH.

Compared to this analogue, the developed software has the following advantages:

– speaker-independence of the system is achieved through better methods of recognition;

– higher recognition accuracy, which is also related to the use of the proposed recognition methods;

– significantly less development costs for this product.

Therefore, the main purpose of this master's thesis will be to solve one of the above problems by developing methods, algorithms and software for describing speech signal with dictation-independent features that can be used to identify speech patterns in the development of dictation-independent speech recognition systems in intelligent automation systems.

1.2.2 Comparative analysis of the proposed version of the dictator-independent description with the existing ones

A joint analysis of hearing models and speech models allows us to formulate the following basic points regarding the methods of isolation and processing of speech signal parameters [21]:

1) The most informative are the conversion of the speech signal in the form of frequency-time-energy.

2) The frequency range of the spectrum is sufficient to effectively describe all the sounds of the speech, and takes a band of approximately 200 to 6400 Hz.

3) The invariance of the characteristic description to changes in the tone of the speech signal can be ensured by placing the center frequencies of the spectral analyzer subbands according to the logarithmic law.

4) The spectral analyzer frequency channels should be selected in proportion to the center channel frequency and approximately equal to the width of the corresponding critical hearing band.

5) The invariance of the energy traits to the change in the intensity of the speech signal can be ensured by the introduction of nonlinear devices for controlling the amplification of the signals.

6) The main auditory features of speech patterns are spectral and temporal enveloping sounds, and they are similar in the form of formant frequencies and their amplitudes, and the dynamics of changes in formant frequencies.

Despite the known progress that has been made in solving the problem of reducing the characteristic space when used, to describe speechsegments, spectral sections, correlation coefficients, linear prediction coefficients, such descriptions are quite voluminous [22]. For example, when speechis represented by digital readings at 6000 Hz. with 10 bits of memory, which takes up a 1-second speech signal, is:

$$Q = 6000 \cdot 10 = 60 \text{ kBit /s.} \quad (1.1)$$

When describing a speech segment of the same length by spectral slices, the amount of memory will be:

$$Q_S = \frac{1}{\tau_S} \cdot N_{АД} \cdot П_{АД}, \quad (1.2)$$

whereτ_S — the duration of the primary segment, τ_S =10 msec.;

$N_{АД}$ —digit ofADT;

$n_ф$ —the number of bandpass filters in the spectral analyzer,

most often $n_\phi \approx 12$.

$$Q_S = \frac{1}{0.01} \cdot 10 \cdot 12 = 12 кБіт/сек. \quad (1.3)$$

Therefore, in the master's thesis, the second method proposed will be a method of describing a speech signal, which is based on the principle of "quasi-frequency" detection of sounds from a speech signal. In this approach, the speech path is represented by a three-channel frequency modulator, in which the first channel corresponds to the band of frequencies of the I formants, the second channel - the band of frequencies of the II formats, and the third channel - the band of frequencies of the III formats. In this case, each frequency channel is coded in three values of modulation frequency - low, medium, high, which correspond to the position of the frequency moment, which requires two bits to display. In addition, the energy ratio in these three channels, which requires two more bits, is coded to determine the global spectrum of the signal. Thus, the description of the primary speechsegment using such a "quasi-frequency" detector is an eight-bit binary word. In order to form a fungible function, this description is supplemented by an eight-bit binary word describing the differential of the spectrum in discrete time.

Thus, the amount of memory required to describe a speechsegment lasting 1 second by this method is equal to:

$$Q = \frac{1}{\tau} \cdot 16 = \frac{1}{0.01} \cdot 16 = 1{,}6 \frac{Кбіт}{сек}, \quad (1.4)$$

which is below the methods discussed above. Thus, the chosen method gives a gain in hardware costs when building a developed device.

1.3 Overall conclusion about the feasibility, relevance and cost-effectiveness of the new development

As a result, of the analysis of the existing methods of recognition of speech signals and systems, it was found that speech recognition is a very relevant topic today, which scientists and engineers in many countries of the world are working on.

It uses different approaches to recognizing speech patterns and highlighting informative features. Also identified were the main problems and shortcomings inherent in the recognition methods themselves, as well as problems encountered in the operation of these systems.

Existing recognition systems have already achieved sufficiently high authentication accuracy. The main trends for developing new systems are:

– increasing the speed of the recognition process that can be achieved by using a neural network approach;

– increasing the vocabulary of words to be recognized, which requires a new method of patterns encoding;

– optimization of the vector of speech signal features used in recognition, that is, creating a new model of signal description;

– the transition from single word recognition to fusion speechrecognition that can be achieved by using formal grammar theory.

Nevertheless, the existing methods are imperfect and not clear enough. Therefore, the main purpose of this thesis is to develop a more sophisticated method of automatic selection of informative features of speech patterns and dictation-independent description method based on them.

In order to solve the problem of this work, we propose to use a combined fuzzy-logic method and the method of "quasi-frequency detection", implemented by a software product, which will allow implementing the process of processing a speech signal with the purpose of its segmentation into separate sound types and isolation of informative features of these sound types.

2 ANALYSIS OF METHODS OF AUTOMATED WORD RECOGNITION AND LABEL DESCRIPTION OF SPEECH

2.1 Development and justification of the structural scheme of the system

The problem of constructing an effective feature system is related to the solution of two main problems: choosing a short set of informative speech signal parameters that are used to construct the feature space, and choosing a rational representation (description) of the features obtained.

Let's present the simplest block diagram of the pattern recognition system shown in Figure 2.1:

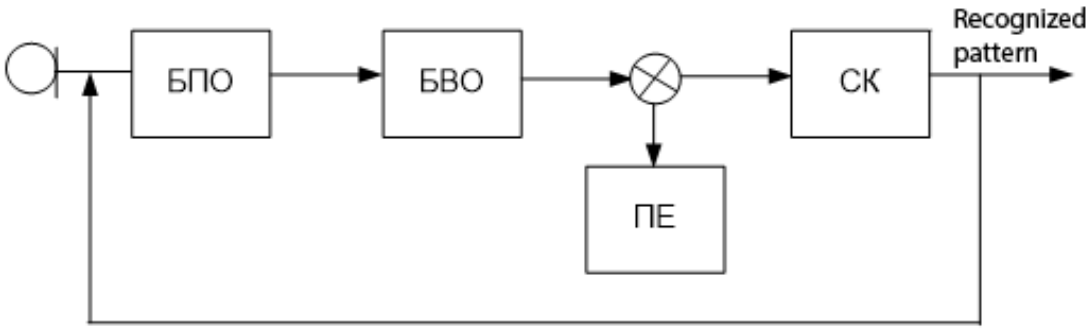


Figure 2.1 - Structural diagram of the pattern recognition system

The following abbreviations have been used in this block diagram:

БПО - signal pre-processing unit;

БВО - block of selection of features;

СК - comparison scheme;

ПЕ - the standard memory;

СК is a classifier scheme.

The input of this circuit, that is, into the microphone, is a signal that is subjected to processing (filtering, recognition) using the БВО, which then goes to the БПО, where it is allocated the necessary features. Then with the help of the

comparison scheme, some features are compared to the memory of the standard, and some come to the classifier, where we get the pattern we need.

When it comes to speech recognition, it is not a standardized system [22, 23, 24]. Speech recognition systems have many different uses. Figure 2.2 shows the structure of possible differentiation of such systems.

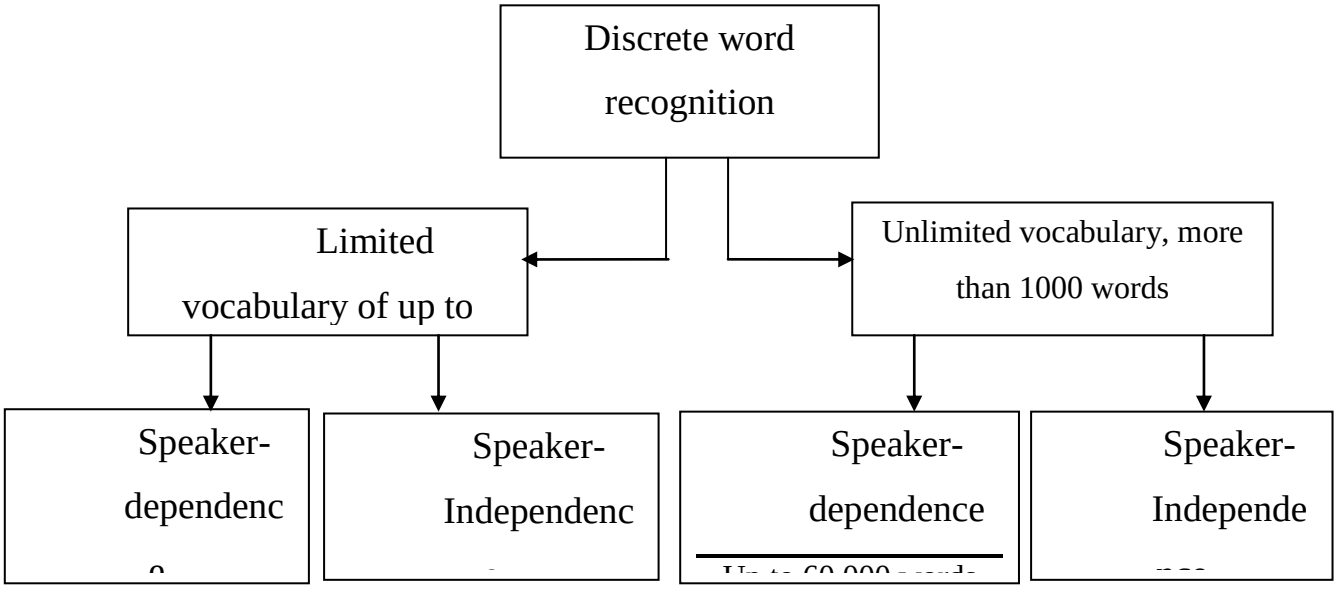


Figure 2.2 - Structure of discrete speech recognition systems

Isolated word recognition systems are not suitable for everyday use in various offices, hospitals or law firms. Such customers place the basic requirement on an automated system as a large (possibly unlimited) vocabulary of words that has a specific speech context [1]. It is easier for a person to recognize words individually, but for computers it is much more difficult to solve this task, since it is necessary to structure a continuous flow of speech or to split whole words into sound types. Due to the rapid growth of microprocessor technology in recent decades, some laboratories have already set the task. Figure 2.3 shows the structure of automatic continuous speech recognition systems.

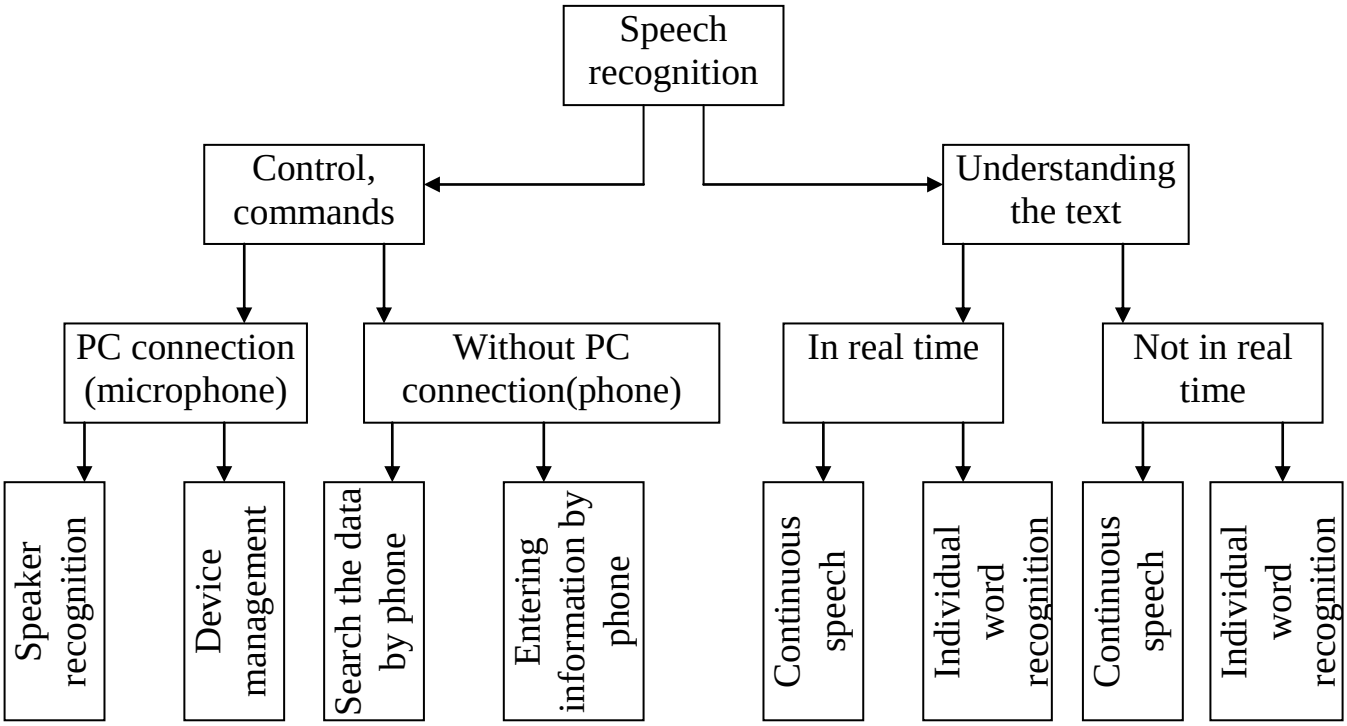


Figure 2.3 - Differentiation of speech recognition

All existing speech recognition systems are based on two ideas that are accepted by most existing quantitative and qualitative theories of speech[25]. The first fundamental hypothesis: the direct form of information transmission by a speech signal is realized in temporal changes of its short-term amplitude spectrum. It is also generally accepted that some of the information is transmitted by time intervals between speech acts and other prosodic features, such as pitch and sound intensity.

In existing speech recognition systems, the spectrum is usually specified by Fourier coefficients, the estimate of the number of intersections by zero, or the parameters of some local signal model, such as the linear prediction coefficients or their conversion. With such approaches, it was found that such features are unstable in the presence of noise, and also depend on the speaker and the transmission channel.

The second type of parameter that is often used to describe a speech signal is the linear prediction coefficients that are determined by the linear prediction model. A number of parameters obtained from non-spectral transformations, such as the

number of zero cross-sections of a signal, reflect its frequency properties and also correspond to the named models of speech.

In most of the speech recognition systems considered, the characteristic description is obtained by constructing a vector space, the coordinate axes of which are selected by various subsets of the above parameters of the speech signal [26].

The most important problems that need to be addressed when creating speech recognition systems are speaker normalization, development and research of analog and digital signal conversion and feature isolation, more accurate segmentation and phonetic markup, noise control, real or near expression processing time.

The main purpose of this thesis is to solve one of the above problems - to develop a software for the pre-processing of the speech signal, which will provide the isolation of the informative features of the speech signal and convert this information to a form that is accepted for further processing.

2.2 Theoretical research

2.2.1 Parametric description of speech signals

The speech signal has a dual nature - on the one hand, it is a normal acoustic signal, which is the process of propagating the energy of acoustic vibrations in an elastic medium. Like any acoustic signal, it can be represented in the form of sound waves, representing the propagation of processes of compression and discharge of particles of the environment, the shapes of the edges of which depend on the properties of the source and propagation conditions. Therefore, like other acoustic signals, speechis characterized by a certain set of objective characteristics: dependence of sound pressure on time (temporal structure of the sound wave), duration of sound, spectral composition, location of the source in space, etc. [27].

On the other hand, speech as a physical phenomenon causes certain subjective auditory sensations (loudness, altitude, timbre, localization, masking, etc.).

The speech signal undergoes the same processing procedure in the auditory system as any other acoustic signal, that is, on the basis of its analysis the same

auditory sensations are formed, for example, the perception of speech in a completely unfamiliar speechis no different from the perception of the surrounding acoustic information - noise, whistling, etc. However, if a person perceives words in a speechin which he or she has been previously taught, then, along with the processing of purely acoustic information (loudness, pitch, timbre, etc.), there is a phonetic, followed by semantic decoding of information [28].

One method of describing speech signals is parametric.

Most speechprocessing methods suggest that the properties of the speech signal are slowly changing over time. This assumption leads to methods of short-term analysis in which segments of the speech signal are separated and processed as if they were short sections of individual sounds with different properties. The procedure is repeated as often as required. Segments, sometimes called intervals (frames) of analysis, usually intersect. The result of processing at each interval is a number or a set of numbers. Consequently, such processing results in a new time-dependent sequence that can be a characteristic of the speech signal [29].

Most methods of short-term analysis can be described by expression:

$$Q_n = \sum_{m=-\infty}^{\infty} T[x(m)]\varpi(n-m). \qquad (2.1)$$

The speech signal is converted to T [•], linear or nonlinear, which may depend on some control parameter or set. The resulting sequence is then incremented by a sequence of values of a time window positioned in time according to the index n. The results are then summed by all non-zero values. Of course, though not always, the sequence of values of a temporary window has a finite length.

The value of Qn is thus the "weighted" average value of the sequence T[x (m)].

Short-term energy is determined by expression:

$$E_n = \sum_{m=n-N+1}^{n} x^2(m)\,. \qquad (2.2)$$

Thus, the short-term energy at time n is simply the sum of squares of N frames from n-N+1 to n. It can be seen from (1.1) that in (1.2) T [•] is simply a squaring operation, and

$$\varpi(n) = \begin{cases} 1, 0 \le n \le N-1, \\ 0, \text{ otherwise} \end{cases} \quad (2.3)$$

When processing signals in discrete time, consider that if two consecutive readings have different features, then the transition through zero occurred. The frequency of occurrence of zeros in the signal may serve as the simplest characteristic of its spectral properties. This is most true for narrowband signals [30]. For example, a sinusoidal signal with frequency Fo subjected to sampling with frequency F_s has F_s/Fo counts over a period. Each period contains two zero transitions, thus the average number of zero transitions over a large interval of time.

$$z=2F_0 /F_s. \quad (2.4)$$

The average number of zero transitions can be taken as a suitable estimate of the frequency of sinusoidal oscillations [30].

The speechsignal is broadband and therefore the interpretation of the mean number of zero transitions is less obvious. However, it is possible to obtain rough estimates of the spectral properties of a signal based on the use of the mean zero-crossing function for the speech signal; consider the method of calculating this value [30]. Determine the average number of conversions after zero:

$$Z_n = \sum_{m=-\infty}^{\infty} \left|\text{sgn}[x(m)] - \text{sgn}[x(m-1)]\right| \varpi(n-m), \quad (2.5)$$

where

$$\text{sgn}[x(n)] = \begin{cases} 1, & x(n) \ge 0, \\ -1, & x(n) < 0 \end{cases} \quad (2.6)$$

and

$$\varpi(n) = \begin{cases} 1/2N, 0 \le n \le N-1, \\ 0, \text{ otherwise} \end{cases} \quad (2.7)$$

The operations included in (1.2) are presented as a block diagram in Figure 2.4.

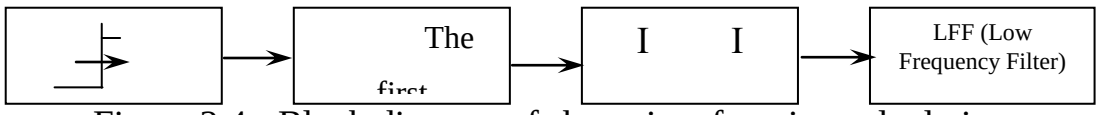


Figure 2.4 - Block diagram of short-time function calculation

average zero crossing

Such a representation shows that the function of the average number of transitions through zero has the same general properties as the functions of energy and mean [30].

The amplitude of the speech signal changes significantly over time. In particular, the amplitude of the un vocalized segments of the speech signal is much smaller than the amplitude of the vocalized segments. Such changes in amplitude are well described by the short-wave energy function. In the general case, the energy function can be defined as

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)\varpi(n-m)]^2 . \qquad (2.8)$$

This expression can be written as:

$$E_n = \sum_{m=-\infty}^{\infty} x^2(m)h(n-m), \qquad (2.9)$$

where

$$h(n) = \varpi^2(n). \qquad (2.10)$$

Equation (2.9) can be interpreted according to Figs. 2.5 a. In this case, the signal $x^2(n)$ is filtered by a linear impulse response system h(n) [30].

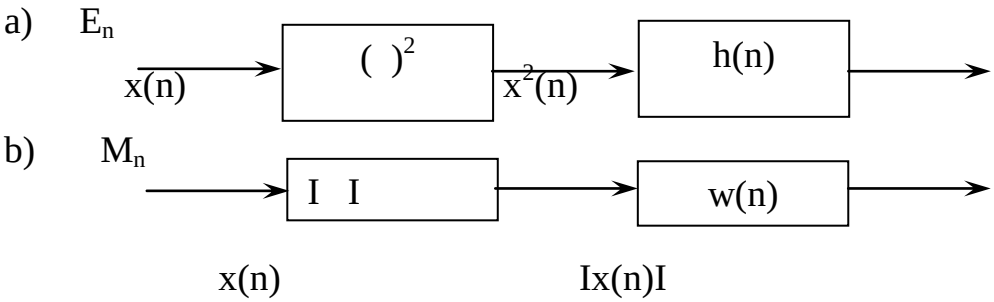


Figure 2.5 - Block diagram for the function of: a) short-lived energy; b) short-term average

The choice of the impulse response h(n) or the window forms the basis for the description of the signal by means of the energy function. To understand how the choice of a window affects the function of a short-lived signal energy, suppose that h(n) is sufficiently long and has a constant amplitude; the value of E_n will thus

change slightly over time. Such a window is equivalent to a low pass filter with a narrow bandwidth [30]. The low pass filter band should not be so narrow that the output signal is constant, in other words, the band must be chosen so that the energy function reflects the change in the amplitude of the speech signal. The described situation expresses contradictions, which often arises when studying the short-term characteristics of speech signals. Its essence is that in order to describe the rapid changes in amplitude it is desirable to have a narrow window (short impulse response), but too small window width may lead to insufficient averaging and, therefore, insufficient smoothing of the energy function [30].

The effect of window type on the calculation of the time-varying signal energy can be illustrated by using the two most common windows:

- rectangular

$$h(n) = \begin{cases} 1, 0 \le n \le N-1, \\ 0, \text{ otherwise} \end{cases} \tag{2.11}$$

- Heming windows

$$h(n) = \begin{cases} 0{,}54 - 0{,}46 \cdot \cos(2\pi n/(N-1)), 0 \le n \le N-1, \\ 0, \text{ otherwise} \end{cases} \tag{2.12}$$

A rectangular window corresponds to the case when all the counts in the interval from (n-N+1) to n are assigned the same weight. The frequency response of a rectangular window (impulse response (2.11)) is equal to

$$H(e^{i\Omega T}) = \frac{\sin(\Omega NT/2)}{\sin(\Omega T/2)} e^{-i\Omega T(N-1)/2}. \tag{2.13}$$

One of the drawbacks of the short-time energy function due to expression (1.9) is that it is sensitive to large signal levels (since in (2.9) each count is squared. As a result, the correlation between the values of the sequence x (n) is significantly distorted. is the transition to determining the mean function in the form

$$M_n = \sum |x(m)| \varpi(n-m), \tag{2.14}$$

where instead of the sum of squares, the sum of the absolute values is calculated. Figure 2.5b shows how (2.14) can be represented by linear sequence filtering $|x(n)|$. The exclusion of the quadratic operation simplifies the arithmetic calculations [30].

The analysis of the existing methods of recognition of speech signals and systems, the work of which is based on the use of these methods, revealed the main problems and shortcomings inherent in the methods of recognition, as well as problems that arise in the process of operation of these systems.

