




РАЗДЕЛ I. БОЛЬШИЕ ЯЗЫКОВЫЕ МОДЕЛИ И ПРОМПТ-ИНЖИНИРИНГ
В ИССЛЕДОВАНИЯХ ЯЗЫКОВОГО ПОВЕДЕНИЯ
В МАШИННО-ГЕНЕРИРУЕМЫХ СРЕДАХ
SECTION I. LARGE LANGUAGE MODELS AND PROMPT ENGINEERING
IN THE STUDY OF HUMAN LANGUAGE BEHAVIOUR
IN MACHINE-GENERATED ENVIRONMENTS

UDC 004.93

DOI: 10.18413/2313-8912-2024-10-4-0-2

Tatiana N. Balabanova¹ 
Diana I. Gaivoronskaya² 
Anna N. Doborovich³ 

Using neural network technologies in determining
the emotional state of a person in oral communication

¹ Belgorod State National Research University,
85 Pobedy St., Belgorod, 308015, Russia
E-mail: sozonova@bsu.edu.ru
ORCID: 0000-0003-3547-3433

² Belgorod State National Research University,
85 Pobedy St., Belgorod, 308015, Russia
E-mail: trubitsyna@bsu.edu.ru
ORCID: 0009-0001-4441-565X

³ Belgorod State National Research University,
85 Pobedy St., Belgorod, 308015, Russia
E-mail: doborovich@bsu.edu.ru
ORCID: 0000-0002-7546-8447

Received 09 July 2024; accepted 15 December 2024; published 30 December 2024

Abstract: Human oral speech often has an emotional connotation; this is due to the fact that emotions and our mood influence the physiology of the vocal tract and, as a result, speech. When a person is happy, worried, sad or angry, it is reflected in various characteristics of the voice, the pace of speech and its intonation. However, assessing a person's emotional state through speech can have a beneficial effect on various areas of life, for example, medicine, psychology, criminology, marketing and education, etc. In medicine, the use of assessing emotions by speech can help in the diagnosis and treatment of mental disorders, as well as in monitoring the emotional state of the patient, identifying diseases such as Alzheimer's in its early stages, diagnosing autism, etc. In psychology, this method can be useful for studying emotional reactions to various stimuli and situations. In criminology, speech analysis and emotion detection can be used to detect false statements and deception. In marketing and advertising, it can help understand consumer reactions to a product or advertising campaign. In education, assessing emotions from speech can be used to analyze the emotional state of students and optimize the educational process.

Thus, automation of the emotion recognition process is a promising area of research, and the use of various machine learning methods and image recognition algorithms can make the process more accurate and efficient.

In order to address the challenge of identifying paralinguistic expressions of emotion in human speech, it is proposed that a neural network approach be employed. This methodology has demonstrated efficacy in addressing complex problems where an exact solution may be elusive. The work presents a neural network of convolutional architecture that allows to recognize four human emotions (sadness, joy, anger, neutral) from spoken speech. Particular attention is paid to the formation of a dataset for training and testing the model, since at present there are practically no open speech databases for the study of paralinguistic phenomena (especially in Russian). This study uses the Dusha emotional speech database.

Mel-spectrograms of the speech signal are used as features for recognizing emotions, which made it possible to increase the percentage of recognition and the speed of operation of the neural network compared to the use of low-level descriptors.

The results of experiments in the test sample showed that the presented neural network helps to recognize human emotions from oral speech in 75% of cases, which is a high result.

Further research involves training and upgrading (if necessary) the presented neural network to recognize paralinguistic phenomena not presented in this study, for example, lies, fatigue, depression, etc.

Keywords: Speech data; Speech databases; Neural networks; Convolutional neural networks; Emotion recognition; Classification; Classification methods

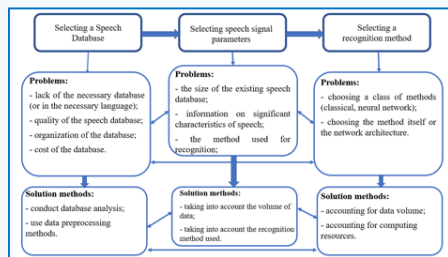
How to cite: Balabanova, T. N., Gaivoronskaya, D. I., Doborovich, A. N. (2024). Using neural network technologies in determining the emotional state of a person in oral communication, *Research Result. Theoretical and Applied Linguistics*, 10 (4), 17–34. DOI: 10.18413/2313-8912-2024-10-4-0-2

USING NEURAL NETWORK TECHNOLOGIES IN DETERMINING THE EMOTIONAL STATE OF A PERSON IN ORAL COMMUNICATION

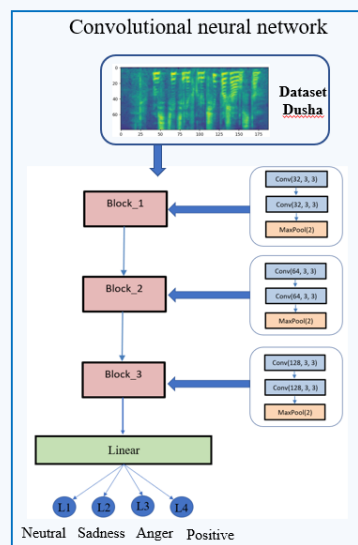
The problem of recognizing emotions from speech

