Ministry of Education and Science of the Republic of Kazakhstan
Suleyman Demirel University

UDC: 004.8

Aisultan Shoiynbek

Automated emotional speech data mining for the speech emotion recognition.

THESIS

Presented in Partial Fulfillment for the
Degree of Doctor of Science in Computing Systems and Software
(degree code: 6D070400)
Department of Computer Sciences
Faculty of Engineering and Natural Sciences

Supervisor: **Assoc Prof. PhD Kanat Kozhakhmet**
Foreign supervisor: **Prof. PhD Paulo Menezes**

Kaskelen, 2020

Contents

Normative References

In this dissertation, references are made to the following standards:

"Instructions for the preparation of a dissertation and author's abstract" VAK MON RK, 377-3zh.

GOST 7.32-2001. Report on research work. Structure and design rules.

GOST 7.1-2003. Bibliographic record. Bibliographic description. General requirements and compilation rules.

GOST 7.32-2017. System of standards on information, librarianship and publishing. Research report. Structure and design rules

Definitions

Artificial Intelligence - the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings.

Audio signal - a representation of sound, typically using either a level of electrical voltage for analog signals, or a series of binary numbers for digital signals.

Dataset - a collection of data in Machine Learning projects.

Emotion - biological states associated with the nervous system brought on by neurophysiological changes variously associated with thoughts, feelings, behavioural responses, and a degree of pleasure or displeasure.

Machine Learning - an application of artificial intelligence that provides systems the ability to automatically learn and improve from experience without being explicitly programmed

Neural networks - The piece of a computing system designed to simulate the way the human brain analyzes and processes information

Sound - a continuous signal a wave with changing amplitude and frequency.

List Of Abbreviations

AI - artificial intelligence
ADC - Analog-to-digital convertation
CNN - convolutional neural network
DAC - Digital-to-analog convertation
DNN - deep neural network
ER - emotion recognition
FER - face emotion recognition
GMM - Gaussian mixture model
GP - Gaussian process
HCI - human-computer interface
HMM - hidden Markov model
MFCC - Mel-frequency cepstrum coefficients
ML -machine learning
NN - neural network
RBF - radial basis function
RNN - recurrent neural networks
SER - speech emotion recognition
SR - speech recognition
SVM - support vector machine
VSS - voiced segment selection

Introduction

Relevance. Emotions take a significant place in interpersonal human interactions and relationships. Emotion affects on our life , because it is a human inside reaction on surrounding and occurring circumstance. It helps us to make intelligent decisions, feel the mood of other person better, convey feelings as feedback for understanding reaction, and adapt to the reality of human behavior. Scientists have established the effect of emotional state on human health and its formation as a person [1]. All these facts motivated scientists to study and measure the level of human emotional states. Considering today's realities, during the development of robotics and AI one of the popular areas for research is the recognition of emotions by a machine with the further ability to reproduce emotion. The primary objective of ER is to aid human to machine interaction. Therefore the ER is significant in AI products since it will make HCI more friendly and improve the user experience.

Recognition of emotions by a person is based on vision and hearing when a person sees the face of another person and hears his voice: timbre, volume, speed, pitch. Of course, the meaning of speech also important and processed by the brain and hear. Recently, scientists in the field of facial recognition of emotions have achieved high results in ER with an accuracy above 98% [2]. According to Albert Mehrabian's 7-38-55 Rule of Personal Communication, speech influence only 7% of perception of affective state. A tone of voice and body language facilitate to 38% and 55% of personal relations, respectively [3]. However, when we recognize emotion by face, we are only covering one channel to estimate an emotional state. It is necessary to use all perception channels of emotions in a complex to improve ER. Compared with other biological signals (e.g., electrocardiogram), speech signals usually can be received more easily and economically. Accordingly, the majority of researchers are interested in SER. The area has received increasing research interest through current years.

According to Paul Ekman, there are seven based face emotions: anger, disgust, fear, happiness, sadness, and surprise. To achieve successful results in SER, we need to consider three main issues, namely, (1) collect emotional speech data or choose already existing database, (2) preprocess the data and extracting useful features, (3) designing robust classification models based on ML algorithms. Many scientists consider that the main issue is to define and explore features with more emotional information and try to discover that area.

Nowadays, usually, MFCC are used for extracting features, and sometimes researchers prefer to use combining feature sets that consist of multiple features containing richer emotional information [4]. One of them is creating new kind of feature, which is more intended to identify emotional content [5].

In recent time, scientists discover and have experimented different classification algorithms to recognize speech emotions, such as SVM [6], [7], [8], [9], [10] RNN [11], [12] HMM [13], NN [14] and Gaussian mixture model GMM [15]. Also it was proposed other types of classifiers such as a modified brain emotional learning model [16] in which the adaptive neuro-fuzzy inference system and multilayer perceptron are merged for SER. Another proposed approach is a multiple kernel GP classification [17], in which two similar meanings in the learning algorithm are presented by combining the linear kernel and RBF kernel. The VSS algorithm also proposed in [18] deals with the voiced signal segment as the texture image-processing feature, which is different from the traditional method.

Consequently it becomes clear that researchers concentrate on improving classifiers of ML algorithms; however, in most researches it was used the emotional speech collected under controlled conditions. The conventional approach may fail when background noise or non-speech filler exists. Almost all existing results have been received from databases, which was recorded by actors imitating and artificially producing emotions in front of a microphone [19]. There are four issues related to the above databases, namely, (1) emotions are not natural, and it means that machine learns to recognize fake emotions. (2) Emotions are limited by quantity and weak in their variety of speaking. The number of records with emotion in those databases will always be not enough for real industrial product, which can endure severe validation tests and work in real-time. (3) There is some language dependency on SER. It has proofed by researchers by comparing the different languages using them SER classifier model [20]. Besides, we have proofed it for Russian and Kazakh in one of the articles [21]. (4) Consequently, each time when researchers want to start work with SER, they need to find a good emotional database on their language.

Approach for skirting these challenges while acquiring naturally occurring emotions in speech is collection of audio and video data, based on real emotions when people are not trying to play them like in movies or TV series. The data can be collected from the news, the interview of people, or videos from the public sources from any events or holidays. However, it is a very complicated and laborious job when it is necessary to choose and prepare the video, to cut and convert to the right format, and the most laborious work is labeling data. It is unable to do manually., In other words, a full process from preparation data to get the right labels. Suppose someone has collected data by hand and marked it by spending enormous effort and time. The problem is still relevant as the assembled database only for one language. Until the issue of collecting high-quality emotional data will not be resolved, the world will not be able to

achieve qualitative progress in this field of science.

Humans use multiple channels to display emotions, and our emotion display is not uniform - sometimes we rely more on the face, other times we display more through our voice, or even our gestures and body movements. Anger is usually displayed more through voice than through face. It might be difficult for a system to recognize anger from just the face, but pairing the facial display with voice can assist the system to detect anger much more reliably. The opposite, it is exact for joy: people usually use their faces to display joy, but they cannot display pronounced joy using voice unless it is an extreme joy and mirth that brings about laughter and a vocally aroused voice. Therefore, in this case, not only voice, as well as a face, will not give a good enough estimate of someone being simply content and happy. It means that we should use both algorithms FER and SER for better recognizing the emotion and automatic labeling.

Aim and objectives of research. The goal of the thesis is developing automatic data collection and labeling methods and algorithms, for SER task with a high level of accuracy, not less than 80%.

Objectives of the study. In accordance with the aim, the following objectives are identified to be solved in this work:

- to study and analyze the exist emotional corpuses
- to study and analyze the current data mining methods
- to study theoretical base of human speech and voice and how to transform to machine language
- to analyze and justify the choice of robust feature transformation method for SER task.
- to analyze and justify the choice of classificatory for SER task.
- to develop the NN model to separate speech and nonspeech audio content.
- to develop the methods and algorithm which extract audio emotional content from video.
- to develop the automated labeling methods and algorithm.

Object of research. The research focuses on data mining automated methods mining for the SER.

Research methods. The objectives assigned were solved by carrying out theoretical and empirical research. As part of the research we used conceptual positions of AI classical ML theories, studies of leading foreign and domestic scientists in the field of ER and SR, probability theory, mathematical statistics, numerical analysis.

Scientific novelty of the work. The novelty of the dissertation is to design an automated method for collecting and labeling speech emotional data. The results obtained in this dissertation will significantly advance the field of AI in recognizing speech emotions. Using the method of collecting emotional data, scientists will be able to collect emotional datasets in all languages of the world.

The following scientific statements are to be defended:

- Methods and Algorithms for emotional data mining

- Methods and algorithms for automated search and download of video containing emotional utterances.

- Methods and algorithms for extracting emotional utterances from videos.

- Methods and algorithms for the classification and labeling of emotional utterances.

- An automated system for collecting and labeling emotional utterances for the task of recognizing speech emotions.

- Designed DNN model for the classification of human speech from various sounds (noises)

Practical significance of the research results. The practical value of the thesis lies in the possibility of qualitative improvement in service of call-centers, in education, banking, insurance, public services, and medicine. The practical significance of the study is determined by the possibility of applying its results and recommendations in the: development recognizing true or false emotion systems, drawing up an emotional portrait of the offender by law enforcement agencies, identification of depressive and suicidal tendencies in order to prevent child suicide.

Acknowledgments. This research was supported by a grant of the program of the Ministry of Education of the Republic of Kazakhstan BR05236699 "Development of a digital adaptive educational environment using Big Data Analytics". I thank my colleagues from Suleyman Demirel University(Kazakhstan) and Coimbra University(Portugal) who provided insight and expertise that greatly assisted the research. I express our hopes that they will agree with the conclusion and findings of this dissertation.

Structure and scope of the dissertation. Thesis consists of introduction, four chapters, conclusion, and references list. It is presented on 125 pages of typewritten text, contains 50 figures, 27 tables, a list of references that includes 115 titles.

In the framework of this thesis, 10 research papers were prepared and published on this topic, including:

- two articles are published in publishing houses that meet the requirements of the higher Attestation Commission of the Ministry of Education of Science of the Republic of Kazakhstan;

- fourth articles are published in the proceedings of international conferences;

- one article in a foreign publication

- three in international peer-reviewed scientific journal.

The first chapter provides an overview of existing research.

In the second chapter, the nature of voice and speech emotions is studied in detail.

In the third chapter, a ML model of deep NN is proposed, as well as, through a comparative analysis, the most effective feature for transforming vocal emotions into a machine form is determined.

In the fourth chapter, a method for the automatic collection of marked

emotional-speech utterances, a classifier of speech and non-speech audio data, a classifier of speech emotions is proposed.

In the fifth chapter and conclusion, the main outcomes of the dissertation were defined based on results of presented study

1. Background of Literature Review

Due to the growing interest into SER field, many scientists, realizing the problem of lack of data, are trying to assemble their datasets. For example, in paper [22], researchers have collected a dataset duration of 187 hours of data from 2,965 subjects. Data includes non-emotional recordings from each subject as well as recordings for five emotions: angry, happy-low-arousal, happy-high-arousal, neutral, and sad. Subjects used their recording equipment, resulting in a data set that contains variation in room acoustics, microphone, etc. Unfortunately, Subjects were prompted to use past emotional experience as the basis for expressing emotion, which means that the actors artificially emulated emotions. The authors of the study [22] did a great job of collecting huge dataset, but they have created faked emotions. Also, it becomes clear that researchers got a satisfactory result based on highlighted results at the end of the article.

In the study [23] authors used speech emotion classifier by authors Berlin[24] and Spanish[25] Dataset, which also indicates the absence of natural emotions. In spite of this, researchers were able to achieve accuracy of 90.94% on the Spanish[25] dataset and 81.1% on Berlin[24] dataset on average using tenfold cross-validation.

The study [26] narrates about creating the three different speech emotional dataset in Chinese. In the first dataset, the emotional speech is recorded by professional actors and actress, six males and six females. The second dataset is unclear to understand how it was collected. It was only mentioned that in dataset fifty-one speakers. The third one is some mix from the first and second datasets. The total amount of utterances is 29000 in all three datasets. The authors have done a tremendous productive job collecting the dataset. Unfortunately, emotions are non-natural, and they collected all manually, spending vast resources.

The researchers in paper [27] have recorded emotions dataset with 16 artists, professional actors/actresses (eight male and eight female) aged from twenty-five to sixty-four, and finally got 560 samples. They have added the Polish Emotional Speech Database (PESD)[28] was collected and shared by the Medical Electronics Division, Lodz University of Technology. The dataset consists of 240 recorded samples. Authors also have added the spontaneous Polish Speech Database (PSSD) [29] consisting of 748 samples containing emotional carrier. Speech utterances were collected from discussions in TV programs, live shows, or

reality shows. The PSSD dataset consists of natural emotions, but 748 samples are too less. As a result, researchers reached the accuracy of 86,14%.

In the paper[30], researchers have collected and labeled the dataset which consists of 2 hours of spontaneous emotional segments extracted from 219 speakers. Twenty-five films, one television series, and twenty-two talk shows were taken to collect the dataset. Despite the fact that films and series are not a source of natural emotions since the actors play a role, the authors claim that this is the first Chinese-language dataset with natural emotions. The segmentation process and labeling process were done by annotations manually.

Two years later, authors of [30] have published the paper[31], where, based on the collected dataset[30], they proposed the method based on the multimodal concept using a FER and automatic SR(ASR) separately. Researchers have divided the dataset[30] between a training set, validation set, and testing set, containing 1981, 243 and 628 clips, respectively. Scientists design CNN model to predict emotions on facial expressions based on Static Facial Expression[32] and FER2013[33] datasets. In a final result, authors have achieved the accuracy in a video 36.56%, in-text 33.84%.

Collection a spontaneous, multi-modal, rich-annotated emotion database is a challenging issue. In conclusion, the most existing datasets were recorded in 'lab controlled' environment and were collected manually. Further will be briefly described existing speech emotion databases.

English datasets.

RAVDESS [34]: 1440 English utterances by 24 professional actors (12 male, 12 female). The dataset includes the states of anger, happiness, sadness, fear, disgust, surprise, calmness and neutral;

SAVEE [35]: 480 English utterances by four non-professional male actors. It includes the states of anger, disgust, fear, happiness, sadness, surprise, and neutral.

Dataset [36]. The dataset was recorded at the Faculty of Electrical Engineering and Computer Science, University of Maribor, Slovenia. It has emotional speech in 6 emotion categories, such as disgust, surprise, happiness, fear, anger and sadness, Slovenian, English, Spanish, and French were used in all records. The dataset contains 186 utterances per emotion category. These utterances are divided into isolated words, sentences both affirmative and interrogative, and a segment.

Dataset [37] was collected at the Queen's University of Belfast. It has emotional speech in 5 emotional states: neutral anger, happiness, sadness, fear, and anger neutral. There were 40 speakers (20 females, 20 males) aged between 18 to 69 years. The subjects read 7-8 sentences recorded in an appropriate emotional tone and content for each emotional sentence.

The dataset [38] is designed to sample natural emotions and to allow exploration of the emotions through the time. The authors have recorded in two different ways. One of them was recorded in a studio, and the second one

extracted from TV programs. A total of 239 clips with duration from10 to 60 sec are included in the dataset. The studio utterances have two parts. The first part contains speech between students on topics, which provoke strong emotions. The second part has audio-visual records of interviews (face-to-face).

The database [39], Kids' Audio Speech Corpus NSF/ITR, was created at the University of Colorado. The project aimed to collect audio and video data from kids in order to provide the development of visual and auditory recognition systems, which allow interacting face-to-face with electronic teachers. The Kids' audio speech dataset is not oriented to classify emotions. Only 1000 out of 45000 records are emotion-oriented.

The SUSAS dataset[40] contains speech from 32 speakers (13 females and 19 males) with ages from 22 to 76. The total amount of dataset is 16 000 records.

The database [41] consists of 40 utterances said by two speakers in 5 emotional categories. There are two repetitions of these 40 sentences. Thirty-one annotators (18 males and 13 females) rated each record on six Likert intensity scales.

The Database [42] was created in M. Edgington at BT Labs, UK. The purpose of the dataset was to train a voice synthesizer. All six speech emotional categories were recorded by one actress, such as sadness, happiness, anger, fear, boredom, and neutral.

The dataset[43] was recorded at Carnegie Mellon University and had five emotional categories. The sentence length varies from two to twelve words.

Database [44] divided into two parts. The first one composed of a corpus with 700 utterances that were spoken by 30 professional actors. The database consists of 5 emotion categories: happiness, anger, sadness, fear, and neutral. The second part consists of 56 telephone calls. The duration of each utterance is from 15 seconds to 90 seconds.

German datasets.

EmoDB [24] is a German Corpus (Berlin Database of Emotional Speech), which contains about 800 sentences (seven emotion classes * five female and five male actors, ten sentences). All sentences were recorded in an anechoic chamber using high-quality equipment with a sampling frequency of 48kHz and later downsampled do 16kHz (mono).

Database [45] was recorded at the Max-Planck-Institute of Cognitive Neuroscience for medical purposes. The project aimed to relate speech emotions with a location in the human brain. The subject was a woman. The dataset contains speech in 3 emotional categories. Twenty subjects have judged the semantic content and the prosodic feature.

The database[46] aim was to construct a brain map of emotions. The dataset contains emotional speech in six emotional categories. The total amount of records is 4200 nouns and pseudowords.

German Emotional Speech database [47] contains data collected from a German 12 TV. The talk show includes the dialogues and utterances. The rich emotional utterances were segmented as audio files. The created corpus

is a collection of the spontaneous speech extracted overall from the show, the communication between the guests on the intriguing themes discussed; Topics include special issues such as friendships, treasons, romantic relationships, and fatherhood issues. The collected audio files were extracted as a wave file the signal was downsampled to 16 kHz (16 bit). Finally, the dataset has been carried out with 1018 emotional files by 47 speakers with an average of 21.7 sentences per speaker. The speaker's age ranges from 16 to 69 years. The average sentence duration was 3.0 s.

Dutch datasets

The database[48] contains 20 hours of Dutch speech. The dataset is only partially oriented to emotion. An electroglottograph and an orthographic transcrip-tion data were also included. The total number of speakers is 238. They are not actors, and the emotions are forced rather than natural. The dataset consists of monosyllabic words, short sentences, short texts, long vowels, and digits.

The Dataset [49] recorder studies the relationship between speed in speaking and emotion. The database consists of seven emotion categories, such as happiness, sadness, indignation, boredom, fear, anger, and neutral. The Authors of the dataset was involved three speakers who read five sentences with semantically neutral content. Dataset has 315 utterances.

Spanish emotion speech databases

ELRA-S0329 [25]: 6041 Spanish utterances by two professional actors (one male, one female). Includes the states of Anger, Joy, Sadness, Fear, Disgust, Surprise and Neutral;

Database [50] Spanish Emotional Speech database (SES). J. M. Montero and his assistants constructed in 1998 a Spanish emotional speech database [10]. It contains emotional speech in 4 emotion categories, such as sadness, happiness, anger, and neutral. The labeling of the dataset is semi-automatic. The corpus consists of 15 short utterances and 30 single words.

Database [51] was collected at the University of R. L. of Barcelona and contained seven emotion categories. To create the database, authors attracted 8 actors. Actors read two texts in three emotional intensities. After 1054 students were labeled recorded texts.

Multilingual emotion speech database

Database 32 Lost Luggage study. K. Scherer has recorded another emotional speech database [52]. The recordings took place at Geneva International Airport. The subjects are 109 airline passengers waiting in vain for their luggage to arrive on the belt. RML [53]: 720 multi-language utterances by 8 non-professional male actors. Includes the states of Anger, Disgust, Fear, Happiness, Sadness, and Surprise;

Other different emotional datasets.

The Danish emotion speech database [54] contains emotional speech in 5 emotion categories, such as surprise, happiness, anger, sadness, and neutral.

The database consists of 2 words (yes, no), nine sentences, and two passages. Twenty judges (native speakers from 18 to 58 year old) verified the emotions with a scoring rate of 67%.

The Japanese speech emotion corpus[55] contains speech in 8 emotion categories. Emotion corpus content has 100 native speakers (50 males and 50 females) and one professional radio speaker. The radio host read 100 neutral words in 8 emotion states. The other 100 speakers were asked to mimic the manner of the professional actor. The total dataset consists of 80 000 words.

The Hebrew emotional multi-modal database [56] was created at the faculty of Holon Academic Institute of Technology at Israel. The database contains emotional speech in 5 emotion categories. The database consists of emotional speech, electro-myogram of the corrugator (a muscle of the upper face which assists in expressing an emotion), heart rate, and galvanic resistance that is a sweat indicator. The subjects (40 students) were told to recall an emotional situation of their life and speak about that. In his study, N. Amir found that there is not a clear way of discovering the real emotion in speech.

The Sweden emotion speech database [57] contains emotional speech in 9 emotion categories, such as joy, surprise, sadness, fear, shyness, anger, dominance, disgust, and neutral. Different nationality listeners classified the emotional utterances into an emotional state. The listener group consisted of 35 native Swedish speakers, 23 native Spanish speakers, 23 native Finnish speakers, and 12 native English speakers. The non-Swedish listeners were Swedish immigrants, and all knew Swedish, of varying proficiency.

The Chinese emotion speech database [58] contains speech segments from Chinese TV shows in four emotion categories, namely, sadness, anger, neutral, and happiness. Four annotators labeled the 2000 utterances. Each annotator labeled all the sentences. When two or more annotators agreed in their label, the utterance got their valid label. Elsewhere the utterance was deleted. After the annotation process several times, only 721 utterances remained. Russian emotion speech database "Russlana" [59] collected at Meikai University in Japan [40]. The total of utterances is 3660 sentences from 61 (12 male) native Russian speakers age from 16 to 28. Features of speech like energy, pitch, and formats curves are also included.

The first part of the Croatian emotional speech corpus [60] called "real-life emotions" was collected from the Internet, mostly from Croatian reality shows and from different documentaries from the internet. The second part, called "acted emotions" was collected from Croatian movies, TV Shows, and Books-Aloud programs. The collected utterances were normalized and stored in 'WAV' format with 11 kHz sampling frequency, 16 bits per sample, monaural. A total of 714 utterances were collected with durations of 56:22 minutes.

The Serbian emotional speech database [61] consists of the recordings of the following emotions neutral, anger, happiness, sadness, and fear. The data was collected from 6 actors, 3 of each gender. The database was collected in

an anechoic studio. The actors were asked to use their usual way of expressing the emotional states and not that of stage acting. The database consists of 32 isolated words, 30 short semantically neutral sentences, 30 long semantically neutral sentences and one passage with 79 words in size. The speech was recorded with a high-quality microphone at a 44.1 kHz sampling frequency. Lately, the recordings were transferred from DAT to PC with a reduced sampling frequency of 22.050 kHz and stored in.WAV format. The listening test of the Serbian Emotional Speech Database showed correct identification of emotions is 95% and the confusion that occurred between anger and happiness and between neutral and fear.

EMOVO [62]: 588 Italian utterances by six professional actors (three male, three female). Includes the states of Anger, Joy, Sadness, Fear, Disgust, Surprise and Neutral;

2. Emotions of speech

2.1 Language of emotions in life and science

A person in the process of speaking transfers to a listener information of two main categories. First, we learn WHAT the speaking person wants to say and what kind of words are pronounced by the speaking person. This is sense-based or semantic information. As it is expressed in words it is called verbal (i.e. in words) or linguistic information. Secondly, coupled with words based on how a person speaks the listener receives a lot of information about the speaking person and about his or her attitude to the subject matter of the talk, about attitude to the listener and even to the speaker herself/himself. All these types of information do not considerably depend on WHAT the person says and therefore are defined by the term "extra-linguistic or nonverbal information".

One of the most important types of extra-linguistic information is the emotional type which characterizes an emotional state of the speaking person, his attitude to subject matter of the talk, to the listener and etc.

As a rule, the emotional context of the speech is in line with the logical sense of the speech and amplifies it. But it does not depend on the logical sense of the speech and therefore it can even contradict it. At the same time in our routine informal conversation we are inclined to trust more this emotional context rather than the logical sense of words, for example, the warm words uttered with sneer or in anger.

Each of the written words YES and NO have only one sense, one meaning. But each of these words when pronounced can deliver many various senses and meanings including the ones that completely contradict their verbal meaning or even significantly change this meaning.

That is why it is clear that the channel of extra-linguistic information for communication of people and for people to understand this information plays a huge psychological role. But we will see that the role of this channel is also very important for communication of humans with machines.

Emotional information is not the only one type of information in the system of human nonverbal sound communication. According to the developed classification by the authors [63], we can differentiate up to seven types of nonverbal extra-linguistic information in a human voice: 1) esthetic - for example, a voice

that can be pleasant, unpleasant, beautiful, ugly, hoarse in terms of pitch and timbre (a bass, a baritone, a tenor) which is very important in arts and etc.; 2) emotional – the subject matter of our topic; 3) individual and personal – which makes it possible to identify the speaking person by his or her voice; 4) biosocial - gender, age, nationality by accent, etc.; 5) psychological - traits of character, will, temperament, self-estimation, the listener's estimation and etc.; 6) medical, demonstrating the state of health, for example, a hoarse voice, etc.; 7) spatial: indicating a place of location and movements in space of the speaking person in relation to the listener.

Of course, we can't think that the above listed types of nonverbal information exist in a human voice as something isolated from each other and easily separable. The situation is much more complicated. All these components of nonverbal supplements to the word are in the most difficult interaction with each other, on the one hand, and with the word, on the other hand. They have different degrees of expressiveness and significance for the listener and at the same time are perceived by the listener as an integral image of the speaking person and of the sense of the information delivered by the speaking person.

Until recently the researchers were interested only in laws of the verbal speech, i.e. in a word. Emotional and expressive properties of the human speech as well as its other nonverbal features were somehow not taken into account. Linguists, in particular, called them extra language – extra-linguistic and even as prof. Bondarko L.V. writes they considered them as "something negative and complicating the language-based communication". However, in recent years these very extra-linguistic properties of speech, specifically, emotional, which are so "annoyingly break" the laws of phonetics started to attract an ever-increasing interest of specialists including phoneticians.

The reasons of these specialists' interest to the language of emotions and to all other types of extra-linguistic information of the speech are explained by the fact that these properties so normally and naturally perceived by our brain and even helping to understand the sense of a conversation make it seriously difficult to solve the problem of automatic SR, i.e. understanding of our natural verbal speech (sounds based) by machines. Therefore as of today not only linguists and physiologists but also the hardware design engineers want to know how the human brain copes with this enormous task of selection of the necessary information from the most difficult and so fancifully and more over so instantly changing in time sound streams which we call the verbal/sound speech. Specialists consider that only such a bionic approach will allow solving the problem of further improvement of the automatic SR systems.

The human ability to express emotions by voice is considered historically as the most ancient in comparison with the verbal speech. Apparently, our far ancestors knew and handled the language of emotions long before they learned the verbal speech. There are many proofs in favor of this opinion. It is curious that one of the proofs is speaking human beings of our times, to be more precise,

special ways of the human speech development starting from childhood. As for voice, human beings have voice since their birth, as for speech, we know it develops much later, by the age of one and a half years old and even by two years old of a child.

However, long before mastering the speech habits a child already perfectly masters its ability to communicate with people around in the language of emotions. Joy, sadness, grief, anger, fear - these feelings are distinguished in the child's voice not only by a nice ear of a mother but also by ear/hearing of any person. At an infantile age a human being expresses not only its own emotions. An infant perfectly understands the emotional intonation of adults, even during the period of time when it does not understand the logical sense of words. For example, a child smiles in response to tender words and can begin to cry in response to harsh words. But, maybe, it can after all understand words? A simple experiment will answer this question.

Try to utter to a six-months-old child the most terrible and harsh words but in a kind, tender voice and the child will smile. And contrary to that, in response to gentle and tender words pronounced by an angry voice, the child will get startled and can begin to cry. It is clear that it reacts to the emotional coloring of a human's voice. In this regard children react to emotions similarly to the higher animals: a dog, for example, reacts to a voice intonation in the same way. An experiment was conducted where the laughter and crying of an eight months old child was recorded. When the recorded sounds were played back to the child it immediately fell into the corresponding emotional state: laughed when hearing the laughter or cried when hearing its own crying. This phenomenon is often observed in maternity homes and in a daytime nursery.

Thus, children in an early age without knowing the speech yet know the language of emotions, communicate with adults in this language and perfectly understand each other. Science considers this period as the period that reflects a particular stage of the human evolution which preceded emergence of logical speech of ancient people.

According to the American scientist Wane Lee the use of the general principles and regularities of the child's mastering of speech in the course of ontogenesis is a very perspective direction specifically for training of the PCs/hardware in more perfect ways to master automatic voice recognition and speech synthesis. It is possible to say that the scientific data can also be very helpful on development of the human speech in phylogenesis, i.e. in the process of its long historical development of evolution.

When we speak about speech, we usually mean a word, i.e. the second signal system - speech. By means of words humans can express not only any thought, a logical concept, but also a feeling, emotions. For this purpose there are such words as "joy", "grief", "anger", "fear", and thousands of others. The pictures of difficult emotional experiences depicted by writers in literature are verbal pictures or as they call it in science a verbal way of expressing emotions.

Any emotion can be expressed by humans, as we know, by using special intonations of voice and adding bright coloring.

Language of emotions as an absolutely independent channel of communication can function not only in parallel with a word thus defining an emotional context of what has been said but it can "work" without any word at all, for example, in the form of various exclamations, harrumphs, crying, laughter, etc.

Among many theories of speech emergence there are the ones that stand in favor of its origination from emotional exclamations which at first were spontaneously escaping in the form of exclamations of ancient ancestors of human beings in the course of different types of their activity [64].

Gestures most likely should be considered one of the most ancient means of information exchanging which according to a number of specialists preceded the emergence of the sounds based speech [65]. The abundance of gesticulations of ape-men well confirms this theory.

No doubt that the collective and community forms of this activity (hunting, work) causing the increased need of ancient people for exchange of information led, according to Friedrich Engels , to emergence of speech.

According to some theories a concrete push to emergence of a word could be onomatopoeia, i.e. designation of objects and phenomena of the outside world by ancient people by voice imitation (sound) which anyway were emotionally and vividly characterizing this subject or phenomenon [64]. Therefore the similarity of this theory to the theory of emotional exclamations is obvious. A similar way of word formation is observed of a child at its early stage of speech development. The child says " bough-wough" instead of "doggie", "tu-tu" instead of "car", "top- top" instead of "to walk", etc.

These emotional and imaginative voice sounds and intonations which the child's speech is so rich with, which do not disappear in the speech of adults adding to their words some special and sometimes an absolutely different sense is the ancient language of emotions developed by our far ancestors according to Darwin Ch. - the only means of communication and mutual understanding. In the light of the evolutionary theory the speech mastering process by our ancient ancestors went on in the following way: from the emotional and imaginative description of objects and phenomena of the outside and inner world by the voice sounds characterizing these phenomena (an iconic sign form) to the abstract word- symbol not directly linked to the described phenomenon (a symbolical sign form).

It is difficult to tell at what stage of human evolution the human being passed from the language of emotions to the word-symbols but it is clear that this evolutionary process lasted for long. Nevertheless attempts to understand how and when it happened have never ceased. Scientists, for example, decided to define how our far ancestors spoke and what were like the sounds of their voice. The anatomist Edmund Krelin from the Yale University reconstructed the structure of the vocal apparatus on the basis of fossilized oddments of bones

of an ancient human. It appeared that the throat of the ancient man was located very high and a pharynx was undeveloped. By the structure of the vocal apparatus he more resembled a newborn child or a chimpanzee [66]. In figure 1 the evolution of articulation organs of the ancient man is presented.

Further under the framework of the research, some modern equipment was used to support the research. A speech specialist Philip Liberman from the Connecticut University measured the resonant cavities of the articulation organs of the reconstructed model, transferred to digital data and received characteristics of the ancient man's speech. It turned out that the ancient man could badly articulate sounds A, E, U while his speech tempo was very slow. A modern human is capable to utter up to 30 phonetic elements per second while an ancient man 10 times less, i.e. he spoke according to our concepts unnaturally, drawling sounds, kind of singing them.

This version is quite in line with the opinions repeatedly expressed in literature that the ancient human learned to sing earlier than to speak [67] , [68], [69].