The main problems that arise with automatic speechrecognition are the following:

- the word can be spoken at any speed. Because of this, the duration of the spoken word does not match the duration of the corresponding standard;

- incorrect setting of word boundaries;

- additional difficulties due to the presence of coarticulation effect in continuous speech;

- background noises may overlap parts of the speech signal, which necessitates a combination of speech recognition systems and interference reduction systems;

- speaker independence of many speechrecognition systems;

- when using large-volume dictionaries, there is a problem with the cost of training and storing models (standards) for recognition.

Existing recognition systems have already achieved sufficiently high authentication accuracy. The main trends for the development of new systems are:

- increasing the speed of the recognition process, which can be achieved by using a neural network approach;

- increasing the vocabulary of words to be recognized, which requires a new method of pattern encoding;

- optimization of the vector of speech signal features used in recognition, ie creation of a new model of signal description;

- the transition from word recognition to fusion speech recognition, which can be achieved through the use of formal grammar theory [28].

But the existing methods of distinguishing features are imperfect and not sufficiently clear. Therefore, the main purpose of this master's thesis is to create a more sophisticated method of automatically distinguishing features of speech patterns[21].

2.2.2 Characteristics of informative features

Automatic speech recognition systems refer to information and measurement systems. Their peculiarity is that the acoustic speech signal to be analyzed contains a great deal of unnecessary information. Shortening the information by simply processing the signal, such as limiting the frequency range or compressing the dynamic range does not allow you to get the results you want. Therefore, to effectively solve the problem of recognition of speech patterns, it is necessary to move from the n-dimensional observation space $Y=\{y_1, \ldots, y_n\}$ of the acoustic signal to the m-dimensional space of features$X=\{x_1, \ldots, x_m\}$ containing only relevant information for the classifier. The requirement to fulfill the condition $m << n$ is understandable. Choosing a set of information features is a key recognition issue, since no classification procedure can compensate for the inadequacy of the feature system. The problem of choosing features must be related to the overall performance of the system. The master's thesis analyzes and classifies the features used in recognition systems, proposes a method of evaluating their information content and discusses the method of selecting a feature set that informs the recognition system the maximum value of the efficiency criterion.

An analysis of existing recognition work shows that the interpretation of the term "sign" is often quite arbitrary, resulting in some cases it is confused with the concept of "the way the sign is presented." As a result, different ways of representing the same properties resulting from different transformations are used as separate features. The definition of the feature of the image has caused great controversy, so in the theory of pattern recognition there is still no clear definition. The reason, obviously, is that the notion of a sign is inextricably linked to the notion of the plural, nor can it be initially defined because of the more general concepts. Therefore, it is

possible to give only its description, thus defining the characteristic features of the sign. Let us give the following formulation of a sign of a speech pattern: a sign is a reflection of x: $T \to X$, where T is the set of time intervals of the observation region of the signal $y(t_n)$, obtained as a result of some transformation $Q[y(t_n)]$; X is the set of values of a measure:

$$\sum_{i=1}^{N} T_i = t_n \;\; \text{i} \;\; X_{T_i} \neq X_{T_j} \;\; \forall j = i \div 1 \qquad (2.15)$$

In other words, a sign is a measure of the property of a phenomenon resulting from some transformation of the observation space; differences in the values of this measure make it possible to divide the phenomenon into many events - patterns. Usually the choice of type of transformation is determined by the model of appearance of the phenomenon, based on a priori knowledge.

In order to give satisfactory results, features for describing speech patternsshould be distinguished by transformations based on the principles of acoustic signal formation in the vocal tract. Ways of presenting featuresshould take into account mechanisms of speech analysis of the auditory system that are invariant to the interference. Currently, two models of speech have become most widespread: the linear speech model and the linear prediction model. Both models provide a speechpath in the form of a linear filter with variable parameters, the transmission characteristic of which is described by the expression:

$$H(z) = \frac{Q_0}{i - \sum_{k=1}^{p} Q_k Z^{-k}}, \qquad (2.16)$$

where Q_i is the impulse response (parameters) of the filter;

Z is a complex exponent.

Changes in filter parameters correspond to changes in the position of the organs of articulation. In the neutral position of the speech tract, its resonant frequencies f_n can be determined by the formula:

$$f_n = (2n-1)\frac{c}{4l}, \qquad (2.17)$$

where n is the resonance number;

c - speed of sound (cm / sec);

l is the average length of the speech tract (cm).

Taking l = 17.5 and c = 35300, we obtain the frequency values in the spectrum of the signal with high energy concentration (formant) f_1=504, f_2=1512, f_3=2524, etc.

When the position of the organs of the articulation of the voice tract and its length change, it corresponds to the change of the filter parameters of the model of speech, the position of the formant and the shape of the spectrum of the output signal changes, and it will correspond to a particular sound of speech. Thus, the transformation of Q of the space of observation of the speech signal $y(t_n)$, adequate models of speech production, is a spectral transformation, and informative features for describing speech patternsmay be the form of the spectrum, the distribution of energy in the frequency ranges, some temporal characteristics of the signal. The means of presentation of the featuresmust satisfy the following requirements:

– to report features of invariance property to the effect of interference of all kinds;

– not to diminish its information content;

– cost of separation of the trait must be within the allowable range.

In part, these requirements can be fulfilled if some principles of operation of the auditory system are used to describe the speech signal, for example, to produce frequency analysis of the signal in the bands corresponding to the critical hearing bands, to distribute energies in different frequency bands in the logarithm of the ratio of these energies, to use the effect disguise. To increase the invariance of featuresof interference, real-time data smoothing with high-speed algorithms and devices is also effective.

It is convenient to classify the features in such an order that is consistent with the strategy of their use when segmenting the speach stream into individual sounds.

The vast majority of work on acoustic analysis uses a sequential strategy of segmentation and phonetic identification of acoustic signals, which first determines the method of sound formation, then the place of formation, and then determines the phonemic identity of the sound. The following groups of acoustic features that can be used in the tree nodes of a sequential recognition strategy can be separated.

Featuresto distinguish between “speach and pause”:

1. Signal energy over the entire frequency range of a signal. It is compared with the noise threshold, which is statistically determined.

2. Intensity of the signal in several frequency bands. The "speech" decision is made when the conditions of exaggerating the signal intensity of the set threshold in each frequency range are jointly fulfilled.

3. A set of features: the magnitude of the logarithm of the signal energy, the number of zero signal intersections for analyzing the interval, the normalized autocorrelation coefficient, the vector of linear prediction coefficients. The discriminant function for patterns space is based on a statistical approach.

Featuresfor distinguishing "tonal - non-tonal":

1. Signal energy in the frequency range 100 Hz – 300 Hz. If it exceeds a certain threshold, then the decision is "tonal".

2. The ratio of the energy of the high-frequency range (usually in the band above 3 kHz or from 5 kHz to 10 kHz) to the energy of the low-frequency range (usually in the band 300-1000 Hz). The decision is made according to the statistical threshold established.

3. Signal space: signal energy, normalized signal autocorrelation coefficient, number of zero crossings, linear prediction error. Classification in space is done using the discriminant function.

4. Tone frequency. Its presence makes the decision "tonal". up to 300 Hz to signal energy in the range of 500 Hz - 2800 Hz - to separate the ringing beams from the loud ones. If it exceeds the set threshold, then the decision is "ringing".

Featuresfor distinguishing "slot - non-slot":

1. The ratio of the energy of the high frequency signal range (usually 3 kHz to 5 kHz) to the energy of the low frequency range (typically 100 Hz to 900 Hz). If it exceeds the set threshold, then the decision is made “slot”.

2. Number of zero crossings. The slot decision is made when the threshold number is exceeded.

3. Smooth differential energy values in the 3400 Hz - 5000 Hz range. If it is large, then the decision is made "slot".

Featuresfor distinguishing "vowel - sonorous":

1. Smooth signal energy values over the entire frequency range and in the range of approximately 600 Hz - 2800 Hz.

2. If they are less than the corresponding values for the vowel by the experimentally determined value, then the decision is "sonorous", otherwise - "vowel".

3. Frequency value of the first formant and its width, value of the frequency of the third formant. The low first formant and its greater width or low third formant determine sonar sounds (nasal and flowing).

Featuresfor classification of slot segments:

1. The logarithm of the signal energy ratio in the range (1.5 - 8) kHz and energy in the range (3 - 8) kHz. If it is less than the set threshold, then the decision is “weak slot” / f, θ, v, δ /, otherwise - “strong slot” / s, z, *ƒ*, ξ /.

2. Spectrum energy of the gap segment, center of gravity of its spectrum. If the center of gravity frequency $f_{кр}$ = 2.5 kHz and the energy is small, then the phoneme class is identified / *ƒ*, θ /: if $f_{кр}$ = 2.0 kHz, the maximum energy is identified by the phonem class /*ƒ*, ξ/; if $f_{кр}$ = 4.0 kHz and the energy has an average value, then the phonemic class / s, z / is identified. The frequency of the amplitude maximum f_m in the spectrum, the amplitude of this maximum, the asymmetry of the spectrum. For example, if f_m> 3.2 kHz, then the phonemic class is identified, /*ƒ*, ξ/, if f_m<3.2 kHz, then the phonemic class /s, z/ is identified. Frequency of the first formant and frequency of the first anti-resonance. The difference between /*ƒ*/ from /s/ is performed by a discriminant function in the space of these two traits.

Features for the classification of adjacent segments:

1. Call time $t_{нз}$ - to separate ringing binary /b, d, g/ from non-ringing /p, t, k/. For example, if $t_{нз}$> 35 msec, then the class of non-ringing bells is identified, if $t_{нз}$> 25 msec, then ring bells are identified.

2. Frequency of amplitude maximum in the burst spectrum f_b, relative value of the energy of the burst spectrum. The decision on the place of creation of the bow is made on the basis of experimentally set thresholds. For example, f_bє (900-2100) Hz corresponds to the given shear /k, g/ in front of the vowels of the back row /a, o, u/, and f_bє (3,0-5,0) kHz corresponds to the tooth shear /t, d/ before back row vowels.

3. Sign of the slope of the first and second formants at the transitional sections.

Featuresfor classification of vowels:

1. The values of the first, second and third formants. Recognition is carried out in the space of these two or three features at the minimum distance of implementation from the standard.

2. The number of zero crossings in the first and second band is for separating the front row vowels from the back row vowels.

3. Frequency of the first formant and frequency center of gravity of the second, third and fourth formant - to identify all the vowels, frequency center of gravity of the first two formants - to identify the back vocal numbers.

Featuresfor sonar classification:

Sonar is the most difficult class to identify. To distinguish between individual consonants in the middle of this class, it is necessary to use both statistical and dynamic features, such as the amplitude of the first three formants, the trajectories of these formants, the distances between them and others. The classification using these features is based on the heuristic solutions found in the statistical analysis of a large number of linguistic material. For example, the low first formant and the small value of the ratio of energy in the second formant band to the energy in the first formant band corresponds to nasal / n, m, η /. The low third formant corresponds to the retroflex /r/, the low first and second formant and the high third formant correspond to smooth /l/.

2.2.3 Method of segmentation of audio signal into “speech-pause” sound types

The problem of speech isolation from ambient noise is quite complicated, except in the case of very high signal-to-noise ratios, in particular in the case of high-quality recordings made in a muted or enclosed soundproof room.

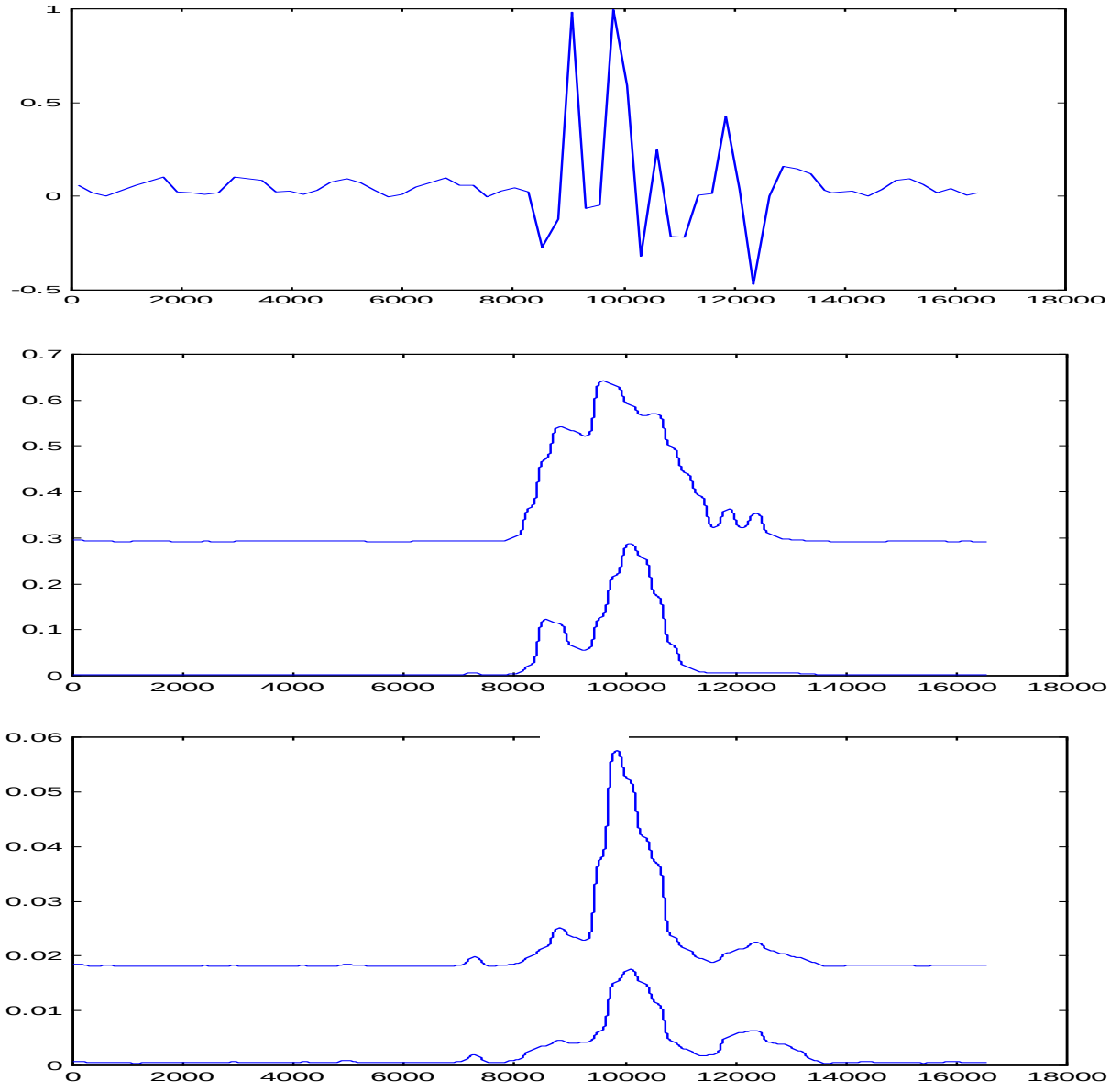


Figure 2.6 - Pause speech separation: a) speech signal; b) signal energy in the 1st and 2nd formats; c) signal energy in III and IV formats.

In this situation, the energy of even the weakest sounds of speech (fricative consonants) exceeds the noise energy and, thus, it is sufficient to measure the energy

of the signal. But such recording conditions are usually not found in real situations. Solving the problem of speech recognition from the audio signal is the first step in the implementation of phonetic speech recognition. With this procedure, we get rid of redundancy as a storage medium. To do this, you must use the sign of the presence of signal energy over the entire frequency range. Figure 2.6 shows that the presence of the amplitude of the signal itself, the value of which would exceed a given noise threshold, is not sufficient.

Therefore, to solve this problem it is proposed to use such features of the signal as its energy E and the number of zero cross sections ZC for each interval of background noise [30].

To implement this algorithm, it is necessary to have noise statistics obtained by cutting noise intervals from different audio signals.

1. Find the maximum noise energy:

$$E_{шуму\max} = \max(E_j), j = 1..n, \tag{2.18}$$

2.Find the mathematical expectation of noise energy:

$$M_E = \frac{\sum_{j=1}^{n} E_j}{n}; \tag{2.19}$$

3. Find the dispersion of noise energy:

$$D_E = \frac{\left(\sum_{j=1}^{n} E_j\right)^2}{n} - \frac{\sum_{j=1}^{n} E_j^2}{n}; \tag{2.20}$$

4. Find the SLE of the noise energy by the formula:

$$\sigma_E = \sqrt{D_E}; \tag{2.21}$$

5.Determine the detection limits:

$$E_{\max} = E_{шуму\max} + 3\sigma_E; \tag{2.22}$$

$$E_{\min} = M_E + 3\sigma_E; \tag{2.23}$$

$$ZC_{шуму} = \left(\frac{\sum_{j=1}^{n} ZC_j}{n} \right) \cdot \frac{t_{інт\_мови}}{t_{інт\_шуму}}, \quad (2.24)$$

where $t_{інт\_мови}$ - the length of the intervals into which the speechis broken (10 ms);

$t_{інт\_шуму}$ - the duration of the intervals at which noise is broken (100 ms).

After receiving audio data intervals of 10 msec, the energy characteristic for each interval and the number of transitions after zero are determined.

$$M_k = \frac{1}{m} \sum_{i=1}^{m} a_i , \quad (3.25)$$

$$E_k = \frac{1}{m} \sum_{i=1}^{m} a_i^2 , \quad (2.26)$$

$$E_k' = E_k - M_k^2, \quad (2.27)$$

where *m* is the number of amplitudes in the interval (for 11 kHz - 128).

The algorithm below is based on the measurement of two simple characteristics - energies and the number of transitions through zero.

Consider the algorithm using figure 2.7.

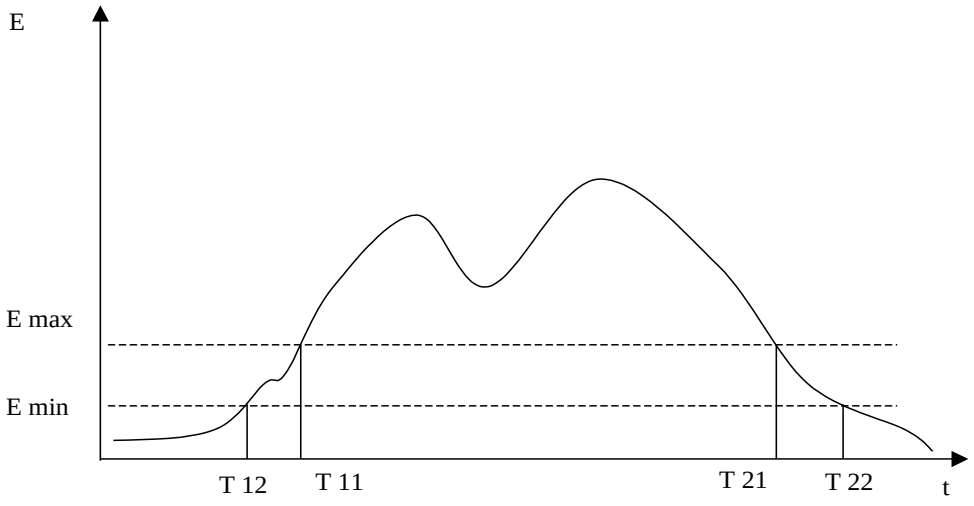


Figure 2.7 - Selecting a speech from an audio signal

Moving from the beginning of the sound signal to the right (see Figure 2.7), we expect an interval at which $E_k > E_{\max}$. We pre-set the speechstart (see T_{11} in Figure 2.7).

Considering the intervals from T_{11} to the right along the time axis, we are waiting for the interval at which $E_k \le E_{\max}$. This is point T_{21} in Figure 2.7. If for some time T_{sp} there has not been a period from T_{21} ($T_{sp} = 400$), at which $E_k > E_{\max,}$ we set the pre-termination of the speech at T_{21}. Otherwise, we continue our search for point T_{21}.

Thus, we have defined the beginning and ending of the speech.

It should be considered that $T_{21} - T_{11} < T_{\min}$, where $T_{\min} = 100$мсек- the minimum duration of the speech signal, it is necessary to consider the found gap is not in the speechand it is necessary to continue the search for point T_{11}.

Moving to the right along the time axis from point T_{11}, we look for the first interval at which the condition is fulfilled $E_k \le E_{\min.}$ This is point T_{12} (see Figure 2.7). After that, to the left along the time axis from point T_{12}, we look for the first interval at which the condition is satisfied $\mu \cdot ZC_k \ge ZC_{noise}$, $\mu = 1.$

The search for such a gap should not be further from the T_{12} point for a time greater than T_{offset} (200 ms). If we move to T_{offset}, then we do not move and consider this point the beginning of speech.

Moving to the right along the time axis from point T_{21}, we look for the first interval at which the condition is fulfilled $E_k \le E_{\min.}$. This is point T_{22}. Then to the right along the time axis from the point T_{22} we look for the first interval at which the condition is satisfied $\mu \cdot ZC_k \ge ZC_{noise}$. The search for such a gap should not be further from the point T_{22} for a time greater than T_{offset} (200 ms). If we have moved on T_{offset}, then we do not move further and consider this point as the end of the speech.

To create a model of speech isolation from the audio signal was collected statistics of different background noises (different speakers, spoken volume) by cutting the necessary intervals from the audio files. The statistical sample was 37 values.

As a result of the above calculations, the following values of the basic parameters were obtained:

$$E_{noise\ \ max} = 1.695\ E - 05;$$
$$M_E = 4.469\ E - 06;$$
$$D_E = 6.991\ E - 10;$$
$$\sigma_{Emax} = 2.644\ E - 05;$$
$$E_{max} = 9.628\ E - 05;$$
$$E_{min} = 8.379\ E - 05.$$
$$ZC_{noise} = 5.991$$

To illustrate the operation of the algorithm, we present the result of the model for the word "four". Figure 2.8 shows the audio signal for this word:

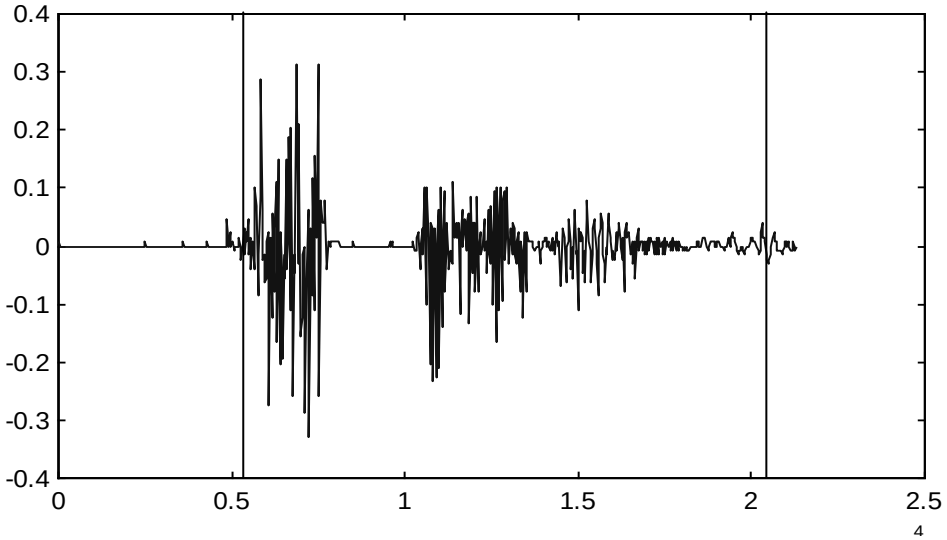


Figure 2.8 - The sound of the word "чотири"

As a result, the audio signal was reduced by 1.56 times.

The start and end marks of the audio signal were placed at fixed intervals of numbers 84 and 161, respectively.

Figure 2.9 shows the energy characteristic of the signal.