- ✓ Selecting a model of a person's emotional state
- ✓ Ambiguity in the manifestation of emotions in speech
- ✓ Insufficiency of computing resources

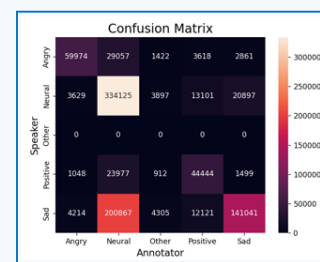
Problems of SER construction



Solution



Result






№	Recognition Method	Metric	Average
1	Human Recognition	Precision	0,7234
		Recall	0,6289
2	CNN (VGG based)	Precision	0,7521
		Recall	0,7513



УДК 004.93

DOI: 10.18413/2313-8912-2024-10-4-0-2

Балабанова Т. Н.¹
Гайворонская Д. И.²
Доборович А. Н.³

Распознавание эмоций по устной речи
с использованием нейросетевого подхода

¹ Белгородский государственный национальный исследовательский университет
ул. Победы, д. 85, г. Белгород, 308015, Россия
E-mail: sozonova@bsu.edu.ru
ORCID: 0000-0003-3547-3433

² Белгородский государственный национальный исследовательский университет
ул. Победы, д. 85, г. Белгород, 308015, Россия
E-mail: trubitsyna@bsu.edu.ru
ORCID: 0009-0001-4441-565X

³ Белгородский государственный национальный исследовательский университет
ул. Победы, д. 85, г. Белгород, 308015, Россия
E-mail: doborovich@bsu.edu.ru
ORCID: 0000-0002-7546-8447

*Статья поступила 09 июля 2024 г.; принята 15 декабря 2024 г.;
опубликована 30 декабря 2024 г.*

Аннотация: Устная речь человека всегда имеет эмоциональную окраску, это может быть обусловлено тем, что наши эмоции и наше настроение влияют на нашу речь. Когда мы рады, волнуемся, грустим или злимся, это отражается в нашем голосе, темпе и интонации. Невозможно говорить без эмоций, так как они являются неотъемлемой частью нашей личности и сопровождают нас повсюду. Наша устная речь становится еще богаче и выразительнее, когда мы передаем свои эмоции и чувства через слова. Однако оценка эмоционального состояния человека по речи может благотворно влиять на различные области жизнедеятельности, например, такие как медицина, психология, криминология, маркетинг и образование и многое другое. В медицине использование оценки эмоций по речи может помочь в диагностике и лечении психических расстройств, а также в мониторинге эмоционального состояния пациента, выявление на ранних стадиях таких болезней как Альцгеймер. В психологии этот метод может быть полезен для изучения эмоциональных реакций на различные стимулы и ситуации. В криминологии анализ речи и определение эмоций может использоваться для выявления ложных показаний и обмана. В маркетинге и рекламе это может помочь понять реакцию аудитории на продукт или рекламную кампанию. В образовании оценка эмоций по речи может быть использована для анализа эмоционального состояния студентов и оптимизации образовательного процесса. Таким образом, автоматизация процесса распознавания эмоций является перспективным направлением исследований, а применение различных методов машинного обучения и алгоритмов распознавания образов, можно сделать процесс более точным и эффективным. В качестве инструмента для решения задачи распознавания паралингвистических явлений в виде эмоций по устной речи человека

предлагается использовать нейросетевой подход, который показывает свою эффективность при решении задач в том случае, когда точное решение найти сложно. В работе представлена нейронная сеть сверточной архитектуры, позволяющая распознавать по устной речи четыре эмоции человека (грусть, радость, гнев, нейтраль). Особое внимание уделено формированию датасета для тренировки и тестирования модели, поскольку в настоящее время открытых баз речевых данных для исследования паралингвистических явлений (особенно на русском языке) практически нет. В данном исследовании используется база эмоциональной речи Dusha.

В качестве признаков для распознавания эмоций используются мел-спектрограммы речевого сигнала, что позволило увеличить процент распознавания и скорость работы нейронной сети по сравнению с использованием низкоуровневых дескрипторов.

Результаты экспериментов на тестовой выборке показали, что представленная нейронная сеть позволяет распознавать эмоции человека по устной речи в 75% случаев, что является высоким результатом.

В качестве дальнейших исследований предполагается тренировка и модернизация (при необходимости) представленной нейронной сети для распознавания паралингвистических явлений, не представленных в данном исследовании, например, таких как ложь, усталость, депрессия и др.