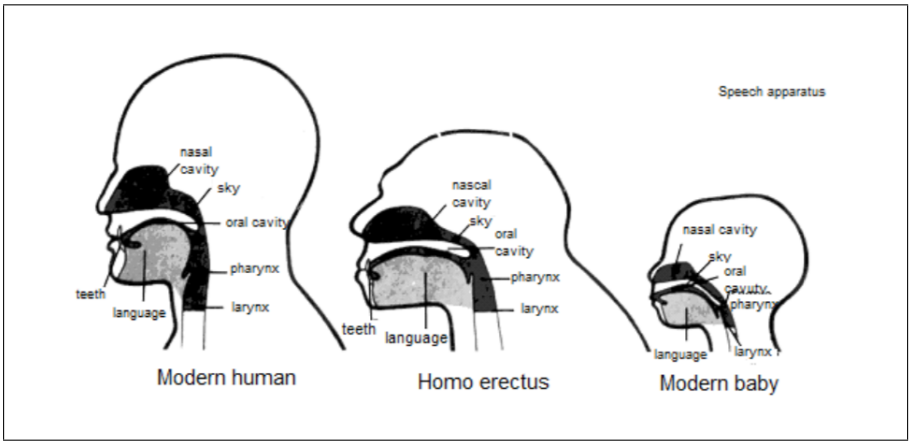

Figure 1. Evolution of human articulation organs

As we see, the language of emotions went through a surprising evolution in the course of millions of years of its existence. We find its origin in sounds of animals [67]. Millions of years it served as the only means of communication of our ancient ancestors. With emergence of the abstract and logical verbal speech and the second signal system as the evolutionarily more progressive form of communication and thinking, however it did not disappear. For modern people the language of emotions is the most important channel for information exchange in the course of their daily and vitally important communication. Moreover, in the art of stage speech and singing this ancient language certainly has definitely changed according to the requirements of the contemporary art forms and regained again its dominant position. Because if you dare to deprive any art of its emotional expressiveness it will stop being the art.

The general intelligibility of auditory signals and universality of the language of emotions of all people in the world in their speech in the sphere of arts that extends even to fauna are based on as shown in figure [70] natural linkage

between acoustic means of expression of emotions by voice and a physiological condition of the organism while experiencing this or that emotion.

Such is the curious fate of the language of emotions for millions of years of its existence. But studying of it is not finished yet, it has just begun.

2.1.1 *Rational and emotional brain*

The characteristic feature of the language of emotions - its independence from words. The emotional intonation can strengthen the words sense, contradict it and exist in general without words. How to explain this independent functioning of these two channels of communication - verbal and emotional and figurative/imaginative?

The reason is found: it is in special characteristics of our brain. It is known that it consists of two halves- the right-hand and the left-hand, as it would seem absolutely symmetric ones. Researches, however, showed that these halves by nature are equipped with different functions. The left-hand hemisphere of brain actually is called a speech or a dominant hemisphere which is responsible for ability of humans to speak logically and coherently (Broca's zone), and to perceive the logic of words (Wernicke's zone). Violations of these brain zones such as traumas and a tumor can lead to loss by a human of the corresponding abilities to coherently and logically speak and to grasp the logics of speech.

In the right half of the human brain there are zones responsible both for forming of emotional and figurative properties of speech and for perception of them in the other person's voice. It is considered that humans are obliged by their musical abilities to development of these corresponding zones of the right half of the human brain. Their violations can lead to tonaphasia/amusia, i.e. inability to understand music and emotional and figurative intonation of speech.

More wide-ranging studies showed that the left-hand half of brain is responsible not only for speech logics but also in general for all forms of the abstract-logical thinking and understanding of the external and internal world of the humans. It is a rational brain operating with logic, signs, symbols and abstractions.

The right half is the emotional brain perceiving all phenomena of the outside and inner world in a concrete, figurative and, perhaps, in an artistic form. Between the brain hemispheres there is a thick "telephone cable" of millions of nerve fibers by which both halves of brain communicate based on any slightest reason. That is why humans perceive all objects and phenomena of the outside world completely: figurative-emotionally and at the same time logically. However, different people have either the figurative and emotional or rational and logical form of thinking that prevails and of perception of the world [71]] which apparently is related to dominance of functions of the right or left-hand hemispheres of brain.

It would be a mistake to think that great artists, writers, actors, composers

have a poorly developed brain function of abstract logics. On the contrary, a great artist is always a big thinker. An artistic creation can reach perfection in any kind of art only when it rises to the level of social and even philosophical comprehension and generalization of the real world. Therefore a true and big art gives not only emotional and esthetic pleasure but also makes you think, suffer, be indignant, and strive for perfection and achievement of truth.

Similarly it would be a mistake to think that creativity of a scientist in its best form is the maximum activity of the rational and logical properties of brain at complete passivity of the emotional and figurative function. Such a model is peculiar to a robot or to pathologic cases in the brain activity.

For harmonious human development both of his brain functions - abstract and logical and emotional and figurative - have to be proportionally "active or loaded". The human brain and health suffer especially strongly from an overload of an abstract-logic function, especially in childhood. The overloaded school programs are based on introduction of information to a child's brain mainly through an abstract-logical half of the brain. But in childhood, especially at an early age of the child this function which is not well developed at all and inefficient, does not provide a larger channel capacity for information and more over it is especially friable and vulnerable. Certainly, the given function needs to be developed and no doubt that it is not permissible to overload it. There are some scientific articles which prove that many diseases and health disorders of children's organism are caused by this fact. We need to consider and employ this fact especially in the movements to reform school programs that started to develop at present.

2.1.2 *Emotional human hearing*

At estimation of the actors' ability to express emotions by voice it was found out that not only actors and vocalists have this ability to certain degrees of perfection but listeners also. It appears that listeners too hear differently the emotional coloring of the speaking person's voice or especially of a singing person.

Human hearing as an ability for perception of sound vibrations has many kinds. So, for example, there is a speech or phonemic hearing as ability to perception and comprehension of speech sounds. A thin tone hearing is identified by scientists and the ear for music of virtuoso performers is well known to everybody. The ear for music in its turn has variations: absolute, relative, timbre, melodic and etc. At the same time singers have vocal hearing which is not limited by the musical hearing because there are great musicians who do not have a vocal ear.

At last, we for the first time named the term "the emotional hearing/ear" which means an absolutely special type of phenomenon. It is clear that emotional hearing is not something related to speech as there are people who poorly hear

emotional intonations or are deaf in hearing of emotions. To the contrary, fine emotional hearing is quite often combined with modest speech abilities. Emotional hearing is in more close relationship with ear for music: musicians and singers have well developed emotional hearing but in different ways. There are musicians with perfect, mathematically exact musicality/musical talent who however lack emotionality. But these are also "formal musicians" and for them it is difficult to win the heart of their listeners.

We found out that not only the ability of people to express emotions by voice varies but also abilities of listeners to correctly perceive what they hear.

There is a concept in science that there are people of artistic type. A famous physiologist academician Pavlov I.P. suggested separating people in two types on the basis of their psychological type: thinking tympanum and artistic [71]. Unfortunately, Pavlov I.P. did not leave any methods allowing defining to what type this or that person belongs. At present such methods are developed in the Institute of Psychology of the Academy of Sciences jointly with experts on vocal and musical art. One of the key indicators of the person's artistic type as not only researchers think but also many experts in the area of arts think is the person's emotional sphere developed which is expressed in degree of subtlety of emotional hearing.

How is the degree of the person's emotional hearing development estimated? For this purpose the tests are applied already known to the reader consisting of a set of emotionally colored speech and vocal phrases. Having listened to the sounding of emotionally colored phrases where only a number of them contains a weak hint to joy, grief, discontent, fear or either no emotion at all, each of listeners has to write down his or her judgment of the nature of the emotional intonation of the actor's voice in particular columns of special forms. The degree of correctness of these records is expressed further in percentage of the correct estimates by the listener of all phrases and is expressed based on a 100-scores scale.

As it is stated by the majority of researches on emotional hearing [72] the degree of its development in people of different ages and of professional categories varies within a wide range from 20-30 to 95-98 conventional units by the 100-scores scale. Researchers think that a nice emotional hearing/ear is the human's natural quality and to a considerable extent is peculiar to people of the artistic nature. At the same time emotional hearing can be developed through special exercises.

It is obvious that emotional hearing as an indication of a person's artistic talent is important, first of all for persons engaged in the sphere of artistic professions. As scientific observations show owners of nicely developed emotional hearing make great success in tutoring.

But emotional hearing is necessary not only in arts. There are such types of important works at present the realization of which depends on the general and emotional state of a person, for example, astronauts, test pilots and etc. As soon

as the only communication channel, for example, with astronauts is the speech channel, we can judge about their emotional state only by voice. But, as we already know, not everybody is capable to apprehend subtle emotional changes in the voice of a speaking person. According to authors [72], only 2-3 persons per 100 people of the total population have an especially nice emotional hearing. They give around 95-98% of the correct answers whereas average persons do not make mistakes only in 70% of cases. The tests and procedures developed by researchers [72] allow finding such people who can be entrusted to control an emotional state of other people in cases when especial responsibility is needed.

2.1.3 *Acoustic alphabet of the language of emotions*

People use a particular way of information coding to transfer the sense of speech, i.e. by using a well- known alphabet. Words and phrases of the coherent speech which convey a particular sense are formed from elements of this alphabet – phonemes. A question arises: is there any similar alphabet for language of emotions? Many specialists ask this question in recent years. Authors [72] tried to solve the problem based on studying the artistic vocal speech or singing which are especially full of emotions and therefore can be the right object for research.

The following task was assigned: if our ear/hearing is capable to find in human voice particular emotional colorings, for example, emotions of grief, joy, anger, etc., then, apparently, there have to be some objective acoustic signs which are responsible for transfer of these emotions to listeners. What signs are they and what acoustic properties of a sound they bear? Whether loudness of a voice, pitch, timbre, etc.? By means of the method of expert evaluations several dozens of phrases were selected from various vocal works filled with absolutely particular emotional sense (joy, grief, anger, fear).

Studying of the acoustic structure of these phrases showed that it significantly differs depending on what emotion is expressed by the singer. It came to light especially when the singer was asked by the experimenters to sing the same vocal phrase several times but every time with a different emotional coloring. Experienced singers perfectly coped with this task. Some of them were even capable to express any emotion while singing melodies without words, the so-called vocalizations and even through singing only one vowel at one musical note. As soon as in these experiences the phonetic structure of phrases is constant, all changes in the acoustic structure of a phrase could be explained by a change of the emotional sense. Figure 2 displays the probability of the correct perception by listeners of different emotions.

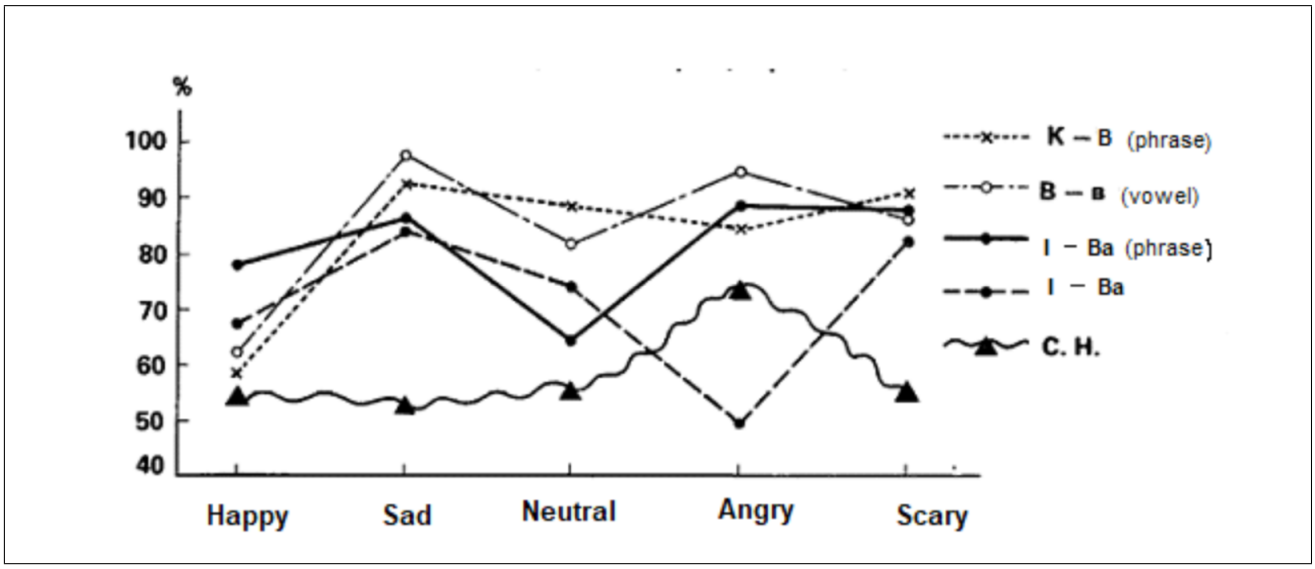

Figure 2. Probability of the correct perception by listeners of different emotions (vocal speech, violin).

Language of emotions is independent from the sense of words. One of proofs of this assumption is the human ability of an actor, singer or musician to express emotions not only when pronouncing or singing a phrase with words but through vocalizations without words, when singing one vowel monotonously at one music note and even by means of the sound of a violin. In the latter case the violinist was given a task to express joy, anger, fear and etc. while playing a piece of music from "The Rondo capriccioso" of Sen-Sans. On a vertical scale - the probability of the listeners' correct perceptions of different emotions is demonstrated (%).

The sophisticated electro-acoustic processing with application, with the samples of voices received based on this method, expressing different emotions showed that each of the listed emotions - joy, grief, anger, fear is expressed by a change not only of one property of a sound but practically by a change of all its properties: force, pitch, timbre and tempo-rhythmic characteristics. These changes were related even to those characteristics of a vocal phrase which were set by the composer, i.e. sound pitch-rise and tempo-rhythmic patterns of a melody. Thus, while expressing this or that emotion, the singer deviates to a certain extent from the instructions of a music sheet and this is a factor that defines the emotional coloring of his voice.

And it turned out that each kind of emotions has its own set of distinctive acoustic signs of voice relevant to it. So, for grief, the longest duration of a syllable, slow increase and dropping of the sound's power/loudness, distinctive "ups" and "downs" in the sound pitch of the music sheets creating a crying intonation and etc. are typical. Anger, on the contrary, is characterized by abrupt, "chopped" fronts and scraps of a sound, by a big volume of a voice, by ominous ringing or hissing timbre. For fear, abrupt differences in the voice volume, a serious violation of a melody rhythm and a sudden increase in pauses turned out to be a distinctive feature.

Each performer certainly used his/her individual voice means for expression of emotions. The singer had some specificities in variation of these means, for example, used not all signs but only part of them, however, none of them chose, for example, for expression of grief the signs that are characteristic of joy or anger, etc.

Statistical processing of acoustic signs of different emotions confirmed that they rather differ from each other. In other words, they are related to the emotional state. It gives the grounds to consider the listed objective signs or acoustic means of emotional expressiveness of the singer's voice to be the elements of the alphabet of the language of emotions in singing [70]. But is it only for singing?

2.1.4 *The language comprehensible to everybody on the earth*

We have used the term "language of emotions" until recently more likely in the imaginative/figurative sense rather than in its exact scientific sense. Now we have definite objective proofs that this is a communication system similar to a certain extent to the phonetic system of the logical speech. But at the same time it is incomparably more universal and claims to be comprehensible to everybody.

Acoustic means of emotions expression by voice strike us with the infinite variety of coloring, shades and specific features. At the same time this infinite variety and quantity of sound colorings submits to some internal regularities. In the beginning the researchers [72] were surprised by the fact that the main acoustic means for expression of emotions in a voice of modern singers were in fact the same as the ones used by Shalyapin despite the originality and distinction of different voices. But it became clear later that if there is language of emotions it cannot be otherwise. The infinite variety of speech features of people does not withhold the existence of the phonetic alphabet uniform for all (within this given language). Apparently, there is an acoustic alphabet of emotions uniform for all people. But if Shalyapin or any other singer would find some other alphabet of the language of emotions then his performance would be unclear for the listeners sitting in the concert-hall. Communication by means of language assumes knowledge of this language both by speaking and hearing people.

Scientists Galunov V.I., Manerov V. and etc. studied the acoustic means of emotional expressiveness of the routine speech. Comparison of the research results [73] showed that the speaking and singing persons in fact use certain general means to express emotions - the common alphabet despite essential differences between speech and singing. It seems that this commonality extends also to the actor's speech: talented actors and singers find voice colorings from

life to express emotions, i.e. from speech. Therefore they differ in truthfulness and common intelligibility of auditory signals.

How the language of emotions originated and what its universality is based on?

The psychophysiological basis of such universality and common intelligibility of auditory signals of the language of emotions most likely lies in dependence of a character of the sound pronounced by a person from a physiological condition of his/her body which at that moment experiences this or that emotion and in particular on condition of the vocal organs. Thus, for instance, a person in anger has all his/her muscular systems including vocal chords and the respiratory system strongly strained which inevitably affects the character of a sound. The total muscular relaxation of a person who is "heart-broken" also leads to the distinctive changes of a voice. At expression of joy it is felt that a person speaks or sings as if smiling to something. And really, his/her face is lit with smile which has an immediate impact on the physiological properties of the sounds. The acoustic theory says that expansion of an oral opening leads to the shift of formant frequencies to the more high-frequency area which is observed in conversations or singing based on "a smile" (see ranges of Shalyapin's voice in figure 3). Formants are groups of the reinforced overtones in a range of a human voice defining the phonetic sound quality or its characteristic timbre features

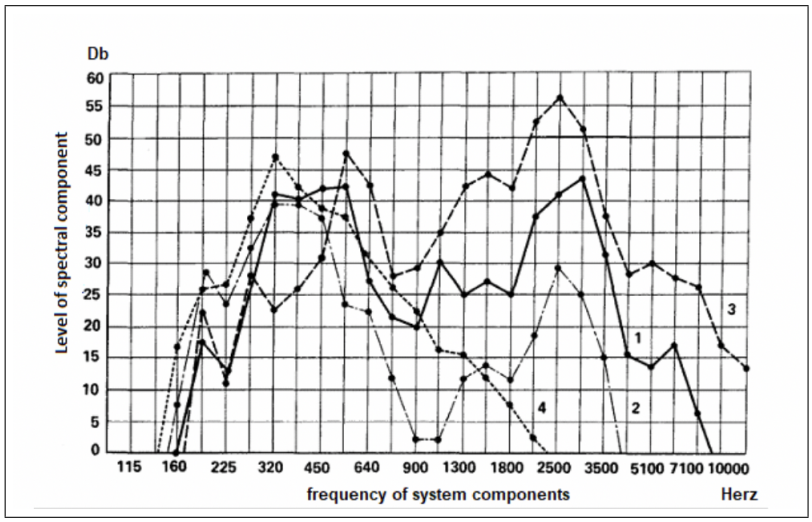

Figure 3. The integral ranges of FedorShalyapin's voice at singing some pieces from the vocal works full of various emotional contents show big differences in the level and frequency of high overtones of voice at expression of joy - 1, grief - 2, anger - 3, fear - 4. These distinctions determine peculiar changes in the actor's timbre of voice at expression of different emotions

Thus, the acoustic sign of emotions in a voice is bound or better to say dictated by a physiological sign of the state in which the singer or his vocal organs are. It depends not only on the unity of voice means to express emotions but apparently on perception of emotions by the listener. Listening to an emotionally charged sound of a voice, we kind of mentally imagine in what way and how it could be made. This process, fast and subconscious, strengthened in

us by millions of years of evolution makes us able to comprehend the language of emotions (figure 2). A similar mechanism is known in science under the term "Analysis through synthesis". It found reflection in the so-called motor theory of speech perception.

Essence of "the motor theory" is the following: a person listening to the speech of the interlocutor kind of mentally repeats his words. It is proven by micro movements of the listener are quite often obviously expressed and noticeable from aside, by his lips, tongue and all articulation bodies movement.

The characteristic feature of the language of emotions - involuntariness and subconsciousness - is manifested both in its perception and decoding and when forming emotional tints of a voice (could it be the reason therefore of that the sound of voice quite often gives away a person: that a person tells not what he or she really thinks). Try to reproduce intonation of your own emotional exclamation and if you are not an actor, then you will be convinced of an absolute falsity of the copy. It is not for nothing that the theory of acting skills [74] developed by Stanislavsky K.S. and which found its physiological justification in P.V. Simonov's [75] works requires an indispensable participation of the elements of a subconscious mind in management of speech intonation, including the actor's behavior on the stage.

In the light of the provided data the universality of language of emotions is based on similarity of the basic physiological processes that follow emotions and are typical to all people let it be ordinary persons, actors or seamen. These assumptions point to the famous commonality of voice means for expression of emotions by humans and the higher animals. Darwin Ch. was the first to pay attention to this fact in the treatise [67].

There are all grounds to consider that the means of musical expressiveness or rather emotionality of the instrumental music also have a common nature with the means of emotional expressiveness of speech and singing.

It is not by chance that the language of music and singing which is understandable to everybody on the Earth-the language of emotions sounds at all international festivals as the language beyond borders and thus promotes mutual understanding and friendship between people.

2.1.5 *The hindward language of Lada of Samsonova*

Imagine that you see a smiling pretty girl, the owner of a unique ability: she knows "hindward language". The point of it is that Lada can pronounce the words of any phrase backwards, i.e. kind of starts to read them from the last letter to the first. At the same time she does it so quickly and surely that makes people around amazed. Such hindward procedures are done by Lada easily not only with familiar words but also with unfamiliar words which she hears for the first time.

Lada Samsonova was invited to the laboratory of nonverbal communication

of the Institute of Psychology of the Academy of Sciences of the USSR to learn in more details about her unusual abilities. For a number of years to be more precise since 1987 we are engaged in studying of abilities of people to perceive sense of words and meaning of different types of nonverbal extra-linguistic information in the conditions of its inverse, i.e. inverted in time presentation to the listener. Practically it is reached by scrolling of a recorder tape with the recorded speech and other sounds in the opposite direction.

Already in the sixties a French researcher Mol A. showed that the method of temporary inverting destroys the semantic information for the listener, i.e. actually the speech, verbal information. He included in the esthetic information based on his statements all types of nonverbal information of imaginative/figurative and subject (iconic) nature. However, Mol did not carry out any detailed research on different types of extra-linguistic information of speech.

According to Morozov's hypothesis there is a fundamental difference in the brain's mechanisms of processing of actual speech (verbal) and nonverbal (extra-linguistic) information. These distinctions were to be manifested in the specificities of perception by brain both of actual speech information and of different types of extra-linguistic information.

The first thing the authors did [72] was that they tried to scroll backwards for a group of participants of the test or listeners a tape with the recorded emotionally colored phrases well-known to readers and once recited by Basilashvili O. and also some tapes with the recorded singers' voices emotionally colored in different tones.

The results showed that the hindward perception on the contrary practically does not deprive of a possibility for listeners to correctly perceive the emotional coloring in comparison with the norm. A little bit lower total percentage of the correct perceptions can be explained by unusualness of "hindward" sounding of speech and singing. A bit of training can practically smooth off these distinctions.

Thus, anybody without understanding a single word of the "hindward" speech and singing will recognize perfectly and immediately (without any training) the language of emotions in the inverted in time option.

An ordinary listener is capable to distinguish in "the hindward" sounding the other type of extra-linguistic information, for instance age of the speaking person. For this purpose the same emotionally colored phrases of Basilashvili O. were used and the participants of the experiment were given a task to guess only the age of the speaking person and nothing else. The results showed that the listeners in general were close to the truth about Basilashvili's age. But at the same time a curious detail came out – by estimation of the listeners the actor's age was defined mostly depending on the emotional coloring of his voice (!): emotions of anger and fear gave the maximum age (about 50 years old) and emotions of joy and neutral sounding led to the minimum estimates (about 39 years old). As the difference by age estimates is considerable, about 10 years, then for people who are eager to look younger, and it refers first of all to the

fair sex (women) there is a chance "to look younger" using the knowledge of this psychological regularity.

We observe here a very important law of interaction and interference of different types of extra-linguistic information at the level of its perception and processing by human brain. These data can turn out to be useful for criminal investigations in case of consideration of a testimony with identification of a person's age by his/her voice.

Can any person recognize his or her acquaintance if his/her voice sounds differently?

The experiments conducted [72] showed that the listeners, 26 people involved, perfectly coped with this task and determined by voice (on average with accuracy of 86%) 20 acquaintances (10 men and 10 women) and distinguished them from two strangers whose voices were included intentionally in the tape test recordings scrolled hindwards. The gender of the announcer was wrongly defined only in 1% of cases and that corresponds to accuracy of defining in these conditions of gender of the announcer by his voice with reliability of 99%!

The inquiry of the participants-listeners showed that when determining based on "hindward" method the voices of their acquaintances and emotions they were mainly guided by the timbre features of the voice that our hearing abilities can preserve under these conditions. Theoretical calculations as well as their experimental checks show that when inverted in time the integral ranges of a voice responsible for individual and emotional features of a timbre remain preserved. It gives all the grounds to state that the ability of listeners to perception of nonverbal characteristics of a voice is based on the integral or the averaged characteristics of the range for a definite time. According to the modern scientific insights such a mechanism of a complete perception is relevant to the work of the right hemisphere of human brain.

The extra-linguistic information of a voice appears to be more noise proof in comparison with the linguistic information not only in respect to action of noise but also in relation to the range of frequency restriction. The graph shows that restriction of high frequencies up to 300 Hz completely destroys the linguistic information. Definition of emotions in such a signal as well as recognition of the announcer can be preserved.

As already mentioned the listener does not clearly understand a speech in case of its hindward sounding and that speaks about absolutely different mechanisms of human brain's speech decoding. It is considered that this mechanism involves a subtle and detailed in terms of time, a segment by segment (a phoneme by phoneme and syllable by syllable) analysis of dynamics of the formant structure of a speech signal. As this dynamics is distorted/inverted in time in "hindward" sounding, naturally in these conditions the human brain cannot decode a voice information/verbal message.

2.1.6 *Two-channel nature of voice communication/ verbal communication*

One of the most important principles of brain work that distinguishes it from many technical systems in particular is the principle of parallel processing of many different types of information delivered by different analyzer channels (hearing, vision, skin and tactile, muscle sensing, etc.) and even through one and the same channel. As for the sound speech the brain can be considered a two-channel system despite of a seeming single channality of the speech acoustic signal.

Thus, the traditional single-channel scheme of voice communication/verbal communication needs a serious correction. In the light of these data and other modern insights provided in this section the two-channel nature of the sound voice communication/verbal communication is illustrated in figure 4 where the specified channels are designated by the terms "linguistic information" and "extra-linguistic information". The extra-linguistic channel in turn consists of a number of sub-channels by nature of different types of this sort of information.

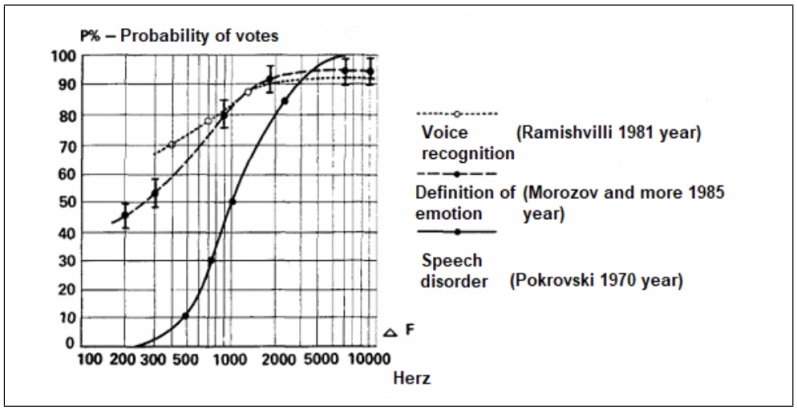

Figure 4. Dependence of the correct estimates of different types of voice information/verbal message on limitation of the high frequencies range

Functional division of these channels happens in a human brain according to the principles of information processing: the left-hand (linguistic) hemisphere carries out a segment by segment analysis of a speech signal, being guided by a subtle dynamics of its formant structure in micro intervals of time. The right hand hemisphere (extra-linguistic) uses the holistic principle of analysis on the basis of comparison of an integral acoustic image of a signal with the patterns (standards) of examples of this type of information stored in memory.

The two-channel principle of brain work is shown not only in conditions of a speech perception but also in the process of forming (generation) of a speech statement in the form of absolutely different functions of larger hemispheres of brain in this process. Figure 5 shows the probability of the correct perception of different types of voice information/verbal message at increase in a noise/signal ratio.

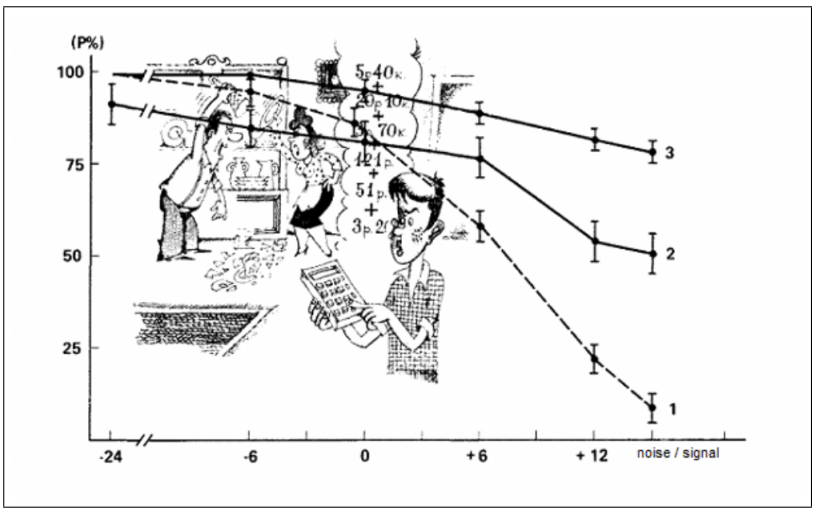

Figure 5. Probability (P, %) of the correct perception of different types of voice information/verbal message at increase in the ratio noise/signal (dB): 1 - linguistic, 2 - emotional, 3 - gender of the announcer

The noise proof property of extra-linguistic types of information (emotional and about gender of the announcer) is much higher versus the linguistic type (speech). At action of strong noise (at a ratio signal/noise=+16 dB) the linguistic information is completely disrupted (listeners cannot understand any word) but reception of emotions is possible with probability above 50%, of the announcer‘s gender - more than 80%.

An objective reason for division by brain of linguistic and extra-linguistic communication channels is the differences in acoustic means and principles of coding of these two types of voice information/verbal message. If for verbal information the dynamics of formant structure of a signal is more determinative then for the extra-linguistic communication a special role is played by dynamics of the main tone of voice and other features of the prosodial speech organization. Thus, linguistic and extra-linguistic channels are isolated (by a number of criteria) in all links of the verbal communication system. As for impact of noise this isolation is manifested in different noise proof degrees of linguistic and extra-linguistic information: the noise proof feature of extra-linguistic information is higher (figure 6).

There are obvious differences between the named channels in the evolutionary and historical aspect: extra-linguistic communication is much more ancient versus linguistic communication. Emergence of a word in the course of evolution as of a very perfect means for delivery of any kinds of information did not lead however to diminishing of the role of an evolutionarily ancient form of extra-linguistic communication. It continues to coexist alongside with a word, significantly supplementing it and altering its sense, and in many cases claims for its independence. In many situations of verbal communication it is less important WHAT somebody says because what matters is HOW somebody speaks and WHO speaks. The domineering role of extra-linguistic information seems to be obvious in such specific kinds of verbal/voice communication as the

art of scenic speech and singing. The main and almost not studied property of a two-channel system of voice communication/verbal communication is interaction of channels of the linguistic and extra-linguistic information appearing in all links of this system and at all stages of brain processing of a verbal message.