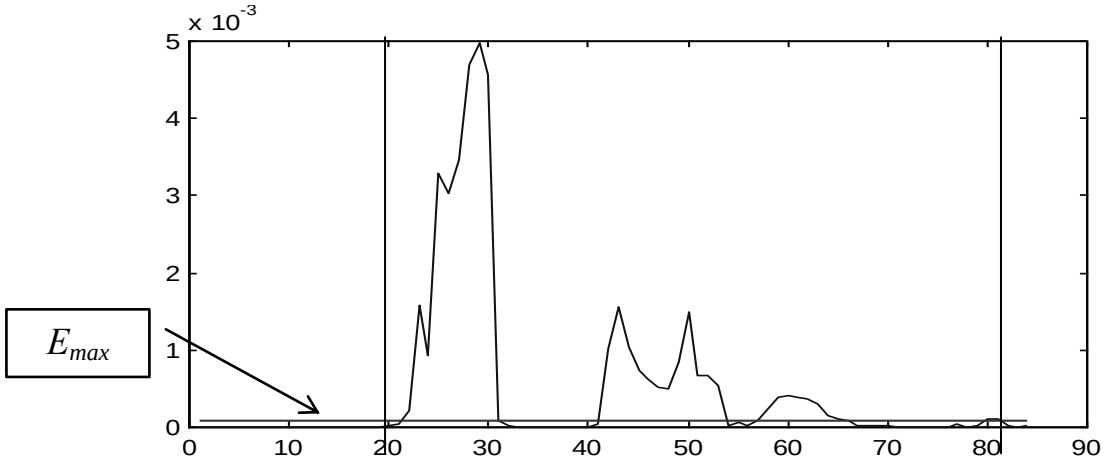


Figure 2.9 - Energy characteristic of the audio signal

According to the results of the study of the noise levels conducted in this section and the spectral characteristics of the Ukrainian language sound types, it was concluded in [22, 23] that the non-invariance of the distinguished features to the change of the acoustic noise level of the environment, to the change of the distance of the acoustic receiver from the speech source signal and change of speakers, which leads to increased reliability of the selection of features of the phonetic classes of sounds, can be achieved by recording voice signals using two acoustic sensors, one of which located at a fixed distance.

2.2.4 Fuzzy methods for distinguishing features

When formalizing knowledge, it is quite common to find qualitative knowledge, such as high fever, low glow of the filament, young diplomat, etc. For the formal presentation of such qualitative knowledge, an American mathematician, professor of computer science at the University of Berkeley (California), Lofty A. Zadi (Iran), proposed in 1965 a formal apparatus of fuzzy logic.

The fuzzy subset N of the set M is defined as the set of ordered pairs N = {μN(x)/x}, where $\mu_N(x)$ is a characteristic membership function (or simply a membership function) that takes values in the interval [0, 1] and the degree, indicating (or level) the membership of the element x in the subset N [31]. Thus, the fuzzy set N can be written as:

$$N = \sum_{i=1}^{n} (\mu(X_i) / X_i), \quad (2.28)$$

where Xi is the i-th value of the base scale, and the sign "+" is not a designation of the addition operation, but makes sense of association.

We define a speech variable (SV) as a variable whose value is determined by a set of verbal characteristics of a property. For example, “age” SV may have the following meaning:

SV = МлВ, ДВ, ОВ, ЮВ, МВ, ЗВ, ПВ, СВ,

which means infant, infant, infant, adolescent, young, mature, aged and old, respectively. The set M is the scale of a person lived for years [0..120]. The membership function determines how confident we are that a given number of years lived can be attributed to a given SV value. Suppose that an expert at a young age refers to people aged 20 years with a confidence level of 0.8, aged 25 years with a confidence level of 0.95, aged 30 years with a confidence level of 0.95, and aged 35 years with a probability of 0 , 7. Therefore:

$$\mu(X_1)=0,8;\ \mu(X_2)=0,95;\ \mu(X_3)=0,95;\ \mu(X_4)=0,7; \quad (2.29)$$

The value of SV = MB can be written:

$$MB = \mu(X_1) / X_1 + \mu(X_2) / X_2 + \mu(X_3) / X_3 + \mu(X_4) / X_4 =$$
$$= 0,8/ X_1 + 0,95/ X_2 + 0,95/ X_3 + 0,7/ X_4 . \quad (2.30)$$

Thus, fuzzy sets allow to take into account the subjective opinions of individual experts.

For operations with fuzzy sets, there are different operations, for example, the operation "fuzzy OR" (otherwise) is set in the Zade logic:

$$\mu(x)=max(\mu_1(x), \mu_2(x)) \quad (2.31)$$

and in the probable approach:

$$\mu(x)=\mu_1(x)+\mu_2(x)-\mu_1(x) \cdot \mu_2(x). \quad 2.32)$$

There are other operations on fuzzy numbers, such as extended binary arithmetic operations (addition, multiplication, etc.) for fuzzy numbers, specified through corresponding operations for fuzzy numbers using the principle of generalization, etc.

As we will see later, fuzzy sets (otherwise called soft calculations) are very commonly used in expert systems. Fuzzy logic is used as a convenient tool for controlling technological and industrial processes, for intelligent household and entertainment electronics, in fault detection systems and other expert systems. Special fuzzy inference tools have been developed, such as the Fuzzy CLIPS tool. Fuzzy logic was invented in the United States, and now the rapid growth of this technology has begun in Japan, Europe and has now reached the United States again.

The development of this direction is the realization of non-factors in the systems of knowledge representation: incompleteness, inaccuracy, uncertainty, ambiguity, incorrectness, etc. Knowledge representation systems and know-how technologies continue to evolve. In addition to the traditional languages (LISP, PROLOG, SMALLTALK, REFAL) and tools (LOOPS, KEE, ART), new web-based versions of IS are currently emerging as JAVA tools: Exsys Corvid, JESS . HTML has become the basis for presenting knowledge in the Internet.

Consider the neural network model used in the formation of speech patterns [32]. Figure 2.10 presents a neural network consisting of three layers - the hi input, which corresponds to auditory cell-related perceptron neurons, analyzing ha and responding ho, which models the decision-making process (pattern recognition).

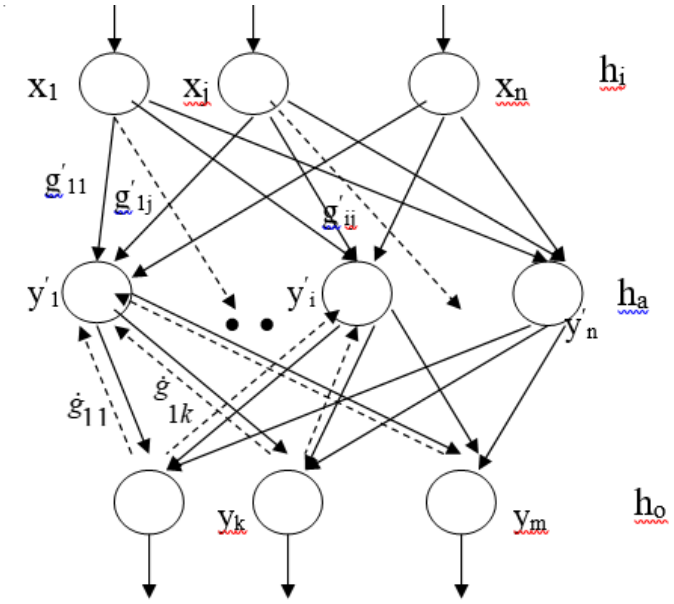


Figure 2.10 - Model of neural network for recognition of speech patterns

The output signal of the i-th neuron y′$_i$ of the analyzing layer is obtained by integrating the following differential equation:

$$\frac{dy_i'}{dt} = \sum_{j=1}^{n} g_{ij}' x_j - \gamma(y_i'), \tag{2.33}$$

where $\overline{X} = \{x_1, x_2, ..., x_j, ..., x_n\}$ - input pattern;

g_{ij}' - the connection between the ith neuron of the analyzing layer and the jth neuron of the input layer;

y_i' - the output signal of the ith neuron of the analyzing layer;

$\gamma(y_i')$ - a component that renders the transfer characteristic of a neuron nonlinear and takes into account nonlinear effects. It is believed that there is γan inverse function f_G, namely $f_G = \gamma^{-1}$.

Permanent integration of a neural cell is in the order of several tens of ms. Given that at this interval, the xi output signals for the stationary speechcan be written:

$$\frac{dy_i'}{dt} \approx 0, \tag{2.34}$$

where

$$\gamma(y_i') = \sum_{j=1}^{n} g_{ij}' - x_j, \tag{2.35}$$

and finally

$$y_i' = f_G(\sum_{j=1}^{n} g_{ij}' - x_j). \tag{2.36}$$

In accordance with equations (2.33) and (2.36), when applying to the input of auditory auditory neurons of the spectrum $\overline{X}$ of the speech signal, the value of the initial excitation of the ith neuron of the analyzing layer is proportional to the product:

$$\overline{g}^{T}{}_i \cdot \overline{x}, \tag{2.37}$$

where $\bar{g}_i^T = (g_{i_1}, g_{i_2}, ..., g_{i_n})^T$ - the vector column of the weights of the connections of the i-th neuron;

$\bar{X} = \{x_1, x_2, ..., x_j, ..., x_n\}$ - vector of the spectral parameters of the speech signal.

The core of activity in the analyzing layer is formed near those neurons for which the product (2.37) is maximum. The speech distribution of communication weights that can be recognized can be established as a result of training by presenting them to a neural network input.

According to Hebb's law, the change in the weight $\Delta g'_{ij}$ of the connections at each iteration step must be proportional to the input and output neurons:

$$\Delta g'_{ik} = \alpha y_i x_k , \qquad (2.38)$$

where α is a constant that determines the magnitude of the increment step.

It is known that the transfer function of a neuron has a sigmoidal shape with lower and upper saturation levels. In the area of active neuron training, one can approximate a linear function with some slope k_f. Then formula (2.14) can be rewritten as:

$$\Delta g'_{ik} = \alpha y_i x_k = \alpha k_f \sum_{j=1}^{n} g'_{ij} x_j x_k . \qquad (2.39)$$

The linkage weights g'_{ij} of the analyzing neural layer after training determine the structure of the neural network that corresponds to the stored image. Therefore, (2.15) is a sign for converting the initial vector of speech signal parameters into a speech pattern.

2.3 Development of a mathematical model of the fuzzy-logic method

In this master project, the minimization of the logical function by the Quine-McClasky method will be applied.

When a logical function has more than five arguments, the Wichy-Carnot diagrams are rendered unobtrusive. Therefore, in this case, to minimize the logical functions depicted in the DDNF, it is more convenient to use the Quinn-Mak-Klasky method, which involves performing two main steps.

In the first one, we move from the DNF to the shortened DNF logicalfunctions. To do this, all conjunctions are first divided into groups, up toeach of which attracts conjunctions with the same number of variableswithout objection: This group includes conjunctions containing variables without negation, i = 0, 1, 2, ...,n.. For example, let's minimize the following function:

$$f = \overline{x}_1 \cdot \overbrace{\overline{x}_2 \cdot x_3 \cdot x_4 + \overline{x}_1}^{1} \cdot \overbrace{x_2 \cdot \overline{x}_3 \cdot \overline{x}_4 + \overline{x}_1}^{2} \cdot \overbrace{x_2 \cdot \overline{x}_3 \cdot x_4 + \overline{x}_1}^{3} \cdot \overbrace{x_2 \cdot x_3 \cdot x_4 +}^{4}$$

$$+ \overbrace{x_1 \cdot \overline{x}_2 \cdot \overline{x}_3 \cdot x_4}^{5} + \overbrace{x_1 \cdot \overline{x}_2 \cdot x_3 \cdot x_4}^{6} + \overbrace{x_1 \cdot x_2 \cdot \overline{x}_3 \cdot \overline{x}_4}^{7} + \overbrace{x_1 \cdot x_2 \cdot \overline{x}_3 \cdot x_4}^{8} \quad (2.40)$$

Then it can be divided into groups as follows:

$$f = (\underline{\overline{x}_1 \cdot x_2{}^2 \cdot \overline{x}_3 \cdot \overline{x}_4}) + (\underline{\overline{x}_1 \cdot \overline{x}_2{}^1 \cdot \overline{x}_3 \cdot \overline{x}_4} + \underline{\overline{x}_1 \cdot x_2{}^3 \cdot \overline{x}_3 \cdot x_4} + \underline{x_1 \cdot \overline{x}_2 \cdot \overline{x}_3{}^5 \cdot x_4} + \underline{x_1 \cdot x_2{}^7 \cdot \overline{x}_3 \cdot \overline{x}_4}) +$$

$$+ (\overbrace{\overline{x}_1 \cdot x_2{}^4 \cdot x_3 \cdot x_4 + x_1 \cdot \overline{x}_2{}^6 \cdot x_3 \cdot x_4 + x_1 \cdot x_2{}^8 \cdot \overline{x}_3 \cdot x_4}^{3}) . \quad (2.41)$$

After such a breakdown, we take the first step from the DDNF to the shortened DNF by gluing the conjunctions of n-thrank.

The action of this should be to compare pairs of elements adjacent to the number of groups, since glued conjunctions of the form can only contain adjacent groups. For example, for entry (12.7) one has to compare the elements of groups 1 and 2, 2 and 3. The conjunctions for which they were glued denote (for example, underline). Thus, for function (2.17), the bonding operation is performed between conjunctions 2 and 3 for the variable x_4, 2 and 7 for x_1, 1 and 6 for x_1, 1 and 4 for x_2, 3 and 4 for x_3, 3 and 8 for x_1, 5 and 6 for x_3 5 and 8 for x_2, 7 and 8 for x_4.

As a result, conjunctions of the first rank, which attract shortened DNF, are formed. Unlabeled conjunctions (those that are not glued together) are simple

implicants and are also involved in shortened DNF. For example, after performing the first reduction step, function (2.17) will look:

$$f = (\overline{x}_1 \cdot x^1{}_2 \cdot \overline{x}_3 + x^2{}_2 \cdot \overline{x}_3 \cdot \overbrace{\overline{x}_4}^{1} +$$

$$+(\overbrace{\overline{x}_2 \cdot x^3{}_3 \cdot x_4 + x_2 \cdot \overline{x}_3^4 \cdot x_4 + \overline{x}_1 \cdot x^5{}_2 \cdot x_4 + x_1 \cdot \overline{x}_2^6 \cdot x_4 + x_1 \cdot \overline{x}_3^7 \cdot x_4 + x_1 \cdot x^8{}_2 \cdot \overline{x}_3 + \overline{x}_1 \cdot x^9{}_3 \cdot x_4}^{2})$$

(2.42)

The following procedure is repeated, with conjunctions obtained of $(n-1)$-th rank, etc., until the DNF contains only simple implants. For example, to obtain a shortened DNF for a given function, it is sufficient to perform one more reduction step to the entry (2.18), thereby merging conjunctions 1 and 8 for the variable and x_1 and 2 and 4 for x_4.. Obtain the following shortened DNF initial logic function:

$$f = \overbrace{\overline{x}_1 x_3 x_4}^{1} + \overbrace{\overline{x}_2 x_3 x_4}^{2} + \overbrace{\overline{x}_1 x_2 x_4}^{3} + \overbrace{x_1 \overline{x}_2 x_4}^{4} + \overbrace{x_1 \overline{x}_3 \overline{x}_4} + \overbrace{x_2 \overline{x}_6}^{56}. \quad (2.43)$$

In the second step, we switch from shortened DNF to deaf DNF and choose among them MDNF. Deaf is called DNF, which comes from the shortened, when it is removed from the extra simple implants. The members of the abbreviated DNF, whose exclusion does not cause changes in the values of the logical function on the possible sets of its arguments, are considered to be superfluous implicants.

In order to get deafDNF, we construct an implicit matrix (see Table 2.1), the rows of which contain simple implicants included in the shortened DNF, and the columns are constituents of the unit of which the DBNF consists.

Table 2.1 –Implicit matrix

Simple implicants		The constituents of the prime							
		1 $\bar{x}_1\bar{x}_2x_3x_4$	2 $\bar{x}_1x_2\bar{x}_3\bar{x}_4$	3 $\bar{x}_1x_2\bar{x}_3x_4$	4 $\bar{x}_1x_2x_3x_4$	5 $x_1\bar{x}_2\bar{x}_3x_4$	6 $x_1\bar{x}_2x_3x_4$	7 $x_1x_2\bar{x}_3\bar{x}_4$	8 $x_1x_2\bar{x}_3x_4$
1	$\bar{x}_1x_3x_4$	x			x				
2	$\bar{x}_2x_3x_4$	x							
3	$\bar{x}_1x_2x_4$			x	x				
4	$x_1\bar{x}_2x_4$					x	x		
5	$x_1\bar{x}_3x_4$					x			x
6	$x_2\bar{x}_3$		x	x				x	x

In a row against each prime, the implicants mark an "x" under those constituents of the unit on which it gains a value of 1; the corresponding constituents are covered (absorbed) by a given implicant. From the total number of simple implicants, we need to select their minimum subset, which implicants cover all the single values of a given, logical function; the rest of the implants will be superfluous.

Deaf forms begin to form when revealing obligatory simple onesimplicants, which (and they alone) cover the corresponding constituents of a unit of logical function.

In Table 2.1, we see that implicant 6 is mandatory because it only covers the single values of the logical function on the 2nd and 7th sets (columns 2 and 7; row 6). In addition, it also covered the constituents of Units 3 and 8. The remaining uncovered constituents 1, 4, 5, 6 can be covered with different combinations of implicants 1-5, resulting in many deaf forms from which to choose MDNF:

$$\begin{cases} f = x_2 \cdot \overline{x}_3 + \overline{x}_2 \cdot x_3 \cdot x_4 + x_1 \cdot \overline{x}_3 \cdot x_4 + \overline{x}_1 \cdot x_3 \cdot x_4; \\ f = x_2 \cdot \overline{x}_3 + \overline{x}_1 \cdot x_3 \cdot x_4 + x_1 \cdot \overline{x}_2 \cdot x_4; \\ f = x_2 \cdot \overline{x}_3 + \overline{x}_1 \cdot x_2 \cdot x_4 + \overline{x}_2 \cdot x_3 \cdot x_4 + x_1 \cdot \overline{x}_3 \cdot x_4; \\ f = x_2 \cdot \overline{x}_3 + \overline{x}_2 \cdot x_3 \cdot x_4 + x_1 \cdot \overline{x}_2 \cdot x_4 + \overline{x}_1 \cdot x_3 \cdot x_4. \end{cases} \quad (2.44)$$

The minimum of these deaf DNFs is the second expression:

$$f_{мін} = x_2 \cdot \overline{x}_3 + \overline{x}_1 \cdot x_3 \cdot x_4 + x_1 \cdot \overline{x}_2 \cdot x_4 \quad (2.45)$$

Thus, the Quinn-McClaskey method makes it easy to minimize the logical functions depicted in the DDNF.

One of the most important and poorly formalized speechrecognition problems today is the task of automating the process of extracting informative parameters to describe speech patterns. It becomes especially relevant when parameters are searching in the form of signal energies in different frequency bands, since it is necessary to analyze a very large number of variants of the frequency range of the speech signal into subbands.

To automate the process of finding information parameters, the authors propose a combined fuzzy-logic approach. His idea came from an analysis of a model of the human hearing system, supplemented by a neural network. This approach consists of the following steps. Initially, an initial partitioning of the entire frequency range is performed by a certain number of frequency bands (for example, according to the human hearing system model). The temporal energy functions in the selected bands are considered to be fuzzy functions of assigning certain time intervals to known sounds. The next step is the operation of exacerbation of these functions by finding $n(n-1)/2$ their pairwise correlations, where n is the number of bands in which the frequency analysis of the speech signal. After that, the obtained fuzzy functions are transformed into logical ones by applying to them α- the level, which is determined by the value of the second after the global maximum of this function.

The logical choice of informative features is made by constructing an implicit matrix for those single values of functions that do not intersect in time. These single time intervals correspond to some sounds and are numbered to indicate the columns of the implicit matrix. Matrix rows are denoted by pairs of frequency band numbers that form a correlation. The implicit matrix denotes the cells that stand at the intersection of some single time interval and the correlation that generated it. The next step is the logical selection of informative features on the implicit matrix, which is carried out similarly to the Quine-McClaskey minimization procedure: all dead-end sets of correlation pairs are formed and one of them is chosen minimal.

In addition to determining the information frequency of the primary frequency bands for a particular sound, this approach allows you to determine how the frequency range of the speech signal is divided into the most informative subbands. To do this, define all the tapes of the implicit matrix, labeled with the same column number. If the band numbers in these correlations are adjacent, then they are combined into one subband (s-norm operation is performed), if not, then the value of the sub bands remains the previous one and the sign is formed by correlating the energy of the frequency band data signals (t-norm operation is performed) .

Therefore, the purpose of the study is to develop a method for automating the selection of informative features for known patterns (in this case, spectral components for given speech sounds).

Hypothesis: Consider the time functions of energy $E(f_i)$in the selected frequency ranges as fuzzy functions of the belonging of some time intervals to known patterns of w_j (sounds).

By the operation of exacerbation of these functions (by finding the correlations E (f1) · E (f2), E (f1) · E (f3),…, E (f1) · E (fn), E (f2) · E (f3),…, $E(f_2)\cdot E(f_n)$,…, $E(f_{k-1})\cdot E(f_n)$) determine the importance of these spectral components for different patterns (if the output of the energy of these spectral components increases the energy in the time interval of a given sound then they are informative for it).

Then, for each sharpened fuzzy membership function (correlation), we allocate logical values "1" at time intervals by applying the α-level. The α-level is selected according to the geometric illustration of Figure 2.11.

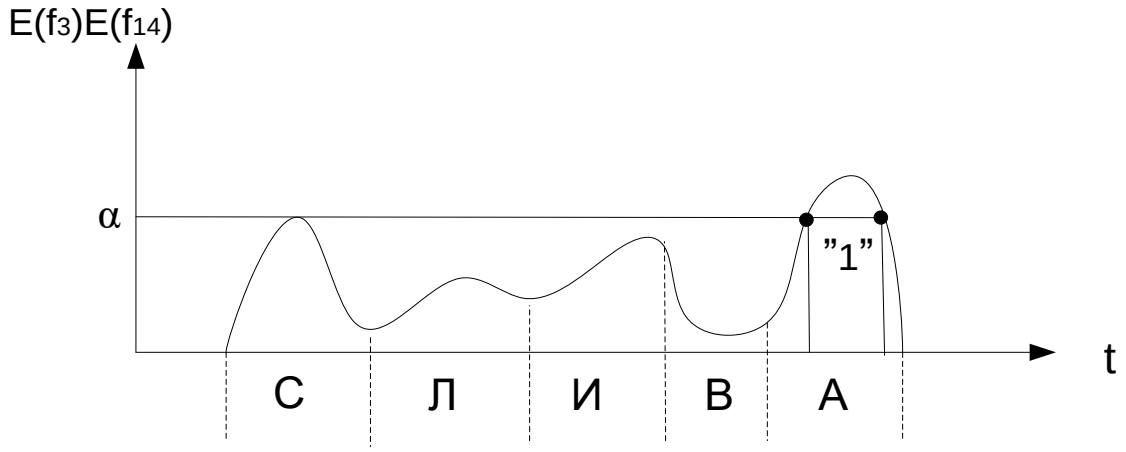


Figure 2.11 - Allocation of logical values "1" for correlation $E(f_3)$ $E(f_{14})$

For example, for the correlation $E(f_1)$ $E(f_3)$, the α-level will be determined as shown in Figure 2.12:

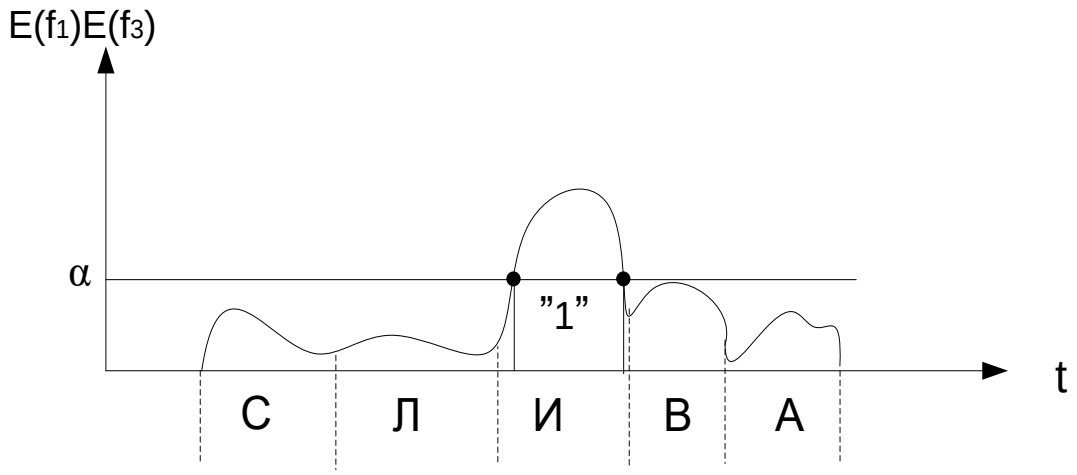


Figure 2.12 - Allocation of logical values "1" for correlation $E(f_1)$ $E(f_3)$

And, accordingly, for the correlation $E(f_5)$ $E(f_{11})$ Figure 2.13.