Ключевые слова: Речевые данные; Речевые базы данных; Нейронные сети; Сверточные нейронные сети; Распознавание эмоций; Классификация; Методы классификации

Информация для цитирования: Балабанова Т. Н., Гайворонская Д. И., Доборович А. Н. Распознавание эмоций по устной речи с использованием нейросетевого подхода // Научный результат. Вопросы теоретической и прикладной лингвистики. 2024. Т. 10. № 4. С. 17–34. DOI: 10.18413/2313-8912-2024-10-4-0-2

РАСПОЗНАВАНИЕ ЭМОЦИЙ ПО УСТНОЙ РЕЧИ С ИСПОЛЬЗОВАНИЕМ НЕЙРОСЕТЕВОГО ПОДХОДА

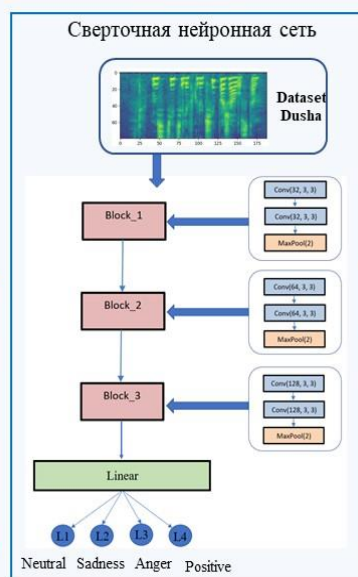
Проблема распознавания эмоций по речи

- ✓ Выбор модели эмоционального состояния человека
- ✓ Неоднозначность проявления эмоций в речи
- ✓ Недостаточность вычислительных ресурсов

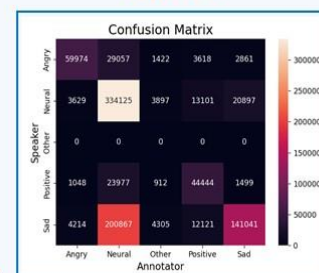
Проблемы построения SER



Решение



Результат



№	Метод распознавания	Metric	Average
1	Распознавание человеком	Precision Recall	0,7234 0,6289
2	CNN (VGG based)	Precision Recall	0,7521 0,7513



Introduction

The investigation of emotional manifestations of oral speech is one of the most complicated problems of modern humanities – not only linguistics itself, but also neurolinguistics, psycholinguistics, and, finally, cognitive science. This is due to several reasons: 1) the complexity of attributing a certain emotion as not always a clearly expressed mental phenomenon; 2) the integrated nature of its transmission by paralinguistic means as through the characteristics of the voice (laryngophony and pitch), speech rate, timbre, pausation and accentuation; 3) the multiplicity of methods and approaches to study emotionality, due to the above-mentioned “borderline” nature of the object of study (Balabanova, Abramov, 2023; Velichko et al., 2022; Santos et al., 2021; Chen et al., 2012).

Undoubtedly, the issue of determining a person’s emotional state does not only have a serious theoretical but also an obvious applied significance. Emotions of different modalities and degrees of intensity underlie the motive of any human activity, but assessing the emotional state of a linguistic personality is especially important for the areas of life activity that have a subject-object nature, where the main object and, simultaneously, the subject is a person, for example, for medicine, pedagogy, psychology, criminology, marketing, etc. Determining the emotional state can be used for various purposes in long-term communication: in medicine – for diagnosing and treating mental disorders in the early stages, in psychology – for studying emotional reactions to certain stimuli and situations, in pedagogy – for analyzing the emotional state of schoolchildren and students and optimizing the educational process. It is also important to study emotions in “one-time” communication: in criminology, for example, – to identify false testimony and deception, in marketing and advertising to determine the consumer’s reaction to a product or advertising campaign (Balabanova et al., 2023; Dellaert et al., 1996).

There is a serious need in the above-mentioned cases to increase the emotion recognition process and automate it, while the use of various machine learning methods and image recognition algorithms can make the process more accurate and efficient. The purpose of this work is to develop a neural network that allows to recognize human emotions from a speech signal. It is worth noting that there is an abundance of research in this area abroad but in Russia it is limited to a small number of such articles. When Russian researchers conduct work on recognizing emotions from a speech signal, English-language datasets are usually used. This article sets the task of training a neural network to recognize emotions from Russian and English speech, which allows to study a wider scope of application of the proposed solution in comparison with monolingual systems.

1. On the model of human emotional state

Speech emotion recognition (SER) is the determination of a person’s emotional state based on a speech signal without taking into account the semantic content. A person in the process of communication solves this problem quite effectively. However, at present, automatic classification of the speaker’s emotional state based on a speech signal is still relevant in various studies (Dvoynikova, Karpov, 2020; Fedotov et al., 2018; Shakhovskiy, 2009).

One of the key points in creating SER is the choice of a model of a person’s emotional state. Currently, psychologists have developed many classifications of human emotions (Gorshkov, Dorofeev 2003; Grimm et al., 2007; Maysak, 2010).

Many scientists believe that the diversity of human emotions can be adequately represented by the model of emotions developed by James Russell (Russell et al., 2005).