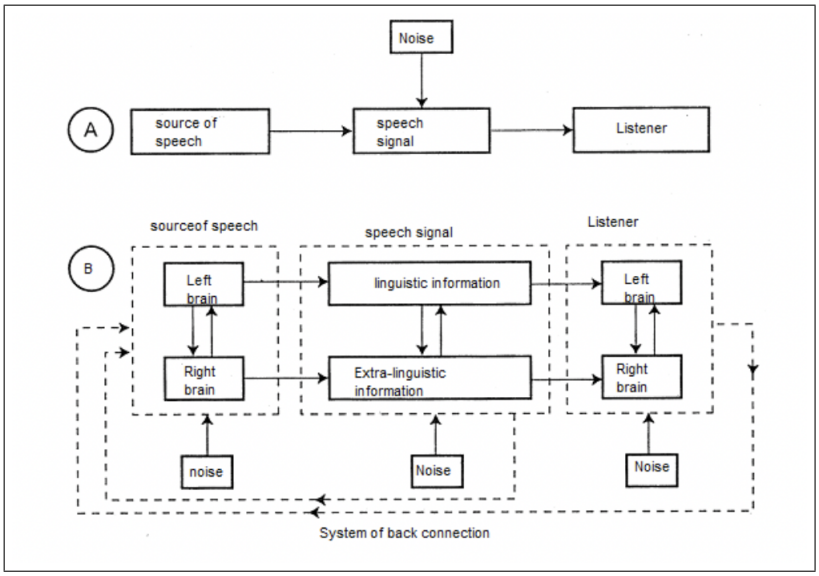

Figure 6. The traditional scheme of verbal communication provided by one channel (A) and the scheme of verbal communication which highlights the two-channel nature of it

2.1.7 *Humans and the machines - problems of mutual understanding*

The most convenient and appropriate solution of the given problem from the human's point of view would be introduction of information to the hardware not by the keyboard but directly from the human voice, i.e. through a natural verbal communication of people. However, as of today the hardware does not understand human speech to the due extent and reliability, hence the user needs to resort to services of the whole staff of "translators" (software developers, operators) who enter information in machine in a special and clear to machine language and also decode the information generated by the machine.

Creation of the 5th generation hardware which reliably understands anybody's speech and also of the speaking hardware is a global task for science which is successfully solved by Japan, USA and France. However, the task turned out so difficult that it has not been completely solved so far. For example, the hardware will easily recognize a speech of one person or of several announcers known to the hardware but it does not want to distinguish the speech of unfamiliar persons, understands the speech of adults and does not want to understand children's speech. If the number of announcers can be increased the volume of the vocabulary needs to be decreased. Even powerful PCs are not able to solve a small problem which any child can solve as to write a letter dictated verbally. Even if a familiar to the hardware announcer reads

the text familiar to the PC but in a hoarse voice, not distinctly or with noise background around the hardware, the machine will not understand the text.

A specialist who works over this problem Doctor of Engineering Sulukvadze M. from the Institute of Control Systems of the Academy of Sciences of the Georgian SSR considers that "automatic SR should be considered one of most complex problems in engineering cybernetics. We are not sure that it will be completely solved even in 20 years. The term "completely" means the level of perception and understanding of speech of a human being under real conditions of his/her verbal interaction with other people".

The reason of this obstinate "unwillingness" of the hardware to learn to understand perfectly human speech is in specific and emotional features of people's speech who considerably distort the phonetic structure of a standard speech signal which the hardware is tuned to recognize in advance. It is known that voice information/verbal message is coded by a formant structure (frequency of formants and their dynamics). But the frequency especially of the first formants significantly depends on the frequency of the main tone of a voice: increases at growing in a voice and decreases when the main tone of speech goes down. Time histories of the main tone of a voice are a major means of emotional expressiveness (voice intonation) which happen in connection with emotions in the range of one, one and a half and even two octaves as we observe n cases of strong emotional exaltations. It leads to the strongest deformation of the total spectral structure of speech and misunderstanding by the hardware of speech. According to Fanta G. [76], female and children's voices that have a higher frequency of the main tone in comparison with men's are characterized also by increased formant frequencies (on average for 17-25%).

Apparently, there is a mechanism in a brain considering information about increase in the average frequencies of formants due to an increase in the main tone of a voice (voice pitch). Therefore it makes no difference for us at what pitch of the main tone words are pronounced: by male, female or children's voices, intelligibility and clearness of the speech are ensured.

But the above listed difficulties make only a small part of all difficulties thus getting in the way of the PCs' training to understand voice speech. Figuratively speaking, all types of voice information-linguistic and extra-linguistic-"are kind of dissolved" in a human voice sound. Our hearing does not have any difficulties with separating and considering them while the hardware feels "at a loss". Therefore we hope that further studying of individual and emotional features of voice speech and the mechanisms which our hearing and brain depend on while separating the sounds will allow us to solve the problem of automatic recognition of human speech at last. Perhaps, the data will be useful that for perception and processing of logical and emotional information of speech in a human brain there are two available specialized and at the same time interacting parts: the left-hand hemisphere of brain - for logic, the right - for emotions. By the way, in one of works under edition of the most prominent American

specialist in automatic SR Wane Lee the algorithm and the speech recognizer are described based on the principles of the right brain hemisphere's work (i.e. the holistic but not segment by segment analysis with consideration of prosodic and extra-linguistic characteristics of a speech signal).

Human interaction with the hardware assumes creation of the speaking robots on the basis of PCs. We all heard over the radio in popular scientific programs the voice of robots deprived of any emotions. The lack of emotionality is a characteristic feature obvious in its voice. But why anyway the robot and its voice have to sound unemotional? Is it impossible "to make alive" its voice and to enrich it with intonations of human speech? Emotional coloring of a voice makes it sound esthetically more pleasant, psychologically compatible to human perception and more over it not a useless acoustic accompaniment to the speech. It conveys very important information, for example, about the degree of importance of the message, about an imminent danger and etc.

To inhale emotions in a robot's heartless brain is one of the acute problems in cybernetic science linked with the issues of selection and formalization of the invariant acoustic signs responsible for emotional coloring of a sound.

It is obvious that "making alive" of the robotic speech is one of many technical tasks which cannot be solved without knowledge of the acoustic language of emotions alphabet. But to integrate this alphabet in the robot's AIit is necessary to identify at first and then to formalize the signs responsible for voice emotionality.

Thus, to solve the problem of complete mutual understanding between a human being and machine we need to emotionalize the latter. It is necessary that a robot as well as a human being after they hear a phrase "I am very glad to see you!" pronounced not in a neutral manner but in a mocking and derisive voice could not only apprehend the mocking appropriately but also could understand the emotional intonation which contradicts the real sense of these words. The robot has to understand our language of emotions.

It is difficult to overestimate the practical value of such an automatic machine, for example, for monitoring of a psychological state of astronauts, test pilots and many other operators who by nature of their work are in difficult and extreme situations and success of business depends so much on accuracy of their actions. Some attempts have already been made to create the machine responding to emotions in a human voice. One of them based on assessment of tempo-rhythmic characteristics of speech belongs to the group of the engineers working in cooperation with a phonetician Nosenko E.L. and they reported about it at the symposium "Speech, emotions and a personality". Leningrad scientists were awarded the copyright certificate [73] for a similar device but on the basis of dynamics of the main tone. Our understanding of voice communication/verbal communication as of a two-channel system is undoubtedly fruitful in a bionic sense, i.e. for creation of new, more perfect systems of the automatic analysis and speech synthesis by means of the hardware. It is possible to state that further

achievements in solution of the problem of the automatic analysis and speech synthesis will depend on how fully using technical means we can model the principle of paired work of the human cerebral hemispheres taking into account the functional specialization of each of them for perception and processing of different types of voice information: actual and extra-linguistic voice speech.

2.2 Vocal speech as language of emotions. Experimental research

The acoustic signal of the vocal speech transfers much more emotional information in comparison with the normal speech. This circumstance is predetermined by nature of the vocal speech and does not need any special proofs. In this regard vocal speech seems to be an exclusively convenient object for studying the emotional-expression means of a human voice.

2.2.1 *Main problems*

The main tasks of the authors involved in the given research [72] included: 1) development of a method of quantitative assessment of emotional expressiveness of different performers singing; 2) quantitative assessment of listeners' abilities to perception of emotional expressiveness of vocal speech; 3) analysis of the acoustic signs of vocal and speech signal preconditioning transfer to the listener of the emotional content of singing.

2.2.2 *Method of an emotional and semantic divergence*

Four main methods are applied for studying of emotional and expressive function of normal speech: 1) analysis of examples of speech of a person who is in a natural emotional state caused by stressful conditions; 2) the clinical method based on use of painful mental conditions of the person when the speech gains emotional character; 3) the method of hypnotic infusion of an emotional state; 4) method of an actor's transformation.

In the work of authors [72] the method of an actor's transformation was used in two options: 1) the method of studying of the natural vocal speech, i.e. the analysis of phrases from the vocal works which are obviously bearing this or that emotional information determined by the content of the work and abilities of the performer; this method analyzed acoustic means of expression of emotions in F. Shalyapin's voice; 2) the method is called by the authors [72] as the method of an emotional and semantic divergence. The advantage of the first method is its naturalness. The advantage of the second is in an opportunity to mark out acoustic signs of emotions in the clearest way.

It would seem that the first method will be enough to solve this objective, i.e. selection and analysis of the phrases bearing particular emotional information, from various vocal works performed by different singers. However, selection of emotional acoustic characteristics from amongst of similar selections encounters obvious difficulties in view of inhomogeneity of musical and lexical (text) material that can be a cause of changes in the acoustic structure of phrases. The method of emotional and semantic inversion offered by the authors [72] has the advantages where all structural acoustic changes are referredonly to a change in an emotional context.

The main point of the method of emotional and semantic divergence consists in performance by a singer of the one and the same vocal phrase with different emotional shades (contexts), selection of the most successfully performed pieces and their subsequent acoustic analysis for the purpose of selection of the physical properties of a sound that define the emotional content of a phrase.

Many qualified vocalists if trained and based on pre-training are capable to enrich a vocal phrase practically with any emotional sense: joy, anger, grief, fear etc. The authors noted that this ability of singers is reflected not only at performing of a vocal and speech phrase but also at vocalization of this phrase without words on one vowel. Moreover, some vocalists were capable to rather brightly express this or that emotional state when singing only one vowel and on one music note! These abilities of vocalists were also used when developing the method called as the method of emotional and semantic inversion because its essence is in alteration – of inversion of emotional content of a vocal phrase in relation to its semantic (sense) content.

The procedure of the research was as follows. Eleven singers - soloists of musical theaters and students of the conservatory senior years were given a task to sing a phrase from any vocal piece several times, every time adding different emotional senses in their performance: 1) joy, 2) grief, 3) anger, 4) fear . For comparison a performer was also offered to sing a phrase without adding any emotional context, i.e. neutral and unemotional singing. At the same time the performer had to try to keep under all conditions the melody and metro-rhythmic structure of a phrase. The head of the department of solo singing of the Leningrad Conservatory Professor BarsovYu.A. participated in this part of the research.

Each of phrases was sung by each performer 10 times and the emotional coloring of a phrase varied 5 times: joy, grief, anger, fear (randomly). In all cases the singers' voices were recorded on the tape recorder for a subsequent listening to the recorded tapes, selection and acoustic analysis.

The method of an emotional and semantic divergence offered by the group of authors [72], despite of the well-known artificiality is not so unusual and alien to the vocal art as it can seem at first sight. Discrepancy elements between the semantic and emotional content are sometimes intentionally introduced by the composer in the musical creation of the vocal piece and by the singer while

performance. So, at performance of Schuman's romance "I am not angry" when the singer sings: I am not angry, let my heart be hurt though you betrayed me ... etc. In the music accompaniment and in the singer's voice we don't hear any meekness at all.

In vocal music we quite often encounter similar examples of opposition or contrast to increase the artistic impact of music on the listener and in particular for disclosure of an internal inconsistency of the character, etc.

2.2.3 *The quantitative estimation of emotional expressiveness of singing by different performers*

The method of an emotional and semantic divergence allows solving several important scientific and practical tasks. One of them is measurement of emotional expressiveness of performance. For this purpose the tape recordings of the performed programs received by the method of emotional and semantic inversion were presented to the audience of listeners (15-20 people) from among students of the Conservatory and vocal teachers. The listeners had to make notes on special forms about their impressions of what emotional state is expressed by the singer per each condition listed in the program on the basis of the definitions of emotions which the singer was guided by in his or her performance following the program. The quantitative estimation of emotional expressiveness of the performance was made by the formula (NX /N0) * 100% where N0 is a total number of the conditions provided to the audience, Nx is the number of correctly identified ones. The named approach gave a chance to quantitatively estimate the degree of emotional expressiveness of each performer and secondly dependence of expressiveness on emotional context.

In figure 7 the average data are provided on degree of emotional expressiveness of 20 programs performed obtained by the authors [72] on the basis of estimates of these programs by the listeners. These data demonstrate that different performers have different abilities to express different emotional states by voice.

It was established that the degree of emotional expressiveness practically does not depend on the text and the melody of a phrase. Moreover, according to the data in figure 7, the singer can express rather expressively by voice this or that emotional state and in absence of the text, i.e. while exercising in vocalization (program No. 17 and 19) and even in absence of both of them: a text and a melody, i.e. while vocalization of separate vowels (program No. 12).

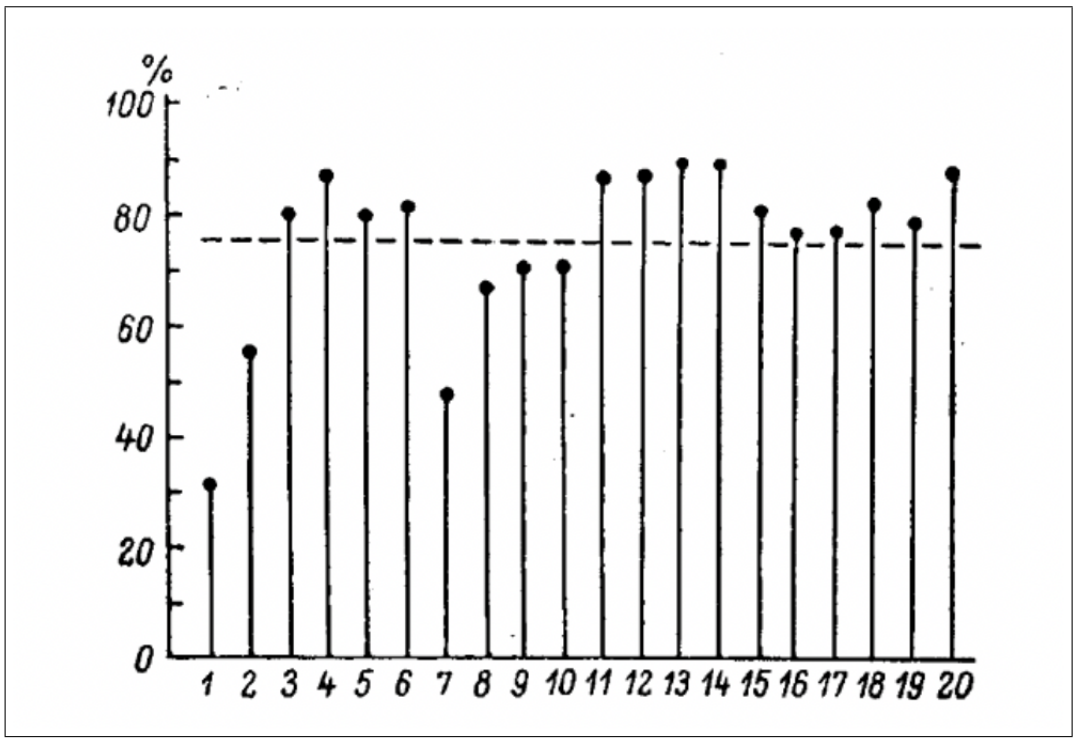

Figure 7. Assessment of emotional expressiveness of singing of different performers

Horizontal line-performers: 1 - bass A. S-n, *Farewell, joy, my life;* 2 - tenor B. A-y, *We will go with you, we will cheer up;* 3 - bass A. Yu-v, *I am not angry, let my heart be hurt;* 4 - mezzo-soprano Gr. P-k, *Here up to what I lived, Grigory;*5 - soprano. L. A-va, *Oh, my mother told me;* 6 - soprano. L. B-va, *Do not sing, my beauty, to me;* 7 - baritone A. M-v, separate vowels; 8 - baritone A. M-v, *Farewell, joy, my life;* 9 - soprano. L. B-va, *Oh, my darling;* 10 - soprano. L. A-va, *Oh, my darling;* 11– soprano. L. B-va, *Oh, my mother told me;* 12 - tenor V. V-v, separate vowels; 13 - baritone A. K-v, *Sleep, my child; 14* - soprano. L. B-va, *Sleep, my child; 15* - soprano, R. Sh-ya *Sleep, my child; 16* - bass. G. K-v, *Sleep, my child; 17* - baritone V.P-v, a vocalese on a phrase melody *"You wroteto me, do not deny it"; 18* - baritone B. P-v, *You wrote to me, do not deny it; 19* – soprano S. Ya-va, a vocalese on a phrase melody *"Sleep, my child"; 20* – soprano S. Ya-va, *Sleep, my child.* ***Vertical line*** -number of the correct estimates (%).

The obtained data allowed the authors [72] to estimate degree of expressiveness of singing depending on an emotional context. For this purpose the estimates of each of "emotions" by listeners were processed statistically. The results of the analysis are presented in figure 8. It turned out that the most expressively expressed emotions were grief, anger, fear and apathy (the number of the correct estimates of conditions with these emotional contexts exceeds 80%). Much lower is the percentage of the correct estimates of joy (56%). The analysis of distribution of estimates based on the terms of all programs showed that joy is expressed worse in comparison with other emotional states by practically almost all singers engaged in all programs. For us this fact was

unexpected because joyful moods and joyful tints are used quite often in vocal singing. It is better to look for explanation of this fact in the distinctions of perception by listeners of emotional contexts and the most important in the research of acoustic signs of emotions expressed in singing.

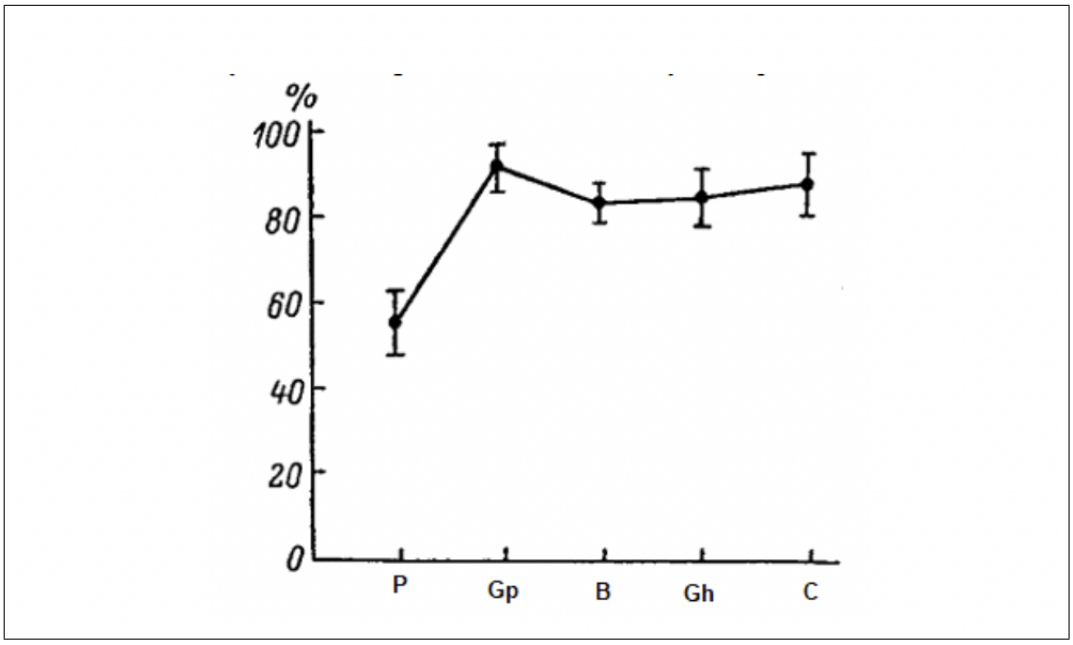

Figure 8. Accuracy of expression of an emotional context in singing depending on nature of emotions

Horizontal line - categories of emotions: *J* - joy, *Gr* - grief, *Ap* – apathy, *Ang* - anger, *F* - fear; *Vertical line* - number of the correct estimates (%). The average data for a group of singers (13 people). Vertical scales indicate credible intervals for the level of significance 0.05

2.2.4 *Perception of an emotional context of the vocal speech by listeners of different categories*

The second possibility of the method of emotional and semantic divergence is assessment of emotional perceptivity/susceptibility of listeners. In this series of works the researchers [72] used two most successfully performed programs. One of them was sung by the singer of the Leningrad Philharmonic Hall Kiselyov A. (baritone), the second – by a student of the Leningrad Conservatory Barysheva L. (soprano). Both programs included a phrase *Sleep, my child* (from Tchaikovsky's P.I. "Lullaby").

155 people representing the following categories of listeners were involved in listening and assessment of these programs:

1 - Pupils of the 1st grade of the comprehensive school (7 years old, 28 people);

2 - Pupils of the 3rd grade of the comprehensive school (9 years old, 31 people);

3 - Adults who don't dedicate themselves to practice music (from 20 to 35 years old, 24 people);

4 - Pupils of the 5th grade of the comprehensive school (11 years old, 33 people);

5 - Pupils of the 1st grade of children's music school (7-8 years old, 15 people);

6 – A voice band "Tonica" (19-20 years old, 5 people);

7 - Students of the vocal faculty of the Conservatory (from 20 to 28 years old, 19 people).

Special attention was paid to see if these children especially of a younger school age correctly understand the emotional states expressed by singers based on the programs. The talk with them showed that the 7-year olds truly comprehend the content of all definitions of emotions offered to them. Answers of each category of listeners were statistically processed and the results are presented in figure 9. An order of representation of the listeners' categories in the graph is similar to the above-mentioned.

The analysis of the listeners' answers shows that the ability to correctly perceive emotional contexts of the vocal speech improves with age but at the same time there are big specific differences between groups. The statistical processing shows that this ability is explained by *age* and people's *level of developed vocal abilities and aptitude for music.*

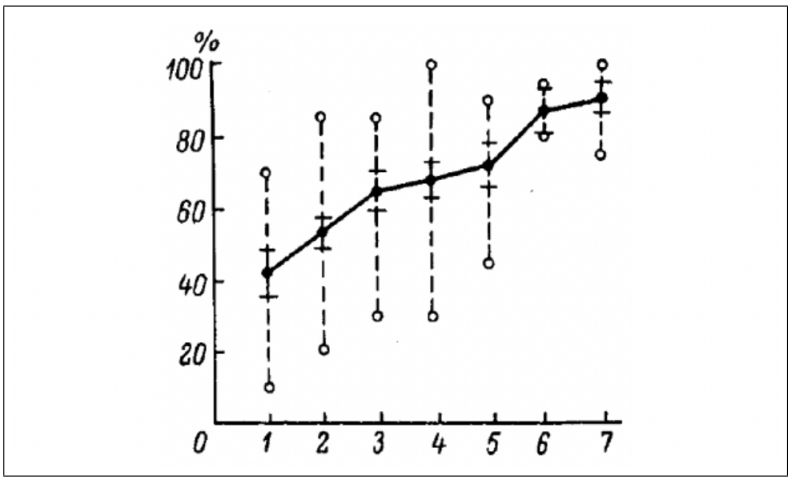

Figure 9. Accuracy of perception of an emotional context of the vocal speech by listeners of different categories

Horizontal line - categories of listeners (see above the order of their representation); vertical line - number of the correct estimates in (%). Dotted vertical lines designate the range of the maximum changes in answers of the participants in the research (individual maximums/limits); horizontal hyphens marked credible intervals for the level of significance 0.05.

In figure 10 the data are provided on distribution of the listeners' correct answers by "emotions". An order of representation of the listeners' categories is similar to figure 9.

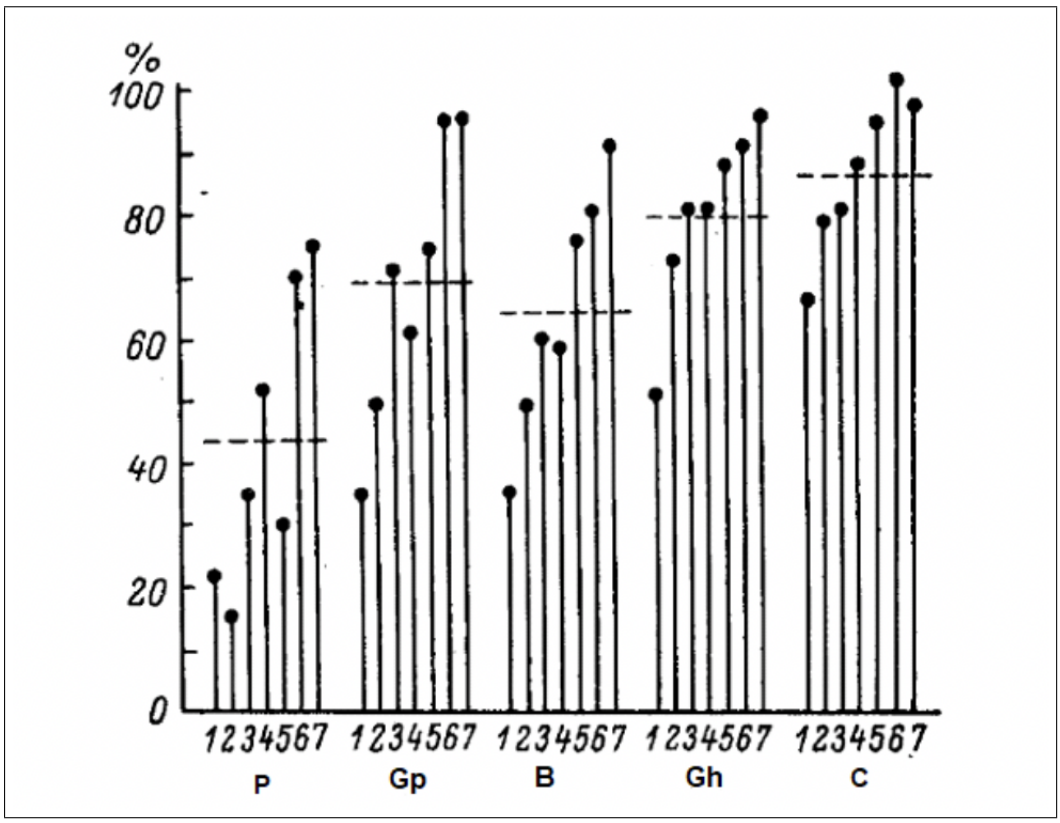

Figure 10. Identification of an emotional context of various characters by listeners of different categories

The same designations as in figures 8 and 9 The data provided in the graph demonstrate that all listeners' categories show a various degree of accuracy for perception of different "emotions". *Fear* (86%), *anger* (79%) is precisely perceived, *grief* (68%) slightly worse, indifference/ *apathy* (64%) less well expressed and badly *joy* (42%). It is important that this order of ranging for the numbers of "emotions" identification is observed practically in all groups of listeners.

2.3 A research on the acoustic parameters of the vocal speech preconditioning transfer of emotional information

The initial material for the acoustic analysis was received by scientists [72] from those phrases of the programs the emotional context of which was correctly estimated by experts with the probability not less than 70%. In this case it was considered that the singer really conveyed and the listener apprehended the conveyed emotional content which was defined by the task and therefore these phrases have acoustic signs which are typical for this emotion.

2.3.1 *Intonation characteristics and vibrato*

Frequency of the main tone in the work [72] was registered by means of the special electro-acoustic equipment: MIK-5 condenser microphone, tape recorder MEZ-28, the main voice frequency meter (intonation meter) and oscillograph chart K-115. 40 conditions were used for analysis of intonation characteristics. Typical oscillograph charts of a curve are presented in figure 11 to measure frequency of change in the main tone of a voice at various emotional contexts of vocal phrases. Periodic changes in the curve pitch on the oscillograph charts reflect a frequent vibrato of the singer's voice.

The below named emotions can be referred to the distinctive changes in the melody structure of vocal phrases. Intonation rising of a sound is observed to express *joy*, for *grief* – intonation- based "ups" and "downs" are typical. To express *apathy*, the melody seeks to flatten its pattern through lowering of the upper tones and increasing of the lower tones. For expression *of fear* the melody pattern looks the least stable while the accuracy of intoning is obviously broken.

2.3.2 *Integral characteristics of a range*

In the work of authors [72] the integral characteristics of the range were measured using the spectral integrator. As soon as duration of sounding of phrases with different emotional contexts varies the readings taken from the integrator were reduced to the unit of time (1 c). It gave a chance to compare the ranges of different phrases of various emotional contexts performed by different singers. 45 conditions of the 9 most successfully executed programs were analyzed.

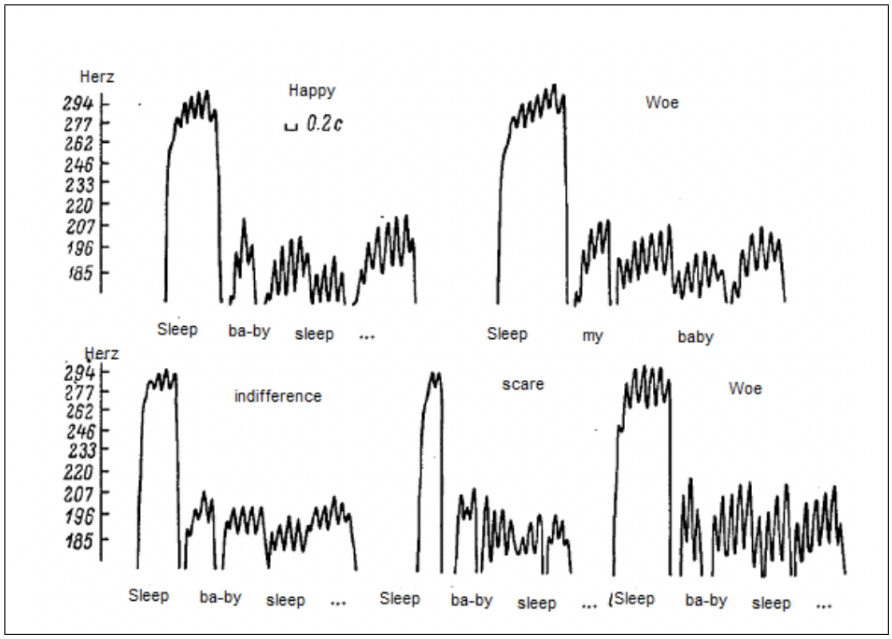

Figure 11. Oscillograph charts on frequency change of the main tone. A phrase "Sleep, my child" performed by baritone A. K-va

1 - Joy, 2 - grief, 3 - apathy, 4 - anger, 5 - fear. Vertical line - pitch of the main tone of a voice (Hz) The vibrato when expressing *grief* and especially *apathy* reduces its range. In some cases *apathyis* followed by almost total absence of the frequency vibrato, i.e. the sound becomes "direct". At joy and to a bigger extent while expressing *anger* the range of a vibrato considerably increases. Violation of periodicity of a vibrato curve is typical of fear. The least frequency of vibrato is observed for grief and *apathy* (5-6 Hz), for joy and *anger* it increases a little (6.0-6.5 Hz) and for expression *of fear* can abruptly increase (up to 8 Hz) which perhaps makes an impression of a "trembling" voice. Change of an emotional context causes also a change of phase ratios between an amplitude and frequency vibrato. This phenomenon substantially predetermines the character of a voice sounding and its timbre qualities [77] , [78].

The most typical spectrograms are provided in figure 12. First of all, the spectrograms show very noticeable changes in the high singing formant (HSF) by the level and frequency position. Changes of intensity and frequency position of HSF are more manifested during exercising vocalizations and singing separate vowels, i.e. when the singer is forced to resort to timbre characteristics of his/her voice as to the emotionally expressive means because the number of means for expression of emotions reduces.