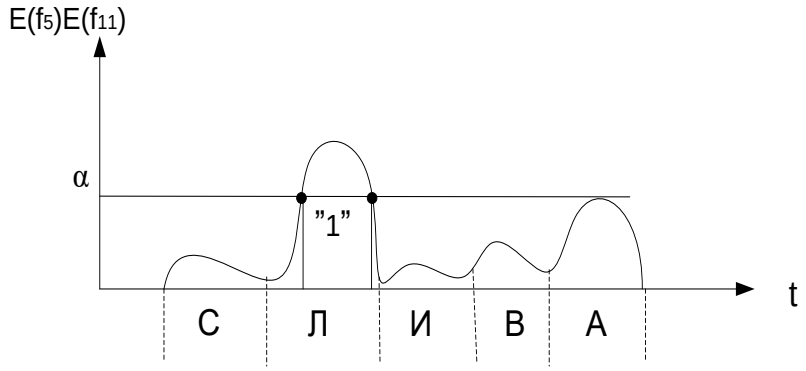


Figure 2.13 - Selection of logical values "1"for correlation $E(f_5)$ $E(f_{11})$

Then the implicit matrix for single intervals is constructed as follows: all single intervals for a given pattern (eg sound (a)) are searched for all correlations are denoted * below it (Table 2.2).

After that, all the correlations relevant to the sound are selected and combined (t-norm operations are performed).

Table 2.2 - The implicitmatrix

$E(f_i)$ $E(f_j)$\ω	а	о	у	е	і	б	в	...
$E_1 \cdot E_2$	*							
$E_1 \cdot E_3$			*					
$E_1 \cdot E_4$			*	*				
…				*				
$E_1 \cdot E_n$	*							
$E_2 \cdot E_3$			*					
…								
…	*							
$E_{13} \cdot E_{14}$				*				

The merging should be done in such a way that the energy of the frequencies in succession is combined in the same band.

Therefore, the selection of informative features is carried out by the following algorithm:

1. Formation of frequency ranges: their number, center frequencies, bandwidth, filter type are selected.

2. Filtration.

3. Calculation of energy

$$E_i(t)=\sqrt{\sum_{k=1}^{N} e_k^2}$$

4. Correlation

5. Finding α-levels for each correlation and forming "1" intervals.

6. Finding those "1" overlapping intervals.

7. Formation of an implicit matrix for them.

8. Formation on the implicit matrix of ranges, the energies of which will be informative features(by the operation of a conorm).

The exacerbation operation narrows the sound interval from the actual one. The operation of the conorm, on the contrary, expands and brings it closer to the real.The scheme of algorithm of selection of informative features is presented in Figure 3.4:

Thus, the use of a fuzzy approach at the initial stage makes it possible to increase the informativeness of the parameters of speech patterns, and the use of a logical approach at the final stage makes it possible to automate the process of their selection and formation. For example, the use of the proposed method to select informative features of phonotypes "sonorous", "fricative", "pause" allowed to select three informative frequency bands and to obtain their recognition accuracy of 98.5%.

2.4 Method for describing a speech signal based on a mathematical model of a “quasi-frequency” detector

A disadvantage of known speechrecognition devices that use the principle of frequency description is that the spectral description is made in narrow bands. This

description is unstable to change the speaker, and therefore when changing the speaker, you need to update the standards by retraining.

The quasi-frequency detection principle is applied to eliminate this drawback. The "quasi-frequency" speech-forming model is used to carry out the segmentation of a recognizable speech signal in the pre-processing stage for sound types. To perform a previous or "rough" segmentation, you can conditionally divide all natural speechsounds into a small number of classes.

"Segmentation" means the objective division of a speech message into units that are related to elements of the accepted alphabet. Thus for the Ukrainian language it is proposed to use the partition presented in table. 2.3 (softened consonants should be attributed to the same phonetic classes as their solid pairs).

Table 2.3 - Classification of sounds of the Ukrainian language

The name of the phonetic class	Sounds
Vowels	а, о, у, е, и, і
Noisy	з, с, ж, ш, ч, х, ц, ф
Sonoric	л, р, м, н, в, й
Explosive	п, т, к, б, д, г

In this section, we consider a model of speech formation that describes the signal by positioning the frequency moments of the energy of the signal in wide formant ranges, which allows to reduce the variation of features due to spectral variations. Binary coding of the position of these moments based on the ratio of energies in the frequency subbands avoids the influence of amplitude variations.

The analysis of the dependence of the informative properties of the speech sounds on their frequency-energy parameters shows that the basic information of the speech signal is encoded in the first three formative bands, so in this work we propose a model of speechformation based on the "quasi-frequency" modulation of

the vocal tract. In this model, the voice tract is considered to be a source of information (speech) signal, the encoding of information in which is carried out by modulating three carrier frequencies - frequencies of the 1st formant, frequencies of the 2nd formant and frequencies of the 3rd formant. The position of the frequency in the formant bands is determined by the position of the frequency moments of the signal:

$$M_{kf} = \frac{\int_{F_{K-1}}^{F_K} A_f \cdot fdf}{\int_{F_{K-1}}^{F_K} fdf}, \tag{2.46}$$

where A_f is the spectral density of the speech signal for the frequency band df;

f - the current value of the signal frequency;

k is the frequency channel number, k = 1,2,3.

Preliminary analysis of the speech signal in the formant bands is performed by means of bandpass filters, with the expression (2.46) for the frequency moment taking the form:

$$M_{kf} = \frac{\sum_{i=l}^{l+m} A_i \cdot f_i}{\sum_{i=l}^{l+m} f_i}, \tag{2.47}$$

where A_i is the amplitude of the output signal and the i-th filter; f_i is the center frequency of the bandpass filter; l is the number of the first bandpass filter in the k-th frequency channel; m is the number of filters in the k-th channel.

By encoding several positions of the frequency moment in each frequency band, it is possible to move from the description of the speech signal in a continuous three-dimensional space to a discrete description in the space of binary values of the frequency-detecting function. In each of the selected frequency channels it is possible to consider the following three forms of the spectrum (position of frequency moments):

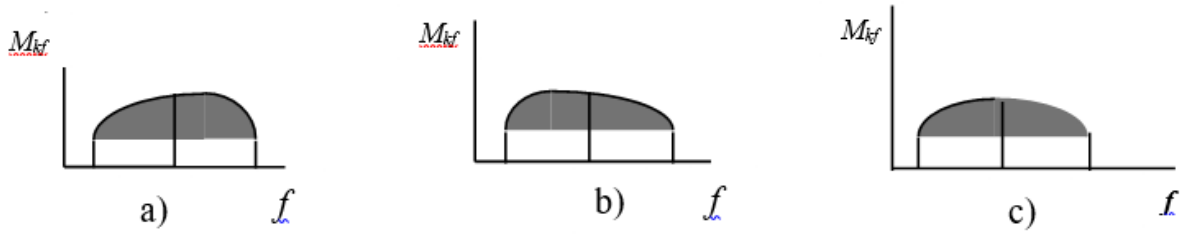


Figure 2.14 - Coded positions of frequency moments in frequency channels: a-deviation towards high frequencies; b-deviation toward low frequencies; c - zero deviation

For binary coding of these positions of frequency moments, each frequency channel is divided into 3 subbands, thus in one channel 2 bits of frequency-detecting function are obtained $\theta_{\partial i}$ by the comparative method:

$$\theta_{\partial i} = \bigcup_{i=1}^{2} \sigma(M_k^i \alpha M_k^{i+1}), \qquad (2.48)$$

where σ- unitary function, $\sigma(M_k^i \alpha M_k^{i+1})$= 1, if $M_k > M_k^{i+1}$, and equals to 0 otherwise; α- ratio of domination.

With this definition, the frequency-detecting function for the first distinguished case of frequency deviation ($\theta_{\partial 1}$, $\theta_{\partial 2}$) = (0, 0), for the second - ($\theta_{\partial 1}$, $\theta_{\partial 2}$) = (1, 1), for the third - ($\theta_{\partial 1}$, $\theta_{\partial 2}$) = (0, 1).

The analysis of the spectra shows that the use of the values of the frequency-detecting function in the three channels is effective for describing the loud sounds of the speechand the sound consonants (м,н,р,л). To distinguish between noisy and erupting consonant sounds, the global characteristics of the spectrum (spectrum form) of the signal are essential. To determine them, the ratio of the signal energies in the three adjacent bands ΔF_1 , ΔF_2 and ΔF_3 is used.

Thus, the initial description of the speech signal using the frequency-detecting function looks like an eight-bit binary word.

For example, for sound [a], the binary description looks like [a] = (0,1,1,0,0,0,1,1). The value of this function is calculated for each τ-th primary segment of the speech signal whose duration is chosen to be t_s = 5 ms (this corresponds to the stationarity interval of the shortest, discontinuous, speech sounds). For each pair of time-adjacent values of the frequency-detecting function, the values of the segmenting function are calculated by the formula $\theta_s^{\tau}=\theta_d^{\tau}\oplus\theta_d^{\tau-1}$, where the symbol ⊕ denotes the logical operation “the sum of the modulus two”. The values of the segmenting function are used to process the segmentation of the speech signal into separate sounds by the rule: $:\tau_s^j = \tau$, if $\Sigma^8_{i=1}(\theta_{si}^{\tau} * \rho_i) \geq \delta_п$, where τ_s^j - is the *j*-th segment label and ρ_i - the weight of the i-th binary digit of the segmenting function, $\delta_п$ is the threshold value, which is determined by the iterative “gradient descent” procedure in the process of training with the “teacher”. Thus, in this algorithm, the process of segmentation into sounds takes place in parallel with the process of their classification.

The values of the average frequencies for each of the three channels fnc are determined by the values of the formant frequencies in the neutral position of the vocal tract:

$$f_{нк} = (2k-1)\frac{c}{4l'}$$

where*c* - speed of sound in [cm/s], *c*= 35300;

l ‘- the average length of the speech tract,*l ‘* = 17,5 cm [4 - 5].

then$f_{н1}$ =504 Hz,$f_{н2}$ =1512 Hz,$f_{н3}$ = 2524 Hz.

When the position of the articulatory organs is changed, the shape of the voice tract and its length change, which corresponds to the modulation parameters, the formant position and the shape of the spectrum of the output signal change, and it corresponds to a particular audio speech.

The frequency ranges of channels of the 1st, 2nd and 3rd formants can be determined by statistics on the values of the formant bands ΔF_1 , ΔF_2iΔF_3[4 , 6]:

$$\Delta F_1 = 250 - 1000 \text{ Hz},$$

ΔF_2= 800 - 2200 Hz,

ΔF_3 = 1780-3560 Hz .

In the spectral analysis of the speech signal by bandpass filters with center frequencies arranged according to the logarithmic law along the frequency axis, in the bands ΔF_1 ,ΔF_2iΔF_3there will be 12 frequency bands distributed along the formant channels as follows: ΔF_1 - (252-317), (317-400), 400-504), (504-635),(635-800),(800-1008); ΔF_2-(1008-1270), (1270-1600), (1600-2016), (2016-2540); ΔF_3 - (2016-2540), (2540-3200), (3200-4032).

The segmenting function is intended for segmentation of a speech signal into separate sound types. Its values are combinations of eight bits that change over time. It is logical to assume that if the values of the digits in the words of the segmenting function change, the energy of the signal from one frequency range to another transitions, which indicates a transition from one sound type to another in a given time interval. Of course, the transitions between different phoneme combinations are different. This complicates the task of constructing an automatic classifier, which must recognize non-stationary (transient) intervals of the speech signal, that is, to put in these places the appropriate transition marks.

The construction of the classifier can be performed using a mathematical apparatus of artificial neural networks. The structure of the automatic classifier for segmentation of the speech signal in this case will be as shown in Figure 2.15.

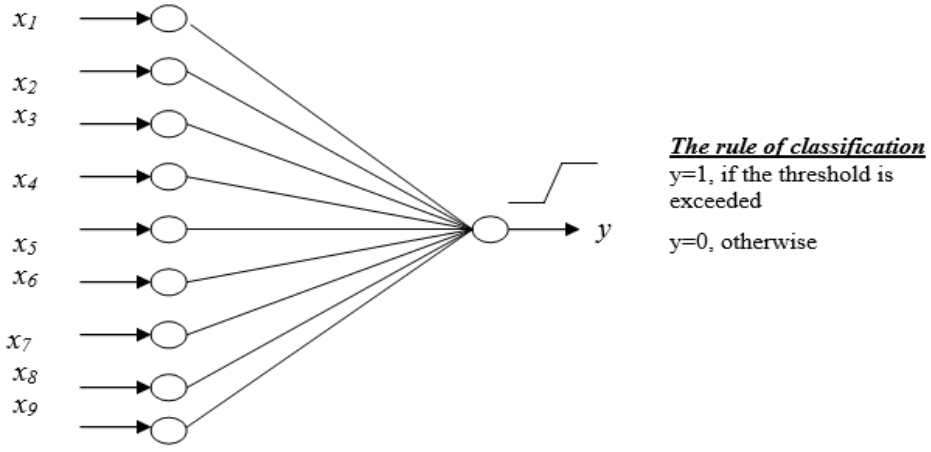


Figure 2.15 - Structure of the classifier

For the purpose of training it will be necessary to form a training sample of forty words of the Ukrainian language. The learning algorithm for each word should consist of the following steps:

- definition of word boundaries;
- filtering the signal in the specified frequency bands of the formant bands;
- calculation of frequency-detecting function;
- calculation of the segmenting function.

The detection of non-stationary areas of the speech signal will be done manually using the spectrogram of the speech signal. Figure 2.26 presents the spectrograms of command words “менше”.

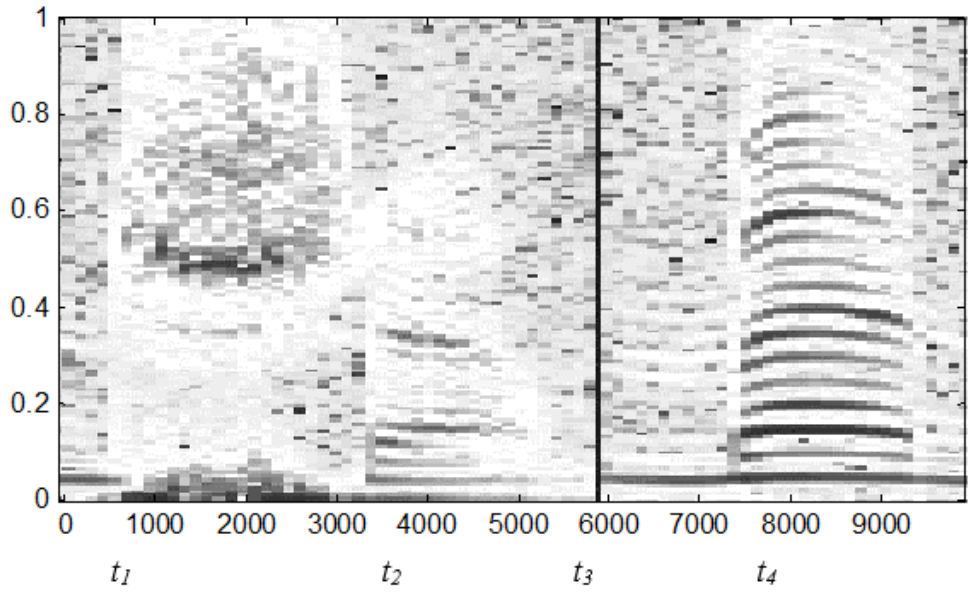


Figure 2.16 - Spectrogram of the word "менше"

Changing the energy of the signal at different frequencies, as shown in Figure 3, shows the intervals of the non-stationarity of the signal (t1,…, t4).

In total, training was conducted on a sample of 1500 combinations of segmenting function. To achieve dictatorial learning outcomes, vocabulary words will be spoken by three speakers six times.

The construction of an automatic neural network classifier with two layers of neurons (input and output) corresponds to the construction of the distribution surface between the two classes of patterns, which is known in classical recognition theory.

Knowing the value of the coupling weights between the input and output layer neurons, we can make the appropriate distribution surface equation:

$$w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5 + w_6x_6 + w_7x_7 + w_8x_8 + 1 = 0. \qquad (2.49)$$

The scheme of SegmUch algorithm for determining the coefficients of the decision function (2.49) on the basis of the above-mentioned sample data is shown in the graphical part, and the results of machine recognition are shown in Figure 3.11 and the graphical part.

3 DEVELOPMENT OF SOFTWARE

3.1 Variant analysis of software platforms

The software developed in the diploma project is implemented in MATLAB and in the VisualC ++ programming language.

The software of the first four programs is developed in Matlab 6.0 mathematical package, as this package has a special module for working with audio information, which allows you to perform all the necessary operations with audio files. Also, this mathematical package allows you to see the spectral characteristic of a language word that cannot be implemented in other software environments. The developed software in MATLAB is formed as one program and three external functions: SIGNAL_PROCESSING, FILTER, CORRELATIONand SPECGRAM.

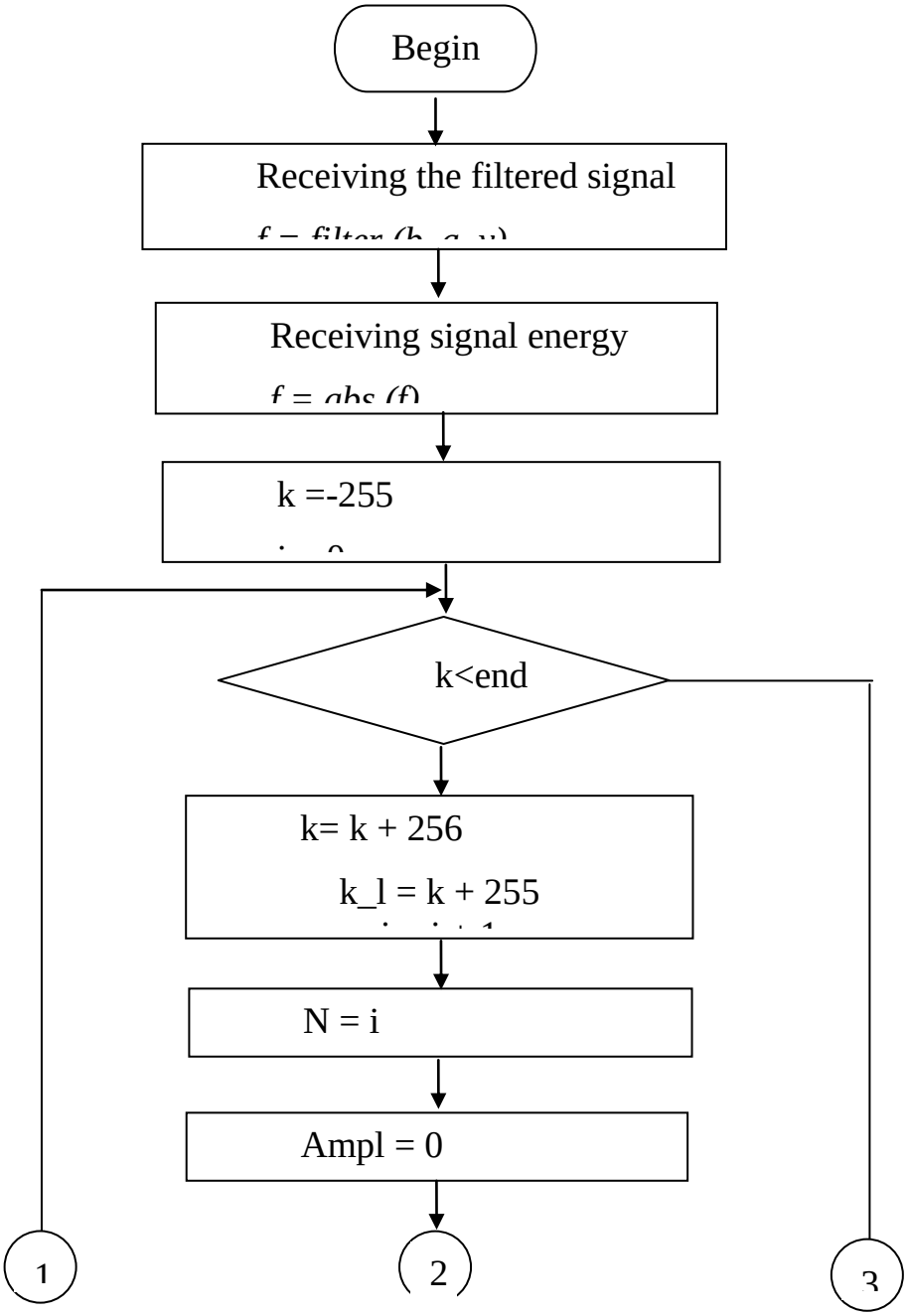
The main one is FEATURE_SELECTION, written in VisualC++ programming language, because in this software environment it is easy to implement a user interface in the form of menus and dialog that cannot be implemented in MATLAB. This program is implemented in this environment, because it is the simplest and easiest to implement the procedures that it must perform, and, most importantly, a user-friendly and user-friendly interface. Visual C++ provides a complete set of structural programming, object-oriented and visual programming. Visual C++ offers an unusually large set of operations. Many Visual C++ operations respond to machine commands and therefore allow live translation into machine code. The variety of operations allows you to choose their different sets to minimize the resulting code. It is through VisualC++ that the user interface for the program in the Visual C++ visual design environment is developed on the basis of dialogue.

The FEATURE_SELECTIONprogram uses data obtained from the programs in MATLAB and performs appropriate operations with them to find informative features of speech patterns.

3.2 Development of speech signal description algorithms and programs

We describe each of the implemented functions and programs.