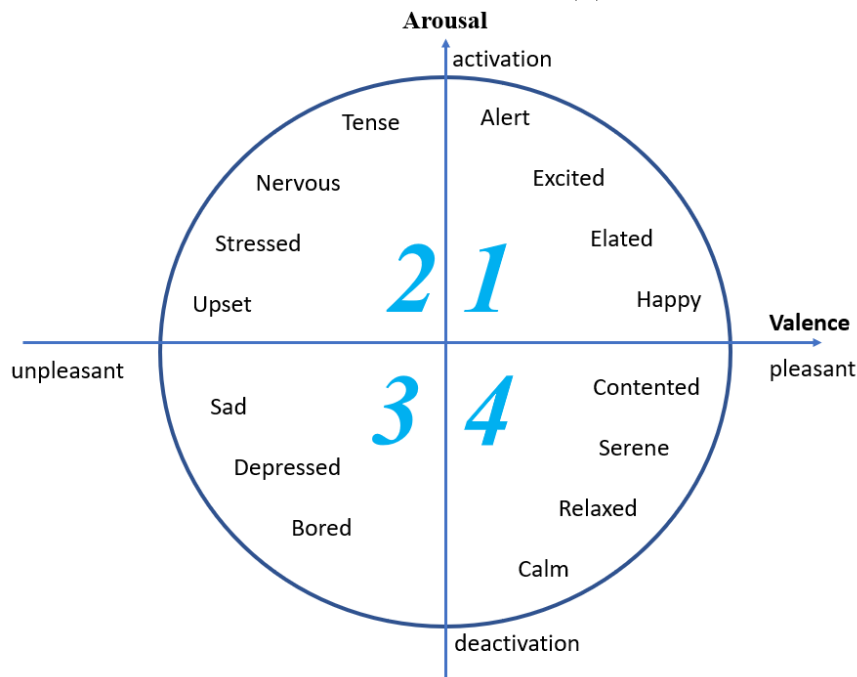
James Russell developed a model based on subjective feelings. He used a statistical method to group emotion ratings based on

positive correlations – essentially grouping similar words about emotions in a circle. This multidimensional scaling analysis revealed two bipolar dimensions – valence and arousal (Russell et al., 2005).

Thus, any emotion can be described by using the unpleasantness/pleasantness dimension (valence) and the high/low arousal dimension. One of the variations of Russell’s model is shown in Figure 1 (Russell et al., 2005).

Figure 1. J. Russell’s model of the human emotional state

Рисунок 1. Модель эмоционального состояния человека Д. Рассела



Russell’s (1980) model proposes that valence and arousal are independent bipolar dimensions. Independence means that valence and arousal are uncorrelated. Bipolarity means that opposite emotion terms represent each of the opposite poles of valence and arousal. For example, in Figure 1 above, “happy” and “sad” are shown at the opposite poles of the valence dimension. Similarly, Figure 1 shows that “excited” and “bored” are shown at the opposite ends of the arousal dimension (Russell et al., 2005).

Thus, a person cannot be excited and bored at the same time. Finally, according to this model, mixed emotions are similar in their subjective experience.

Therefore, a mixed emotion cannot consist of feelings that differ greatly in valence or arousal, such as happiness and sadness. In Figure 1, mixed emotional

experiences are emotions that are located next to each other in the same quadrant.

2. On the algorithm for recognizing human emotions

The generalized algorithm for recognizing human emotions from a speech signal, shown in Figure 2, includes the following stages:

- Pre-processing of the audio signal: before starting to analyze the speech signal, it is necessary to pre-process the data, such as noise filtering, volume normalization and feature extraction from the audio file, etc.

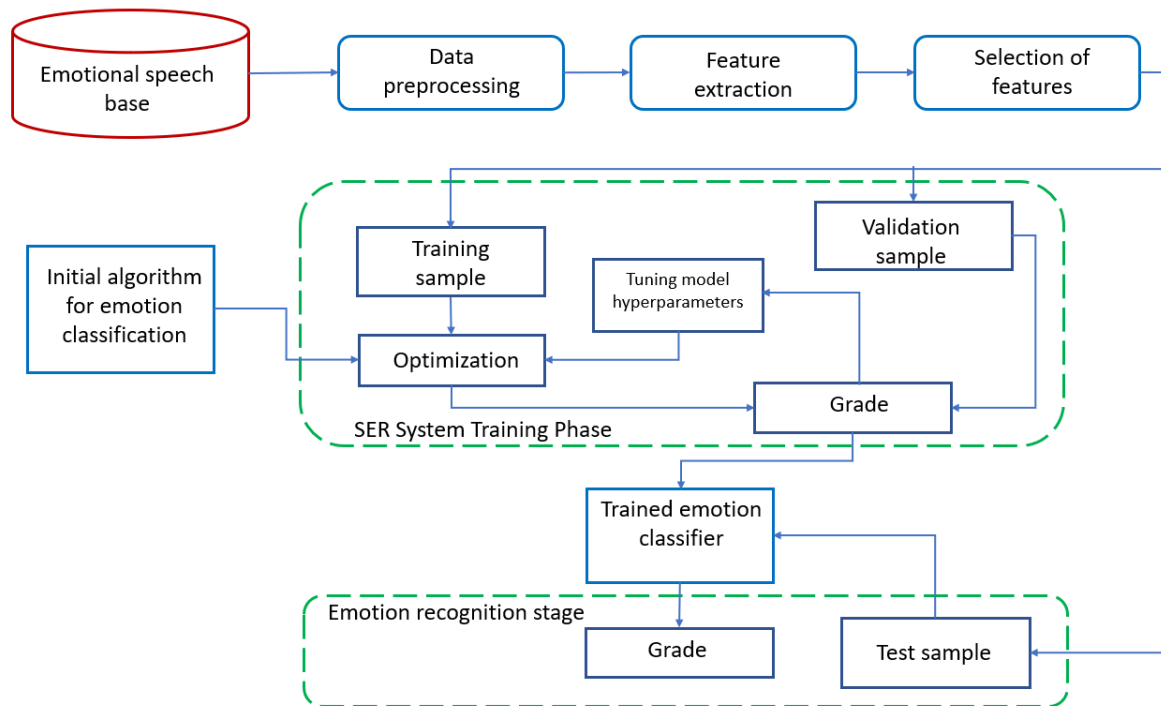
- Feature extraction: to analyze a person’s emotional state from speech, various features are used, such as voice frequency, intonation, speech rate, phrase duration and others. These features can be extracted using special signal processing algorithms.