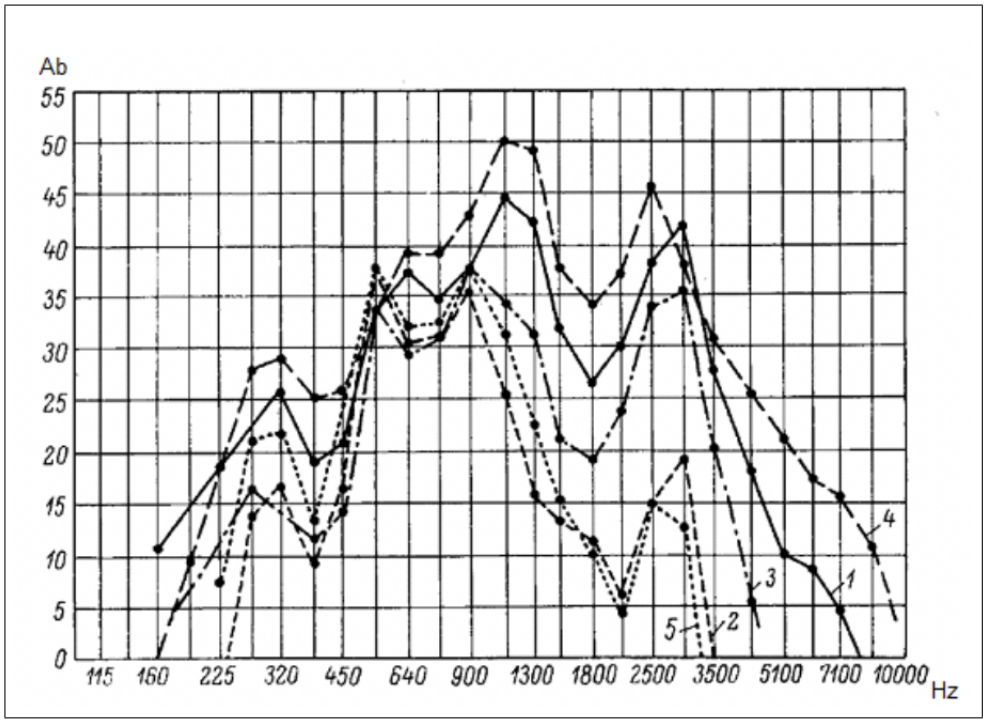

Figure 12. Change of the integral characteristics of a range depending on an emotional context

Vowel A - pitch is 311 Hz, tenor V. V-v. Horizontal line - average frequencies of a spectrometer filters (Hz); vertical line - level of spectral components (dB): *1 - joy, 2 - grief, 3 - apathy, 4 - anger, 5 - fear*

There is an interesting fact that in case of expressing joy the HSF top shifts to the highest frequency area. Aurally it is perceived as a mild and "light" sounding of a voice. While expressing grief and anger the HSF frequency position slides down. Apparently, it explains such a change in a timbre of voice as "darkening".

As for *anger*: It is not only that HSF occupies the lowest frequency area but a typical significant increase in the sound intensity is observed and in the result of it an increase in intensity of extremely high components. The level of a low singing formant increases as well. The sound becomes *more dense, massive* and *colored* with "metal tints".

At fear and grief the intensity of high frequencies in a sound goes down and expressiveness of HSF decreases. In extreme cases at expression *of fear* the spectrogram loses typical "vocal" features and by its outlines is almost similar to spectrograms of a speech signal. The voice completely loses its glare and sonority but gains a deaf and constrained voice character.

2.3.3 *Temporary and dynamic characteristics*

Temporary and dynamic characteristics were investigated by the authors [72] according to oscillograph charts of 137 vocal phrases and vocalisations which are selected from 16 programs for the method stated above.

On the oscillograph charts (the readings taken by oscillograph chart K-115), in addition to an audible signal by means of the linear amplitude detector the curve was registered of a change of the mean effective value of tension of the studied signal (amplitude envelope). For a more precise reproducing of a form of the amplitude envelope the frequency of the HF filter cut-off at the output of the detector was chosen at the level of 40 Hz.

Typical oscillograph charts of a phrase Sleep, my child, performed with various emotional colorings are presented in figure 13. We can see that temporary and dynamic characteristics of a phrase significantly change depending on an emotional context.

The following parameters were measured when processing oscillograph charts: 1) the average duration of a syllable in a phrase (T,ms); 2) a variation factor of duration of syllables in a phrase (VT, %); 3) the relative duration of pauses in a phrase in relation to the total duration of a phrase (t, %); 4) the average level of sound intensity of syllables in a phrase (P, dB); 5) a voice force variation factor by syllables (Vp, %); 6) the average duration of the front (increase in a sound) syllables (τ, ms); 7) average duration of going down (decrease of a sound) of syllables (τc, ms) [4].

[4]VT= (σT/T) • 100 %, where σT – standard deviation of syllable duration from average duration, – average syllable duration. VP=(σP/P) • 100 %, whereσP – standard deviation of syllable level from average level of sound strength in a phrase, – average level of syllable sound strength in a phrase

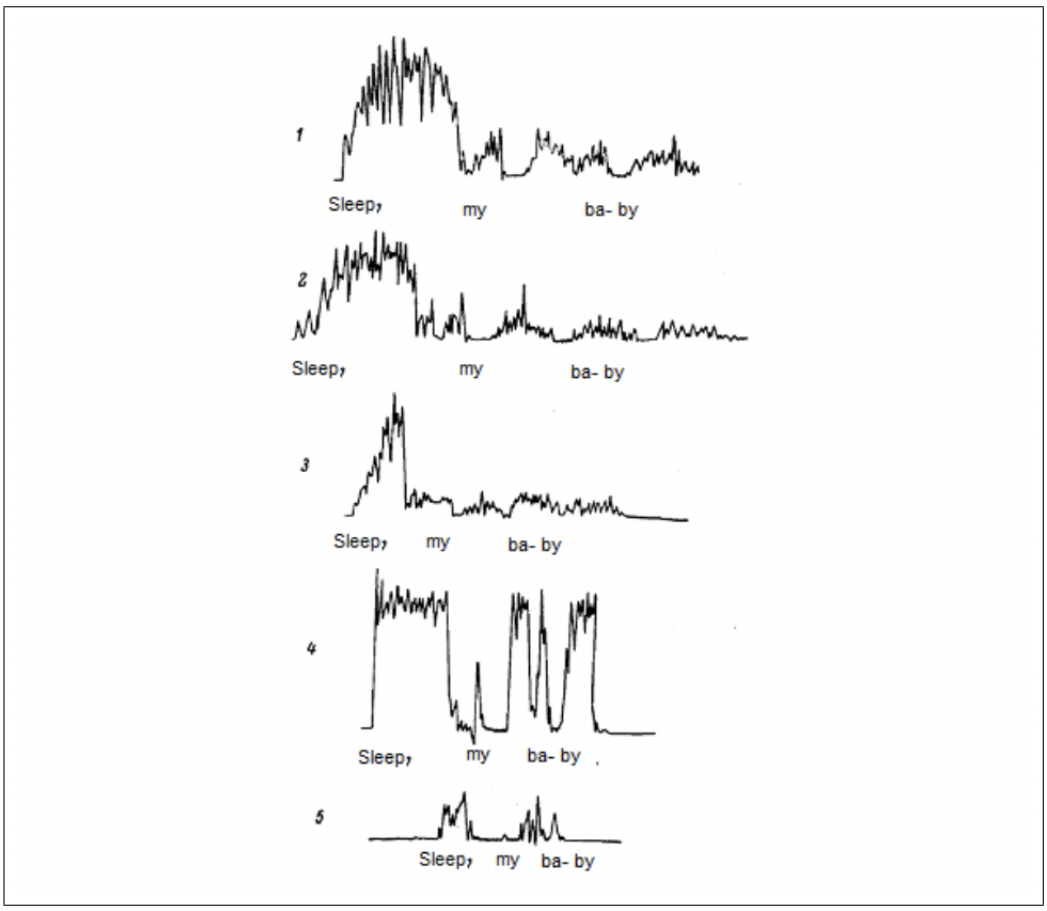

Figure 13. Oscillograph charts of a phrase Sleep, my child, performed by soprano L. B-va in a different emotional context

1 - Joy, 2 - grief, 3 - apathy, 4 - anger, 5 - fear. The oscillograph charts amplitude reflects changes in the force of the singer's voice

Statistical processing of the obtained data was done on PC Nairi-3. The received results are presented graphically in figures 120-125 (borders of credible intervals at the significance level P=0.05 are indicated in all graphs by vertical lines and scales).

The data are provided in figure 14 of a change in the average duration of a syllable depending on an emotional context of a phrase. The largest duration of a syllable is observed in phrases at expression of grief (1240 ms) and the least - at expression of fear (540 ms), i.e. the average rate of alternation of syllables at expression of grief is two times slower than of fear.

In figure 15 differences are shown in the relative duration of pauses in relation to the total duration of phrases. The largest size of pauses is observed at expression of fear (12.6%) and the least - at apathy (2.1%), i.e. at expression of fear the voice is interrupted while expressing apathy it sounds almost continuously.

Figure 16 provides the data about a change of the average force of a voice depending on an emotional context. As the figure demonstrates the strongest sound intensity is observed when the singer's voice expresses anger (100 dB) and the least when fear (92 dB) and apathy (93 dB) are expressed. As for the

sound intensity at expression of joy and grief, it takes anintermediate position.

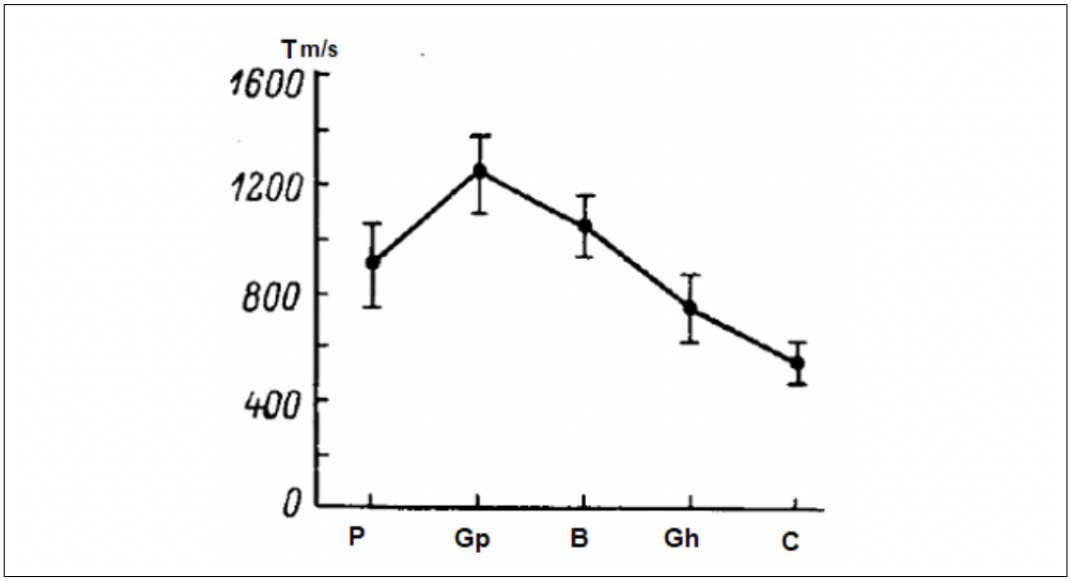

Figure 14. Change of the average duration of a syllable of a vocal phrase depending on an emotional context

Horizontal line - categories of emotions: J - joy, Gr - grief, Ap- apathy, Ang - anger, F - fear; vertical line - duration of a syllable (ms)

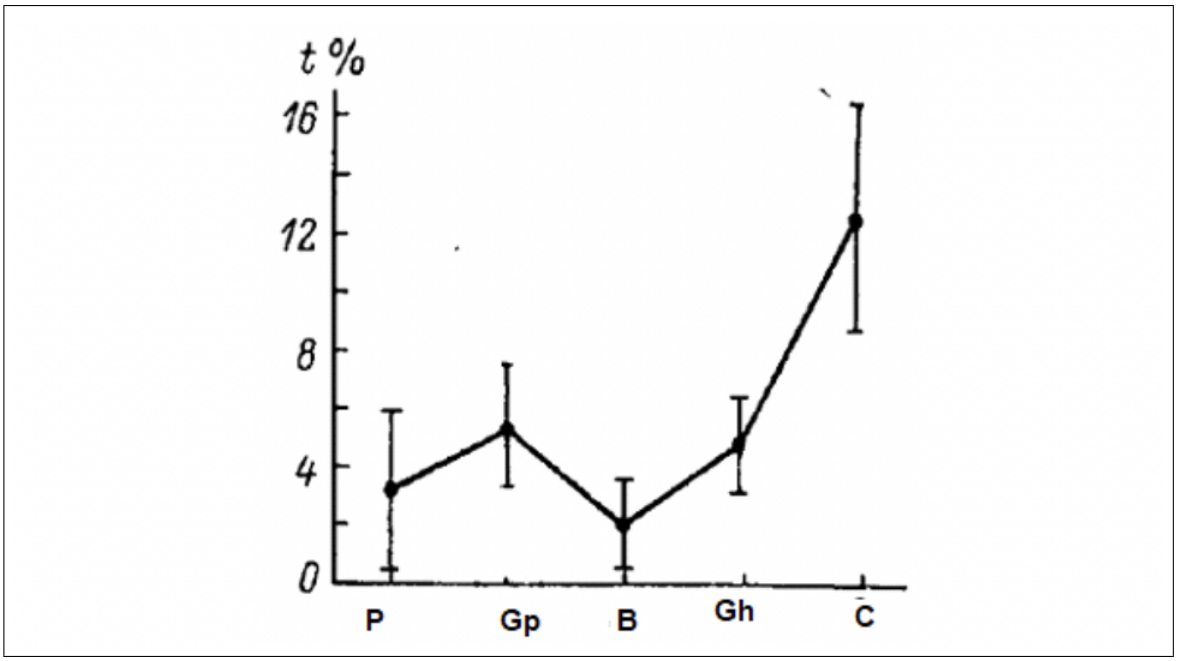

Figure 15. Change of the relative duration of pauses between syllables in phrases depending on emotional context

Vertical line - the relative duration of pauses (t, %). Same designations as in figure 8

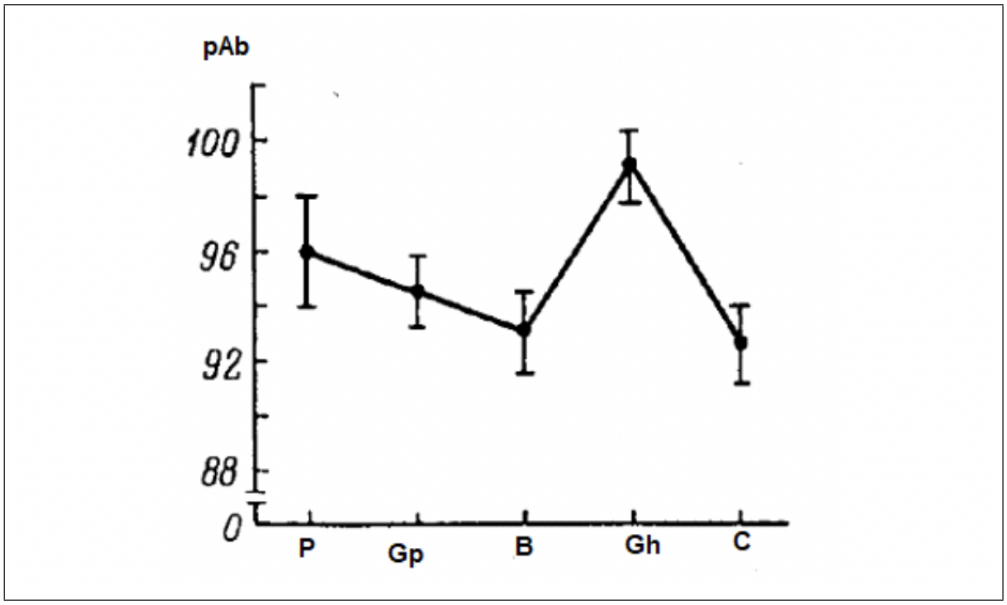

Figure 16. Change of a voice average force depending on an emotional phrase context

Vertical line- voice sound level (J, dB). Same designations as in figure 14

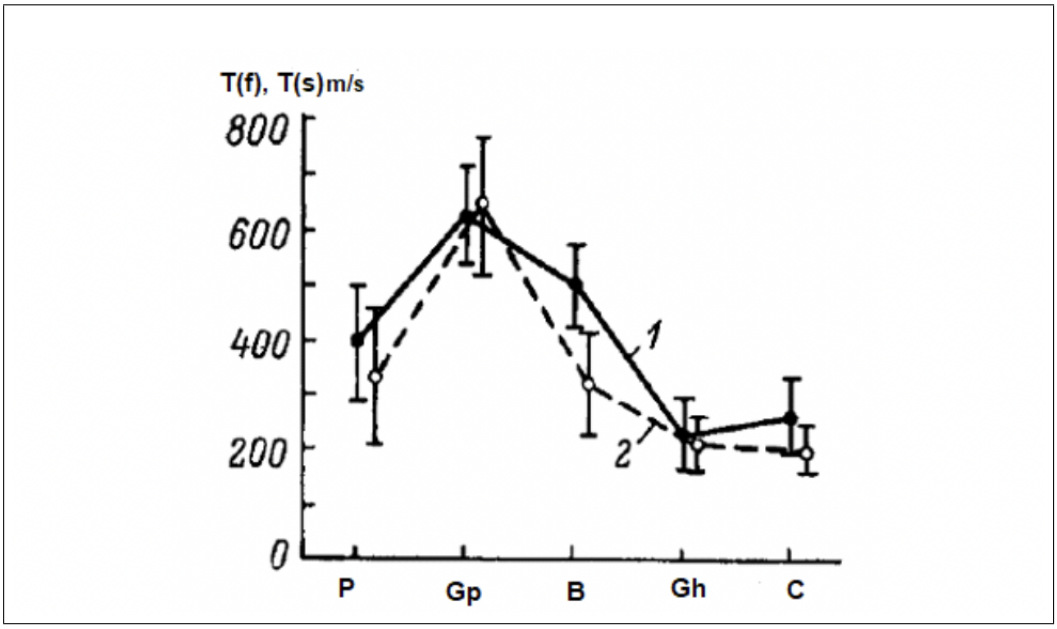

Figure 17. Change of a voice average force depending on an emotional phrase context

Vertically - duration (ms) of increase (τ) and recession (τ). Same designations as in figure 8

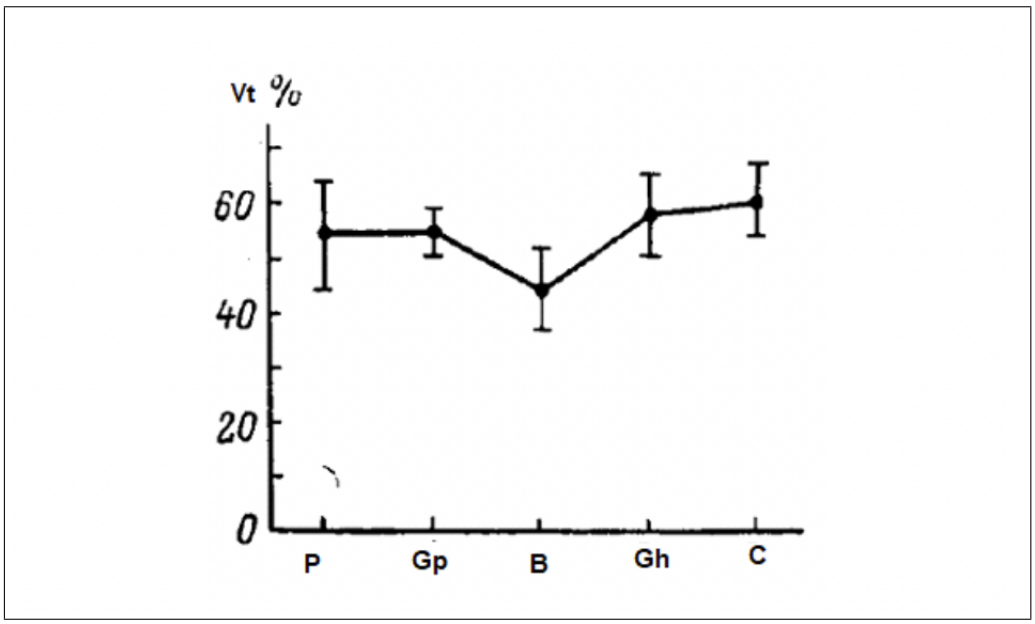

Figure 18. Change of a variation factor of syllables duration of vocal phrases at change of their emotional context

Vertical - value of a variation factor of syllables duration (VT, %). Same designations as in figure 8

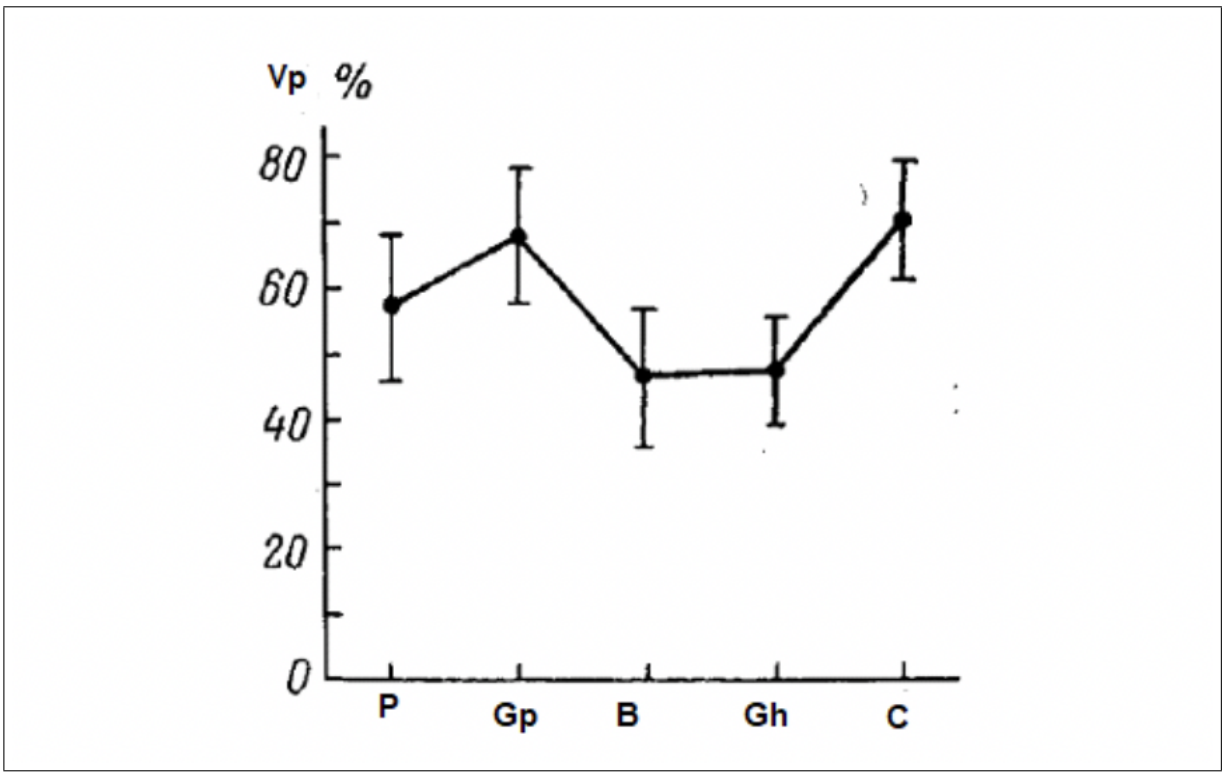

Figure 19. Change of a variation factor of syllables sound intensity of vocal phrases at change of their emotional context. Vertical line - value of a variation factor of syllables sound intensity (VP, %). Same designations as in figure 14

Two parameters are given in figure 17: the average duration of the front and the average duration of recession of a syllables' sound. Maximal duration of the front and recession of a syllables' sound fall on the phrases expressing grief (626

and 641 ms, respectively). Aurally it is perceived as a very soft attack and messa di voce. Minimum duration of fronts and recessions of a sound are observed at expression of anger (218 and 210 ms) which is subjectively estimated as a very abrupt beginning and ending of a sound.

The variation factor of syllables' duration indicates of a character of the rhythmic pattern of a melody. If the rhythm in a melody is interrupted/dotted and abrupt then a variation factor of syllables' duration is bigger; if the melody in relation to rhythm is quiet and weak, then a variation factor of syllables' duration is small; if syllables in a phrase have an absolutely identical duration then the factor equals to zero. Values of a variation factor of syllables depending on different emotional contexts are presented in figure 18. It turned out that at expression of apathy phrases have the least value of a variation factor of syllables' duration (44.5%). Phrases that have various emotional contexts significantly differ by a larger variation factor of syllables' duration (on average from 55 to 61%).

The variation factor of the syllables sound intensity is characterized by smoothness of the sounds managing in a phrase. If this indicator is big then it means that the sound by force is not even in syllables; if the variation factor decreases then it demonstrates that the sound of syllables became more even and smooth by its level. The data obtained by scientists [72] in figure 19 show that the variation factor of syllables sound intensity increases at expression of grief (67.5%) and fear (70.1%) and decreases at anger (46.7%) and apathy (46.1%). At joy this parameter is intermediate (56.4%).

Thus, the acoustic analysis of the vocal phrases, vocalizations and separate vowels performed with various emotional contexts showed that despite the instruction not to change the melodic, metro-rhythmic and other characteristics of the singing piece at expression of different emotional states the singers had to resort to these changes. We have the right to consider that all the listed acoustic changes in singing happen only due a change in the emotional content of phrases because the lexical and phonetic structure of a phrase remained unchanged at expression of different emotions.

2.4 *About an important role of temporary and dynamic characteristics of the vocal speech as means of coding of its emotional content*

The authors [72] made the totaling of the analyzed temporary and dynamic indicators adhering to the following principle: the informational content of an index was estimated by the sign "plus" (informative sign) if its size got an extreme value, i.e. maximal or minimum, and by sign "minus" (not informative

indicator) if the size of this sign among similar ones held any intermediate position. A similar assessment of signs in comparison with the size of the correct estimates of phrases by listeners is given in table 1.

We can see that emotions of fear (5), grief (4) and anger (4) have the greatest number of informative signs, less by apathy (3) while joy does not possess any of studied by the authors [72] acoustic informative signs of this type.

It should be noted that the number of informative signs which the phrase contains is in direct correlation with the number of the correct estimates of this phrase by the listeners. It suggests that the temporary and dynamic characteristics studied play an essential role in processes of perception and correct recognition and discernment by listeners of emotional context of the vocal speech.

Table 1. Assessment of informational content of dynamic characteristics of the vocal speech in terms of emotions transfer

Emotion	Acoustic signs							All sign	number of correct grades (%)
	T	V_T	t	P	V_P	τ_Φ	τ_C		
Happienest	–	–	–	–	–	–	–	0	55,3
Sadness	+	–	–	–	+	+	+	4	90,6
indifference	–	+	+	–	+	–	–	3	83,0
Woe	–	–	–	+	+	+	+	4	84,0
SCARE	+	+	+	+	+	–	–	5	86,6

For confirmation of this assumption the researchers [72] carried out an additional experiment based on the following technique. A group of listeners (11 people) was given an artificial signal - the tone modulated by an amplitude envelope of the vocal and speech signal (AM-tone), i.e. a signal which did not have such informative signs as intonation and spectral characteristics. The listeners had to determine a possible emotional content of the vocal phrases delivered to them in this transformed form. Table 2 illustrates the results of the listeners' assessment of 40 performances of a phrase Sleep, my child by four singers.

It is important to note that despite a low percentage of discernment of AM-signals which is quite natural, there exists a correlation in discernment of these signals with identification of undistorted phrases and also with a set of the informative signs summarized in table 1.

Thus, the received results show that temporary and dynamic characteristics fall into the number of very important means of emotional expressiveness of the vocal speech. The obtained data helps to answer the question why the emotion of joy is identified by all listeners more badly than any other emotional context

[70]. Not any typical temporary and dynamic signs were found for this emotion and it looks like its discernment is based only on spectral and intonation signs.

Table 2. Identification of an emotional context of vocal phrases before and after their acoustic transformation

Emotion	Number of right answers	
	voice	**AM**- tone
Happy	65	8
Woe	93	66
indifference	87	24
woe	97	60
SCARE	98	80

2.5 About psychophysiological bases of origin of acoustic means for expression of emotions by voice

The researches described in the book [72] revealed a variety of acoustic transmission means to convey emotional information. Practically any of acoustic parameters of a sound (force, frequency, a range, a vibrato, temporary parameters) can fit these purposes. At the same time it turned out that each of emotions has its own set of distinctive acoustic signs relevant only to it (see item 6).

The results received by the authors [72] give the grounds to express a view in favor of an evolutionary and historical approach to solution of the issue about the nature of acoustic means of emotional expressiveness of singing. In a general view this idea was stated by Darwin Ch. who repeatedly turned to human singing as to the striking example illustrating means of expression of human emotions and a regularity in their forming.

There are all grounds to believe that even in a such special form of sound communication between people as singing, the informational content, expressiveness and reliability of the acoustic code of this or that emotion were clearly defined apparently by a biological significance (in the course of evolution) of the emotional state expressed by voice. In this respect the biological significance of emotions of fear, anger and grief perhaps was bigger than of joy. Therefore, these emotions and especially fear were shown and were characterized by a large number of distinctive acoustic signs in comparison with the emotion of joy (table 1).

There are all grounds to relate the features of acoustic means for expression of emotions with special features of a physiological condition of the human body that experiences these emotions.So, for example, at expression of grief there happens a lowering of physical and mental activity including a depressed state. All this finds its reflection in a sound structure: decreasing of its force, increase in the fronts and recessions of a sound, "crying" intonations, etc. At anger expression, on the contrary, due to the general activation of mental and physiological activity and increase in a muscular tone the sound of a voice becomes strong and harsh (fronts and recessions are shortened), like "the ringing metal" due to a rise in intensity of high spectral components.

Finally, a strong increase in pauses between words notable to fear is apparently caused from the evolutionary point of view by a need not only to make signals in fear but also to perceive signals from a possible object that causes fear (a condition of awareness and listening).

Presence of the internal natural linkage between the character of a voice sound expressing this or that emotion and a physiological condition of the human body going through this or that emotional state is apparently a physiological basis for overall verbal intelligibility and kind of universality of the main means for expression of emotions by voice despite of a huge variety of both emotions and acoustic means for their expression.

2.6 Features of emotional hearing of the Chinese and Koreans

The psychological role of emotional expressivity lies in that it can strengthen considerably the sense of words of a person speaking or on the contrary can weaken up to the contrary effect, as for example, in a phrase "I am glad to see you!" said in the ironic or angry tone.

This is an important and practical value of ability for recognition by a communicant of the emotional expressivity of speech and in particular of the Russian speech by foreigners.

The results of the experimental research on ability of the Chinese and Koreans to recognize and discern emotional expressivity of the Russian speech are given in this work [79].

*Note.*Due to the fact that the term "emotional hearing" means in fact ability to discern emotional expressivity of speech (ES) the study of ES was conducted in 2010 and 2011. In 2010 two groups of listeners were examined [79]:

1. A group of the Russian students-vocalists (15 people)

2. A group of the Chinese students (15 people).

The studies were conducted with application of a speech kind of test of Morozov V.P. for emotional hearing which represented a series of the emotionally colored phrases of the Russian speech received by a method of actor's modeling of emotional expressivity: 30 phrases in which emotions of "joy", "grief", "anger", "fear" and also "neutrally" are repeated 6 times, each emotion randomly.

The psychometric analysis showed that this emotional hearing test has rather a high differentiating ability, a satisfactory test-retest reliability and validity [79]. The listeners were given a task to estimate the character of emotions of each phrase and to write down the answers in the listener's special form.

The ability to assess emotional expressivity was estimated by a number of correctly defined emotions in percentage to all listened emotionally colored phrases that characterizes the degree of emotional hearing ability developed.