The FILTER function calculates the parameters of a digital bandpass filter, transmits the input signal and generates an energy vector of the audio signal with a sampling period of 25 ms. The subroutine uses the following data: input vector, upper and lower bandwidth limits, filter number. The scheme of the FILTER algorithm is shown in Figure 3.1:



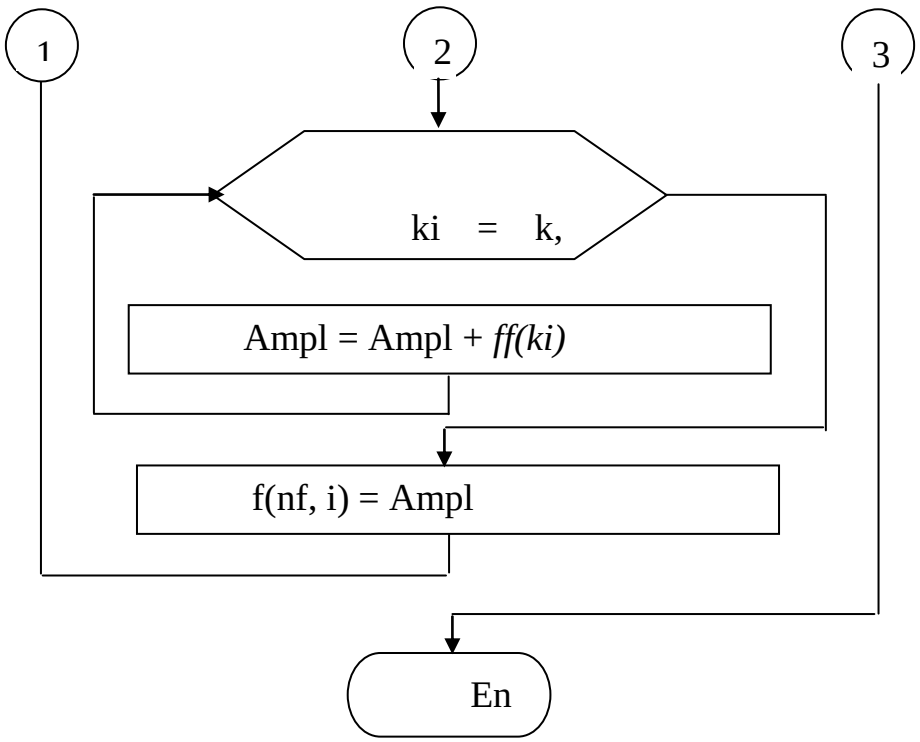


Figure 3.1 – Schematic diagram of the FILTER routine

The result of the subroutine operation will be a ribbon of the matrix, which is an energy vector of the audio signal with a sampling period of 25 ms, which is allocated by a specific bandpass filter. The bandpass number corresponds to the ribbon number of the original matrix.

The second external subroutine, SIGNAL_PROCESSING, has the following features: download an audio file and run it for each of the fifteen FILTER filters. The subroutine uses the following data: input vector, generates upper and lower bandwidth limits of bandpass filters, number of filters.

The frequency ranges of the channels of the 1st, 2nd and 3rd formants can be determined by statistics on the values of the formant bands ΔF_1 , ΔF_2 , ΔF_3:

ΔF_1 = 252 - 1008 Гц,

ΔF_2= 1008 - 2540 Гц,

ΔF_3 = 2540-4032 Гц .

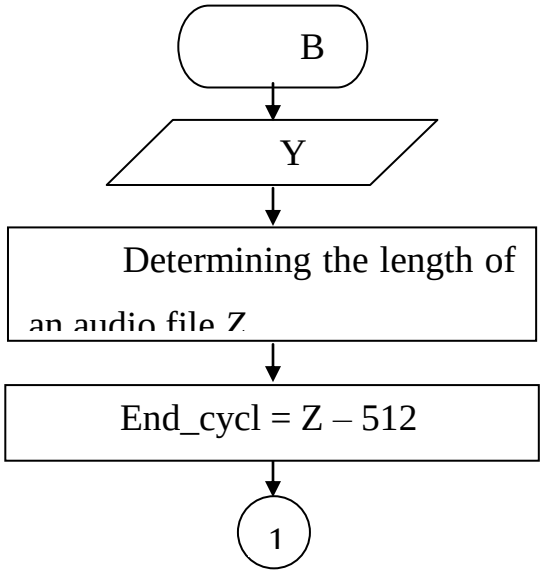
In the spectral analysis of the signal of the speech by bandpass filters with center frequencies arranged according to the logarithmic law along the frequency

axis, the bands ΔF_1 , ΔF_2 i ΔF_3will find the frequency subbands indicated in (Table 3.1), according to which the boundaries of the bands will be distributed SIGNAL_PROCESSING.

Table3.1 – Frequency band limits

Frequency channel number	Frequency range, Hz	Filter number, Hz	Filter bandwidth, Hz	Frequency range Δf_{ki}, Hz	Average subband frequency Δf_{ki0}, Hz
I	ΔF_1 252 – 1008	1	252 – 317	Δf_{11} = 252 – 400	317
		2	317 – 400		
		3	400 – 504	Δf_{12} = 400 – 635	504
		4	504 – 635		
		5	635 – 800	Δf_{13} = 635 – 1008	800
		6	800 – 1008		
II	ΔF_2 1008 – 2540	7	1008 – 1270	Δf_{21} = 1008 – 1600	1270
		8	1270 – 1600	Δf_{22} = 1270 – 2016	1600
		9	1600 – 2016	Δf_{23} = 1600 – 2540	2016
III	ΔF_3 2540 – 4032	10	2016 – 2540	Δf_{31} = 2016 – 2540	2262
		11	2540 – 3200	Δf_{32} = 2540 – 3200	2850
		12	3200 – 4032	Δf_{33} = 3200 – 4032	3592

The schematic of the algorithm routine SIGNAL_PROCESSINGis presented in Figure 3.2:



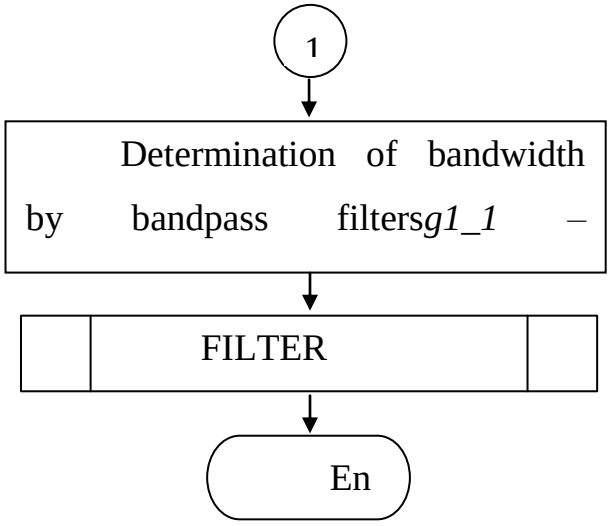
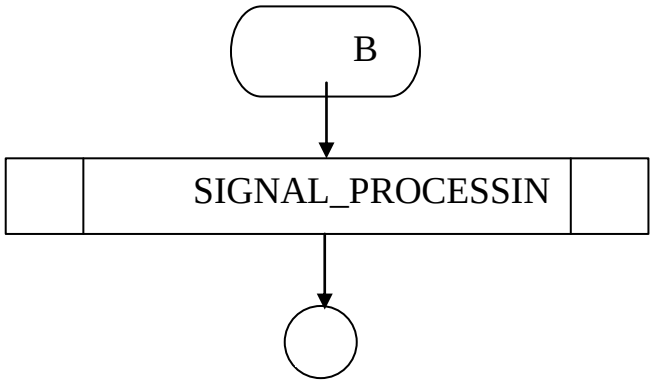


Figure3.2 – Schematic diagram of the subroutine SIGNAL_PROCESSING

The subroutine will result in a matrix containing the values of the sound energy emitted by the fifteen bandpass filters, with a sampling period of 25 ms.

The CORRELATION software module implements speechsignal representation in the space of correlation features. The program uses the data obtained in the SIGNAL_PROCESSINGmodule, namely, the energy matrix of the speech signal decomposed by bandpass filters.

The result of the speech signal description program in the space of correlation features is a correlation three-dimensional matrix of size [N, I, J], where N depends on the size of the input audio file, I = 15, J = 15. Each element of the correlation matrix characterizes the output of signals from the two bandpass filters at a predetermined interval of 25 ms. A diagram of the CORRELATIONsubroutine algorithm is presented in Figure 3.3:



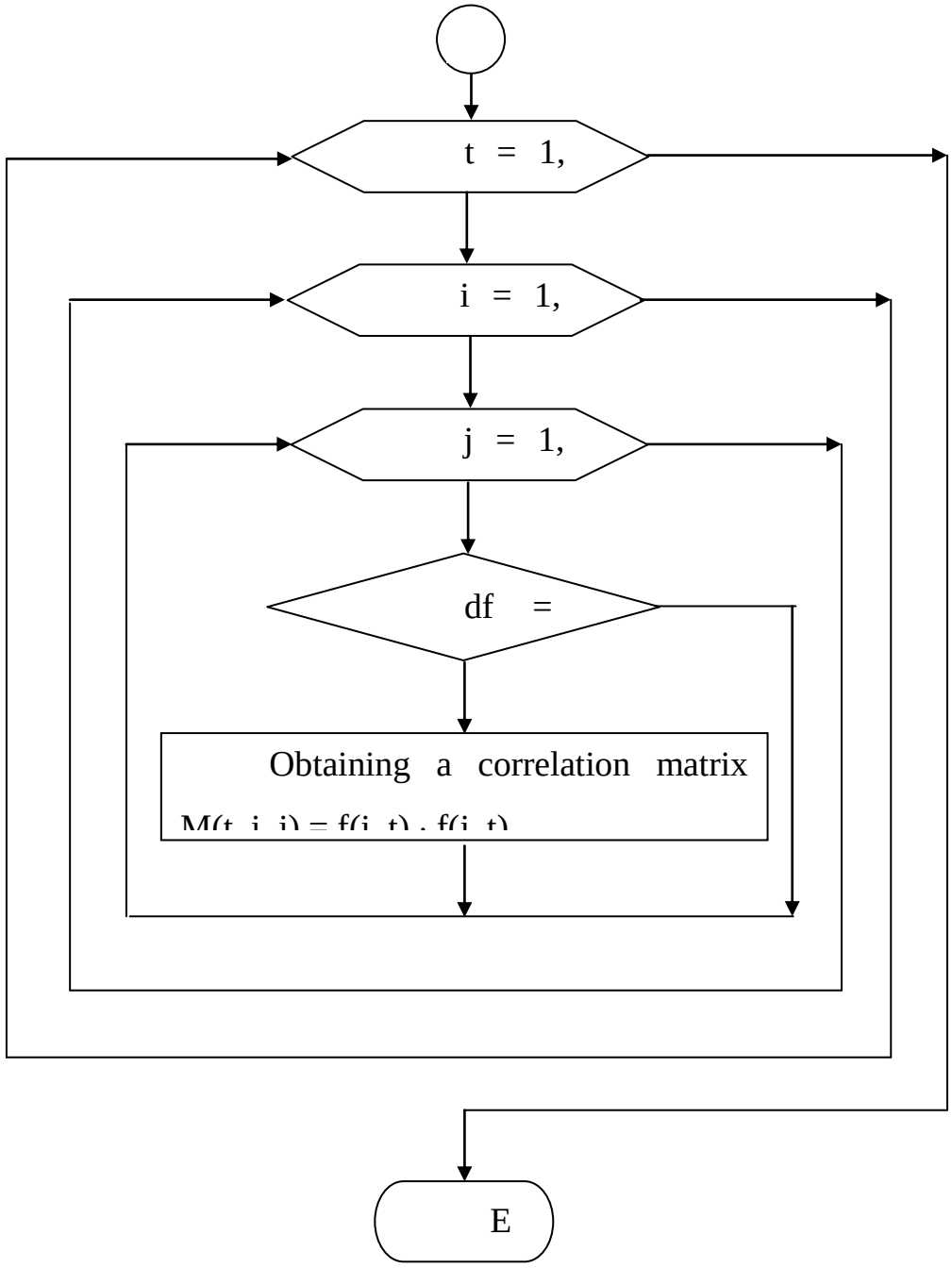
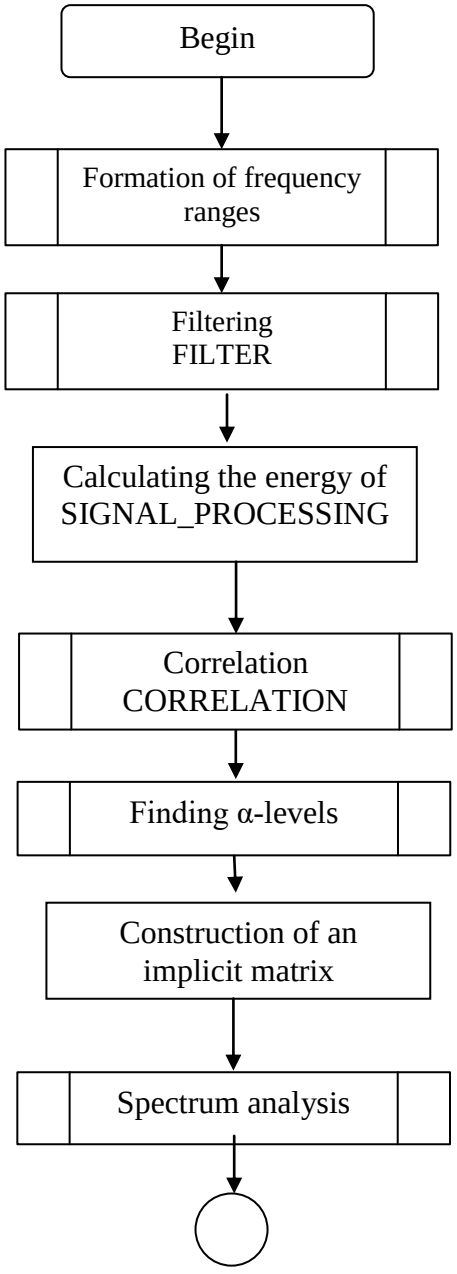


Figure 3.3 – Scheme of CORRELATION subroutine algorithm

The SPECGRAM subroutine shows a speech signal spectrum by which we extract the approximate boundaries of each speech sound to find informative features for each of them.

The FEATURE_SELECTION program implements feature-based feature extraction in the VisualC++ programming language. The program uses the data

obtained from the CORRELATIONmodule, namely a correlation matrix, which is a representation of a speech signal in the space of correlation features and a spectrogram of a speech word. This program converts fuzzy functions into logical ones by applying them to them - a level that is determined by the value of the second after the global maximum of this function. An implicit matrix is then constructed for single intervals. Next, we set the boundaries of each sound of the speech word on the spectrogram to form informative features for them. Then there is the formation on the implicit matrix of ranges for each sound, the energies of which will be informative features(by the operation of a conorm). The result of the program is the selection of informative features for all sounds of the word. The scheme of the algorithm of the program FEATURE_SELECTIONis presented in Figure 3.4:



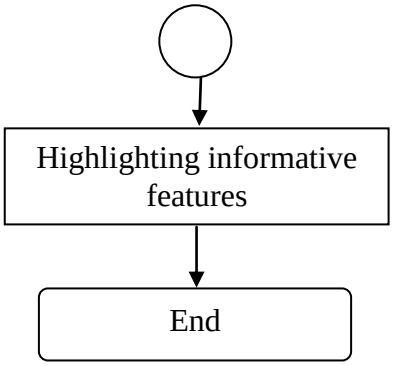


Figure 3.4 – Schematic diagram of the FEATURE_SELECTION program algorithm

Therefore, the FEATURE_SELECTIONsoftware module eventually implements feature extraction based on a fuzzy-logic method. The developed program allows you to segment words automatically into sound ranges, that is, to find ranges of sounds that correspond to certain sounds and informative featuresfor these sounds.

3.3 Development of the structure of input, output and intermediate data

Initial data that is downloaded for signal processing to the program is audio files that must be in wav format, with a sound file length limitation of 2000 (0.2s) to 120,000 (12s).

The audio file is loaded into the signal processing routine, where it is filtered, its energy is calculated and correlation is made.

The output of the whole system is informative, that is, time intervals that correspond to specific sounds and correlations on them, the energies of which are informative for the data intervals of the audio signal.

We describe the structure of all the data that outputs all functions and programs.

The Filter subroutine and SIGNAL_PROCESSINGfunction uses the following data: input vector, upper and lower bandwidth limits, filter number. The result is a

matrix containing the values of the sound energy emitted by the fifteen bandpass filters, with a sampling period of 25 ms. The data structure is presented in Table 3.2:

Table 3.2 – Data structure of Filter subroutine and SIGNAL_PROCESSINGfunctions

№	Name	Description	Comments
1	y	Audio file	Input file with extension "wav"
2	f	The energy values of the audio signal	Matrix 36x15 36 – the number of time frames
3	pr1	the upper bandwidth limits of the bandpass filter	An elliptical filter is used
4	pr2	the lower bandwidth limits of the bandpass filter	An elliptical filter is used
5	nf	filter number	Number of filters –15

The CORRELATIONfunction implements the representation of a speech signal in the space of correlation features. The program uses the data obtained from the SIGNAL_PROCESSINGmodule, namely, the energy matrix of the speech signal decomposed by bandpass filters. The result of the speech signal description program in the space of correlation features is a correlation three-dimensional matrix of size [N, I, J], where N depends on the size of the input audio file, I = 15, J = 15. The data structure represented by the CORRELATIONsubroutine is presented in Table 3.3:

Table3.3 – The data structure of the CORRELATIONroutine

№	Name	Description	Comment
1	M(:,:,1)	Correlation matrixM($E_{1-15}E_{15}$)by columns from time t(1-36)	Matrix 15x36, where 15 – number of correlations, 36 –number of time frames
2	M(:,:,15)	Correlation matrix M($E_{1-15}E_{15}$) by time	Matrix 15x36

There are 15 such matrices, but only the first and the last are presented in Table 3.3.

The FEATURE_SELECTIONprogram uses the data obtained from the CORRELATIONmodule, namely a correlation matrix, which is a representation of a speech signal in the space of correlation features. The result of the program is the selection of informative features, that is, time intervals and correlations on them, the energies of which are informative for the data intervals of the audio signal.

The data structure in FEATURE_SELECTIONis shown in Table 3.4:

Table3.4 – The data structure of the FEATURE_SELECTIONroutine

№	Name	Description	Comment
1	New2	Correlations of all energies of the audio signal over time	A text file with the extension "txt"
2	Test	Implicitmatrix	Out file
3	Logical2	Informative features for timing	Out file

Therefore, all the data are interconnected and together with the features and programs described above allow the development of an automatic feature extraction system.

3.4 Development of program documentation

3.4.1 Description of the logical structure of the program

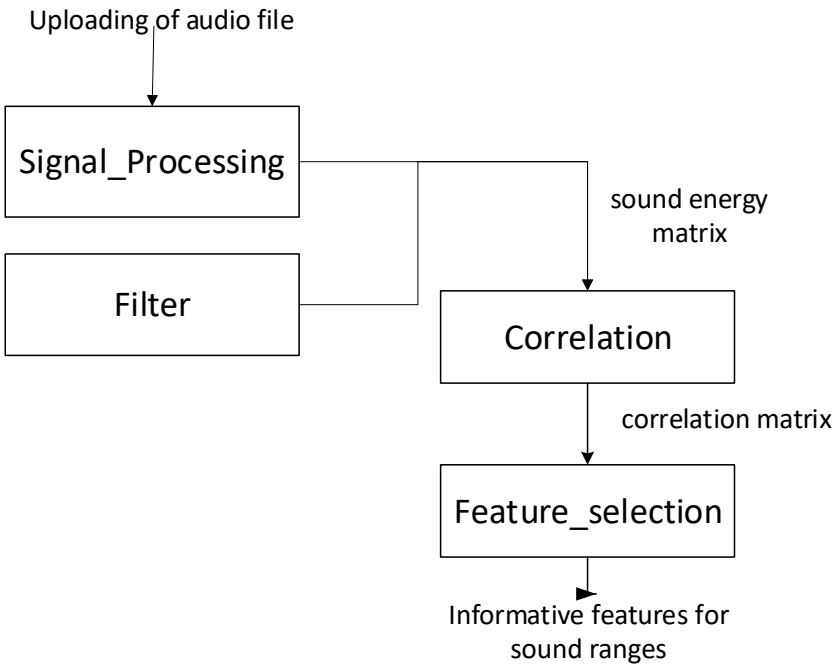


Figure3.5 – Scheme of interaction of software modules

We justify this scheme. First, the audio file is loaded using SIGNAL_PROCESSING, then it is filtered by the Filter module, and as a result we get a filter filtered by 15 filters, which is a matrix of energies of the audio signal from the time frames with a sampling period of 25ms. Next, we do the energy exacerbation with the help of the CORRELATION function, which as a result produces a correlation matrix for all energies and time counts. This data is used by the main FEATURE_SELECTION program, which as a result extracts informative features for sound ranges.

The FEATURE_SELECTION software module implements feature extraction based on a fuzzy-logic method. The program uses the FILTER module, namely, filters the input signal and generates an energy vector of the audio signal. The program uses the data obtained in the SIGNAL_PROCESSING module, namely, the energy matrix of the speech signal decomposed by bandpass filters. The CORRELATION program is then used, namely, a correlation matrix, which is a representation of a speech signal in the space of correlation features. After that, the obtained fuzzy functions are transformed into logical ones by applying to them - the level, which is determined by the value of the second after the global maximum of

this function. An implicit matrix is then constructed for single intervals. The following is the formation on the implicit matrix of ranges, the energies of which will be informative features(by the operation of a conorm).

The result of the program is the selection of informative features.

Therefore, the program is formed in the form of two subroutines and two external functions. External functions are universal and are called by both programs.

3.4.2 Operator manuals

The program has a size of 3 KB. Runs the file FEATURE_SELECTION.exe.

In order to work with this program you need:

1. Run SIGNAL_PROCESSINGto execute it to download the audio file and filter the input using Filter.

2. Run CORRELATION to create an energy correlation matrix. The correlation energies will be stored in the New2.txt file.

3. Run FEATURE_SELECTION to find bands whose energies will be informative.

4. In the window that appears, you can see all the results that the program has completed:

- Correlations of all energies by time intervals can be viewed by using the buttons in the program interface: "Out", "Correlations of energy"

- Implicit matrix values. Can be viewed by clicking the Output, Implicit Matrix buttons sequentially.

- Selection of informative featuresat intervals of time can be pressed "Out", "Informative features".

5. You can see the progress of the program from the menu item "Help", "About the program"

6. Exit the program by clicking on the menu item "Exit"

3.5 Results of the machine experiment

The result of the operation of the program of description of the speech signal by the method of quasi-frequency detector is a set of bytes. To describe a segment of a speech signal two bytes. The first byte is the expression of the frequency-detecting function, the second is the segmentation function. This pair of bytes is generated every 20 ms. and characterizes a twenty-millisecond time span. Using a set of bytes, that is, the description of the speech signal in the machine codes, allows you to achieve the required performance of the entire recognition system, since the data presented in the machine codes are most quickly processed by a digital computer. Appendix B presents an excerpt from the complete set of bytes of description of the speech signal for three speakers. When analyzing the results of a speech signal description program, it can be concluded that this technique can be used to create a speaker independent automatic speech recognition system.

Let's explore two words: "Меньше" and "Машина". To work with the FEATURE_SELECTION.exe informative feature extractor program, we first process the speech signal by determining its length and the energy values of the signal in the time domain using the SIGNAL_PROCESSINGprogram, which in turn will call the Filter function and filter it. Then, using CORRELATION, we get correlated energy values at intervals. Having stored the data in a file with the extension .txt, we can find informative featuresfor this speechusing FEATURE_SELECTION.exe. This program has a simple and user-friendly interface: a window with a menu bar.

You can find the correlated energy values of any word using the Open File menu item in the main program. Let's first examine the information features of the word "Less", for which we open a file with the correlated energy values of this word, which were obtained from MatLab.

To view with the help of the energy correlation program that the program will work with, select the menu item "Output / Correlations".