- Machine learning: at this stage of emotion recognition, machine learning methods are used to build a classifier.
- Emotion assessment: having a training model, it is possible to recognize the

speaker's emotional state from a speech signal based on feature analysis.

- Interpretation of results: it involves analyzing the results and using them for a specific task. For example, in control systems or improving human-machine interaction.

Figure 2. Generalized algorithm for recognizing emotions from speech
Рисунок 2. Обобщенный алгоритм распознавания эмоций по речи



In general, speech emotion recognition systems are analyzed from the point of view of pattern recognition in three areas (Lemaev, Lukashevich, 2024):

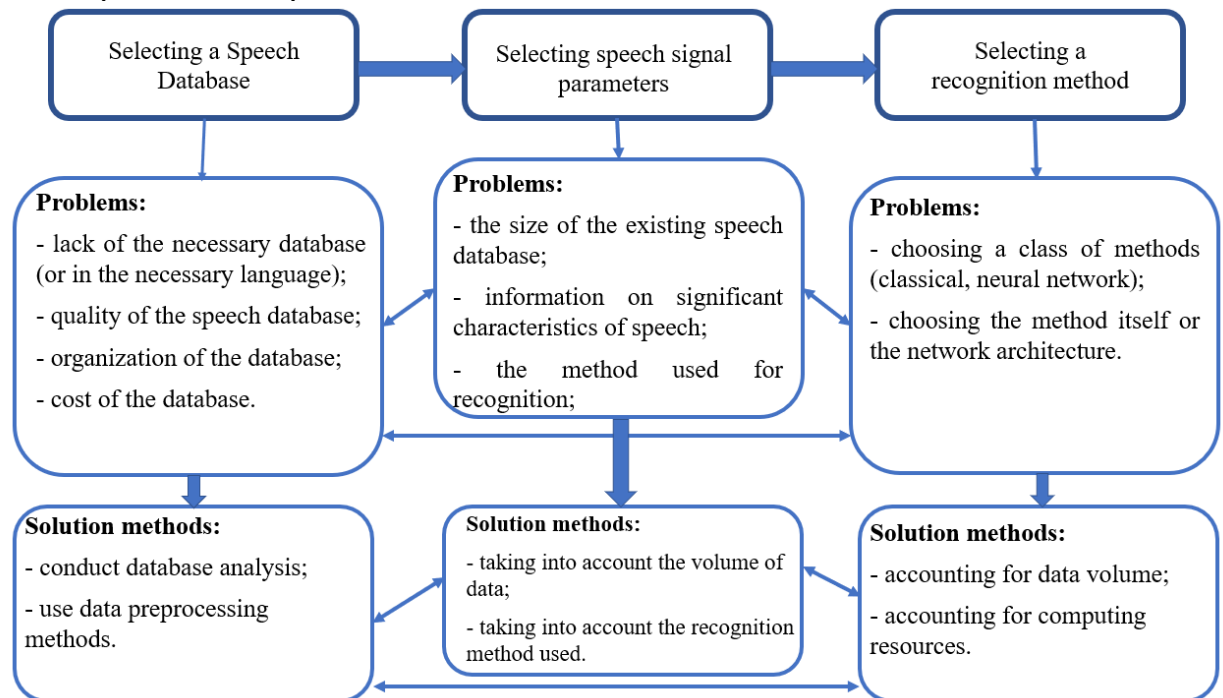
- 1) selection of an emotional speech database,
- 2) extraction of effective features,
- 3) development of reliable classifiers using machine learning algorithms.

However, it should be noted that data preprocessing algorithms can also

significantly increase the quality of SER work.

The quality of solving the problem of emotion recognition by speech largely depends on the correctness of the approach to solving each of the above stages. However, the solution to each of the above subtasks is associated with a number of problems. Schematically, the problems of each stage and the ways to solve them are shown in Figure 3 (Fedotov et al., 2018; Dvoynikova, Karpov, 2020; Velichko et al., 2022).