The results showed that the emotional hearing (EH) of the Chinese auditors/listeners in the group in general makes 63.8% or for 8.9% lower in comparison with the Russian listeners whose ES totaled to 72.7%. Histograms in figure 20 show that ES of the Chinese group has a maximum in the area of the average ES (60-69%) and a maximum of ES of the Russians in the area of good ES (70-79%). The difference is statistically reliable (p =0.0027).

The analysis of ES based on the categories of different emotions (figure 21) shows that the Chinese perception of anger is considerably lower (43.3%), in comparison with the Russians (67.8%), at p =0.0000 and also emotions of joy (p =0.0323) and neutral intonation (p =0.0081).

Degree of distinctions and criteria of reliability are provided in table 3.

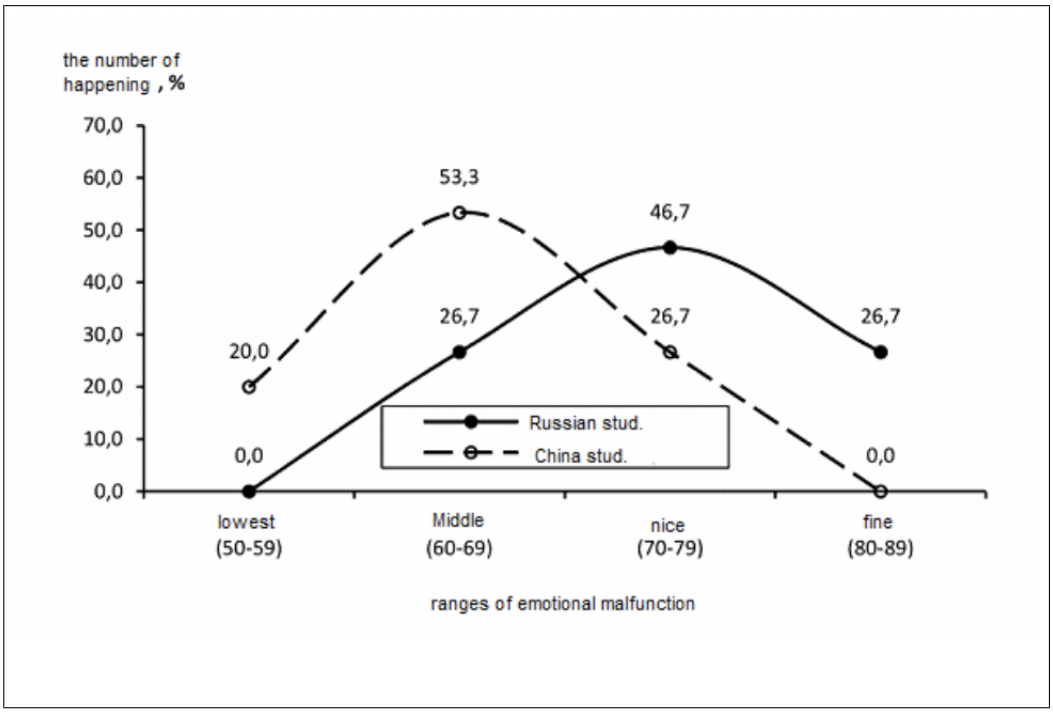

Figure 20. Comparative histograms of the groups' distribution of the Chinese and Russian students based on the ES

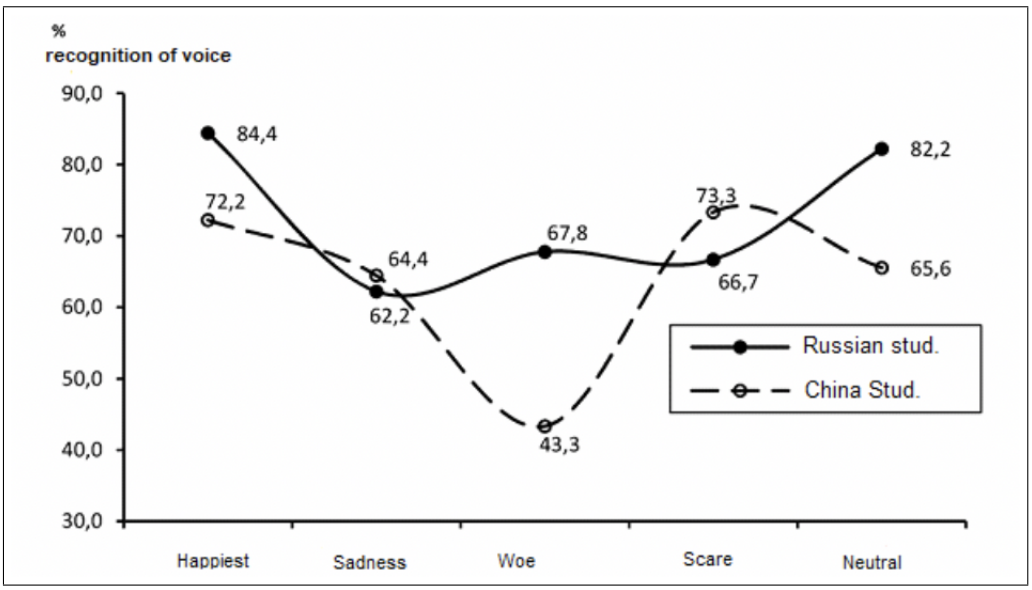

Figure 21. Discernment of emotions by the Russian and Chinese students

Table 3. Differences between the Russian and Chinese students

	ES	Happiest	sadness	woe	scare	Neutral
Russian student	**72,7**	84,4	62,2	67,8	66,7	82,2
st.div.	**7,5**	17,2	14,7	13,3	14,1	14,7
China student	**63,8**	72,2	64,4	43,3	73,3	65,6
st.div.	**7,3**	12,1	22,6	13,8	16,4	17,2
P difference criterion	0,0027	0,0323	0,7520	0,0000	0,2428	0,0081

The authors [79] examined three groups of students of the Moscow Conservatory in 2011:

1. A group of the Russian students-vocalists (22 people).
2. A group of the Korean students-instrumentalists who do not speak the Russian language (12 people).
3. A group of the Chinese students-instrumentalists who do not speak the Russian language (8 people).

The ability to discern emotional expressivity of speech was studied by the authors [79] similarly to the above described method using tests on emotional hearing. In view of the fact that the Korean and Chinese students did not know the Russian language the instruction to them was translated in the Korean and

Chinese languages. A follow-up task was explained by translators from among the senior year students.

Results. The basic statistical data processing of 2011 showed:

- ES of the Korean students made 66.1% (at min=53.3 of %, max = 80.0% and St. Dev. =8.4);
- ES of the Chinese students made 65.2% (at min=56.7 of %, max = 76.7% and St. Dev. =7.3);
- ES of the Russian students made 75.6% (at min=53.3 of %, max = 93.3% and St. Dev. =11.2).

The results showed that ES of the Korean and Chinese students of 2011 on average is lower than of the Russian students for 9.5% and for 10.4% accordingly (distinctions are reliable at probability of zero-hypothesis at p =0.0182 and p =0.0159, respectively).

Comparative histograms of the Korean, Chinese and Russian-speaking groups have maximums in the ES ranges: average ES (60-69%), average and good ES (60-79%) and excellent ES (80-89%), respectively (figure 22).

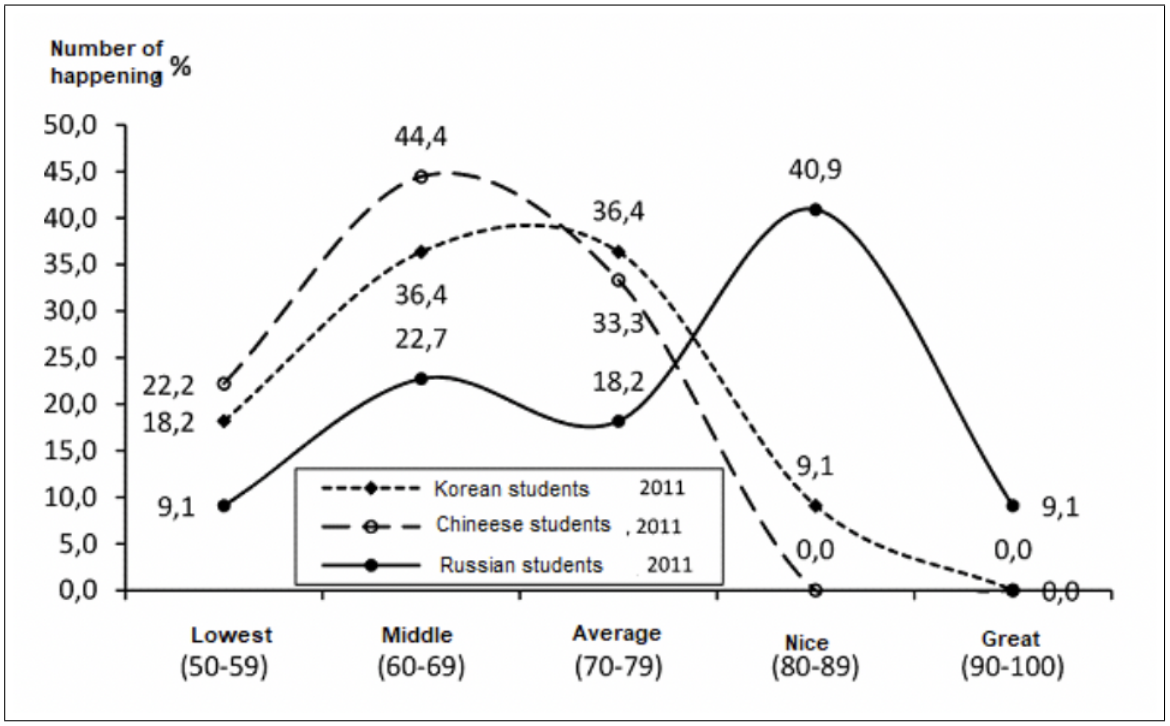

Figure 22. Comparative histograms of distribution by ES groups of the Chinese, Korean and Russian students

The analysis of discernment of various emotions (joy, grief, anger, fear, neutrality) showed that both the Koreans and the Chinese students' discernment of anger is lower (ESang=53.7 and 31.8%, respectively), in comparison with the Russians (65.2%) (Figure 23).

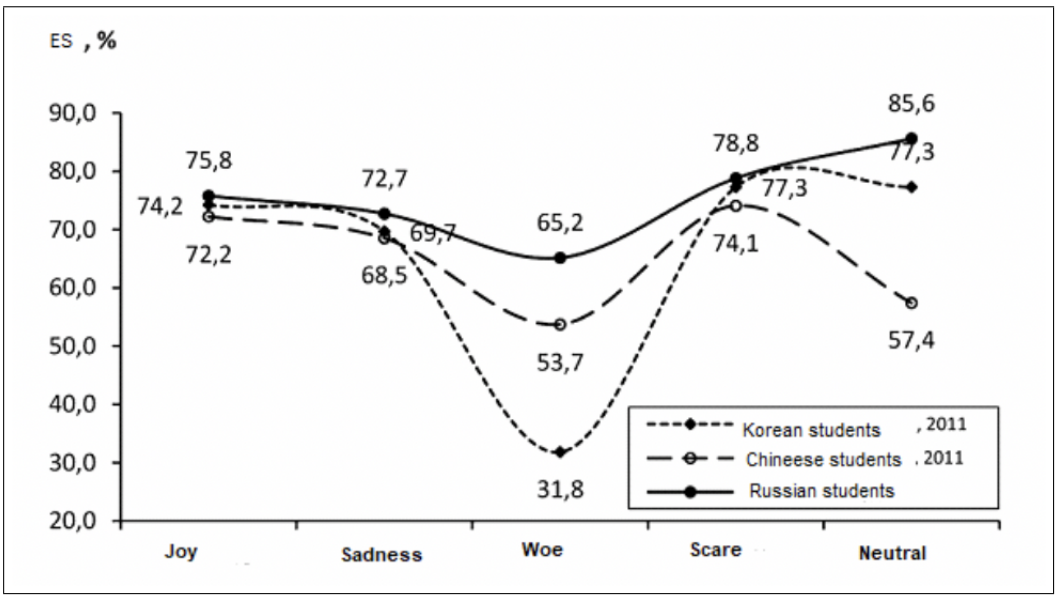

Figure 23. Discernment of emotions by the Chinese, Korean and Russian students (2011)

The authors [79] found out that the reasons of satisfactory ability of the Chinese and Koreans to discern emotional expressivity of the Russian speech is related to first of all universality of the emotional acoustic code of emotions dictated by a psychophysiological state of the speaking person which is peculiar to all people irrespective of their nationality and language competence [70]. And secondly because the Chinese language belongs to the so-called *tone language group* which is characterized by sense differentiating (semantic) function of intonation at pronouncing of phonetic elements of the Chinese speech.

At the same time some essential differences in anger discernment between the Russian and foreign-language students are found: the Koreans and Chinese distinguish anger worse than Russians for 33.4 and 11.5%, accordingly (2011). The similar result was received about the Chinese group in 2010 [79].

The explanation of this phenomenon maybe in that the Chinese and Koreans hear anger in normal intonation of the Russian speech (according to teachers of the department of Russian language of MGK). Therefore, they consider as a norm or any other emotion the apparent to the Russian-speaking listeners signs of anger in the actor's voice (in the test for ES). The analysis of the correct and wrong discernments of anger proves it: the Koreans take for norm 18.5% of intonations of anger (in the test for ES and the Chinese - 21.4 (2010) and 21.2% (2011), in comparison with Russians (13.2%), see figure 24. However, this is only one of the possible assumptions but something different seems quite probable.

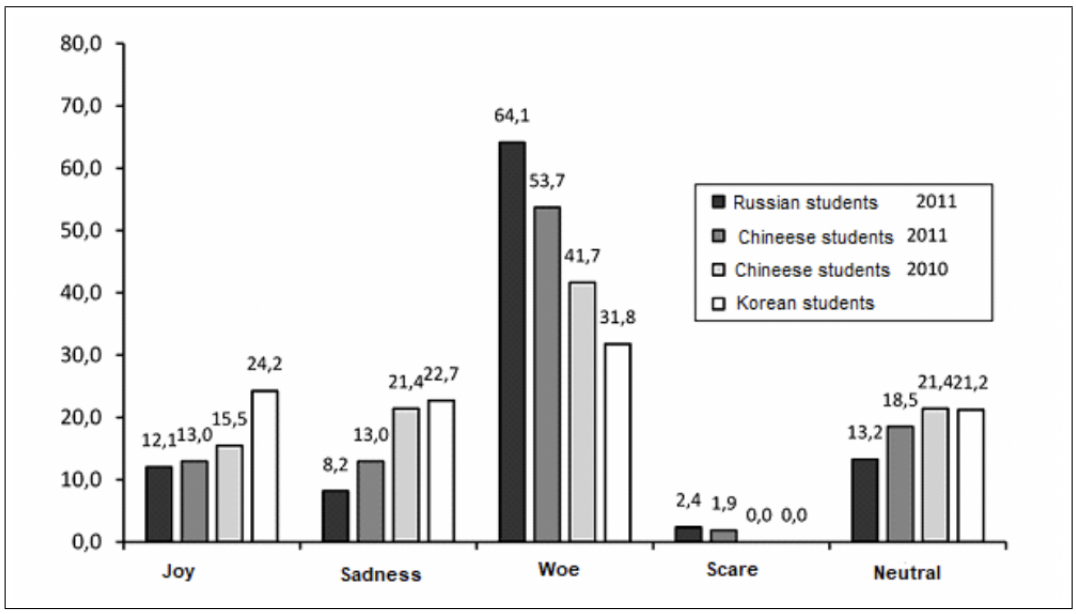

Figure 24. A ratio of the correct discernments of anger (in the center) and errors of taking the intonation of anger for other emotions and neutral emotion (on the right and left) by the Russian, Chinese and Korean students

3. Speech emotion recognition task

3.1 Speech emotion transformation to machine language

Sound is a continuous signal - a sound wave with changing amplitude and frequency. The larger the amplitude of the signal, the louder it is for a person.The higher the signal frequency is, the higher is the tone. The frequency of a sound wave is expressed by the number of oscillations per second and is measured in hertz (Hz, Hz).The human ear is able to perceive sounds in the range from 20 Hz to 20 kHz, which is called sound. Modern sound cards provide 16-, 32- or 64-bit audio encoding depths.When encoding audio information, a continuous signal is replaced by a discrete one, that is, it turns into a sequence of electrical pulses (binary zeros and ones). Figure 25 shows the sound range.

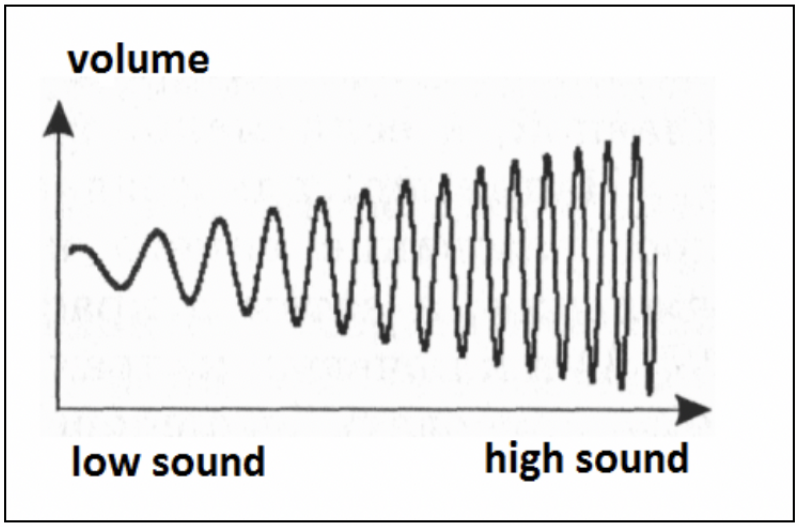

Figure 25. Sound range

The process of translating sound signals from a continuous form of representation to a discrete digital form is called digitization.A critical characteristic when encoding sound is the sampling rate - the number of measurements of signal levels per 1 second:

- (one) measurement per second corresponds to a frequency of 1 Hz;

- 1000 measurements per second correspond to a frequency of 1 kHz

A Sound sampling rate is the number of measurements of sound volume in one second.The number of measurements can lie in the range from 8 kHz to 48 kHz (from the broadcast frequency to a frequency corresponding to the sound quality of musical media).

The higher the frequency and depth of sound sampling, the better the sound of the digitized sound will be. The lowest quality of digitized sound, corresponding to the quality of telephone communication, is obtained at a sampling frequency of 8000 times per second, a sampling depth of 8 bits, and recording of one audio track (mono mode). The highest quality of digitized sound, corresponding to the quality of audio CDs, is achieved with a sampling frequency of 48,000 times per second, a sampling depth of 16 bits, and recording two audio tracks (stereo mode).

When encoding stereo sound, the sampling process is performed separately and independently for the left and right channels, which, accordingly, doubles the volume of the sound file compared to mono sound.

For example, we estimate the informational volume of a digital stereo sound file lasting 1 second with average sound quality (16 bits, 24000 measurements per second). For this, the coding depth must be multiplied by the number of measurements in 1 second and multiplied by 2 (stereo sound):

V = 16 bits * 24000 *2 = 768000 bits = 96000 bytes = 93.75 Kbytes.

Figure 26 presents converting an audio signal into a discrete signal: - an audio signal at the input of the ADC; b - discrete signal at the output of the ADC.

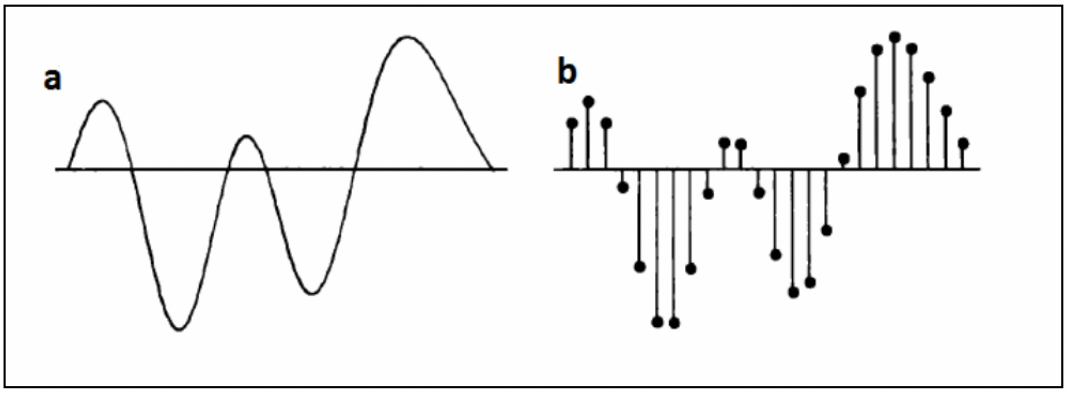

Figure 26. Converting an audio signal into a discrete signal

Figure 27 shows the conversion of a discrete signal into an audio signal: a - a discrete signal at the input of the DAC; b - an audio signal at the output of the DAC.

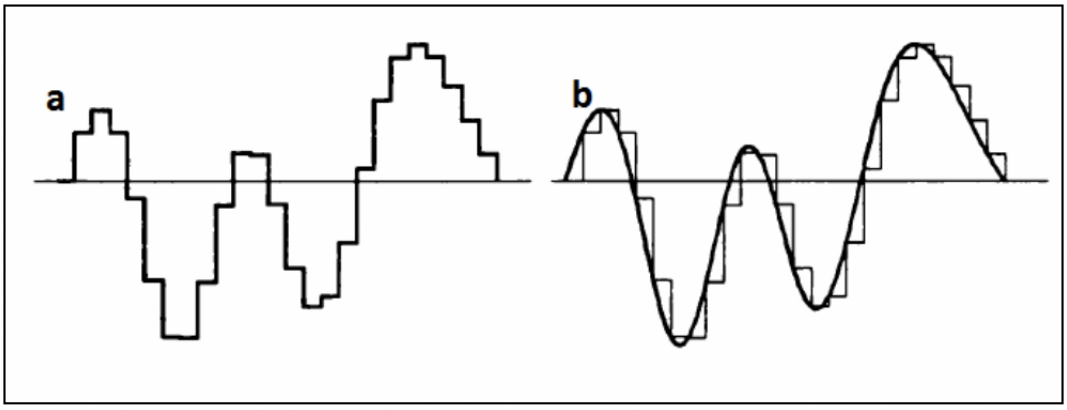

Figure 27. Conversion of a discrete signal into an audio signal

Fourier transform has become a powerful tool used in various scientific fields.The Fourier transform is the mathematical basis for spectral analysis, which relates a temporal or spatial signal to its representation in the frequency domain.Spectral analysis is one of the signal processing methods that allows us to characterize the frequency composition of the measured signal.

There are several spectral features for transformation human speech to machine representation. To extract features from speech, researchers usually use MFCC feature or just Spectrum feature. For example, based on the study [80] where was presented a comparative analysis of spectral features, I can conclude that MFCC is the most popular feature for SER. To prove that,there wasconducted thethe experiment, where I was using each spectral feature type to recognize three speech emotions.

For feature extraction,I have used LibROSA[81]python package for music and audio analysisLibrosa supports 14 different spectral features. Consider each feature in extraction.

1. Chroma_stft. Calculatea chromagram from a waveform or power spectrogram. This implementation is receivedfrom chromagram _ E[82].

2. Chroma _cqt. Constant-Q chromagram

3. Chroma _cens. Calculatesthe chroma variant "Chroma Energy Normalized"
   (CENS), following [83]

4. Melspectogram.

5. MFCC

6. RMSE. Compute root-mean-square (RMS) energy for each frame, either from the audio samples y or from a spectrogram S.S is spectrogram magnitude. Required if y is not input.Computing the energy from audio samples is faster as it does not require an STFT calculation. Using a

spectrogram will give a more preciserepresentation of energy over time because its frames can be windowed, thus prefer using S if it is already available.

7. Spectral _centroid. Calculatesthe spectral centroid, andeach frame of a magnitude spectrogram is normalized and considered as a spreading over frequency bins, from which the mean (centroid) is extracted per frame.

8. Spectral _bandwidth

9. Spectral _contrast. Compute spectral contrast[84]

10. Spectral _flatness. Calculate spectral flatness. Spectral flatness (or tonality coefficient) is a measure to quantify how much noise-like a sound is, as opposed to being tone-like [85]. A top spectral flatness (closer to 1.0) shows the spectrum is like white noise. It is often converted to a decibel.

11. Spectral _rolloff. Calculate roll-off frequency. The roll-off frequency is defined for each frame as the center frequency for a spectrogram bin such that at least roll _percent (0.85 by default) of the energy of the spectrum in this frame is contained in this bin and the bins below. It can be used to, e.g., approximate the maximum (or minimum) frequency by setting roll _percent to a value close to 1 (or 0).

12. Poly _features. Receive coefficients of fitting an nth-order polynomial to the columns of a spectrogram.

13. Tonnetz. Calculates the tonal centroid features (tonnetz), adhering to the methodof [86].

14. Zero-crossing rate. Calculate the zero-crossing rate of an audio time series.

3.1.1 *Dataset*

For the next experiment, the EmoDBdataset [24] was used. This preliminary experiment was conducted on a smallersubset of this corpus containing 271 labeled recordings with a total length of 783 seconds. Because of non-equality betweenclasses and in order to get comparable results with, was usedall sentences from all actors but only from 3 emotional states:angry (127 recordings, 334 seconds), neutral (79 recordings,186 seconds), sad (65 recordings, 263 seconds).

Ihave split dataset intro TRAINING 80,81% (219 files) VALIDATION 9.594% (26 files: 12-angry; neutral-8, sad - 6) and TESTING 9.594% (26 files: 12-angry; neutral-8, sad - 6). The TESTING set (used for testing) was taken from filesthat have not been used either in DNN training or validation.

3.1.2 *DNN architecture*

The chosen DNN architecture contains six fully connected layers with activation function relu[87], and the last layer also fully connected but with activation function softmax. The structure is as follows: 1 – 320 neurons, 2 – 160 neurons, 3- 80 neurons, 4 – 40 neurons, 5- 20 neurons, 6 – 10 neurons. The last layer with activation function softmax contains three neurons. For the regularization of the DNN, we used a 0.2 dropout [88]between the third and the fourth layers and batch normalization before the first layer. All layerswere initially assigned using Glorot uniform initialization [89].Detailed information about the architecture is shown in Figure 28.

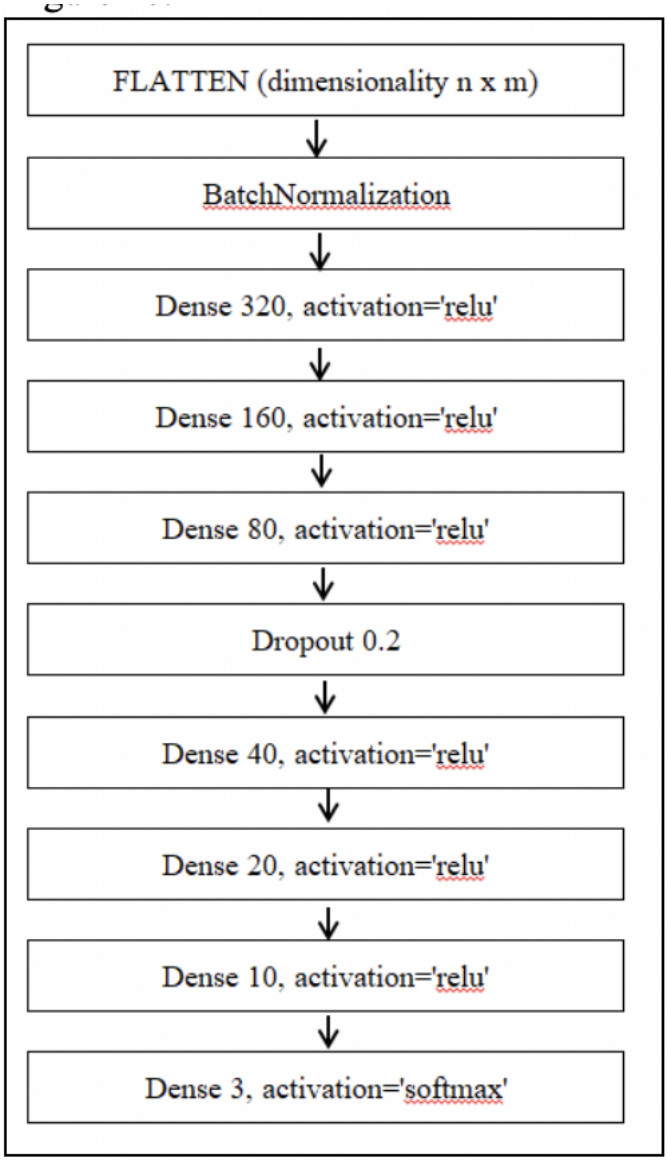

Figure 28. Detailed architecture of proposed DNN

3.1.3 *Experiment*

For training our proposed model, we utilized the StochasticGradient Descent algorithm with a fixed learning rate of 0.11 tooptimize a Binary cross-entropy

loss function, also known aslogloss[90].The metrics of the model are accuracy. The input data were prepared to the DNN in batches ofsize 16 in multiple epochs (iterations).

3.1.4 *Result*

For each feature, a model with a different number of epochs was trained.In this approach, we considered loss and accuracy in the context of 5,10,20,40 epochs. The results are displayed in Table 4.

Table 4. Comparative analysis of spectral features.

№	Feature	Dimensionality	Acc(%) epoch 5	Loss epoch 5	Acc(%) epoch 10	Loss epoch 10	Acc (%) epoch 20	Loss epoch 20	Acc (%) epoch 40	Loss epoch 40
1	Chroma_stft	12x320	57,69	0.99	34.61	2.33	61.53	1.716	61.53	1.716
2	Chroma_cqt	12x320	30.76	1.368	65.38	1.337	53.85	2.488	53.85	1.719
3	Chroma_cens	12x320	73.08	0.850	73.08	0.639	84.61	0.598	76.92	0.522
4	Melspectogram	128x320	80.77	1.18	76.92	2.72	80.77	1.37	84.61	1.768
5	MFCC	20x320	100	0.076	96.15	0.644	96.15	0.621	96.15	0.621
6	RMSE	1x320	69.23	1.047	76.92	0.932	76.92	0.932	76.92	1.404
7	Spectral_centroid	1x320	69.23	1.507	69.23	1.853	76.92	1.455	76.92	1.421
8	Spectral_bandwidth	1x320	50	1.566	53.84	1.101	53.84	2.359	65.38	2.076
9	Spectral_contrast	7x320	38.46	1.775	65.38	1.531	65.38	1.642	69.23	1.688
10	Spectral_flatness	1x320	61.53	1.270	73.07	0.707	76.92	0.729	73.07	0.731
11	Spectral_rolloff	1x320	61.53	0.95	73.07	1.204	73.07	1.08	69.23	1.859
12	Poly_features	2x320	76.92	1.323	57.69	1.864	73.07	1.62	69.23	1.791
13	Tonnetz	6x320	53.85	1.189	50	1.902	57.69	1.714	50	2.499
14	Zero-crossing rate	1x320	84.61	0.435	88.46	0.474	88.46	0.320	73.07	1.876

Acc – is the accuracy of Test set
Loss – is a loss function of the Test set

Table 1 presents the accuracy and losses of the Test set. According to the data of Table 1, features with accuracy greater than 80% are defined, such as Chroma_cens feature and Melspectogram feature achieved 84.61% accuracy on 20 epochs in training and 40 epochs in training respectively. The zero-crossing rate feature got the 88.46% accuracy on 10 and 20 epochs in training. MFCC reached the 100% accuracy on five epochs in training. The MFCC feature with a growing amount of epoch accuracy decreased to 96.66%. It means that training set with the MFCC feature learned faster than other features.

The purpose of this experiment was to study different types of spectral features and identify the most robust type of feature. This part of the dissertation has described the comparison of fourteen spectral features, as described in the

second part of the article. According to the experiment, it was proved that the MFCC feature is a reliable function for the task of recognizing speech emotions.