An illustration of these actions and menu items is shown in Figure 3.5:

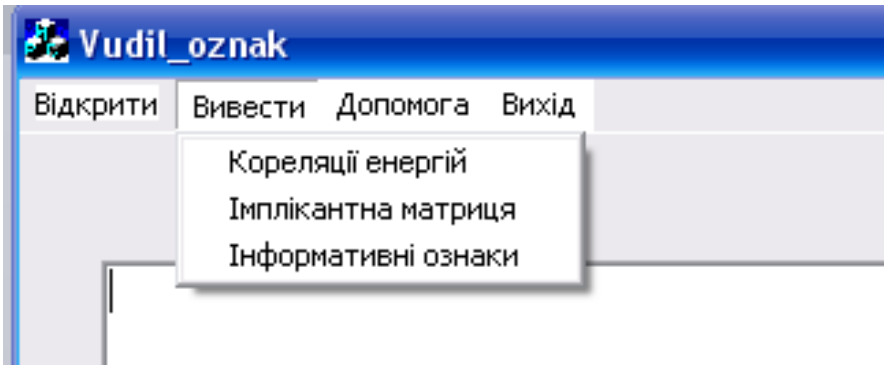


Figure 3.5 – Main program menu items

A matrix of correlations of all energies, from $E_{1\text{-}15}E1$ до $E_{1\text{-}15}E_{15}$, will be displayed, with each band being the value of these correlations at intervals of time every 25 ms. Part of the output of this matrix is shown in Figure 3.6.

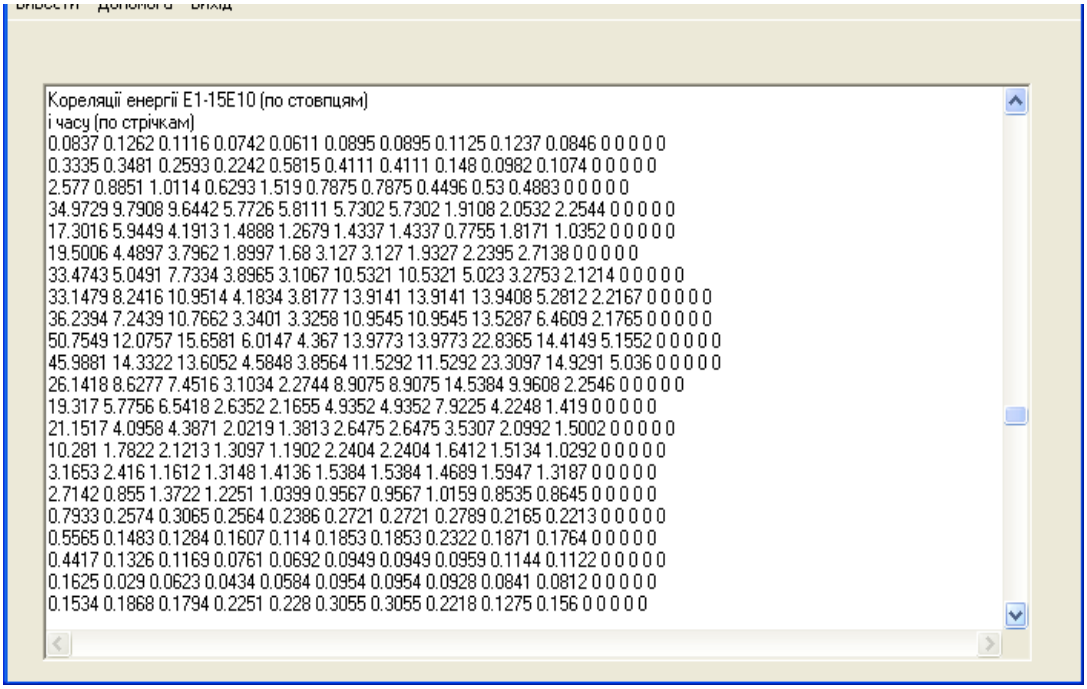


Figure 3.6 – The matrix of energy correlations

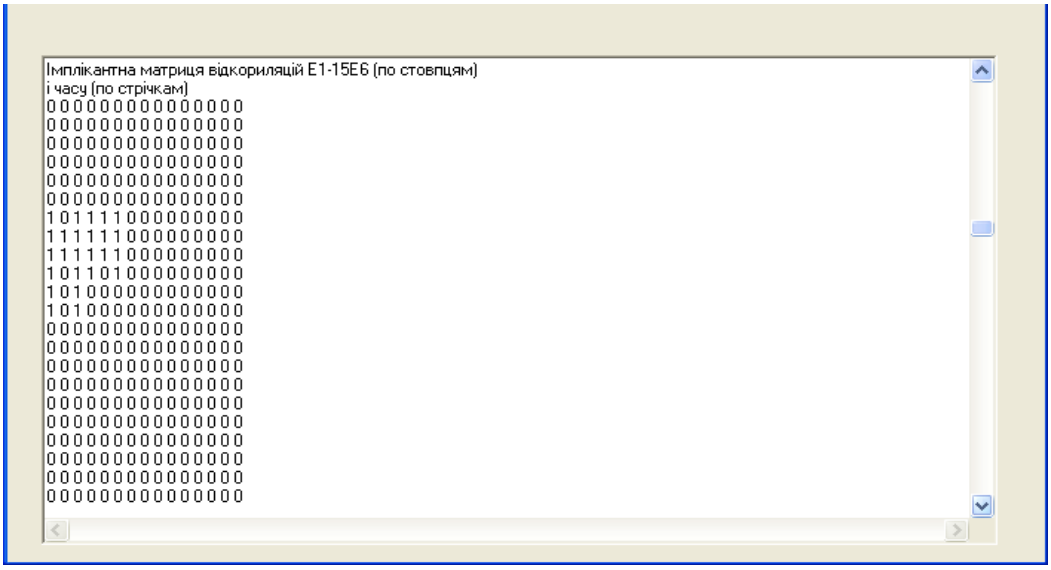


Figure 3.7 – Implicit matrix from $E_{1\text{-}15}E_6$correlations and time

The program builds an implicit matrix, which can be viewed using the "Output / Implicit Matrix" menu item. The image of the implicit matrix is shown in Figure 3.7.

The results of the correct output of the implicit matrix can be viewed using graphs. We construct, for example, a graph of E_1E_6 correlation from time intervals. The implicit matrix in the first column shows that the values of the units should be in the time intervals 7-12, ie below this interval immediately there is a second maximum, above which all values must have the values of logical units. According to the graph in Figure 3.8, this is indeed true. Having done the rest of the similar checks, we can conclude that the implicit matrix is constructed correctly.

Next, to display informative features in the New Range window, we will set the intervals of speech sounds, which will be selected using the spectrogram of the same word. The spectrogram of the word "Менше" is shown in Appendix B.

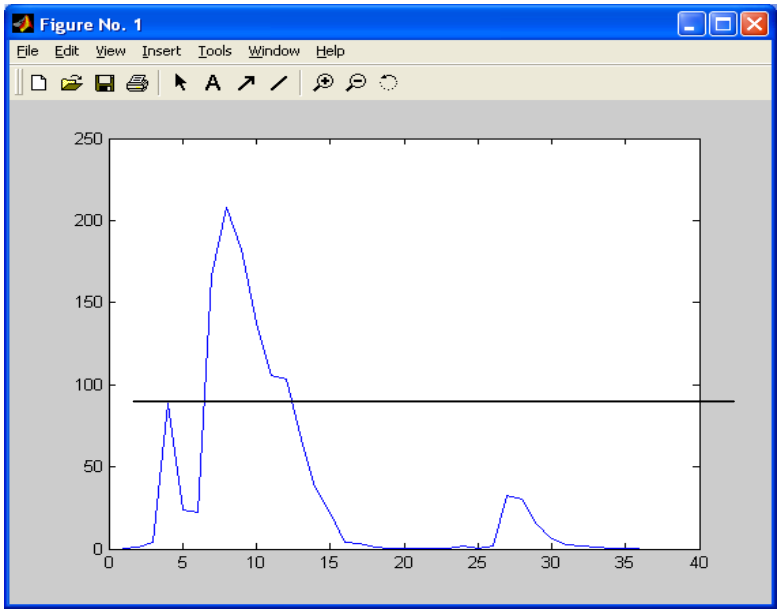


Figure 3.8 – Graph of correlation of E1E6 on time intervals

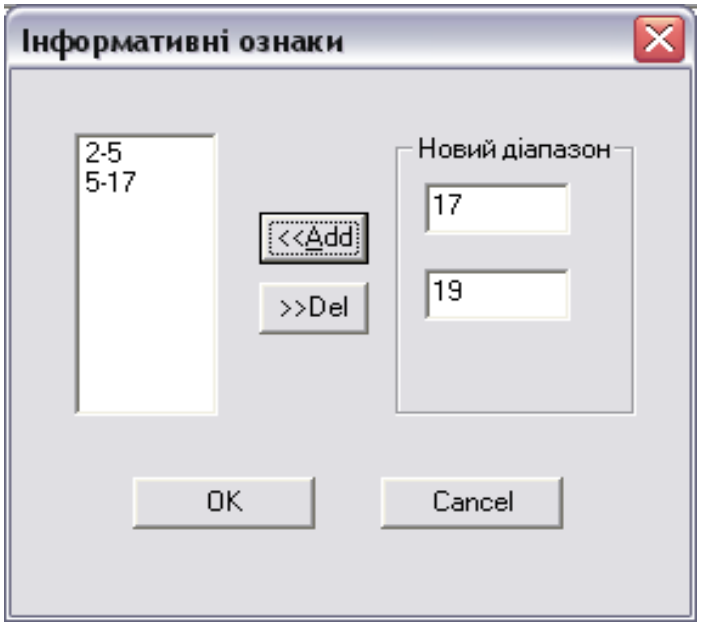


Figure 3.9 – Specify a range of speech sounds

When you click "Ok", the results of the program will be displayed, that is, informative featuresof the sounds, the intervals of which we set. The result is shown in Figure 3.10:

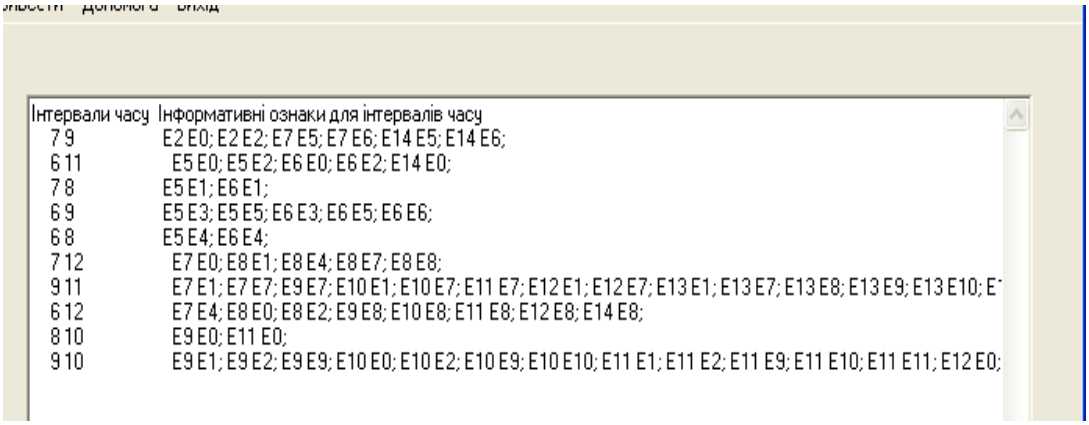


Figure 3.10 – Informative features of speech sounds

Thus, by specifying all the intervals of speech sounds of a word, we obtain informative featuresfor each sound present in the studied word. In order to find informative featuresof other sounds, we open a file with energy correlations, such as the word "System". Similarly, in the "New Range", we set the intervals of sounds, which will also be highlighted in the spectrogram of a given word and the result will be obtained. The syllabus of the word system is given in Appendix B.

Thus, the developed program package allows to obtain informative featuresfor all sounds.

4IMPLEMENTATION AND DISSEMINATION OF THE DEVELOPED SYSTEM

The purpose of this section is to conduct a marketing analysis of the project to determine the feasibility of its market implementation and possible directions for its implementation.

4.1 Development prospects

The main obstacles to the further development of automated speech recognition systems are:

- the need for large volumes of dictionaries;
- noisy speech signal;
- various accents and pronunciations.

The volumes of dictionaries determine the degree of complexity, the requirements for computing power and the reliability of speech recognition systems.

This will solve the problems associated with morphology, accents, pitch, tempo, volume, merging words, articulation, speech information, etc. It is expected that the main direction of development will be language modeling for use in speech recognition systems.

The problem of isolating the speech signal from the noise background has not been finally resolved either. Currently, users of voice recognition systems are forced to work in conditions of minimal noise background.

One of the priority developments in the field of speech recognition is man-machine dialogue systems, which are being worked on in many research laboratories around the world. One of such developments is the technical system of AT&T (USA), which is used for speech recognition in the telephone network: the client can request one of five categories of services using any words; he speaks until one of the

five key words appears in his statement. This system currently serves about a billion calls per year [32].

Such systems “are able” to work with a continuous speech stream and with unknown speakers, understand the meaning of fragments of speech of a limited vocabulary, and take response actions. Systems work in real timeand are able to perform five functions:

1. Speech recognition - the conversion of speech into text consisting of individual words.

2. Understanding - grammar analysis of sentences and recognition of semantic meaning.

3. Information recovery - receiving data from operational sources on the basis of the semantic value obtained.

4.Speech information generation - construction of sentences representing the obtained data in the language chosen by the user.

5. Speech synthesis - the conversion of sentences into computer-synthesized speech [32].

4.2 Areas of application

We outline the main areas of application of speech recognition systems:

1. Automated user interface. Today, for many people, communication with a computer is still difficult. Speech recognition systems overcome these difficulties. A huge advantage of voice recognition systems is that they are much faster than any other types of interfaces. Voice email program allows you to turn on the computer, dictate and send messages without touching the mouse and keyboard. Also, people with physical disabilities will receive a more effective way of interacting with a computer. The most obvious use of continuous speech recognition system is to createautomatic shorthand systems that can replace secretaries when dictating voice texts of letters, notes in a diary, reports. In this case, there is not only savings due to

reduction in the work of the stenographer, but also an increase in the degree of confidentiality of information.

2. Mobile device management. It is known how inconvenient and dangerous it is to use mobile phones with the usual (tactile) way of dialing while driving. Many countries have passed laws banning the use of such phones by drivers in order to reduce the number of accidents. Therefore, lately, mobile phones with voice dialing have been popular, saving the user from having to dial the desired number manually. It is enough to say the name of the subscriber, and the connection will happen automatically. Control and management audio systems are already used in vehicles of some manufacturers. The owner of the car gives voice commands to control the temperature, radio, navigation system, which receive voice and execute commands (DIVO and VoiceCommander).

3. Information services. Modern speech recognition systems are used, for example, for booking airline tickets, viewing news, accessing databases.Voice recognition technology has quickly changed the market for telephone services. Speech recognition systems work in call centers. These systems allow you to automate the dialogue with the client, as a result of which there is no need for a huge number of operators who receive phone calls, and reduced staff costs. In addition, the quality of customer service is improved, since the connection to the machine is carried out almost immediately, saving customers from the long wait for a free operator on the line.

4. Business and professional support. For many years, voice recorder systems designed for representatives of certain professions, for example, doctors and lawyers, can be found in the software market.

Many representatives of these professions use speech recognition systems in everyday work. Voice activated home appliances and appliances have become popular.

5. Combined human-machine interfaces. Over the past decade, the applications of such systems have expanded significantly and will continue to expand. They are used, in particular, to control limited access to an object by recognizing a person’s

face and speech, and performing financial transactions using speech and touch screens of ATMs [32].

4.3Problems of implementation of speech recognition systems

Consider aspects that impede the global solution to the problem of quality speech recognition.

1. The pace of speech varies widely, often several times. At the same time, the various sounds of speech are stretched or squeezed out of proportion. For example, vowels change much more strongly than half-vowels and especially vocal consonants. For the so-called slotted sounds there are patterns.

2. By pronouncing the same word or phrase at different times, under the influence of various factors (mood, state of health, etc.), we generate noticeably different spectral-temporal energy distributions. This is true even for a double-spoken word. This effect is much stronger when compared spectrograms of the same phrase uttered by different people. This effect is usually called the spectral non-stationarity of speech samples (see examples of spectrograms).

3. A change in the rate of speech and clarity of pronunciation is the cause of co-articulation non-stationarity, which means a change in the mutual influence of neighboring sounds from sample to sample.

4. The problem of clustering continuous speech: in a continuous speech stream, it is difficult to recognize speech units due to inaccurate definition of boundaries. Here are just a few reasons that prevent the full implementation of speech recognition systems [32].

Conclusions. Limitations of the use of speech recognition systems within the framework of the most traditional applications allow us to conclude that it is necessary to search for potentially new solutions in the field of speech recognition. In the next decade, the task of recognizing and understanding natural speech, regardless of language and speaker, will occupy a central place in speech technologies.

CONCLUSIONS

This master's thesis deals with the problems of speech recognition, namely, the study of methods for identifying speechsounds.

In the section of feasibility study, the analysis of the existing models of speech making is made, the choice of the optimal variant of the solution of the task set in the technical task is made, and also the approximate calculations of the economic feasibility of the new development are made.

In the special part, a mathematical model based on two modulation methods is developed: the 1st method is based on the model of "quasi-frequency modulation" of the speech signal and is applied at the stage of initial description of the sound types; Method 2 is based on a fuzzy-logic model and is used to describe sounds that correspond to particular phonemes of the speech. The offered methods allow to apply logical procedures of description, segmentation, and identification of sounds and to automate the process of selection of informative parameters. According to the terms of reference, algorithms for the identification of sounds in automated keyword search systems were developed based on the developed mathematical model. Matlab 5.0 and Visual C++ have developed software that implements the relevant algorithms.

As a result of the machine experiment, the results were obtained, which confirm the accuracy of the proposed signal description techniques. The analysis of the machine experiment proves that the applied technique allows obtaining results invariant to the speaker's voice. Therefore, the proposed technique can be used to create automated keyword search systems.

In the business part, was conducted a marketing analysis of the project to determine the feasibility of its market implementation and possible directions for its implementation.

REFERENCES

1. Биков М.М. Розпізнавання мовних образів з використанням нейромережевого підходу /[Биков М.М., Грищук Т.В.] МКІМ – 2002. Міжнародна конференція з індуктивного моделювання, Львів 20 – 25 травня 2002: Праці в 4-х томах. – Львів, ДНДІ інформаційної інфраструктури, 2002. – Т1. ч2. – С.203 – 207

2. Джелинек Ф. Разработка экспериментального устройства, распознающего раздельно произносимые слова / Джелинек Ф. - ТИИЭР. – 1995. - №11. - С. 91-100.

3. Рабинер Л.Р. Цифровая обработка речевых сигналов/ Рабинер Л.Р.Шафер Р.В -Пер. с англ. М.: Радио и связь, 1981. – 496 с.

4. Буриляну К. Характеристика огранниченного словаря отдельно произносимых слов с целью автоматического рапознавания речи. / Буриляну К. М:,1986. - Т. 14. - С. 36-131.

5. Биков М.М. Методи підвищення дикторонезалежності опису і розпізнавання мовної інформації в мережі INTERNET / Биков М.М., Грищук Т.В./ “Інтернет – Освіта – Наука - 2002”, третя міжнародна конференція ІОН – 2002, 8 – 12 жовтня 2002 р. Збірник матеріалів конференції. – Вінниця: УНІВЕРСУМ – Вінниця, 2002. – Том 2. – С. – 329-332.

6. Биков М.М. Використання нейронних мереж для розпізнавання звуків мови / Биков М.М., Грищук Т.В., Янчук О.В. Оптикоелектронні інформаційно-енергетичні технології. –2001. - №2. - С. 92-97

7. Ковтун В.В. Вибір інформативних ознак в задачі ідентифікації диктора // МКІМ – 2002. Міжнародна конференція з індуктивного моделювання. Львів, 20 – 25 травня 2002: Праці в 4-х томах. – Львів, ДНДІ, 2002. – Т1. ч2. С.280 – 287

8. Ротштейн А.П. Интеллектуальные технологии идентификации: нечеткие множества, генетические алгоритмы, нейронные сети/ Ротштейн А.П.- Винница: УНИВЕРСУМ-Вінниця, 1999. – 320 с.

9. Learning Spectral-temporal dependencies using connectionist networks/ Lubensky David.// ICASSP 88: Int. Conf. Acoust., Speech and Signal Process, New York City, Apr. 11-14, 1988. – New York (N. Y.). - 1988- Vol. 1.S. .- P.418-427.

10. A spatio-temporal pattern recognition approach to word recognition/ Tom M. Daniel, Tenorio M. Fernando // IJCNN Int. Joint Conf. Neural Networks, Wachington, 1989. - New York (N. Y.). – 1989. - Vol. 1. – P.351-355.

11. Phonetic Speech recognition system based on a neural network/ Kepuska Veton Z., Gowdy John N. // IEEE SOUTHEASTCON ' 89: Conf. and Exhib. "Energy and Inf. Technol. Southeast.", Columbia, S. C., Apr. 9-12, 1989. – New York (N. Y.). – 1989. - Vol. 2. – P.770-775.

12. Word recognition with neural network / Albesano D., Gemello R, Mana F. // CSELT Techn. Repts. – 1992. – №6. – P.553-559.

13. William C. Newman. Detecting speech with an adaptive neural network. // ED 1990. – 1990. - № 6. - P. 79-90.

14. Быков Н.М. Методы и средства измерения и преобразования информации в системах машинного распознавания речи./ Быков Н.М.- Дис. на соискание уч. ст. канд. техн. наук.-Винница, ВПИ, 1985.-243 с.

15. Чумаков Н.М. Оценка эффективности сложных технических устройств./ Чумаков Н.М., Серебряный Е.И. - М.: Советское радио,1982.-192 с.

16. Кузьмин И.В. Оценка эффективности и оптимизация АСКУ./ Кузьмин И.В. – М.: Советское радио,1971.-296 с.

17. Быков Н.М. - Анализ эффективности средств распознавания в системах принятия решений.- В кн.: Эффективность и надежность автоматизированных систем принятия решений./ Быков Н.М., Данилюк Ю.С., Ильницкий Н.П. - Киев: ИК АН УССР, 1982. - С. 24 - 29.

18. Быков Н.М. Модель потенциальной системы для распознавания речи в СЧМ.- В кн.: Исследование и проектирование систем "человек-машина"./ Быков Н.М., Агеев А.С.-Киев:ИК АН УССР. - 1989.- С.57-61.

19. Сапожков М.А. Речевой сигнал в кибернетике и связи./ Сапожков М.А. – М.: Связьиздат, 1963. – 452 с.

20. Биков М.М. Ідентифікація передаточних характеристик акустичного каналу в системах розпізнавання мови/ Биков М.М., Кузьмін І.В., Грищук Т.В., Ковтун В.В. – В кн.: Матеріали міжнарожної конференції “Автоматика 2002”, Донецьк, 2002. – С.107

21. Фант Г. Акустическая теория речеобразования. – М.: Наука, 1964. – 284с.

22. Харкевич А.А. Спектры и анализ. – М.: Физматиз, 1962. – 320 с.

23. Маркел Д.Д., Грей А.Х. Линейное предсказание речи. – М.: Связь, 1980. – 380с.

24. Бондарко Л.В., Загоруйко Н.Г. и др. Модель восприятия речи человеком. – Новосибирск: Наука, 1968. – 58 с.

25. Галайчук Г.Л., Перваченко В.С. Об одной модели слухового аппарата человека // Вестник КПИ. Серия электроакустики и звукотехники. – 1977. - Т.1. - С.10-11.

26. Лабутин В.К., Молчанов А.П. Модели механизмов слуха. – М.: Энергия, 1973. – 200 с.

27. Кельманов А.В. Решения проблем выделения спектральных кореляционных признаков речевого сигнала – М.: Мир, 1991. – 170 с.

28. Биков М.М., Ковтун В.В., Савінова Н.Г., Пристрій для виділення складових сегментів у мовному сигналі // Патент на корисну модель. – 2007. – №19. –11с.