Figure 3. Problems of SER construction
Рисунок 3. Проблемы построения SER



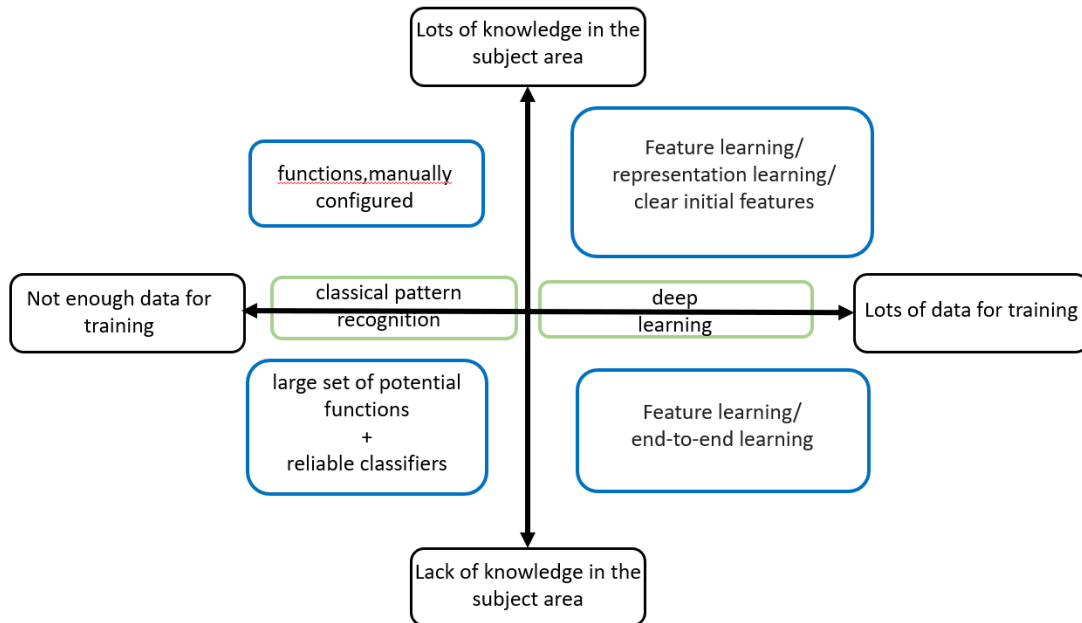
When choosing a dataset for training SER, it is necessary to take into account a number of factors. First of all, it is the quality of speech data in emotional speech. However, it should be noted that SER trained on a monolingual dataset will not provide high-quality recognition of speaker emotions in another language. Another aspect to be considered is the choice of speech parameters when constructing SER. Also, the choice of speech parameters is directly related to the choice of the classification method and vice

versa. Just as the choice of the classification method, the choice of speech parameters is directly related to the amount of data in the dataset used.

Figure 4 schematically shows the strategy for choosing an approach to constructing SER depending on the availability of representative data for training and knowledge of the subject area (Fedotov et al., 2018; Dvoynikova, Karpov, 2020; Velichko et al., 2022).

Figure 4. Basic SER development strategies

Рисунок 4. Основные стратегии разработки SER



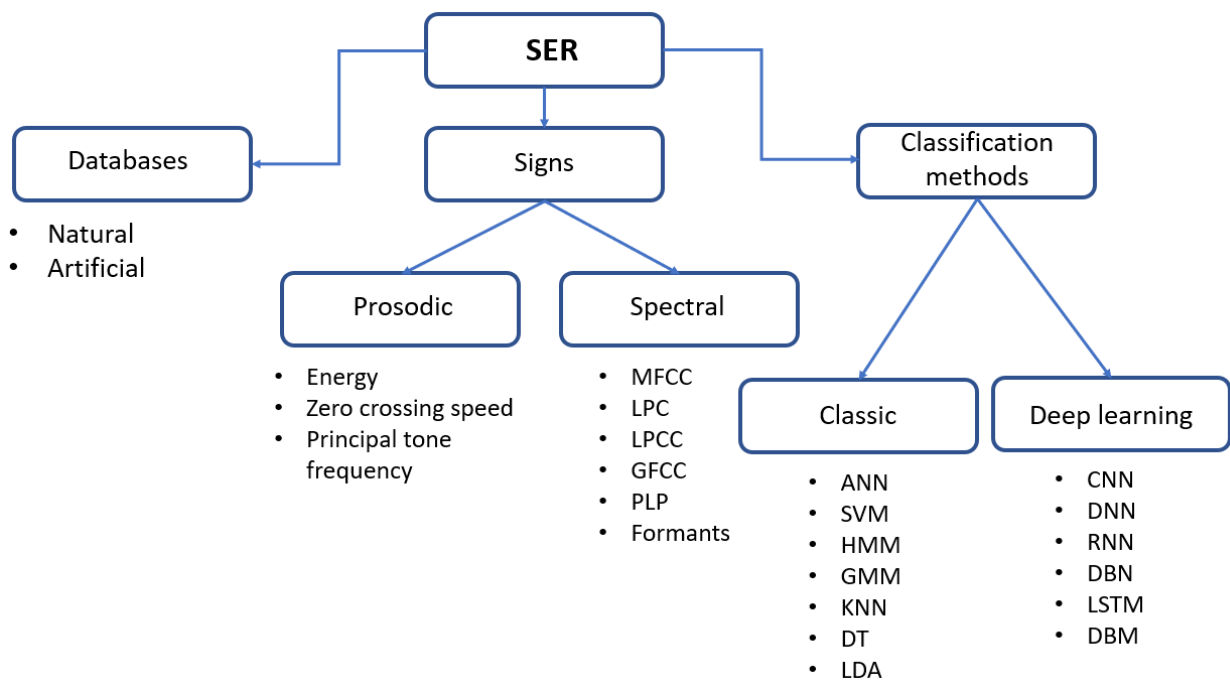
When building systems for recognizing paralinguistic phenomena in general and recognizing a person’s emotional state from speech in particular, the quantity and quality of the data, that can be used for training the model, play a major role. Using a dataset of emotional speech with a relatively small amount of data makes the use of neural networks impractical, since high-quality training of a neural network requires a much larger amount of data in relation to training models built by using classical classifiers. On the other hand, using a large amount of data for training does not imply the use of classical methods due to their large computational power. Another factor is the quality and quantity of the features of the speech signal used to recognize emotions. Thus, using a large number of features, especially if they contain those that do not carry significant information for recognizing emotions, is impractical when using classical classification methods, since unnecessary irrelevant information leads to a decrease in the quality of a classifier. On the contrary, provided there are functions that clearly reflect the features of emotions in speech, the use of classical classification methods is recommended. The neural network approach has proven itself to