3.2 SER language dependency

The SER task is language-sensitive. The study [91] proves emotions recognition dependency. Authors have trained a KNN model on German EMO-DB [24] and have predicted the test set emotions on English, Malay, and Mandarin.As a result, the authors got accuracy of 78.5%, 71.0%, and 72.5%, respectively, when the accuracy of the Test set on German is higher than 95%. However, they used four emotions (sad, happiness, anger, neutral).Consequently it is possible to conclude that speech emotions in Kazakh and Russian will have confusion in recognition.There was considered the dependency of ER on the language, on the example of Russian and Kazakh due to the fact that I am a citizen of Kazakhstan where the official language is Kazakh and Russian is the language of international communication.For this case, there has been considering language dependency on Russian and Kazakh speech emotions recognition. The objective of the next experiment is to identify language dependency on Russian and Kazakh speech emotions recognition and prove the need to develop an automated method for collecting emotions data for each language.

3.2.1 *Dataset*

We used EMO-DB [24] for the training. There were split the dataset into TRAINING 80,7% (276 files) VALIDATION 9.64% (33 files: 12-angry; neutral-8, sad – 6, happiness - 7) and TESTING 9.64% (33 files: 12-angry; neutral-8, sad – 6, happiness - 7). Testing set used for experiment was taken from files that have not been seen by DNN during training or validation.

To reach the aim of the research, we collected the small emotion dataset (FOREIGN Dataset) in Kazakh and Russian languages to compare predictions with TESTING. Emotions were recorded on mobile phones recorder with a different number of samples and later down sampled do 16kHz (mono).

3.2.2 *Participants for Test set and Development set*

Four people participated in the compilation of the FOREIGN dataset, and the average age is 20, two men and two women. They recorded twelve sentences (ten of this in Russian language and a half in Kazakh language) on four emotions (angry; neutral, sad, happiness). All sentences are shown in Table 5.

As a result, we got 320 sentences in Kazakh language and 320 in Russian language. By analogy with TESTING, we selected randomly 33 files from

Russian language and 33 files for Kazakh language (12-angry; neutral-8, sad – 6, happiness –7).

Table 5. Kazakh and Russian sentences for FOREIGN dataset.

№	Kazakh language	Russian Language	Meaning in English
1	Оларғасенмұнықалайістеді н?	Как ты мог так поступить с ними?	How could you do this to them?
2	Олмағанқараптұрды, бірақбайқамады.	Он смотрел на меня в упор, и не замечал.	He stared at me but did not notice.
3	Бүгінмендеемтиханболады, сондықтанменіаландатпаңы з, мен дайындапотырмын	Сегодня у меня экзамен, поэтому не беспокойте меня, я готовлюсь	Today I am having an exam, so do not botherme, I am preparing.
4	Мен кешекешке хат жібердім	Я отправил письмо еще вчера вечером	I sent the letter last night
5	Олар оны жоғарыкөтеріп, ендіқайтадантүсіпкетті	Они просто подняли его наверх, и теперь они снова спускаются	They just lifted him upstairs, and now they are going downstairs again.
6	Демалыскүндері мен әрқашанүйгеоралып, Арсендікөрдім	В выходные дни я всегда возвращался домой и видел Арсена	At weekends I always came back home and saw Arsen
7	Оләрқашан оны сақтайтынжердеболады	Он будет в том месте, где мы всегда храним его.	He will be in the place where we always keep him.
8	Олсәрсенбікүнікеледі.	Она придет в среду.	She will come on Wednesday.
9	Төсекүстелтоңазытқыштаор наласқан.	Скатерть лежит на холодильнике.	The tablecloth is on the fridge.
10	Бүгінкешке мен оғанайтааламын.	Сегодня я мог сказать ему.	Today I could tell him.

3.2.3 *Feature extraction and DNN architecture*

For feature extraction, we used the LibROSA[81]python package and converted all to MFCC feature. Detailed DNN architecture is described in figure 4.

3.2.4 *Experiment*

As mentioned earlier, we created two additional test sets in Kazakh and Russian languages with the same number of files with emotions for comparison. An equal number of files in each emotion will allow comparing the accuracy and identifying the dependence of emotions on language.

3.2.5 *Result*

For each test set, a model with a different number of epochs was trained. In this approach, we considered accuracy in the context of 5,10,20,40 epochs. The results are displayed in Table 6.

Table 6. Comparative analysis of spectral features.

№	Feature	Accuracy(%) epoch 5	Accuracy (%) epoch 10	Accuracy(%) epoch 20	Accuracy(%) epoch 40
1	German_test	84.84	84.84	93.93	93.93
2	Kazkakh_test	27.27	24.24	30.30	30.30
3	Russian_test	39.39	36.36	39.39	39.39

Based on the results shown in table 1, there is a big difference in the accuracy of recognition of emotions in the Kazakh and Russian languages from German. However, Russian was predicted better at 9.39% than the Kazakh language. In Table 7,8,9 is shown a confusion matrix of each language.

Table 7. Confusion matrix of prediction emotion on the German language

German	anger	happiness	neutral	sad
anger	12	0	0	0
happiness	1	6	0	0
neutral	0	0	8	0
sad	0	1	0	5

Table 8. Confusion matrix of prediction emotion on Kazakh language

Kz	anger	happiness	neutral	sad
anger	5	1	0	6
happiness	5	0	0	2
neutral	4	0	0	4
sad	1	0	0	5

Table 9. Confusion matrix of prediction emotion on the Russian language

Ru	anger	happiness	neutral	sad
anger	7	0	0	5
happiness	2	2	0	3
neutral	4	0	0	4
sad	2	0	0	4

In the German TestSet there were confusions, once a happiness emotion with anger, and sad emotion with happiness. Emotions of happiness and neutral in Kazakh language have not once been recognized. Emotion happiness is most often confused with anger. Emotion anger is most often confused with sadness. Emotion of neutral in equal amounts was confused with anger and sadness. As a

result of the analysis, emotion of sadness is best recognized in Kazakh language. Other emotions are recognized poorly.

In Russian, emotions were recognized a little better than in Kazakh language. For example, emotions of anger were recognized two times better but also confused with the emotions of sadness. The emotion of sadness was recognized worse than in Kazakh but also was confused with anger. Emotions of happiness and neutrality are also recognized poorly as in Kazakh language.

The objective of the experiment was to identify the language dependency of Russian and Kazakh SER and to prove the need to develop an automated method for collecting emotions data for each language.The model was fully trained based on German EMO-DB [24].The experiment was shown that prediction of emotion is worse in Kazakh and Russian languages than in German. This conclusion is made based on table 1, where the difference in recognition accuracy is more than 50%. Confusion matrix shows that emotions of happiness and neutral recognized worst of all. As a result, Russian was predicted with accuracy 39.39% better than Kazakh language which was predicted with 30.3% accuracy.

It means that there is a reason to investigate the automated method for collecting emotions data for each language.

4. The architecture of the proposed method for automated collecting and labeling speech emotion data.

In this section we present the architecture for the proposed method of automated SER data mining. Essentially, the method follows a group of back to back objectives, starting with raw audiovisual data at the input and generating labeled emotional speech segments as the output. The objectives of the method, in order of requirement, are the following:

- Download audiovisual data from news, interviews, etc.
- Detect and parse speech segments from the downloaded video.
- Extract the audio file from the video.
- For each speech segment, recognize most common facial emotion from video.
- For each speech segment, recognize most common speech emotion from audio.
- Compare recognized emotional state obtained from FER and SER.
- If the recognize emotion matches between modalities, assign it to that speech segment

Considering how each individual objective presents a completely independent field of study, now follows a detailed description of each component or our architecture separately. As previously mentioned, the method must be automated and allow the researcher to only be required to set initial parameters and provide the system with raw data, in whichever preferred language, as input. The proposed architecture is depicted in Figure 29

4.1 Video parser

The first part of the method requires suitable audiovisual data to be found for processing. Naturally, YouTube was considered as the source of downloadable videos given its immensity. In order to search for and download proper videos, the python module available in [92] was used which provides several options in terms of video specifications. By means of this tool, the user needs only to create a list of keywords with a particular emotional theme in mind (e.g. interview, good news). From these, the tool proceeds to assemble an exemplary playlist of videos based on the provided keywords, which can then be downloaded upon request. The first mode is a tool for playlists search by keyword and gets the links with video to download.

1. The second mode is a tool for playlists search by multiple keywords and gets the links with video to download.
2. The third mode is a tool to get links with videos to download from the target playlist.
3. The fourth mode is a tool for links search with video to download by multiple keywords.
4. The fifth mode is a tool for links search with video to download by one keyword.

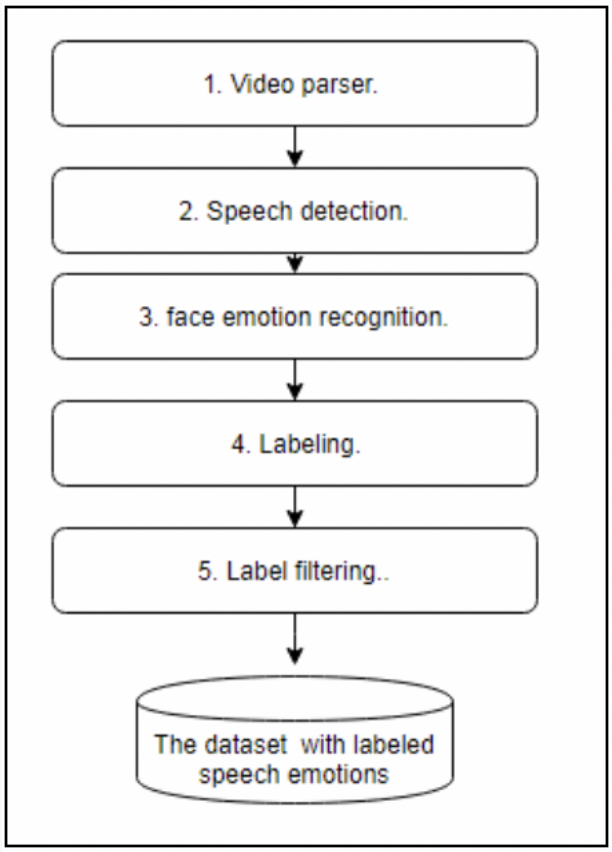

Figure 29. The architecture of the proposed method for collecting and labeling emotion data

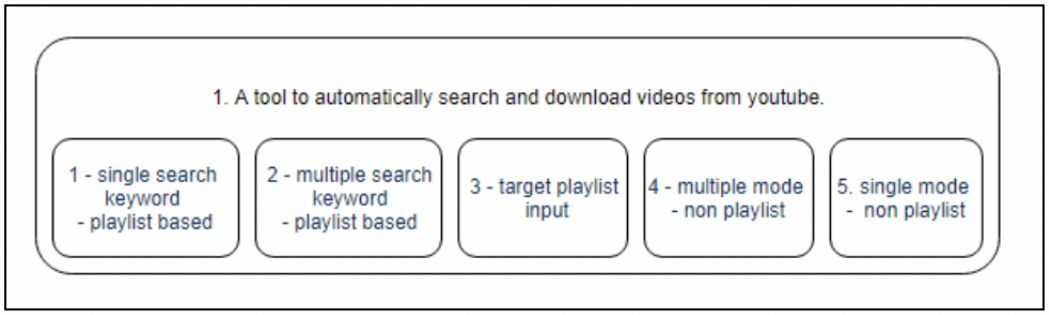

Figure 30. Options of video parser

4.2 Speech detection model based on a fully-connected DNN

To optimize the algorithm and improve the performance the speech detection model was designed and described in this section. Speech detection model is based on ML algorithms to extract parts of a video only with speech. In 5.3section of the dissertation was described the architecture in the details, as mentioned above.

4.2.1 *Related works*

On the Internet, there are quite interesting implementations of voice activity detection(VAD). The most famous VAD implementation is VAD WebRTC[93] by google. However, all of them are weak to recognize real speech. They work as a detector of silence and completely inoperative to detect speech. When we have tested some kind of video parts of news, they have detected noises and music as speech.

Many researchers had the same issue and we have found a few articles where [94]scientists give special focus on the detection of voice segments in music songs. The solution presented extracts of Mel-Frequency Cepstral Coefficients of the sound and uses a Hidden Markov Model to infer if the sound has a voice.They obtained 83.1% accuracy.

In this paper[95], researchers propose to apply object detection methods from the vision domain on the SR domain,by treating audio fragments as objects.Their system is composed of a CNN, with a simple least-mean squares loss function based on the YOLO algorithm[96]. They achieved 77.4% accuracy.

A computational framework for combining different features for emotional speech detection is described in the paper [97]. The statistical fusion is based on the estimation of local posteriori class probabilities and the overall decision employs weighting factors directly related to the duration of the individual speech segments. The best result is obtained by combining segmental and supra-segmental features using Gaussian Mixture Models classifier with a weighting

factor equals to 0.5 revealing a balance between two features; 87.5% accuracy.

This paper[98]shows an improved statistical test for voice activity detection in noise adverse environments. The experimental results showed a high speech/nonspeech discrimination accuracy over a wide range with accuracy on average 92.04%.

All the above results are not enough to quality speech detection for opportunities to move on a pool of objectives with minimal error propagation.Therefore, the goal of the paper is to develop a model to detect speech based on ML with an accuracy of more than 95%.

4.2.2 *Dataset*

To implement a speech detection task we have collected the data for Kazakh, Russian and non-speech audio files with varied sounds and noises.Detailed dataset composition is shown in figure 31.

Kazakh dataset contains three databases. First of them was amply described in the article[99]. The database contains ten sentences in Kazakh and each sentence was recorded by 101 actors in eight different pronunciations appropriate to eight emotions: anger, disgust, happiness, boredom, surprise, sadness, neutral, fear. Totally there were 8080 records in Kazakh. However, after converting all records to the necessary format, we have lost a lot of records and finally, we got 6792 records. To avoid the only above-mentioned sentence recognition we have decreased record sampling three times through random selection. Finally, we have included only 2264 records from the first database to the dateset.

The second Kazakh database has been collected in the base of Suleyman Demirel University for the SR task.The team, consisted of 35people, has given 360 sentences, which have been collected from the famous Kazakh books and news portal.Each person has recorded using Adobe Audition program the utterances and saved it with a corresponding transcription file. Since the size of the collected dataset is extremely low, we have applied some of the augmentation techniques to extend the size of the current audio data. Instead of increasing the size of the dataset using simple augmentation such as changing the pitch, changing the speed, we have applied the modern SpecAugmenttechnique. Finally, we have included the dataset with only 18 000 records from the second database.

The third one is a database that has been collected 12 000 different sentences on the base of Nazarbayev University[100]. However, in open access there are only 100 sentences. *on their website.

Finally, the Kazakh dataset consists of 20 364 audio files.

For the Russian dataset, we have used the Russian Open Speech To Text (STT/ASR) Dataset which is available in free access on github[101]. It is a representation of several free access databases in one source. We have chosen three databases, namely, public_youtube700_val, asr_calls_2_val, voxforge_ru. Those databases were chosen by us firstly because of high crisp speech approximately

99% and secondly many various speakers. The public_youtube700_val database has 7311 mp3 format files, asr_calls_2_val database has 12950 mp3 format files and voxforge_ru database has 8344 mp3 format files.

Additionally, we have an appended database described in the article[99]. The database contains ten sentences in Russian and each sentence was recorded by 101 actors in eight different pronunciations appropriate to eight emotions: anger, disgust, happiness, boredom, surprise, sadness, neutral, fear. Totally there are 8080 records in Russian. In the same way, after converting all records to the necessary format, we have lost many records and finally, we got 6866 records. To avoid the only above-mentioned sentence recognition we have decreased record sampling three times through random selection. We have included the dataset with only 2288 records from the first database. Finally the Russian dataset consists of 30 893 audio files.

For the non-speech dataset, we have used FSDKaggle2018[102] and FSDnoisy18k[103]. Freesound Dataset Kaggle 2018 (or FSDKaggle2018 for short) is an audio dataset containing 11,073 audio files.FSDnoisy18k is an audio dataset collected with the aim of fostering the investigation of label noise in sound event classification. FSDnoisy18k contains 18,532 audio clips in 42.5 hours of audio across 20 sound classes, including a small amount of manually-labeled data and a larger quantity of real-world noisy data.

Finally, the non-speech dataset consists of 29 605 audio files.

4.2.3 *Data preprocessing*

All audio files were converted with FFmpeg codec to wav format, pcm_s16le - 16-bit dual digital channel, mono, sample rate 16000Hz. To remove silence from audio files, all files were filtered in Voice activity detection (VAD) webRTC[93]from Google with mode 2. For extracting the speech from the audio file it is necessary to divide into chunks and after the ML model have to predict each chunk whether it is the speech or not. Consequently, each audio file was split into chunks of 1-second. For example, suppose that the length of the audio file is 2 seconds and 36 milliseconds. It means that we have divided the file into three chunks, the first chunk is from 0 to 1 second, the second chunk is from 1 second to 2 seconds and the last one is remainder from 2 seconds to 2:36 milliseconds. As a result, we got 65 127 chunks of Kazakh, 82 755 chunks of Russian and 154 809 chunks of non-speech. We have mixed Kazakh and Russian datasets to one “Speech” dataset and got in finally 147 882 chunks. In the end, two classes were divided into train set, development (dev) set and test set by proportion 80%, 10%, 10%. All preprocessing steps are shown in figure 32.

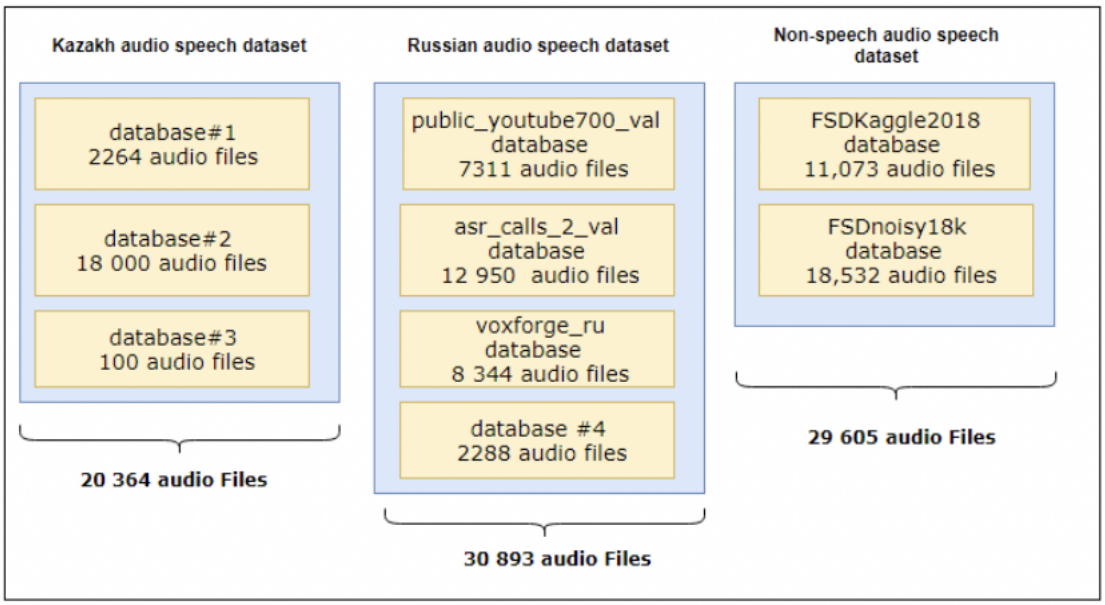

Figure 31. Data content

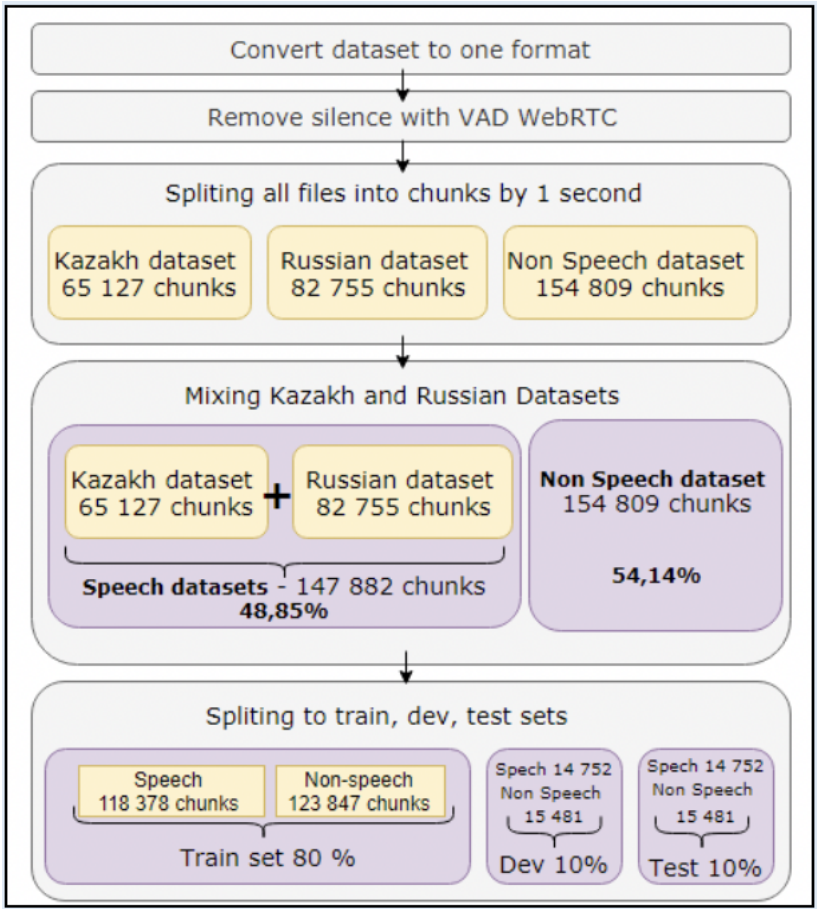

Figure 32. Data preprocessing

4.2.4 *Feature extraction and DNN model*

For feature extraction, we have used LibROSA[81] python package and extract MFCC features. The dimensionality of the MFCC features is 20x49.

The chosen DNN architecture contains four fully connected layers with activation function relu[87], and the last layer is also fully connected but with softmax activation function. The structure is as follows: 1 – 49 neurons, 2 – 25 neurons,3- 13 neurons, 4 – 5 neurons. The last layer with softmax activation function contains 2 neurons. For the regularization of the DNN, we used a 0.2 dropout [88] between the second and the third layers and batch normalization before the first layer. All layers were initialized using Glorot uniform initialization[89]. The detailed information about the architecture is shown in Figure33.

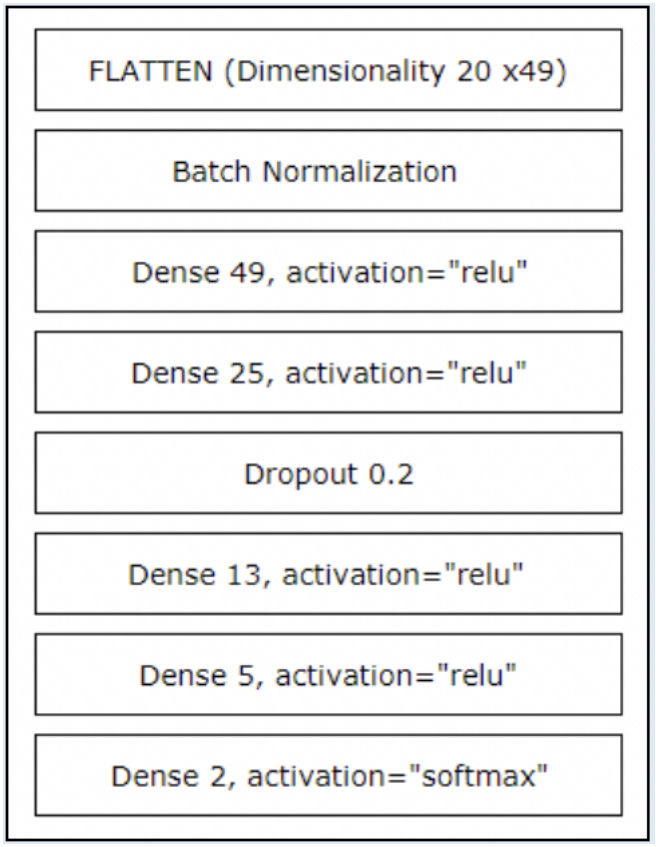

Figure 33. DNN model

For training the proposed model we utilized the Stochastic Gradient Descent algorithm with the fixed learning rate of0.2 to optimize a Binary cross-entropy loss function also known as logloss[90]. Momentum is 0.1. The metrics of model is the accuracy.The input data were presented to the DNN in batches of size 2048 in 30 epochs (iterations).

4.2.5 *Network learning and results*

After the model training, the accuracy of the development set is 97.11% and the accuracy of the test set is 97.31% accuracy. In table 10, there is shown a confusion matrix for the test set.

Table 10. Confusion Matrix (Test set)

	Non-speech	Speech
Non-speech	15 131	350
Speech	463	14 289

The recall, precision, and F1-score are used as evaluation metrics. Based on Table 118, the metrics in the binary classification case are computed as follows:

$$Recall = \frac{TP}{(TP + FN)}$$

$$Precision = \frac{TP}{(TP + FP)}$$

$$F_1 score = 2\frac{(PrecissionRecall)}{(Precission + Recall)}$$

Table 11. Recall, precision, and F1-score in the binary case.

	Predicted Class		
		(+)	(-)
Actual class	(+)	True Positives (TP)	False Negatives(FN)
	(-)	False Positives (FP)	True Negatives(TN)

For the evaluation performance of the model, based on the confusion matrix we have calculated the next metrics, namely, precision, recall, F1 score. Table 12 shows the values precision, recall, F1 score.

Table 12. Recall, Precision, F1 score (Test set)

	Recall	Precision	F1 score
Non-speech	0.9703	0.9774	0.9738
Speech	0.9761	0.9686	0.9723

The reported performance allows us to understand that recognition of Non-speech class better than a recognition of speech class.

We also have trained other classifiers as SVM, Logistic regression, Random forest, K-means, and Decision Tree. The table 13 shows the finally accuracy for training set.

Table 13. The accuracy of classifiers for training the training set

Classifier	Accuracy (%)
SVM	84.21
Logistic regression	70.38
Random Forest	78.35
K-means	68.92
Decision Tree	76.64

4.2.6 *Applying the model to real task*

To check the above results of the model and investigate real efficient of the recognition in natural tasks we have implemented an experiment. The idea of the experiment is trying to get speech from video in Kazakh[104] and one in Russian[105], and analyze the final speech results.

Firstly, we need to extract audio in wav format from mp4 video format. For realizing that objective we have used the "moviepy" library in python.

The next step is to convert to a single format and split the audio file into 1-second chunks. For splitting the audio into the chunks, we have used the "AudioSegment" from "pydub" library in python. After the splitting, we got 214 chunks with Russian news and 211 chunks with Kazakh news and all of them were put to the input to extract MFCC features.

The model of prediction allowed us to get probabilities of belonging seconds of audio file classes.

Having the information about the probability of belonging, we have divided and got two lists with serial numbers of seconds with speech and non-speech classes.Based on the above-mentioned F1 score metrics we have decided that recognition of non-speech class is better than the recognition of speech class. To extract the speech class we have used two different conditions. First, when the probability of non-speech class id more than 0.9 and second when speech class is more than 0.9. It is shown in Appendix A in the python code. As the last step is to compose the new video with speech from the initial video.

4.2.7 *Results*

As the result, we got two videos with speech and non-speech classes for each language news for two cases. It should be clarified that the sound quality in Russian news is more complex since there is a lot of noise in the background of the conversation, and the speech in the Kazakh news is cleaner. Tables 14 and 15show the competitive analysis of statistics with right recognized chunks for the case where the probability of non-speech class more is than 0.9 whereas tables 16 and 17illustrate the case where the probability of speech class is more than 0.9.

Table 14. Confusion matrix of Russian news probability, when non-speech>9.

	Non-speech	Speech
Non-speech	88	6
Speech	14	106

Table 15. Confusion matrix of Kazakh news probability,when non-speech>0.9.

	Non-speech	Speech
Non-speech	19	1
Speech	1	190

Table 16. Confusion matrix of Russian news, when probability speech>0.9.

	Non-speech	Speech
Non-speech	94	0
Speech	92	28

Table 17. Confusion matrix of Kazakh news, when probability speech>0.9.

	Non-speech	Speech
Non-speech	20	0
Speech	42	149

Consequently, we can see from the table that the case with the probability of speech class more than 0.9 works perfect. It means that the model extracts only clear speech without confusion.In general, the results are consolatory since the speech class consists of only really speech chunks and confused chunks, which appeared in non-speech class, have many background noises.

4.3 FER part

The speech audio extraction part is composed of six steps, which are illustrated in Figure 34.The first step is to extract the entire audio from the current video as a single WAV file. Secondly, the extracted audio is converted to a suitable format, such as PCMS16LE - 16 bit dual digital channel, mono and with 16kHz

sampling rate. Third, the converted audio is split into 1-second long chunks for later speech detection. As a fourth step, MFCC features are extracted from each chunk and into a list. Finally, using these features, speech detection is performed for each audio chunk using the model described in 5.2. Each audio chunk is stored with the corresponding initial timestamp for indexing, and label.

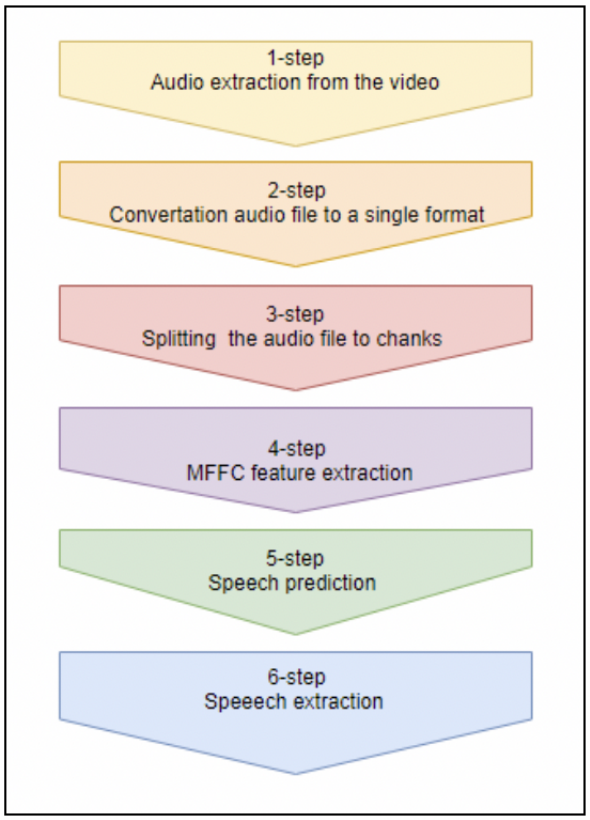

Figure 34. The second part of architecture

The next part of the automated method of collecting and labeling data is frame extraction from segments of video with speech and following FER. Before FER, we should detect the face on a frame. Figure 35 shows the detailed steps of that part of architecture.

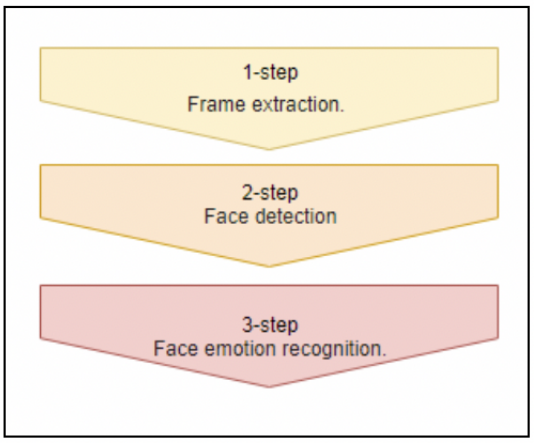

Figure 35. The sequence of the FER process

4.3.1 *Frame extraction*

Video is a quick alternation of pictures. For a person, there should be 24 pictures in one second so that the human mind perceives the ongoing picture smoothly and quickly, turning the interchangeable pictures into movement. The concept of the 25th frame has appeared relatively recently. Currently, the standard is considered 25 frames per second. Each frame has several properties like width and height. At the same time, for video, these values are measured in points. A certain number of dots in height and width set the format of the image. These points are called pixels. Figure 36 shows frame representation in a video clip with height and width.