29. Вачевський М.А. і інші. Маркетинг для менеджера. – Стрий: Просвіта, 1990, – 267 С.

30. Винцюк Т.К. Анализ, распознавания и интерпретация речевых сигналов. – К.: Наукова думка, 1987, – 195 С.

31. Винцюк Т.К. Проблема распознавания речи . – К.: Наукова думка, 1989. – 211 с.

32. Распознавание речи: Этапи развития, современные технологии и перспективы их приминения[Електронний ресурс] – Режим доступу до ресурсу:http://openarchive.nure.ua/bitstream/document/884/1/Bionica-2010-2(73)-164-168.pdf

Appendix A

(Mandatory)

VNTU

APPROVED

Head of the AIIT department

D.Sc., prof. R.N. Kvyetnyy

_____

«04» October 2019 y.

TECHNICAL ASSIGNMENT

of performing a master's qualification work

«Development and research of methods and models of speaker-independent identification of phonotypes in intelligent automation systems»

08.01.MKP.012. 00.000 ТЗ

Supervisor:

Ph.D., prof.M. M. Bykov

“02” ___October__ 2019 y.

Developer: st. of gr. 3ACIT–18m

V. V. Stetsiuk

“___” _____ 2019 y.

Vinnytsia 2019

1. Title and field of application

1.1. Title – Development and research of methods and models of speakerindependent identification of phonotypes in intelligent automation systems.

1.2. Field of application– automation and instrumentation.

2. The basis for the development.

The topic of master's thesis is approved by the order of VNTU № ___ from "___"09 2019 y.

3. Purpose.

The purpose of the master's qualification work is to increase the efficiency of the process of recognition of speech patterns.

4. Initial data for development.

Master's qualification work is performed for the first time. The following documents should be used during development:

1. 1. Ахо А. Теория синтаксического анализа, перевода и компиляции "Синтаксический аналіз" / А. Ахо, Дж. Ульман. – М. : Мир, 1978. – 224 с.

2. Биков М. М. Використання нейронних мереж для розпізнавання звуків мови / М. М. Биков, Т. В. Грищук, А. А. Раїмі // Оптикоелектронні інформаційно-енергетичні технології. – 2001. – №2. – С. 92–97. ISSN: 1681-7893.

3. Винцюк Т. К. Анализ, распознавание и интерпретация речевых сигналов / Винцюк Т. К. – К. : Наукова думка, 1987. – 261с.

4. Винцюк Т. К. О математических моделях речевого сигнала, используемых в распознавании речи / Т. К. Винцюк // Автоматическое распознавание слуховых образов. – Киев: ИК АН УССР, 1982. – С. 34- 37.

5. Development requirements.

5.1. List of main features:

- load audio file;
- filtering the input signal by all filters;

- selection of energy values of filtered signals;
- correlation of all possible combinations of signal energies and finding a correlation matrix, which is a representation of the speech signal in the space of correlation features;
- automation of feature selection to describe predetermined phonotype classes in automatic speech recognition systems.

5.2. Basic technical requirements for development.

5.2.1. Requirements for the software platform:

- WINDOWS 7\8\10;
- Matlab 7.0;
- Microsoft Visual Studio.

5.2.2. Operating conditions of the system:

- work on standard in rooms with standard conditions;
- possibility of round-the-clock functioning of the system;
- the text of the system software is completely closed.

6. Stages of development.

6.1 Explanatory note

1	Feasibility study of selected methods of phonotype identification in intelligent automation systems	02.10.2019y.
2	Development of methods and mathematical models for automated selection of phonotype features	20.10.2019y.
3	Software development and machine experiment	02.11.2019y.
4	Testing the results of the study	22.11.2019y.
5	Publications	
6	Design of explanatory note, graphic material and presentation	10.12.2019 y.
7	Master’s thesis defense	12.12.2019 y.

6.2 Graphic materials:

- structural diagram of the system: «10» December 2019 y.
- data schema: «10» December2019 y.
- diagrams of algorithms of system functioning«11» December2019 y.
- demonstration posters: «11» December2019 y.

7. The order of control and acceptance.

7.1. The course of master's qualification work is controlled by the head of work. The border control should be performed till «07» December 2019 y.

7.2. Certification of the project is carried out on preliminary defense. Preliminary defense of master's qualification work to be carried out till «10» December 2019 y.

7.3. The final decision on the evaluation of the quality of performance of the master's qualification work is made at a meeting of the commision. Defense the master's qualification work«12» December 2019 y.

AppendixB

(Mandatory)

APPROVED
Head of the AIIT department
D.Sc., prof. R.N. Kvyetnyy
_____
« _____» _____2019 y.

LIST OF MATERIALS

to protect the master's qualification work on the subject

DEVELOPMENT AND RESEARCH OF METHODS AND MODELS OF SPEAKER INDEPENDENCE IDENTIFICATION OF PHONOTYPES IN INTELLECTUAL AUTOMATION SYSTEMS

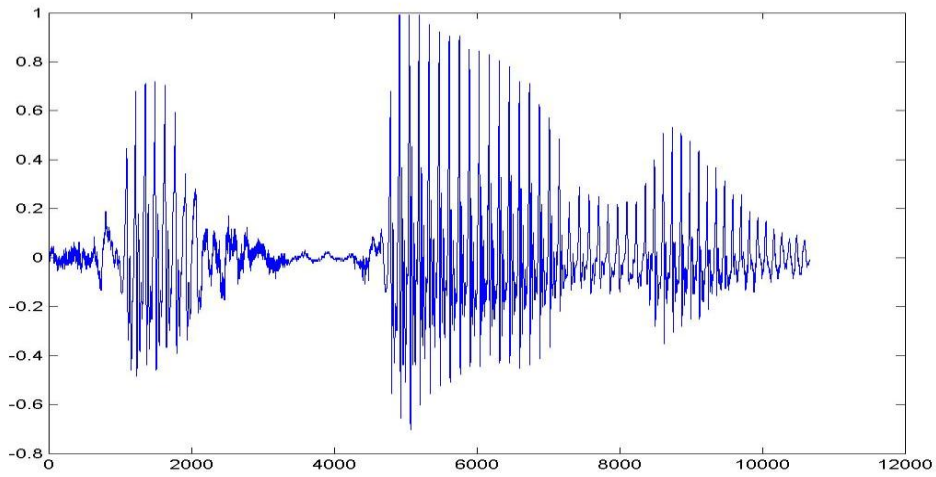
1. Methods of selection of features.
2. Models of quasi-frequency modulator and fuzzy-logic method.
3. Scheme of algorithm of selection of informative features by the fuzzy-logical method.
4. SIGNAL_PROCESSINGalgorithm scheme.
5. Scheme of interaction of software modules for the fuzzy-logical method of feature selection.
6. Use-case UML-diagram
7. Scheme of feature selection algorithm by quasi-frequency method.
8. Scheme of the algorithm of training of the classifier on quasi-frequency characteristics.
9. Experimental studies. Fuzzy_ logical method of feature selection.
10. Experimental studies. The results of the quasi-frequency feature detector.
11. Experimental studies. Results of operation of the frequency-segmenting classifier.

Developed by: StetsiukV.V.
_____ _____
(signature) (date)
Verified by: BykovM. M.
_____ _____
(signature) (date)
Reviwed by: Palamarchuk Y. A.
_____ _____
(signature) (date)

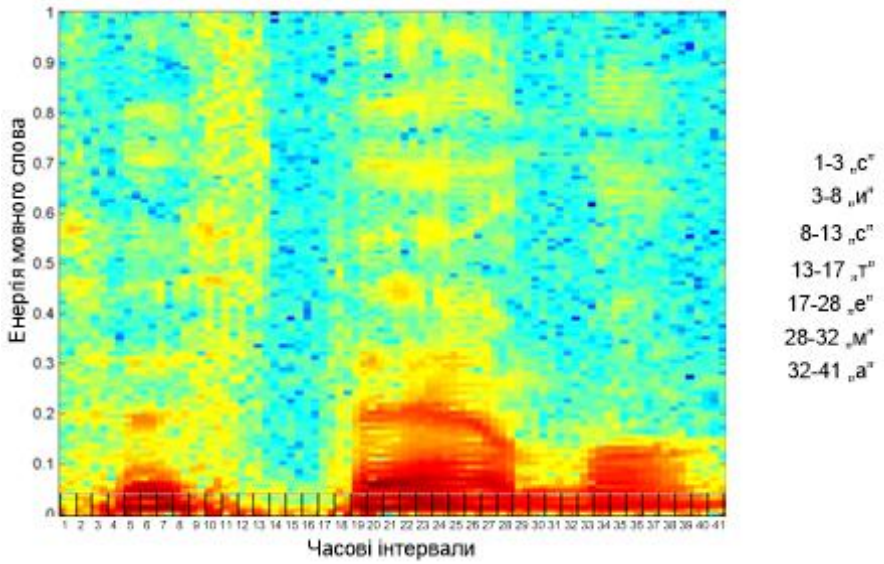
METHODS OF SELECTION OF FEATURES

Method 1 is based on a fuzzy-logic model

Method 2 is based on the model of "quasi-frequency modulation" of the speech signal;



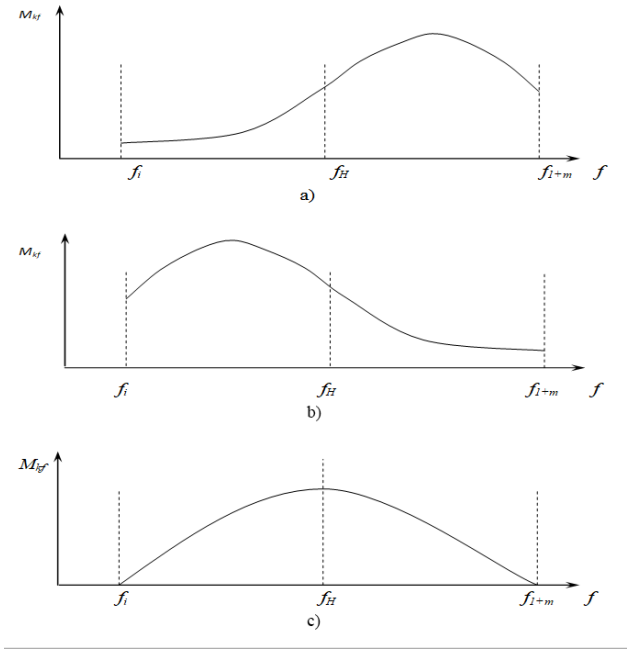
The signal of the word "система" as a function of time



Spectrogram of speech signal

MODEL OF QUASI FREQUENCY MODULATION

The coded formant positions in the frequency channels:



a) high frequency torque value; b) low frequency torque value; c) average frequency torquevalue

State of signal frequency moments:

$$M_{kf} = \frac{\int_{F_{k-1}}^{F_k} A_f \cdot f df}{\int_{F_{k-1}}^{F_k} f df} \qquad M_{kf} = \frac{\sum_{i=1}^{1+m} A_i \cdot f_i}{\sum_{i=1}^{1+m} f_i}$$

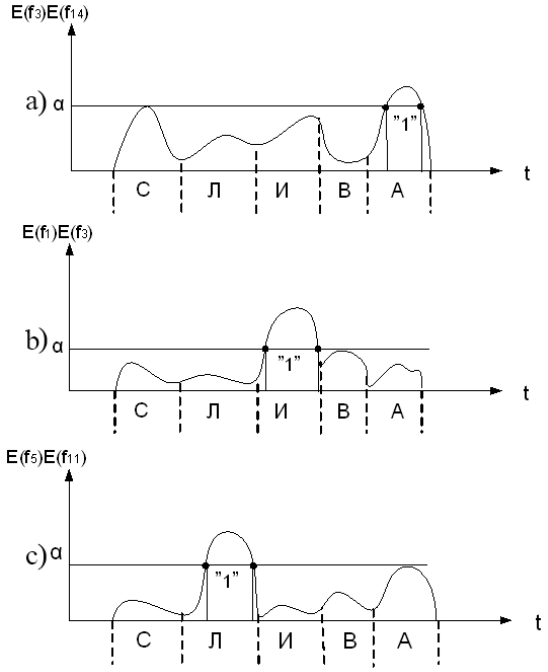
Frequency detection functionQ$_g$:

$$Q_g = Y_{i=1}^{2} \sigma\left(M_k^i \alpha M_k^{i+1}\right) \qquad \sigma\left(M_k^i \alpha M_k^{i+1}\right) = 1, M_k^i > M_k^{i+1};\ 0, M_k^i \le M_k^{i+1}$$

The average frequenciesf$_{нк}$ are determined by the value of the formant frequencies in the neutral position of the path:

$$f_{нж} = (2k-1) \cdot \frac{c}{4 \cdot 1_м}$$

COMBINED METHOD OF AUTOMATED FEATURE EXTRACTION



The logical values "1" are allocated for correlation:

a)selection of logical values "1" for correlationof $E(f_3)E(f_{14})$

b)selection of logical values "1" for correlationof $E(f_1)E(f_3)$

c)selection of logical values "1" for correlationof $E(f_5)E(f_{11})$

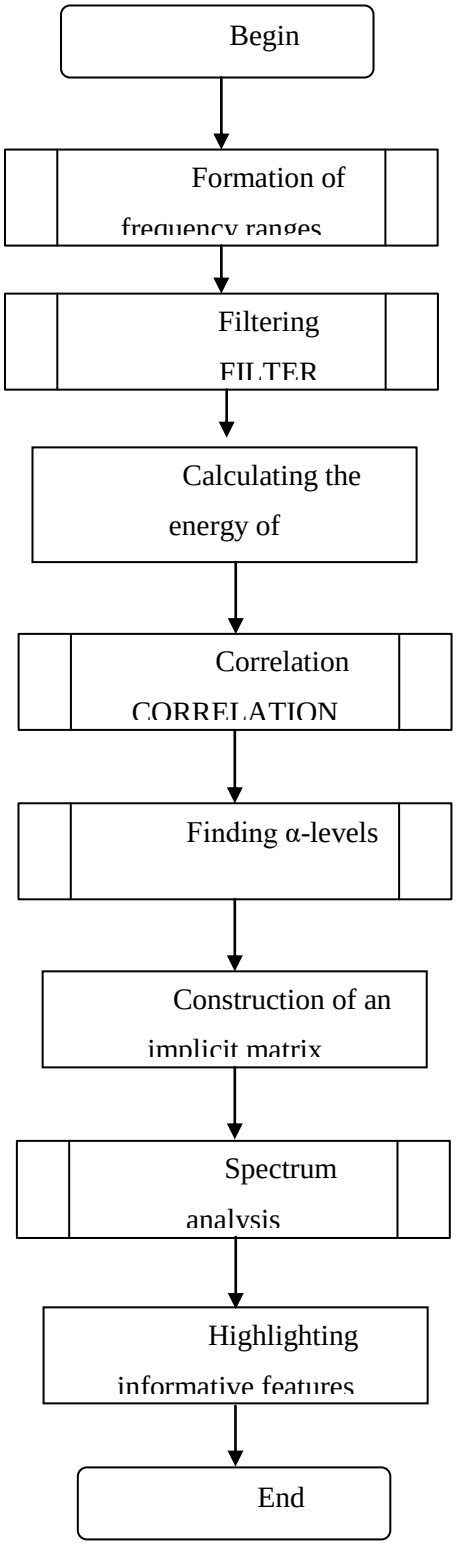
Algorithm:

1. Formation of frequency ranges: their amount, center frequencies, bandwidth, filter type are selected.
2. Filtrating
3. Calculation of the energy

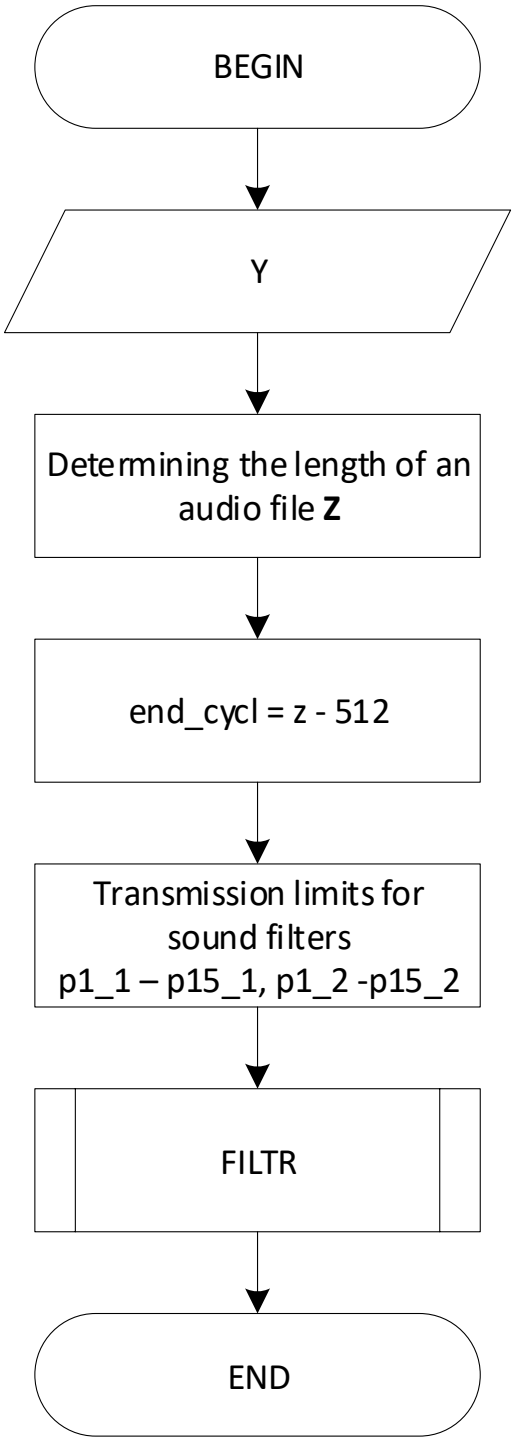
$$E_i(t) = \sqrt{\sum_{k=1}^{N} e_k^2}$$

4. Correlation
5. Finding b-levels for each correlation and forming "1" intervals
6. Finding those "1" overlapping intervals
7. Formation of an implicit matrix for them
8. Formation by implicit matrix of ranges, energy which will be informative features (byconorm operation)

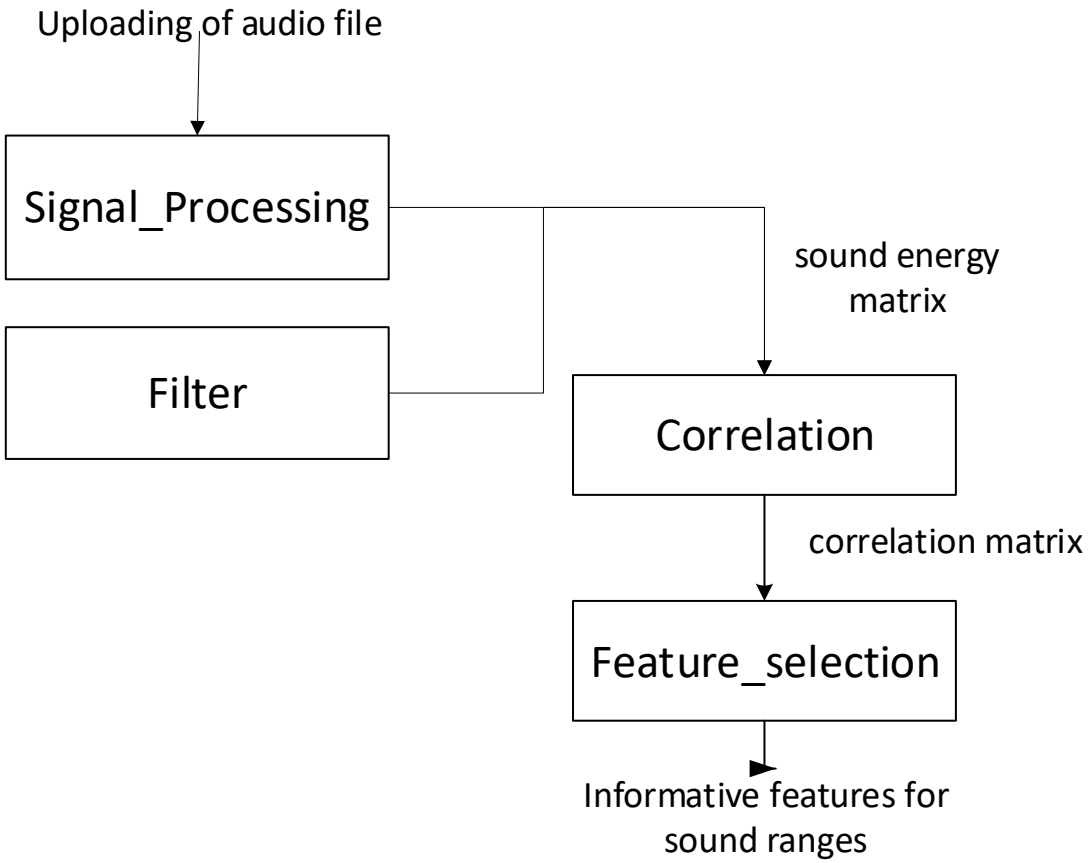
SCHEME OF ALGORITHM OF SELECTION OF INFORMATIVE FEATURES BY THE FUZZY-LOGICAL METHOD