be effective in solving problems of paralinguistic speech analysis, since it demonstrates its effectiveness in solving problems when it is difficult to find an exact solution. Thus, in the presence of a large volume of emotional speech, deep neural networks can be used both to search for useful representations of emotional speech features and to directly recognize emotions. It should also be noted that paralinguistic speech analysis in general and emotion recognition by speech signal in particular is a relatively new area of research from the point of view of applied linguistics and from the point of view of speech signal processing, both in Russian and in world science. Therefore, it is advisable to conduct research on recognizing human emotions by oral speech in various directions (Albornoz et al., 2011; Ayadi et al., 2011; New et al., 2003; Fedotov et al., 2018; Dvoynikova, Karpov, 2020; Velichko et al., 2022).

3. On emotion recognition systems

Figure 5 provides a comprehensive overview of emotion recognition systems based on speech (Dvoynikova, Karpov, 2020; Balabanova et al., 2023; Abramov et al., 2024).

Figure 5. Overview of speech emotion recognition systems
Рисунок 5. Обзор систем распознавания речевых эмоций



Thus, when creating a system that allows to recognize emotions from a speech signal, it is necessary to consider three main aspects: the choice of the emotional speech base, the choice of speech signal features and the choice of a classification method.

I. Selecting a dataset of emotional speech.

Despite intensive research in corpus linguistics in the last decade, only a few speech datasets include emotional speech (Cowie et al., 2001; Sadiq et al., 2021; Sahoo, Routray, 2016; Nogueiras et al., 2001), but most of the existing emotional speech datasets are not publicly available. Therefore, researchers have to turn to proprietary datasets for prosody, emotion recognition studies.

In particular, very few studies on emotional intonation are devoted to the Russian language (Holden, Hogan, 1993; Hozjan, Kačič 2003; Siging, 2009).

However, it appears that the expression of emotions in Russian has both universal and language-specific features. Therefore, it poses a challenge for the theory of emotion

recognition from speech (Makarova, 2000; Fedotov et al., 2018; Dvoynikova, Karpov, 2020). Thus, it seems appropriate to consider Russian-language corpora of emotional speech. Currently, there are three main corpora of Russian emotional speech: RUSLANA, RAMAS, and Dusha.

1) RUSLANA (RUSsian LANguage Affective speech) (Zeiler, Fergus, 2013).

A dataset of affective (emotional) utterances for the Russian language. The RUSLANA dataset contains recordings of 61 speakers (12 men and 49 women) who pronounce ten sentences neutrally (non-emotionally) and express the following five emotional states: surprise, happiness, anger, sadness, and fear. The average age of the speakers is 18.7 years, with a range from 16 to 28 years.

2) RAMAS (Russian Acted Multimodal Affective Set).

RAMAS is the first multimodal corpus in Russian. This dataset contains about 7 hours of high-quality video recordings of the subjects' faces and speech (Perepelkina et al., 2018).

The dataset was created by engaging 10 semi-professional actors (5 men and 5 women) in acting out interactive dyadic scenarios. Each scenario included one of the basic emotions: anger, sadness, disgust, happiness, fear, or surprise, as well as some characteristics of social interaction, such as dominance and submission (Perepelkina et al., 2018).

In order to note the emotions that the subjects actually experienced during the process, the creators of the dataset asked them to fill out short questionnaires (self-reports) after each scenario. The recordings were labeled by 21 annotators (at least five annotators labeled each scenario) (Perepelkina et al., 2018).

RAMAS is an open dataset that provides the scientific community with multimodal data on the relationship between faces, speech, gestures, and physiology. In this paper, the focus is on the speech data recording and its labeling. RAMAS contains recordings of basic emotions: anger, sadness, disgust, happiness, fear and surprise (Perepelkina et al., 2018).

3) Dusha (Lemaev, Lukashovich, 2024).

Dusha is a bimodal corpus suitable for speech emotion recognition (SER) tasks.

This dataset was created by SberDevices and is currently the largest dataset in Russian designed to solve problems of recognizing emotions in spoken language.

The dataset is divided into 2 parts: for the first part, called Crowd, the authors generated texts based on the conversations of real people with a virtual assistant which were then voiced using crowdsourcing (the speakers were given a text and an emotion with which this text should be pronounced, and then the resulting audio recordings were additionally checked by a second group); the second part, called Podcast, contains short (up to 5 words) excerpts from Russian-language podcasts, which were then classified by emotion.