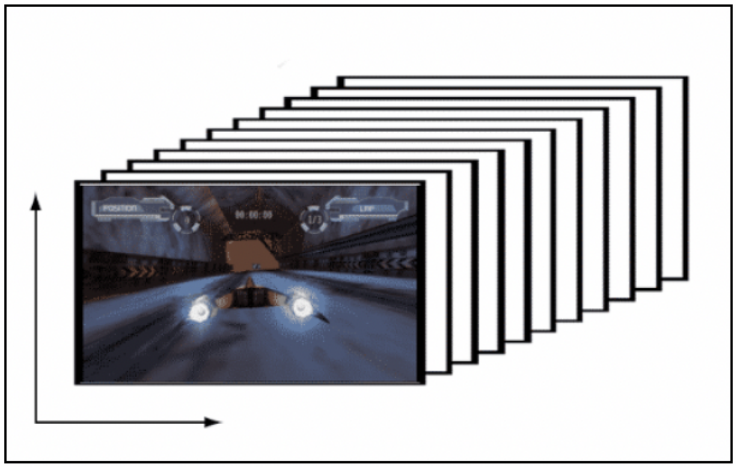

Figure 36. Frame representation in a video clip.

Enlarging any frame from the video,it is possible to see the individual pixels that make up the image frame. As example, figure 37 shows the pixel representation in a frame.

Figure 37. Pixel representation in a frame

Given how the technique must be generalized for all videos considered, even if their frame rate may be different, our technique used a fixed sampling rate of

ten frames per second across all videos. That being said, for each 1 second long segments of audiovisual data 10 frames were extracted.

After getting the list with speech seconds and extracting ten frames from each second, the next issue is detecting face in each frame.

4.3.2 *Face detection*

In order to more accurately perform FER, an initial step of facial detection is required. Many methods have been proposed over the years to address facial detection from static images, such as the Viola-Jones' method which has been applied extensively.In the Viola-Jones method, the basis is Haar primitives, which are a breakdown of a given rectangular region into sets of various rectangular subspaces, which are shown in Figure 38.

Object Detection task using Haar feature-based cascade classifiers is a useful object detection method proposed by Paul Viola and Michael Jones in their paper, "Rapid Object Detection using a Boosted Cascade of Simple Features" in 2001.

In the original version of the Viola-Jones algorithm, only primitives without rotations were used, and to calculate the value of the feature, the sum of the brightness of the pixels of one sub-region was subtracted from the sum of the brightness of the other sub-regions [106].In the development of the method, primitives with an inclination of 45 degrees and asymmetric configurations were proposed. Also, instead of calculating the usual difference, it was proposed to assign specific weights to each region and calculate the values of the attribute as a weighted sum of pixels of different types of regions [106].

This module was integrated into our technique through OpenCV and it was specified that only frames containing a single face would be considered for processing. This is critical given how if more than one face is present in an audiovisual segment the method may not be able to successfully identify which of them is speaking at each time. An example of this is given in Figure 40, a situation in which the proposed technique disregards the segment altogether.

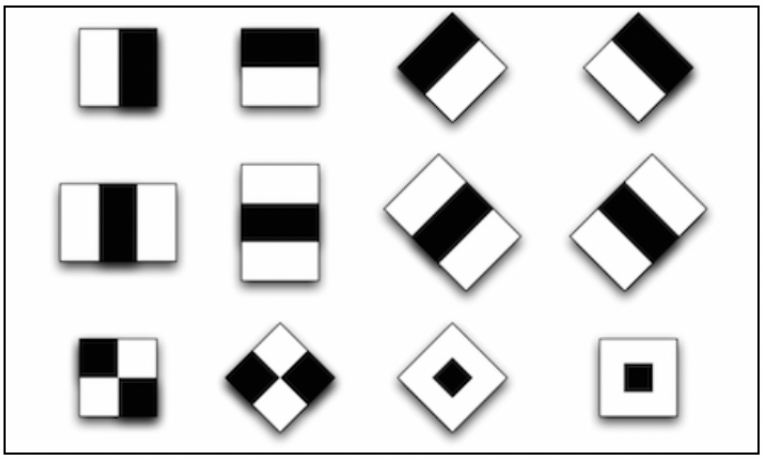

Figure 38. Heterogeneous rectangular subspaces

The main reason of using Haar primitives form as the basis of the method was an attempt to get away from the pixel representation while maintaining the speed of computing the attribute. It is difficult to derive any meaningful information for classification from the values of a pair of pixels, while, for example, the first cascade of a face recognition system that has a significant interpretation is constructed from two Haar signs [107]. Figure 39 shows the search for matches of a rectangular area with the criteria of a human face.More detailed information on how does it works is described in the paper[107].

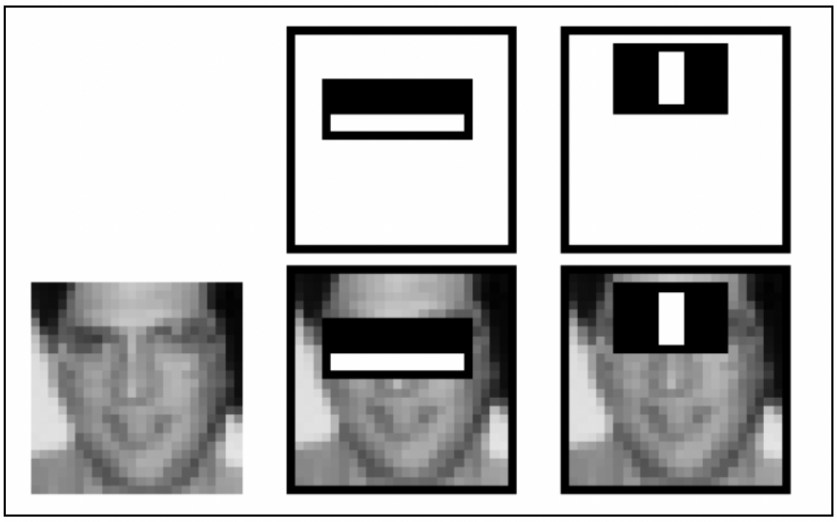

Figure 39. Search for matches of a rectangular area with the criteria of a human face

To implement face detection, the python library "open CV"has been used. Figure 40 shows the example of "open CV" face detection.

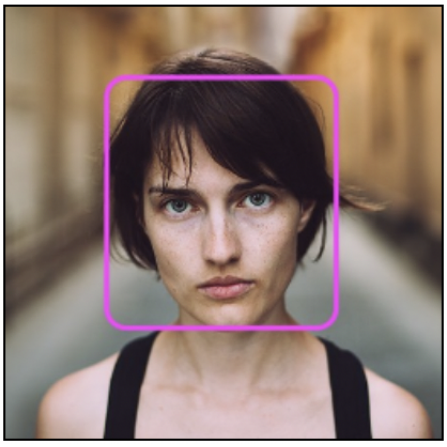

Figure 40. Face detection

There were set parameters to identifier only one face at one frame. This is critical given how if more than one face is present in an audiovisual segment the method may not be able to successfully identify which of them is speaking at each time. An example of this is given in Figure 41, a situation in which the proposed technique disregards the segment altogether.

Following face detection, segmentation is performed so as to remove the detected face's surrounding background and irrelevant information. The segmented face is then ready for FER, such as the example shown in Figure 42. Should no face be detected in a frame, then it is deleted for freeing up memory.

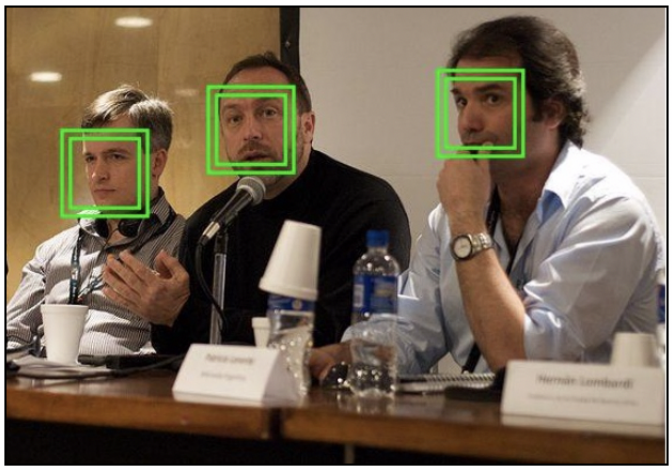

Figure 41. The case when was detected three faces

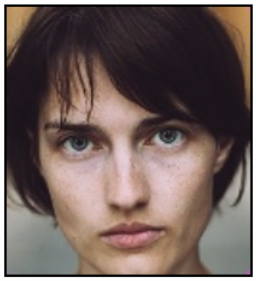

Figure 42. A clipped rectangle with the face

4.3.3 *FER*

The FER is the ML task. For FER issue, the already created model[108] based on CNN have been used. With the detailed, the architecture of the CNN model may familiarize at reference [108].It is necessary to mention that the model was trained on Kaggle competition dataset – FER2013 [33]. In finally, the accuracy of the model on the test set is 66%. In other words, the model correctly recognizes two of three emotions. In general, the accuracy of the model is sufficient considering that we have ten frames per second and the final decision on marking is made by the majority vote method. It is also worth considering that this task is one of the most difficult in the world to recognize face emotion by and the best result in the world is 72% accuracy[109].Figure 43 shows the example of working the recognizing face emotion model.

Continuing the sequence of actions of architecture, if a face was discovered, then clipped rectangle with the face are resized to 64x64 pixel according to the above-mentioned face emotion model conditions. After the prediction, the model returns the probability for seven emotion classes, namely:anger, disgust, scared, happiness, sadness, surprised, neutral.

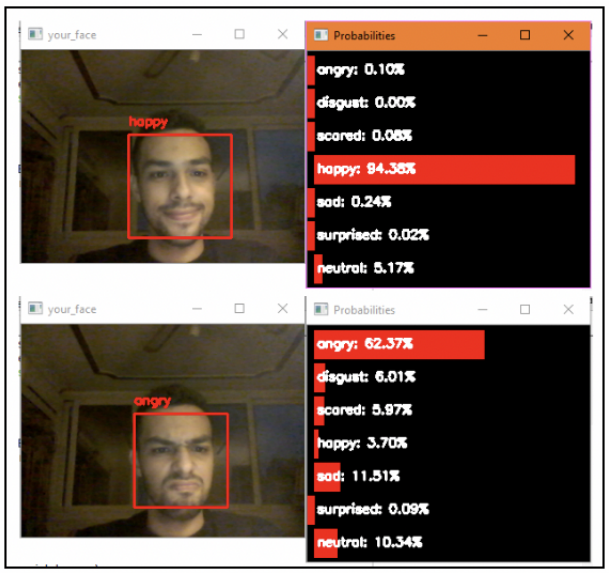

Figure 43. Example of working the recognizing face emotion model

4.4 Labeling

The main objective of the current part of the architecture is to define the video segment with speech and predicted emotion face for extracting an audio segment with an emotion label.There are three sequential steps are shown in figure 44.

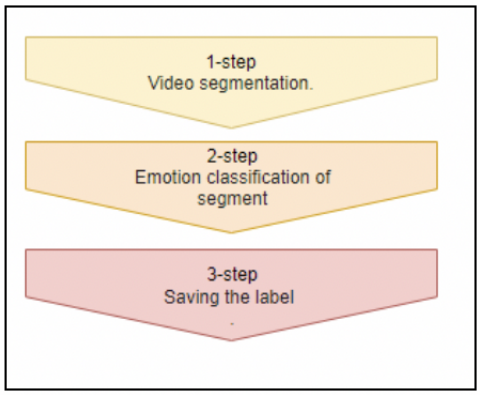

Figure 44. The sequence of a labeling process

4.4.1 *Video segmentation*

The video segmentation is an essential part of the method. That step is necessary to reduce error propagation and define the segment with speech and face inside the video. To define the segment I have used simple logistical operators.If the difference of the next millisecond from the current millisecond is more than ten frames, then the current millisecond as last and next millisecond as the first frame in the next segment was defined. Figure 45 shows the detection of the first and last frames in the segmentation process.

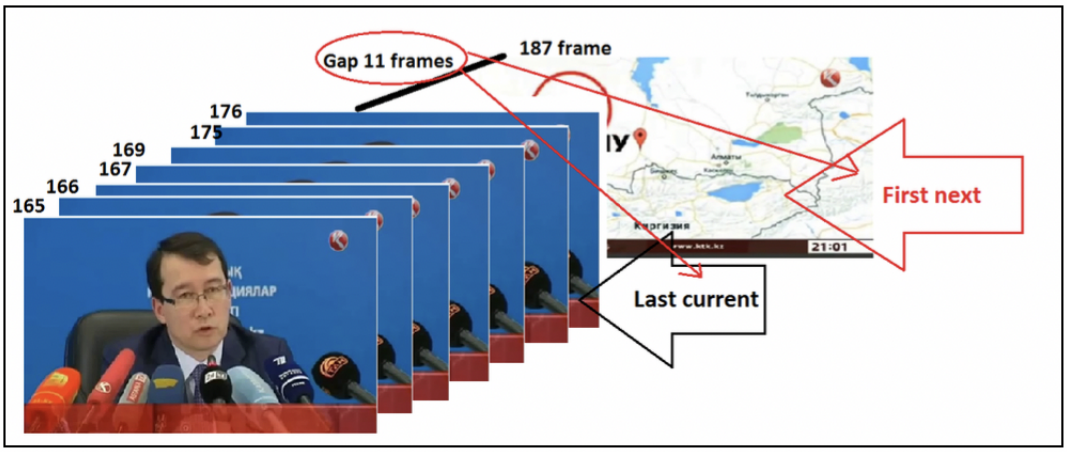

Figure 45. Detection of the first and last frame

4.4.2 *Emotion classification of a segment*

Each segment must be classified, but before classification, if the length of the segment more than three seconds, then we need to split the segment into sub-segments. It is necessary for cases when the duration of speech is long, and during a long speech, the human can have many speech emotions. For example, the duration of a segment is 5 second 300 milliseconds. The dividing of the segment will have two sub-segments: duration of first is three seconds, and the duration of the second is two seconds and 300 milliseconds. After the procedure of splitting segments, each segment and sub-segment classified.

The emotion classification of a segment process based on the majority voting method. It means that each frame in the segment contains the label. The maximum possible amount of labels is 31 because the counting begins from 0. For example, a segment with a duration - two seconds and 300 milliseconds has 24 labeled frames. 17 from them is neutral, three anger, four sadness. The majority voting method will assign a neutral label to the segment

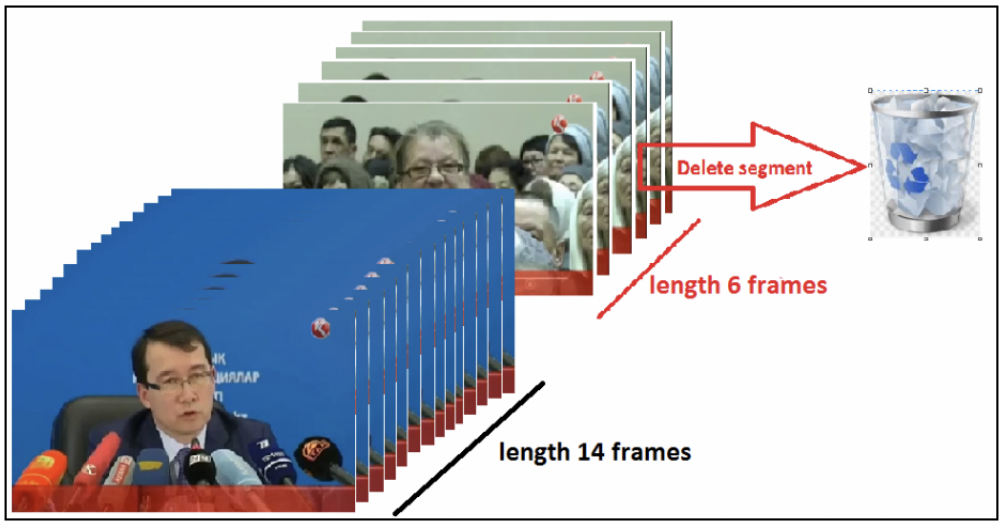

Figure 46. Sub-segmentation process

4.4.3 *Saving the label*

The last process is saving the video segment with the predicted label to an audio file. To this task, the python library "pydub" has been used. All have predicted segments get into precisely in the appropriate class folder.

4.5 Label filtering

Finally, when a dataset is collected, we need to recheck each label to delete confusing labels. It can be happened by the following reasons: first case when face human emotion can be different from speech emotion when a person is speaking. The second case is a voiceover. It means the method for collecting and labeling emotion data detected some speech and detected some face in a frame because the camera looks to person, but the voiceover is asking something at a person. For example, if the person is smiling at the camera when the voiceover is asking him with neutral speech, then a label will be saved as happiness. Figure 47 shows the example of voiceover.

To solve those issues I have used the speech emotion model with accuracy 85,6%[50]. The speech emotion model predicts each audio file. If, after emotion prediction label is the same with initial, then the file was saved. Otherwise, the audio file was deleting. The approach allows the filtering of confusing labels automatically.

The speech emotion model[110] was trained at six standard emotional speech databases which were pooled together, namely, SAVEE [35], EMOVO [62], RML[53], EMODB [24], ELRA-S0329 [25], RAVDESS [34].

Totally in dataset were at around 9000 variable duration utterances, in 8 different languages. To achieve accuracy, 85.6% in speech emotion classification, authors have chosen the mix from the VGGVox[111] model with the support vector machine(SVM)[112] algorithm.

Figure 47. Voiceover

This VGG-like architecture, developed by Nagrani et al. [111] and made up of 12 layers, is ideal for capturing even the slightest of stimuli at specific locations due to its convolutional background. This is taken advantage of by providing the network with spectral representations of the audio signals, from which features are then extracted. These narrow band spectrograms are obtained directly from raw data, in order to retain all information, by using a sliding Hamming window of width 25ms and step 10ms. After this, normalization is performed on mean and variance at every frequency bin of the spectrum. Considering the former, n-second inputs provide 100n frames spectra. No further action is taken on the input data.A high-level diagram of the entire system is provided in Figure 48, to aid in understanding.

The network deals with variable length input through the apool6 layer see [111]), where the filter dimension is adaptable to the corresponding clip's duration. Provided this duration is between 1 and 10 seconds long, and by the stride and padding methods used by the model, the filter dimension takes the same value as that of the input array to the apool6 layer. The clip duration to apool6 layer dimension correspondence is shown in Table 18. Clips, which are 10 seconds in length or longer,are also accepted, though the model will only consider the central 10-second window and disregarded all other encircling audio.

The model as a whole was already trained for speaker classification using the related VoxCeleb1 dataset [113], which is made up of 100,000+ utterances by over 7000 different speakers of different cultures and backgrounds, totaling at more than 2000 hours of audio. That being said, the model is undoubtedly capable of building complex representations of the data it receives, while also encompassing a vast and varied set of speaker-specific cues and prosody mannerisms. As such, it becomes an ideal candidate for performing speaker adaptation in SER systems and preventing premature feature specialization by taking further advantage of more intertwined speech information. The Figure 48 shows a VGGVox model in detail.

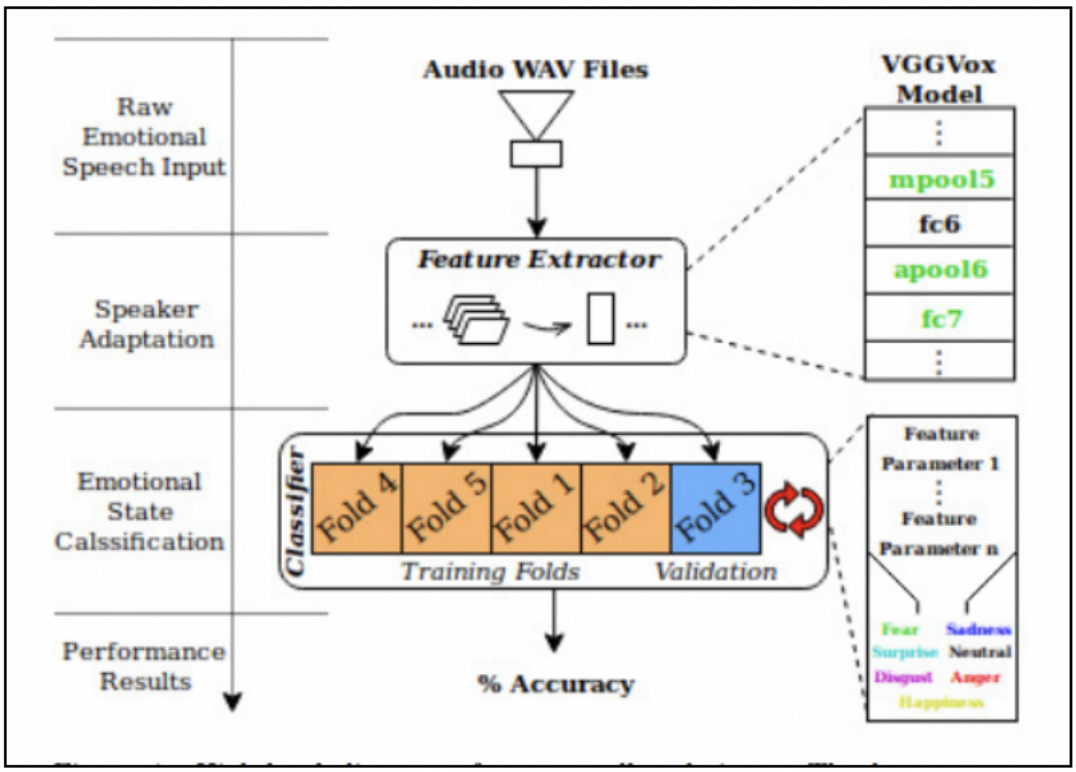

Figure 48. High-level diagram of our overall technique. The layer names in green correspond to the layers from where the features were extracted. Having 5-fold cross-validation been used, each time four folds were used for training (orange) while the remaining fold was used for validation (blue)

Table 18. The apool6 adaptation to the clip's duration

Number of frames	Filter Dimension
100	2
200	5
300	8
400	11
500	14
600	17
700	20
800	23
900	27
1000	30

Finally, all parts and steps in the proposed automated method for collecting and labeling speech emotion data can be described in figure 49.

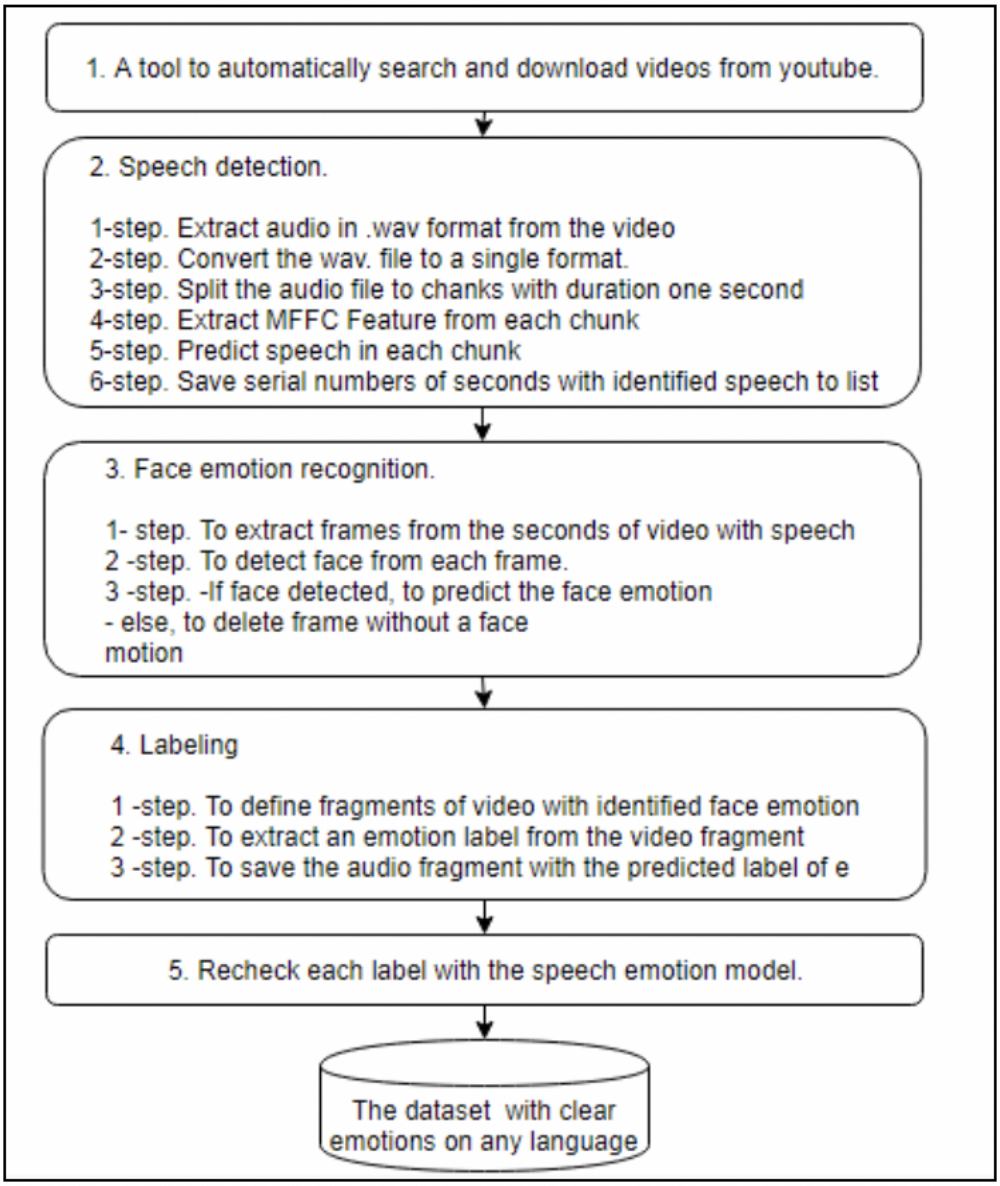

Figure 49. An automated method for collecting and labeling speech emotion data in details.

5. Results

5.1 Principles of data collection using Video Parser

The collected dataset was extracted from1 243video files, 343 of them in Kazakh, and 900 in the Russian language. The video files were downloaded from youtube. The collection of video consist from 5 TV shows on Russian (such as “DNA” –a matter of true fatherhood, “Wait for me” - the long-awaited reunion of family members of the missing, “House-2” - show where participants are trying to build a relationship while living in a country house as a group, “Windows” – scandalous TV show, “Seems groped” – TV show with scarred and exiting emotions where participants must guess the contents inside the black box), 3 TV shows on Kazakh(such as “Sorry me”- TV show with intrigues and gossips, “Literal truth” - the investigations and shocking truth about the man. “What is it” – the analog on Kazakh of “Seems groped” TV show), emotional videos with rude behavior of traffic police officers, and a lot of interview face to face with famous persons. The total duration of videos with Kazakh speech is 452 hours and 753 hours with Russian speech. Many television shows deal with relatively strong emotion.The requirements considered in the selection of these candidate shows are listed below.

1. The shows for the candidate should close to real-life scenarios.
2. The shows should contain real interactions ratherthan acted materials.
3. The shows should contain a lot of impressive and exciting topics with natural emotions.

The movie and serials have not been included due to the fact that actors play their roles without natural emotions.

5.2 Labels extraction

The audio labeled extraction process has taken an enormous time, approximately two weeks. Two computers have been used to a label extraction process. The

performance of the first one was the processor: Intel (R) Core (TM) i7-2630QM the quantity of cores four real and four virtual with 2,00 GHz, RAM – 8 GB, OS – Windows 7 – 64. The second one has the next properties, processor: Intel (R) Core (TM) i7-7700 3.60GHz *8, RAM 47 GB, Graphics GeForce GTX1060 6GB/PCIe/SSE2. The computers have worked parallel using Anaconda python 3, and Jupyter python IDE. The performance of the extraction task can be accelerated using technology Hadoop[114] or Spark[115], if you have the amount of PC more than five, it will be a substantial difference in speed of extraction.

The surprised and disgust emotions very rare was detected by the supposed method. Therefore it is complicated to surprise people in our time. Table 19 shows the amount of extracted audios with emotion labels.

Table 19. Collected datasets on Kazakh and Russian

	Emotion	Kazakh	Russian
	Anger	5 165	14 873
	Disgust	3 947	982
	Happiness	14 343	15 740
	Neutral	40 065	81 072
	Sadness	3 965	18 107
	Fear	4 335	13 613
	Surprised	336	1 816
Total amount:		72 156	146 203

After the filtering process we have got the next amount of labeled data, which are shown in table 20.

Table 20. Emotional collected datasets after the label filtering process.

	Emotion	Dataset 1. Kazakh	Dataset 2. Russian
	Anger	548	1 131
	Disgust	269	123
	Happiness	4 066	3 598
	Neutral	9 019	20 553
	Sadness	286	2 005
	Fear	872	2 900
	Surprised	16	73
Total amount:		15 076	30 383
Total duration:		11h 10min 5sec	22h 5min 29sec

5.3 Datasets

According to table 19, was createdsix datasets. Two of them are Russian and Kazakh dataset, which shown in table 20. The third one is a mix of the first and second datasets. Table 21 shows the dataset 3 in detail.

Table 21. Dataset 3.

	Emotion	Kazakh+Russian
	Anger	1 679
	Disgust	392
	Happiness	7 664
	Neutral	29 572
	Sadness	2 291
	Fear	3 772
	Surprised	89
Total amount:		45 459
Total duration:		33h 15min 34 sec

As can see from tables 15 and 16, the class Surprised and Disgust have very least amount utterances, therefore the datasets not uniform. The dataset 4 consist from dataset 3, but without disgust and surprised emotions. Table 22 shows the dataset 4 in detail.

Table 22. Dataset 4.

	Emotion	Kazakh+Russian
	Anger	1 679
	Happiness	7 664
	Neutral	29 572
	Sadness	2 291
	Fear	3 772
Total amount:		44 978
Total duration:		32h 50min 46sec

Datasets 5 – EmoDB[24] and Datasets 6 – RAVDESS[34] are shown to compare accuracy in recognition with collected databases and shows the superiority of hypotheses that datasets with lots of data much better and improves the accuracy.Table 23,24 shows the dataset 5 and dataset 6 in detail.

Table 23. Dataset 5.

	Emotion	German
	Anger	127
	Boredom	81
	Disgust	46
	Fear	69
	Happiness	71
	Neutral	79
	Sadness	62
Total amount:		535
Total duration:		0h 42min 38sec

Table 24. Dataset 6.

	Emotion	English
	Anger	192
	Disgust	192
	Fear	192
	Happiness	192
	Neutral	96
	Sadness	192
	Surprised	192
Total amount:		1 248
Total duration:		1h 16min 19sec

All datasets were divided into training set, development set and test set by proportion 80%, 10%, 10%. In accordance with the above datasets, in the next sections of the paper experiments will be conducted for each dataset respectively.

5.4 Feature extraction and DNN model

Six feature extraction, VGGVox [39] model was used to extract features. The dimensionality of the features is 1024.

The real duration of the Dataset 5 is 24 min 48 sec. The VGGVox[39] preprocessing model accept the audio files with duration from 3 to 10 sec. A duration of the most audio files was less than 3 sec therefore to everyone audio file was added 2 silent seconds, one in the beginning and one in the end.

We used several classifier for that task such as as SVM, Logistic regression, Random forest, K-means, Decision Tree and DNN. The best result in accuracy we have got with the DNN model. The chosen DNN architecture contains eight fully connected layers with activation function relu [41], and the last layer also fully connected but with activation function softmax. The structure is as follows: 1 – 1024 neurons, 2 – 512 neurons,3- 256 neurons, 4 – 128 neurons, 5- 64 neurons, 6-32 neurons, 7-16 neurons,8-8 neurons. The last layer with activation function softmax contains two neurons. For the regularization of the DNN, a dropout 0.5was used [42] between the fourth and fifth layers and batch normalization before the first layer. All layers were initialized using Glorot uniform initialization[43]. The detailed information about the architecture is shown in Figure 50.