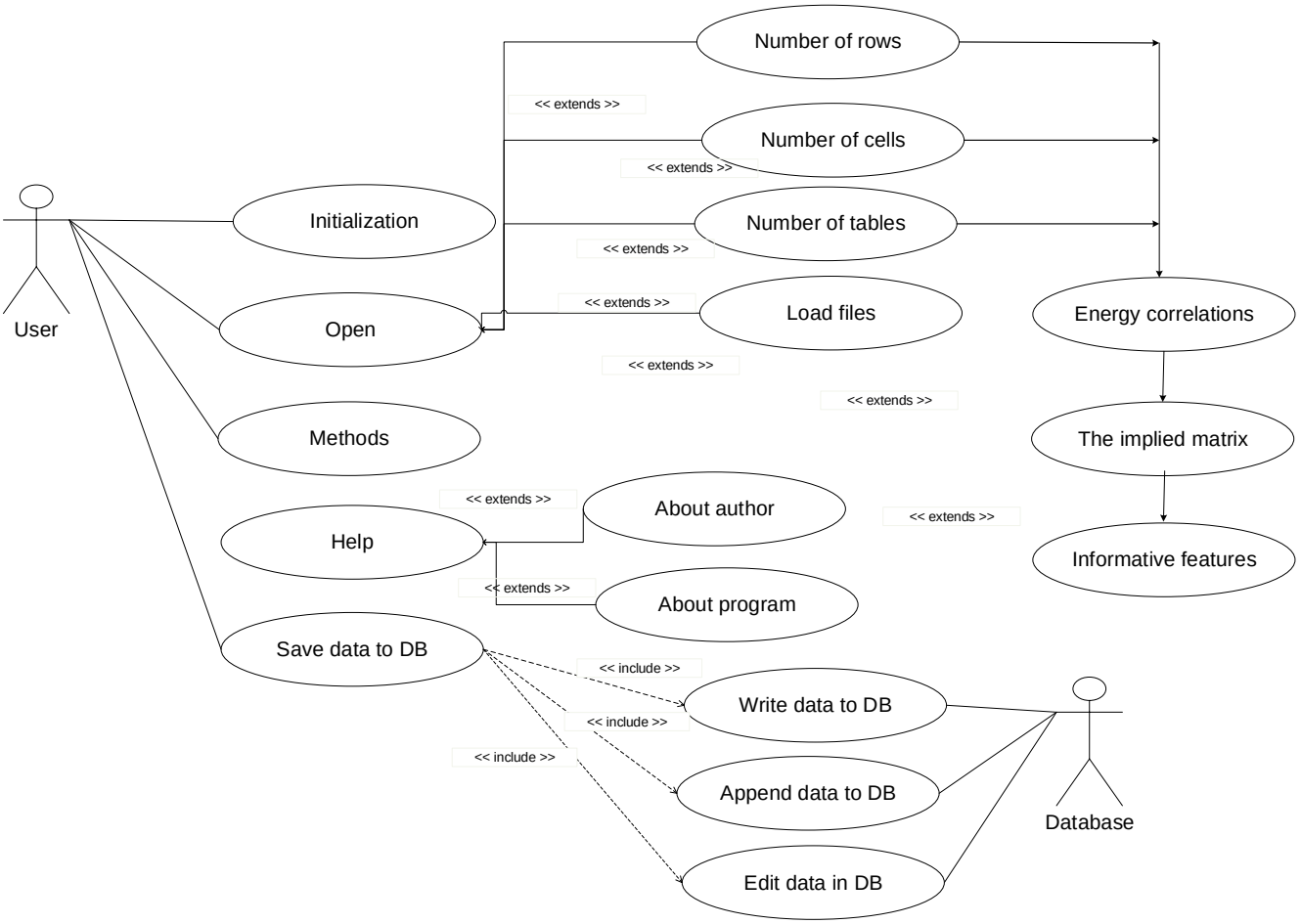
SIGNAL_PROCESSINGALGORITHM SCHEME



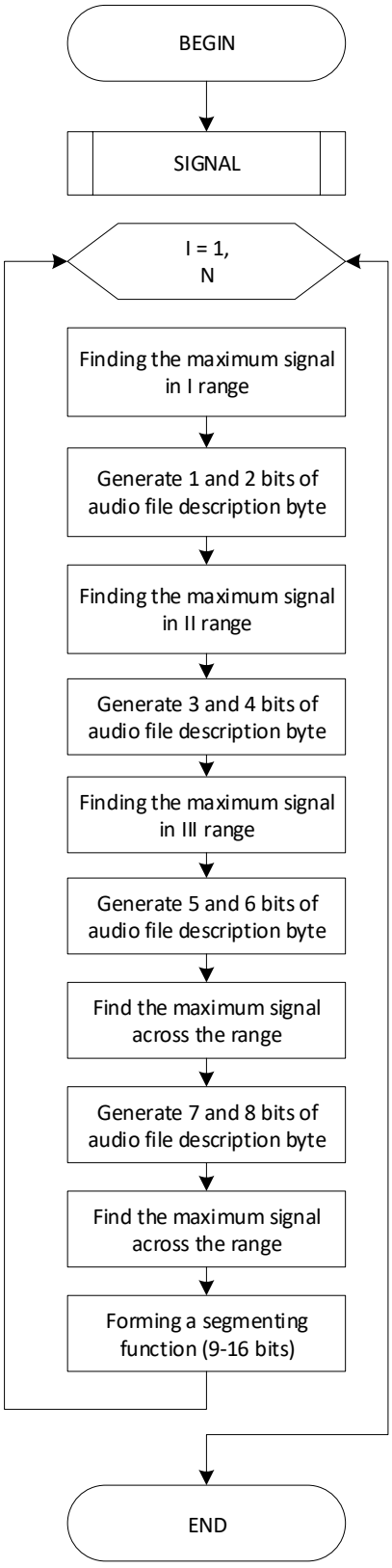
SCHEME OF INTERACTION OF SOFTWARE MODULE FOR FUZZY-LOGICAL METHOD OF SELECTION OF FEATURES



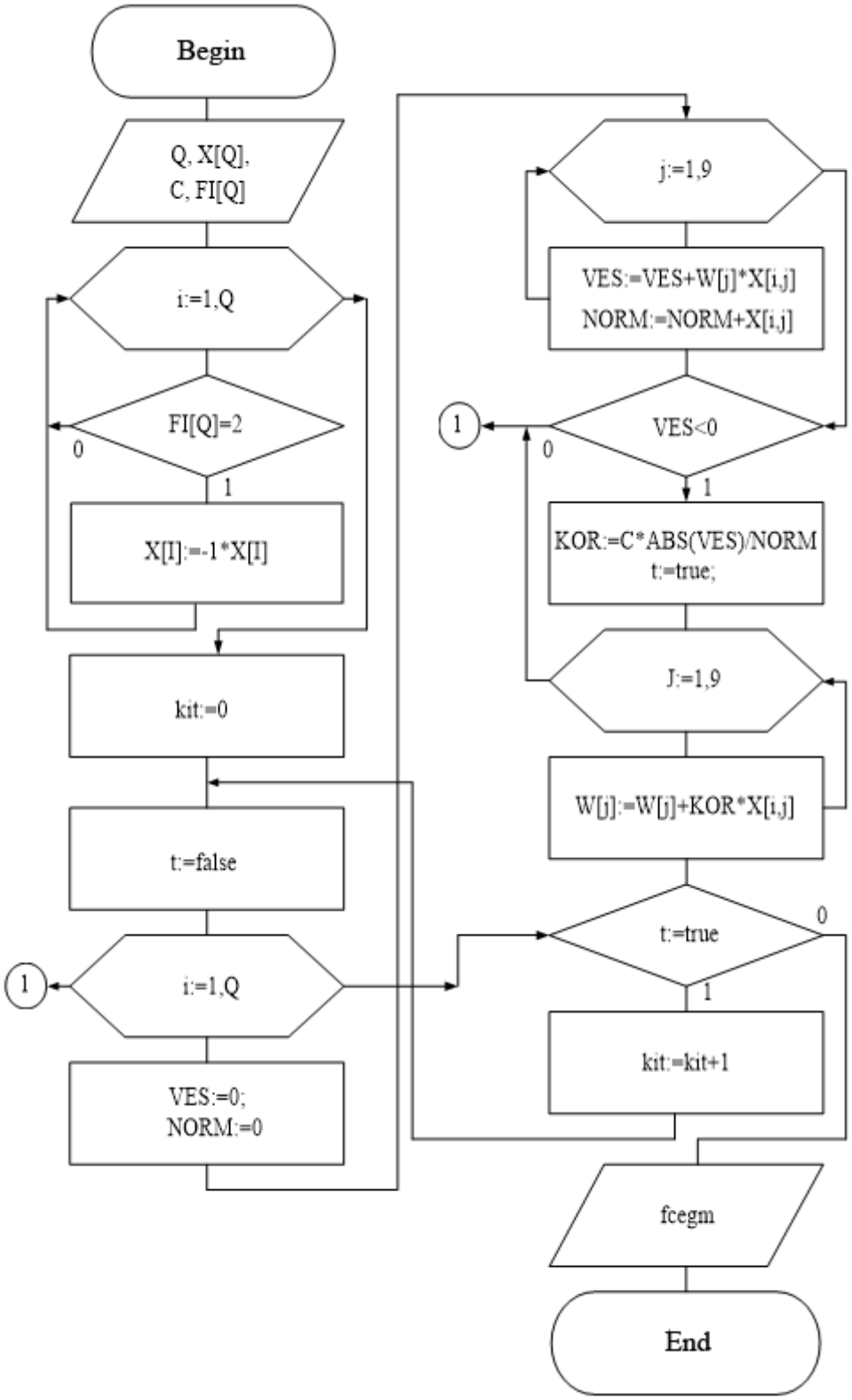
USE-CASE UML DIAGRAM



SCHEME OF ALGORITHM OF SELECTION OF FEATURES BY QUASI-FREQUENCY METHOD

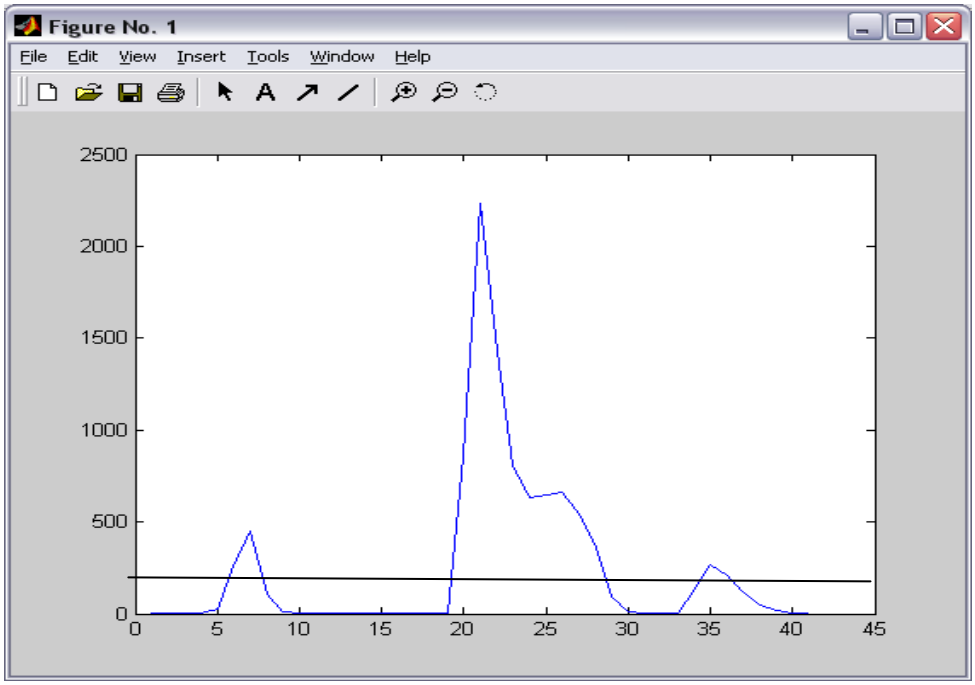


SCHEME OF THE ALGORITHM OF TRAINING OF THE CLASSIFIER ON QUASI-FREQUENCY CHARACTERISTICS

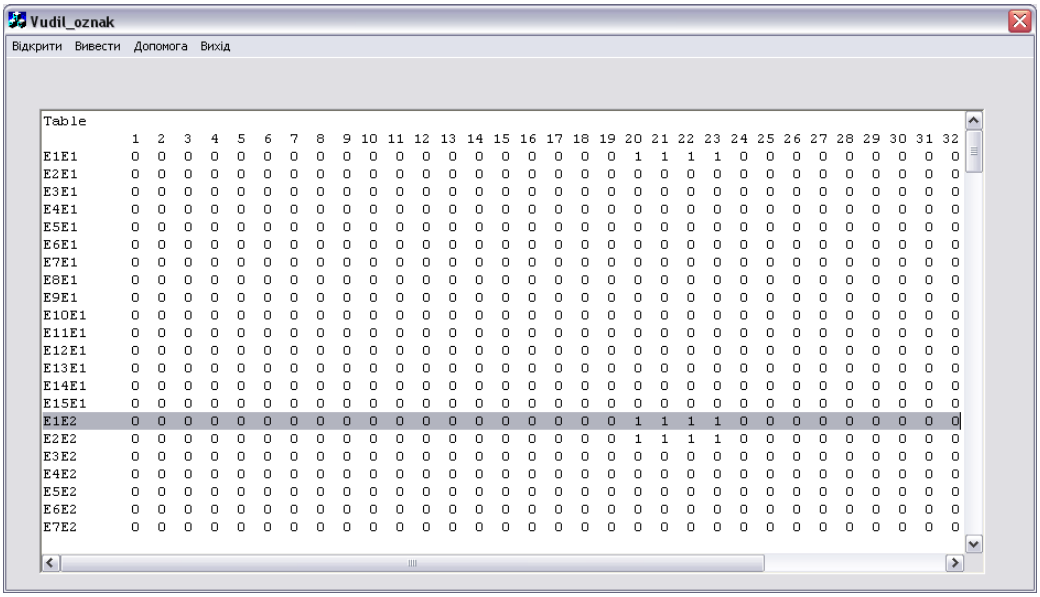


EXPERIMENTAL RESEARCH.

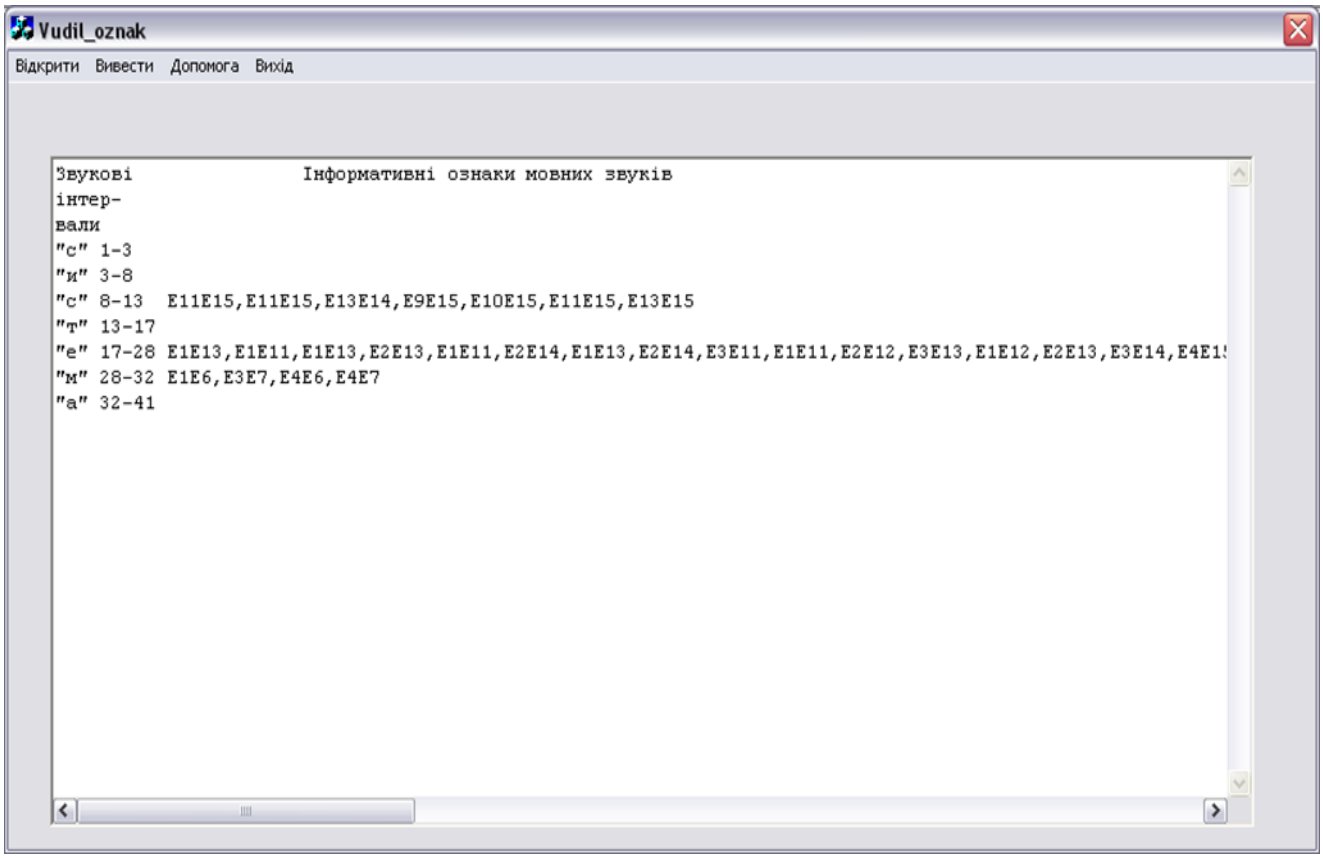
Fuzzy_ logical method of feature selection



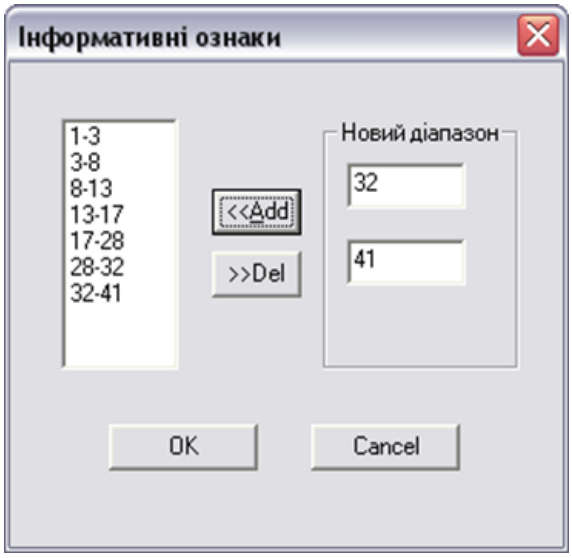
Selection of logical values of "1" for correlation $E(f_1)$ $E(f_2)$



The implicit matrix



Informative features of speech sounds



Results of operation of the frequency-segmenting classifier

EXPERIMENTAL RESEARCH

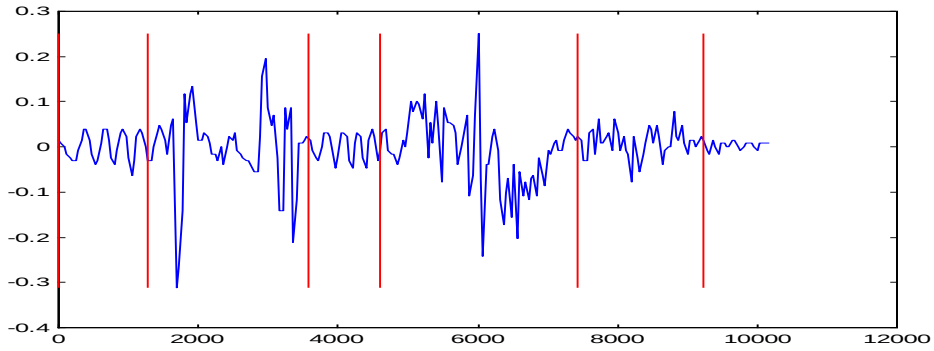
Results of the quasi-frequency feature detector

Announcer 1	Announcer 2	Announcer 3
1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0
1 0 0 1 1 0 1 0 0 0 1 1 0 0 0 0	1 0 1 1 1 0 1 0 0 0 1 0 0 0 0 0	1 0 1 1 1 0 1 0 0 0 1 0 0 0 0 0
1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0	1 0 1 1 1 0 1 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 1 0 0 0 0 0
1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 1 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0
1 0 1 0 1 0 1 0 0 0 1 1 0 0 0 0	1 0 1 0 1 0 1 0 0 0 1 1 0 0 0 0	1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0
1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0	1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0	1 0 1 0 1 0 1 0 0 0 1 1 0 0 0 0
1 0 1 1 1 0 1 0 0 0 0 1 0 0 0 0	1 0 0 1 0 1 1 0 0 0 0 0 1 1 0 0	1 0 1 1 1 1 1 0 0 0 0 1 0 1 0 0
1 0 1 0 1 0 1 0 0 0 0 1 0 0 0 0	1 0 0 1 1 0 1 0 0 0 0 0 1 1 0 0	1 0 1 1 1 0 1 0 0 0 0 0 0 1 0 0
1 0 0 1 1 0 1 0 0 0 1 1 0 0 0 0	1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 1 0 0 0 0 0
1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0
1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0
1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0
1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0
1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0
1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0	1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0
1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0	1 0 0 1 0 1 1 0 0 0 0 0 1 1 0 0	1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0
1 0 1 1 1 0 1 0 0 0 0 1 0 0 0 0	1 0 0 1 0 1 1 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 1 1 0 0 0 0
1 0 1 0 1 0 1 0 0 0 0 1 0 0 0 0	1 0 0 1 0 1 1 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0
1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0	1 0 0 1 0 1 1 0 0 0 0 0 0 0 0 0	1 0 1 1 1 0 1 0 0 0 1 0 0 0 0 0
1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 0 0 1 1 0 0	1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0
1 0 0 1 1 0 1 0 0 0 1 1 0 0 0 0	1 0 1 1 1 0 1 0 0 0 1 0 0 0 0 0	1 0 1 1 1 0 1 0 0 0 1 0 0 0 0 0
1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 1 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 1 0 0 0 0 0
1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0	1 0 1 0 1 0 1 0 0 0 1 1 0 0 0 0	1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0
1 0 1 0 1 0 1 0 0 0 1 1 0 0 0 0	1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0
1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 1 1 0 0 0 0	1 0 1 0 1 0 1 0 0 0 1 1 0 0 0 0
1 0 1 1 1 0 1 0 0 0 0 1 0 0 0 0	1 0 0 1 1 1 1 0 0 0 0 0 0 1 0 0	1 0 1 1 1 1 1 0 0 0 0 1 0 1 0 0
1 0 1 0 1 0 1 0 0 0 0 1 0 0 0 0	1 0 1 1 1 1 1 0 0 0 1 0 0 0 0 0	1 0 1 1 1 0 1 0 0 0 0 0 0 1 0 0
1 0 0 1 1 0 1 0 0 0 1 1 0 0 0 0	1 0 1 1 1 0 1 0 0 0 0 0 0 1 0 0	1 0 0 1 1 0 1 0 0 0 1 0 0 0 0 0
1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 1 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0
1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0
1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0	1 0 1 0 1 0 1 0 0 0 1 1 0 0 0 0	1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0
1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0	1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0
1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0	1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0
1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 1 0 0 0 1 1 0 0 0 0	1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0
1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0	1 0 1 1 1 0 1 0 0 0 1 0 0 0 0 0	1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0
1 0 1 1 1 0 1 0 0 0 0 1 0 0 0 0	1 0 1 1 1 1 1 0 0 0 0 0 0 1 0 0	1 0 0 1 1 0 1 0 0 0 1 1 0 0 0 0
1 0 1 0 1 0 1 0 0 0 0 1 0 0 0 0	1 0 0 1 1 0 1 0 0 0 1 0 0 1 0 0	1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0
1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0		

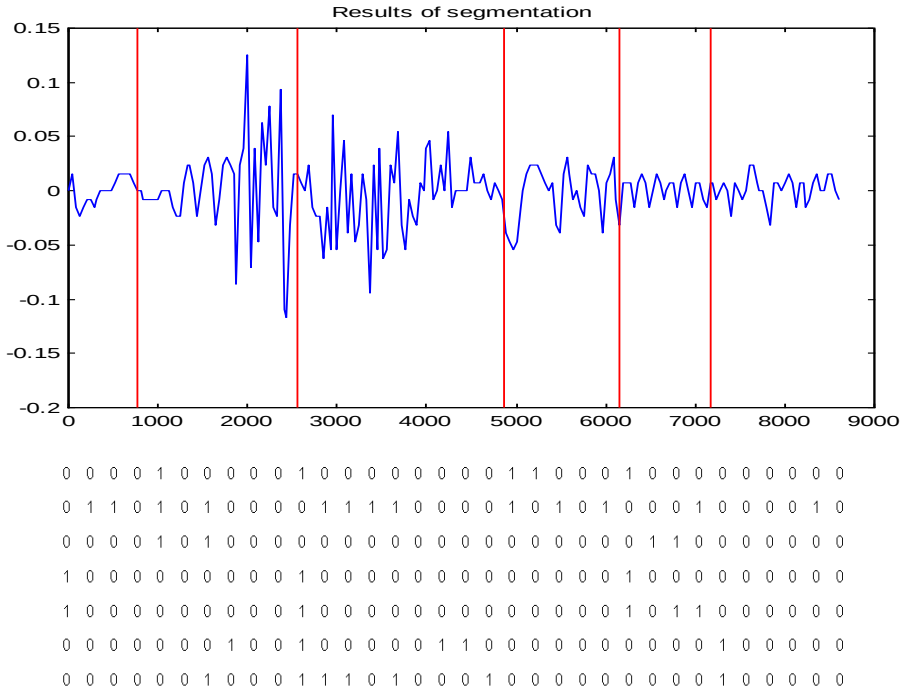
Bytes of audio file description

EXPERIMENTAL RESEARCH

Results of operation of the frequency-segmenting classifier



Segmentation of word “*меньше*”



Segmentation result for word "машина"