For training the proposed model, the Stochastic Gradient Descent algorithm was utilized with a fixed learning rate of0.11 to optimize a Binary cross-entropy loss function, also known as logloss [44]. Momentum is 0.1. The metrics of model is the accuracy.The input data were presented to the DNN in batches of size 256 in 100 epochs (iterations).

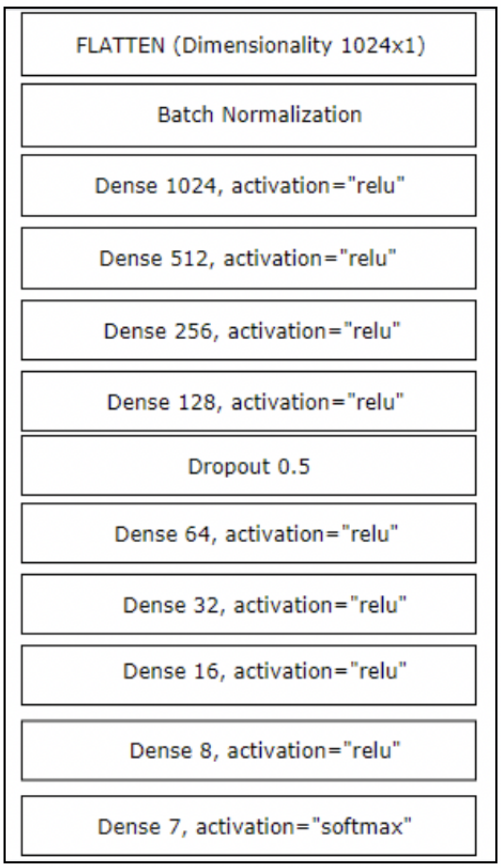

Figure 50. DNN model

5.5 Network learning and results

Four datasets have been trained the DDN and other classifiers and after the model training, have got the following results with shown in table 25. In table 26,27,28,29,30,31 there is shown a confusion matrix, precision, recall, and F1 score for the test set for DNN classifier.

Table 25. The accuracy of test set

Classifier	D1 (%)	D2 (%)	D3 (%)	D4 (%)	D5 (%)	D6 (%)
SVM	86,95	80	82.47	84.38	79.83	71.17
Logistic regression	59.88	61.28	60.37	61.09	59.88	52.27
Random Forest	65.36	67.58	64.23	67.96	64.51	59.26
K-means	58.35	55.43	56.27	56.98	51.29	44.03
Decision Tree	62.93	61.23	59.88	61.97	58.33	49.29
DNN	89	85.2	86.84	88.56	83.33	75

D1 - Dataset 1, D2 - Dataset 2, D3 - Dataset 3, D4 - Dataset 4, D5 - Dataset 5, D6 - Dataset 6. The measure is accuracy in %.

For a more accurate comparison, Dataset 4 is not shown in Figure 51 due to the different number of classes in relation to other datasets.

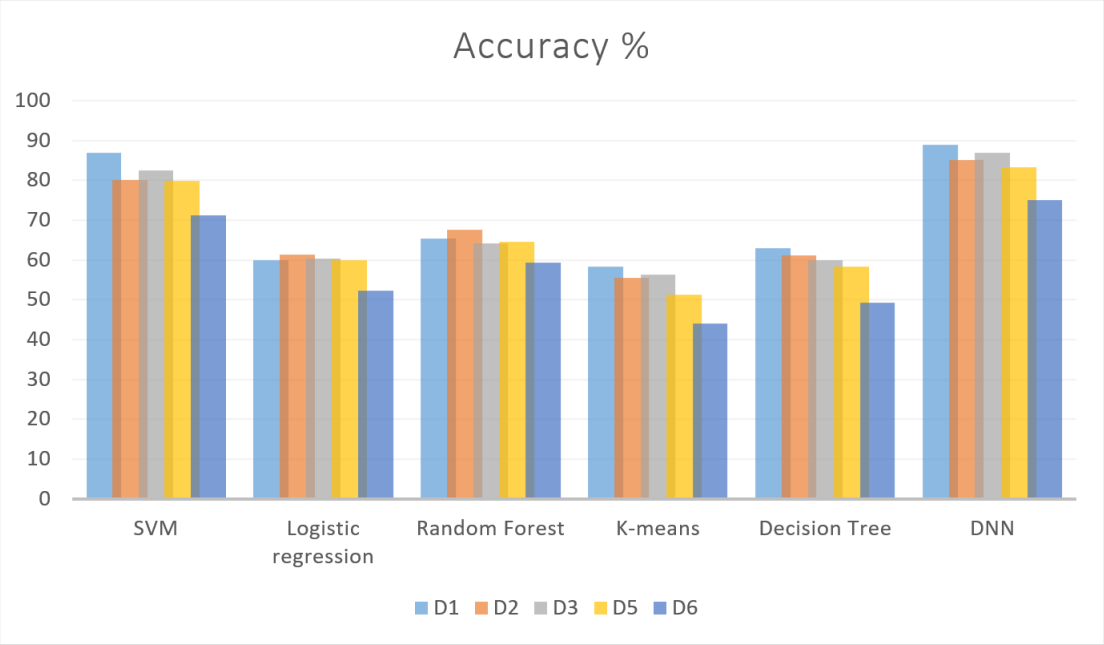

Figure 51. The bar char with accuracy results

D1 - Dataset 1, D2 - Dataset 2, D3 - Dataset 3, D5 - Dataset 5, D6 - Dataset 6. The measure is accuracy in %.

Table 26. Dataset 1. Confusion Matrix (Test set) of DNN classifier

	anger	disgust	happiness	neutral	sadness	fear	surprised	Recall	Precision	F1 score
Anger	32	1	18	3	0	1	0	0.5818	0.6956	0.6339
Disgust	3	12	2	5	1	4	0	0.4615	0.6315	0.5332
happiness	8	2	354	32	2	9	0	0.8697	0.8697	0.8696
Neutral	2	1	15	873	5	6	0	0.9678	0.9407	0.954
sadness	0	3	3	8	11	4	0	0.3793	0.4782	0.4229
Fear	1	0	15	6	4	62	0	0.7126	0.7126	0.7125
surprised	0	0	0	1	0	1	0	0	0	0
Total								0.5675	0.6183	0.5894

Table 27. Dataset 2. Confusion Matrix (Test set) of DNN classifier

	anger	disgust	happiness	neutral	sadness	fear	surprised	Recall	Precision	F1 score
Anger	64	0	15	3	0	31	0	0.5663	0.8205	0.67
disgust	0	0	2	4	3	3	1	0	0	0
happiness	7	1	230	36	2	85	0	0.6371	0.7098	0.6715
neutral	1	0	37	1975	18	25	0	0.9606	0.9440	0.9522
sadness	0	0	1	38	141	21	0	0.7014	0.8057	0.7502
fear	5	0	37	35	10	181	0	0.6753	0.5186	0.5869
surprised	1	0	2	1	1	3	0	0	0	0
total								0.5058	0.5426	0.5186

Table 28. Dataset 3. Confusion Matrix (Test set) of DNN classifier

	anger	disgust	happiness	neutral	sadness	fear	surprised	Recall	Precision	F1 score
anger	102	2	36	12	0	16	0	0.6071	0.7286	0.6623
disgust	1	7	6	16	3	7	0	0.175	0.4375	0.2501
happiness	18	3	608	63	5	70	0	0.7927	0.7845	0.7885
neutral	8	1	34	2772	26	22	0	0.9682	0.9374	0.9526
sadness	0	1	9	49	153	16	2	0.6522	0.75	0.6978
fear	10	2	80	45	13	227	1	0.6005	0.6486	0.6236
surprised	1	0	2	0	4	2	1	0.1	0.25	0.1429
total								0.5565	0.648	0.5882

Table 29. Dataset 4. Confusion Matrix (Test set) of DNN classifier

	anger	happiness	neutral	sadness	fear	Recall	Precision	F1 score
anger	127	29	4	0	8	0.756	0.7938	0.7744
happiness	20	598	59	8	82	0.7797	0.8147	0.7968
neutral	6	33	2762	35	27	0.9647	0.9518	0.9582
sadness	0	7	34	163	26	0.7087	0.7581	0.7329
fear	7	67	43	9	252	0.6666	0.638	0.652
total						0.7751	0.7913	0.7828

Table 30. Dataset 5. Confusion Matrix (Test set) of DNN classifier

	anger	boredom	disgust	fear	happiness	neutral	sadness	Recall	Precision	F1 score
Anger	11	0	0	0	2	0	0	0.846	0.7333	
boredom	0	8	0	0	0	0	1	0.8888	0.8888	
disgust	0	0	4	0	0	0	0	1	1	
fear	0	0	0	6	1	0	0	0.857	1	
happiness	4	0	0	0	3	0	0	0.4286	0.5	
neutral	0	1	0	0	0	7	0	0.875	1	
sadness	0	0	0	0	0	0	6	1	0,857	
total								0.8422	0.8541	0.8482

Table 31. Dataset 6. Confusion Matrix (Test set) of DNN classifier

	anger	disgust	fear	happiness	neutral	sadness	surprised	Recall	Precision	F1 score
Anger	11	2	1	2	1	2	0	0,579	0.9166	
disgust	0	16	0	0	1	1	1	0,8421	0,6666	
fear	0	2	14	2	0	1	0	0.7368	0.8235	
happiness	0	1	2	13	2	0	1	0.6842	0.7222	
neutral	0	0	0	0	6	3	1	0.6	0.6	
sadness	0	3	0	0	0	16	0	0.8421	0.695	
surprised	1	0	0	1	0	0	17	0,8947	0,85	
total								0.7398	0.7534	0.7465

6. Discussion

The F1 score and confusion matrix shows the consequence of the dependence of the amount of data on the accuracy. For example, the recognition of neutral and happiness is the highest in dataset 3, the number of utterances, respectively, more than a number of utterances of other emotions. It means that the success of the SER task depends on the amount of data. With a designed automated method of collecting and labeling data, we can increase the number of other emotions that will make recognition more robust with higher accuracy.

The novelty of the dissertation is to design an automated method for collecting and labeling speech emotional data.The results obtained in this dissertation will significantly advance the field of AI in recognizing speech emotions. Using the method of collecting emotional data, scientists will be able to collect emotional datasets in all languages of the world. Soon, machines will be able to recognize all seven speech emotions in any language of the world with high accuracy.

The obtained results are great practical value since with the recognition of speech emotions, it will become possible to understand human feelings and improve the quality of the services provided, receiving instant feedback.The practical value of the thesis lies in the possibility of qualitative improvement in service, in education, banking, insurance, public services, medicine, law enforcement agencies, and the military.

Having solved the task of recognizing speech emotions,it becomes possible to work in the field of recognizing true or false emotions, as well as recognizing people by voice.

7. Conclusion

The SER task is one of the critically part of AI sphere.The dissertation reveals current issues in SER sphere such as lack of labeled data almost in all languages, difficulties in the labeling process, and creating natural speech emotions. In the dissertation were considered spectral features and practically proved that MFCC features more preferable to SER.Additionally, in science, work was proved the sensitivity of emotion classification on Kazakh and Russian from German, which means exists the dependency in SER from language.

As one of the approaches for solving current problems in the field of SR, an automated method for collecting and labeling data for SER based on FER is proposed. The proposed method consists of five parts that are described in detail. Based on the developed method, a dataset with speech emotions for the Kazakh and Russian languages was collected. The dataset gathered exclusively on the natural emotions of people from videos with emotional components such as personal interviews, reports on tragedies and people who urgently need surgery, scandalous news, and live reports.As a result, 1243 video files from YouTube with a duration of approximately 1058 hours were collected. From which 218 359 emotions were extracted and labeled.After filtering, 45,459 emotions have remained. The duration of the corpus is 33h 15min 34 sec. Based on the collected dataset, a comparative analysis of ML models was carried out where the DNN model reached 86,84% recognition of speech emotion in a test set.

Also, during the development of the proposed method, a model for speech detection was developed on the example of the Kazakh and Russian languages with a recognition accuracy of 97.3%. The developed model was tested on two videos with a duration of 3 to 4 minutes. Test results show that this model is capable of recognizing clear speech.

The collected audio emotion corpus is the biggest and first database with natural emotions that were collected and labeled automatically.All the goals of the thesis were achieved during the research work.

References

[1] *A Shevtsova. INFLUENCE OF EMOTIONS ON HUMAN HEALTH. Students Scientific Community: Interdisciplinary RESEARCH: Sat. Art. bymat. XLII Int. Stud. scientific-practical conf. No. 7 (42).* [Online]. Available: `https://sibac.info/archive/meghdis/7(42).pdf`, (accessed: 10.10.2019).

[2] *Michael Revina , W R Sam Emmanuel. A Survey on Human Face Expression Recognition Techniques. Journal of King Saud University - Computer and Information Sciences, September 2018. ISBN 1319-1578.*

[3] *Mehrabian, Albert. Silent , Implicit Communication of Emotions and Attitudes. — 2nd. — Belmont, CA : Wadsworth, 1981. — ISBN 0-534-00910-7.*

[4] *Wu S. Recognition of human emotion in speech using modulation spectral features and support vector machines [PhD thesis]. 2009.*

[5] *A. Paeschke, W. F. Sendlmeier. Prosodic Characteristics of Emotional Speech: Measurements of Fundamental Frequency Movements. ISCA archive, ITRW on Speech and Emotion Newcastle, Northern Ireland, UK, September 2000.*

[6] *Milton A, Sharmy Roy S, Tamil Selvi S. SVM scheme for speech emotion recognition using MFCC feature. International Journal of Computer Applications. 2013;69.*

[7] *Divya Sree GS, Chandrasekhar P, Venkateshulu B. SVM based speech emotion recognition compared with GMM-UBM and NN. IJESC. 2016.*

[8] *Melki G, Kecman V, Ventura S, Cano A. OLLAWV: Online learning algorithm using worst-violators. Applied Soft Computing. 2018;66:384-393.*

[9] *Pan Y, Shen P, Shen L. Speech emotion recognition using support vector machine. International Journal of Smart Home. 2012;6:101-108.*

[10] *Peipei S, Zhou C, Xiong C. Automatic speech emotion recognition using support vector machine. IEEE. 2011;2:621-625.*

[11] *Alex G, Navdeep J. Towards end-toend speech recognition with recurrent neural networks. In: International Conference on Machine Learnin.; Vol. 32.*, 2014.

[12] *Ingale AB, Chaudhari D. Speech emotion recognition using hidden Markov model and support vector machine. International Journal of Advanced Engineering Research and Studies. 2012:316-318.*

[13] *Sathit P. Improvement of speech emotion recognition with neural network classifier by using speech spectrogram. International Conference on Systems, Signals and Image Processing (IWSSIP). 2015:73-76.*

[14] *Martin V, Robert V. Recognition of emotions in German speech using Gaussian mixture models. LNAI. 2009 5398:256-263.*

[15] *Sara M, Saeed S, Rabiee A. Speech Emotion Recognition Based on a Modified Brain Emotional Learning Model. Biologically inspired cognitive architectures. Elsevier; 2017;19:32-38.*

[16] *Lim W, Jang D, Lee T. Speech emotion recognition using convolutional and recurrent neural networks. Asia Pacific. 2017:1-4.*

[17] *Chen S, Jin Q. Multi-Modal Dimensional Emotion Recognition using Recurrent Neural Networks. Australia: Brisbane; 2015.*

[18] *Yu G, Eric P, Hai-Xiang L, van den HJ. Speech emotion recognition using a voiced segment selection algorithm. ECAI. 2016;285:1682-1683.*

[19] *D. Ververidis and C. Kotropoulos, "A state of the art review on emotional speech databases," in 1st Richmedia Conference, Lausanne, Switzerland, 2003, pp. 109–119.*

[20] *R. Rajoo, C. Chee Aun. Influences of languages in speech emotion recognition: A comparative study using Malay, English, and Mandarin languages. 2016 IEEE Symposium on Computer Applications Industrial Electronics (ISCAIE). Batu Feringghi, Malaysia. September 2016. ISBN 978-1-5090-1543-6.*

[21] *A. Shoiynbek,K. Kozhakhmet, D. Kuanysbay. VARIOUS LANGUAGES IMPACT ON THE PROBLEM OF EMOTION RECOGNITION IN SPEECH.VESTNIK KazNRTU 5 (135). 2019. ISSN:1680-9211.*

[22] *A Jennifer Smith, Andreas Tsiartas, Valerie Wagner, Elizabeth Shriberg, Nikoletta Bassiou. CROWDSOURCING EMOTIONAL SPEECH. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20.*

[23] *Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, Mohamed Ali Mahjoub and Catherine Cleder. Automatic speech emotion recognition using machine learning. Social Media and Machine Learning Publisher: IntechOpen March 25th 2019.*

[24] *Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F. and Weiss, B., 2005, September. A database of German emotional speech. In Interspeech (Vol. 5, pp. 1517-1520).*

[25] *ELRA. Emotional speech synthesis database s0329. catalogue.elra.info/en-us/repository/browse/ELRA-S0329/, 2012.*

[26] *Huang, C.; Liang, R.; Wang, Q.; Xi, J.; Zha, C.; Zhao, L. Practical speech emotion recognition based on online learning: From acted data to elicited data. Math. Probl. Eng.*

[27] *Dorota Kamińska. Emotional Speech Recognition Based on the Committee of Classifiers. Special Issue "Statistical Machine Learning for Human Behaviour Analysis" September 2019.*

[28] *K.Ślot, J.Cichosz, L. Bronakowski. Emotion recognition with Poincare mapping of voiced-speech segments of utterances. In Proceedings of the International Conference on Artificial Intelligence and Soft Computing, Zakopane, Poland, 16–20 June 2019; pp. 886–895.*

[29] *A.Arruti, I.Cearreta, A.Álvarez, E.Lazkano, B.Sierra. Feature Selection for Speech Emotion Recognition in Spanish and Basque: On the Use of Machine Learning to Improve Human-Computer Interaction. PLoS ONE, 2014, 9, e108975.*

[30] *W.Bao, Y.Li, M.Gu, M.Yang, H. Li, L.Chao, J.Tao. Building a Chinese natural emotional audio-visual database. In: 2014 12th International Conference on Signal Processing (ICSP), pp. 583–587 (2014).*

[31] *Y. Li, J. Tao, B. Schuller, S. Shan, D. Jiang, and J. Jia. MEC 2016: The multimodal emotion recognition challenge of ccpr 2016, in Chinese Conference on Pattern Recognition 2016, pp. 667–678. Springer, 2016.*

[32] *A.Dhall, R.Goecke, S.Lucey, T.Gedeon. Static facial expression analysis in tough conditions: data, evaluation protocol and benchmark. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 2106– 2112 (2011).*

[33] *https://www.kaggle.com/c/challenges-in-representation-learning-facial - expression-recognition-challenge/data.*

[34] *Steven R. Livingstone and Frank A. Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in North American English. Plos One, 2018.*

[35] *S. Haq and P.J.B. Jackson. Speaker-dependent audio-visual emotion recognition. Proc. Int. Conf. Auditory-Visual Speech Processing, 2009.*

[36] *D. C. Ambrus, "Collecting and recording of an emotional speech database", Technical Report, Faculty of Electrical Engineering and Computer Science, Institute of Electronics, University of Maribor.*

[37] *R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schroder, "Feeltrace: An instrument for recording perceived emotion in real-time", in Proc. ISCA Workshop (ITRW) on Speech and Emotion: A conceptual framework for research, pp. 19-24, Belfast, 2000.*

[38] *E. Douglas-Cowie, R. Cowie, and M. Schroder, "A New Emotion Database: Considerations, Sources and Scope", in Proc. ISCA (ITWR) Workshop Speech and Emotion: Aconceptual framework for research, pp. 39-44, Belfast 2000.*

[39] *The Center for Spoken Language Research (CSLR), CU Kids' speech corpus. http://cslr.colorado.edu/beginweb/reading/data collection.html.*

[40] *Linguistic Data Consurtium (LDC), http://www.ldc.upenn.edu/.*

[41] *C. Pereira, "Dimensions of emotional meaning in speech", in Proc. ISCA Workshop on Speech and Emotion: A conceptual framework for research, pp. 25-28, Belfast, 2000.*

[42] *M. Edgington. "Investigating the limitations of concatenative synthesis", in Proc. Eurospeech 97, pp 593-596, Rhodes, Greece, September 1997.*

[43] *T. S. Polzin and A. H. Waibel, "Detecting emotions in speech", in Proc. CMC 1998.*

[44] *V. A. Petrushin. "Emotion in speech recognition and application to call centers",in Proc. ANNIE 1999, pp. 7-10, 1999.*

[45] *K. Alter, E. Rank, and S. A. Kotz, "Accentuation and emotions - Two different systems?", in Proc. ISCA Workshop on Speech and Emotion: A conceptual framework for research, Belfast, 2000.*

[46] *B. Wendt and H. Scheich, "The Magdeburger Prosodie-Korpus", in Proc. Speech Prosody Conf. 2002, pp. 699-701, Aix-en-Provence, France, 2002.*

[47] *Michael Grimm, Kristian Kroschel, Shrikanth Narayanan, "The Vera Am mittag German Audio-Visual Emotional Speech Database", IEEE, ICME 2008.*

[48] *European Language Resources Association, (ELRA), www.elra.info.*

[49] *S. J. L. Mozziconacci and D. J. Hermes, "Expression of emotion and attitude through temporal speech variations", in Proc. 2000 Int. Conf. Spoken Language Processing (ICSLP 2000), vol. 2, pp. 373-378, Beijing, China, 2000.*

[50] *J. M. Montero, J. Gutierrez-Arriola, J. Colas, E. Enriquez, and J. M. Pardo, "Analysis and modelling of emotional speech in Spanish", in Proc. ICPhS'99, pp. 957-960, San Francisco 1999.*

[51] *I. Iriondo, R. Guaus, and A. Rodriguez, "Validation of an acoustical modeling of emotional expression in Spanish using speech synthesis techniques", in Proc. ISCA Workshop (ITRW) Speech and Emotion: A conceptual framework for research, pp. 161-166, Belfast, 2002.*

[52] *K. R. Scherer. "Emotion effects on voice and speech: Paradigms and approaches to evaluation", in Proc. ISCA Workshop (ITRW) on Speech and Emotion: A conceptual framework for research, Belfast, 2000.*

[53] *Zhibing Xie and Ling Guan. Multimodal information fusion of audiovisual emotion recognition using novel information theoretic tools. IEEE International Conference on Multimedia and Expo (ICME), 2013.*

[54] *I. S. Engberg, and A. V. Hansen, "Documentation of the Danish Emotional Speech Database (DES)," Internal AAU report, Center for Person Kommunikation, Department of Communication Technology, Institute of Electronic Systems, Aalborg University, Denmark, September 1996.*

[55] *R. Nakatsu, A. Solomides, and N. Tosa, "Emotion recognition and its application to computer agents with spontaneous interactive capabilities", in Proc. IEEE Int. Conf. Multimedia Computing and Systems, vol. 2, pp. 804-808, Florence, Italy, July 1999.*

[56] *N. Amir, S. Ron, and N. Laor, "Analysis of an emotional speech corpus in Hebrew based on objective criteria", in Proc. ISCA Workshop (ITRW) Speech and Emotion: A conceptual framework for research, pp. 29-33, Belfast, 2000.*

[57] *A. Abelin and J. Allwood, "Cross linguistic interpretation of emotional prosody", in Proc. ISCA Workshop (ITRW) on Speech and Emotion: A conceptual framework for research, Belfast, 2000.*

[58] *F. Yu, E. Chang, Y.Q. Xu, and H.Y. Shum, "Emotion detection from speech to enrich multimedia content", in Proc. 2nd IEEE Pacific-Rim Conference on Multimedia 2001, pp.550-557, Beijing, China, October 2001.*

[59] *V. Makarova and V. A. Petrushin, "RUSLANA: A database of Russian Emotional Utterances", in Proc. 2002 Int. Conf. Spoken Language Processing (ICSLP 2002), pp. 2041-2044, Colorado, USA, September 2002.*

[60] *Branimir Dropuljić, Miłosz Thomasz Chmura, Antonio Kolak, Davor Petrinović,"Emotional Speech Corpus of Croatian Language", ISPA-2011.*

[61] *Slobodan T. J., Zorka Kašić, Miodrag Đorđević, Mirjana Rajković, "Serbian emotional speech database: Design, Processing and Evaluation", Specom'2004.*

[62] *Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, and Massimiliano Todisco. Emovo corpus: an italian emotional speech database. In LREC, 2014.*

[63] *V.P. Morozov. Speech perception: Issues of functional brain asymmetry. Nauka, 1988.*

[64] *Linden Yu. Monkeys, man, language. M.: Mir, 1981.*

[65] *V. P. Morozov. Entertaining bioacoustics. Knowledge, 1987.*

[66] *M. Idi. The emergence of man. The missing link. translation from English, Mir, 1977.*

[67] *Ch. Darwin. Expression of emotions in humans and animals. Works. M., 1953a. T. 5. P. 681–920.*

[68] *Rubinstein S. L. Being and consciousness. M., 1957.*

[69] *S. L.Rubinstein. Fundamentals of General Psychology. M., 1946.*

[70] *V. P. Morozov. Biophysical foundations of vocal speech. Nauka, 1977a.*

[71] *I.P.Pavlov. Complete works. Publishing House at the USSR Academy of Sciences, 1951.*

[72] *V.P. Morozov, Language of emotions and emotional hearing. Selected Works 1964-2016 . Institute of Psychology RAS; Moscow. 2017. ISBN 978-5-9270-0346-4.*

[73] *V.I.Galunov, V.X.Manerov. A device for determining the emotional state by a speech signal. AC No. 793575.1981.*

[74] *K. S. Stanislavsky, Collected Works. In 3 vol. M., 1955.*

[75] *P. V.Simonov. The method of K. S. Stanislavsky and the physiology of emotions. Nauka, 1962.*

[76] *G.Fant. Acoustic theory of speech formation. Translation from English M, Nauka, 1964.*

[77] *V.P.Morozov. Sensitivity of human hearing to changes in phase relations between amplitude and frequency modulations in amplitude-frequency modulated sound. Biophysics. 1967a. Vol. 5, pp. 948–950.*

[78] *GM Kotlyar. On the study of singing vibrato. 13th All-Union Acoustic Conference: Abstracts of reports. M., 1973.P. 98.*

[79] *V. P. Morozov Emotional hearing. Research methods and applications. Modern experimental psychology. Publishing House "Institute of Psychology RAS", 2011b. S. 261–283.*

[80] *D.Bitouk, R.Verma, and A. Nenkovab. Class-Level Spectral Features for Emotion Recognition . Author manuscript; available in PMC 2013 Jun 19. Published in final edited form as SpeechCommun. 2010 July-August; 52(7-8): 613–625. Published online 2010 Febrary.*

[81] *Brian McFeek, Colin Raffel, Dawen Liang, Daniel P.W. Ellis , Matt McVicar, Eric Battenberg, Oriol Nietok. librosa. Audio and Music Signal Analysis in Python. PROC. OF THE 14th PYTHON IN SCIENCE CONF. (SCIPY 2015) p. 18.*

[82] *Ellis, Daniel P.W. "Chroma feature analysis and synthesis" 2007/04/21 http://labrosa.ee.columbia.edu/matlab/chroma-ansyn/.*

[83] *Meinard Müller and Sebastian Ewert Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features In Proceedings of the International Conference on Music Information Retrieval (ISMIR), 2011.*

[84] *Jiang, Dan-Ning, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. Music type classification by spectral contrast feature. In Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on, vol. 1, pp. 113-116. IEEE, 2002.*

[85] *Dubnov, Shlomo "Generalization of spectral flatness measure for non-gaussian linear processes" IEEE Signal Processing Letters, 2004, Vol. 11.*

[86] *C. Harte, M.Sandler, M.Gasser. "Detecting Harmonic Change in Musical Audio." In Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia (pp. 21-26). Santa Barbara, CA, USA. 2006 ACM Press. doi:10.1145/1178723.1178727.*

[87] *V.Nair, G.E.Hinton. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (ICML-10) pp. 807-814. 2010.*

[88] *N.Srivastava, G.E.Hinton, A.Krizhevsky, I.Sutskever, R.Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), pp.1929-1958.2014.*

[89] *X.Glorot, Y.Bengio, 2010, May. Understanding the difficulty of training deep feedforward neural networks. In Aistats. Vol. 9, pp. 249- 256.*

[90] *T.Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In Proceedings of the twenty-first international conference on Machine learning (p. 116). 2004, July. ACM.*

[91] *Rajesvary Rajoo , Ching Chee Aun .Influences of Languages in Speech Emotion Recognition: A Comparative Study Using Malay, English and Mandarin languages. 978-1-5090-1543-6/16/31.00 ©2016 IEEE, p.35.*

[92] *https://github.com/spidezad/Youtube-Videos-Search-and-Download.*

[93] *https://webrtc.org/.*

[94] *R.Nóbrega, S.Cavaco. Detecting key features in popular music: case study-singing voice detection. Proc. of the Workshop on Machine Learning and Music of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. 2009.*

[95] *Yael Segal, Tzeviya Sylvia Fuchs, Joseph Keshet. SpeechYOLO: Detection and Localization of Speech Objects. in Proceedings of conference.Interspeech. 2019.*

[96] *J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.*

[97] *A. Mahdhaoui, F. Ringeval, and M. Chetounani. Emotional speech characterization based on multi-features fusion for face-to-face communication. in Proc. Inter. Conf. SCS, Jerba, Tunisia, Nov. 6–8 2009.*

[98] *J. Ramírez, J. Segura, J. Górriz, and L. García. Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition. IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 8, pp. 2177–2189, 2007.*

[99] *Kanat Kozhakhmet, Rakhima Zhumaliyeva Aisultan Shoiynbek and Nazerke Sultanova. Speech Emotion Recognition For Kazakh And Russian Languages. Math. Inf. Sci. p.65-68. USA. 2020. DOI: 14 10.18576/amis/140108.*

[100] *Olzhas Makhambetov, Aibek Makazhanov, Zhandos Yessenbayev, Bakhyt Matkarimov, Islam Sabyrgaliyev, and Anuar Sharafudinov. 2013. Assembling the Kazakh Language Corpus. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1022–1031, Seattle, Washington, USA, October. Association for Computational Linguistics.*

[101] *A. Slizhikova; A. Veysov; D. Nurtdinova; D. Voronin; Y. Baburov. Russian Open Speech To Text (STT/ASR) Dataset. https://github.com/snakers4/open_stt/annotation-methodology.*

[102] *Eduardo Fonseca, Manoj Plakal, Frederic Font, Daniel P. W. Ellis, Xavier Favory, Jordi Pons, Xavier Serra. "General-purpose Tagging of Freesound Audio with AudioSet Labels: Task Description, Dataset, and Baseline". Proceedings of the DCASE 2018 Workshop. 2018.*

[103] *Eduardo Fonseca, Manoj Plakal, Daniel P. W. Ellis, Frederic Font, Xavier Favory, and Xavier Serra. Learning Sound Event Classifiers from Web Audio with Noisy Labels. arXiv preprint arXiv:1901.01189. 2019.*

[104] *https://www.youtube.com/watch?v=w-LR8rs3d44.*

[105] *https://www.youtube.com/watch?v=hRpP5LvXF1Et=111s.*

[106] *P. Viola and M. Jones. Robust real-time face detection. IJCV 57(2), 2004.*

[107] *Lienhart R., Kuranov E., Pisarevsky V.: Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In: PRS 2003, pp. 297-304. 2003.*

[108] *https://github.com/omar178/Emotion-recognition.*

[109] *https://www.kaggle.com/c/challenges-in-representation-learning-facial- expres. recognition-challenge/leaderboard.*

[110] *G. Assunção, F. Perdigão, and P. Menezes. "Premature overspecialization in emotion recognition systems". In Audio Engineering Society Conference: 2019 AES International Conference on Audio Forensics, Jun 2019.*

[111] *A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In INTERSPEECH, 2017.*

[112] *CC Chang and CJ Lin. Libsvm: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, page 7:1–27:27, 2011.*

[113] *A. Nagrani. The voxceleb1 dataset. http://www.robots.ox.ac.uk/ vgg/data/ voxceleb/vox1.html, 2017. Accessed: 2019-03-14.*

[114] *https://hadoop.apache.org/.*

[115] *https://spark.apache.org/.*