In total, 5 classes of emotions are presented: anger, sadness, positive, neutral, and others.

This dataset of emotional speech is the focal point of the research as it contains both

the emotions generated by actors and those obtained in a natural environment.

II. Speech Signal Feature Selection for Emotion Detection (Kerkeni et al., 2020; Kim et al., 2017)

In fact, emotional feature extraction is the core problem in the SER system. Many researchers (Surabhi, Saurabh, 2016; Neiberg et al., 2006) have proposed important speech features that contain emotion information, such as energy, pitch, formant frequency, cepstral coefficients (LPCC and MFCC), and spectral features (Vu et al., 2021). Therefore, most researchers prefer to use a combined feature set consisting of a number of features containing more emotional information (Wu et al., 2011; Hsu et al. 2021). However, using a combined feature set may lead to high dimensionality and redundancy of speech features which complicates the training process for most machine learning algorithms and increases the likelihood of overfitting.

Thus, feature selection is necessary to reduce the redundancy of feature sizes.

Both feature extraction and feature selection can improve training performance, reduce computational complexity, build more generalizable models, as well as reduce the amount of memory required for storage.

III. Methods for recognizing emotions in a speech signal.

The key aspect of emotion recognition in speech is the selection of a classification method. Currently, many methods have been proposed, which are used both individually and in various combinations. New solutions for combining methods and neural network architectures constantly appear in open sources. However, as a first approximation, they are divided into two types in literature: classical methods and neural network methods (Uzdyaev, 2020; Wang et al., 2020).

4. Development of a neural network for emotion recognition

As noted above, the selection of an emotional speech base and its preprocessing is an important stage in building the SER system.

In this paper, the emotional speech dataset Dusha was chosen for our research.

This choice was due to several factors:

1. Since the SER system being developed is intended for Russian-speaking users, the choice of the Dusha speech dataset is obvious, as its language is Russian.

2. This emotional speech dataset contains a significantly larger amount of data compared to other Russian-language emotional speech datasets.

3. The Dusha emotional speech dataset contains two types of records: those obtained in the laboratory (actors playing out emotions) and those obtained from real emotional dialogues.

4. Dusha is a free, open emotional speech dataset.

Preprocessing of speech data for building SER.

A random sample of 201,850 audio files was downloaded from the Dusha dataset (Lemaev, Lukashevich, 2024). During the sample analysis, it was found that preprocessing is needed. The purpose of preprocessing was to obtain high-quality data for training the model, taking into account the available resources for training models. Thus, it was necessary to reduce the number of audio recordings, while leaving those that are most indicative of the recognized emotions.

After preprocessing, the data set contained audio recordings of emotional speech that met the following criteria:

1. Correspondence of the label given by the expert to the emotion ordered by the actor.

Part of the Dusha data set was obtained in the laboratory, that is, using the actors' performance with subsequent labeling by four experts which consisted of determining the emotion contained in the audio signal. Only those audio signals were used where the labels of all four experts and the label of the played emotion were matched. If at least one discrepancy was detected, the audio file was removed.

2. Audio file label presence.

Analyzing the original data set, it was discovered that some of the audio recordings contained the label "Other", meaning – they did not relate to any of the emotions under consideration (neutral, anger, joy, sadness). Those audio recordings were excluded from the resulting dataset.

As a result, the dataset size after preprocessing was 29,698 audio files. The training and test sets were divided the following way: 27,256 audio files in the training dataset and 2,442 audio files in the test dataset.

At the next stage of preprocessing, a balanced dataset was obtained. Balancing was carried out by including in the resulting data set the same number of audio files of each of the emotions under consideration. Thus, the training data set included 6,814 audio recordings of each emotion (neutral, anger, joy, sadness). In this form, the resulting data set was used to solve the problem of recognizing anger using neural network methods. The next stage of preprocessing consisted of audio files size alignment, which was 3 seconds. The choice of such a parameter value is due to the available computing resources and sufficient time to express the emotion. Reduction to a size of 3 seconds was carried out as follows:

- from audio recordings lasting more than 3 seconds, 3-second fragments from the middle of the signal were selected, since, on average, this is the fragment of the most vivid expression of the emotion;

- zero values were added to audio signals lasting less than 3 seconds to the required duration. It should be noted that during the training, it was necessary to increase the recognition accuracy by adding not zeros to shorter audio recordings but by repeating a fragment of the same recording, since the semantic component of speech was not used in the developed emotion recognition algorithm. However, it did not increase the accuracy of emotion recognition and was rejected in further studies, since adding zeros carried a smaller computational load in relation to repeating informative fragments.

Features of a speech signal in recognizing emotions using neural network methods.

In order to recognize a person's emotional state from oral speech using neural network methods, mel-spectrograms were used, which represent the energy characteristic of a speech signal over time. The Mel-scale was chosen, as it most

adequately reflected the psychophysical perception of sound by a person (Abramov et al., 2024; Balabanova, Abramov, 2023).

The parameters used to construct a mel-spectrogram are presented in Table 1.

The illustration of a spectrogram construction is shown in Figures 6 and 7.

Table 1. Parameters for constructing a mel-spectrogram

Таблица 1. Параметры при построении мел-спектрограммы

№	Parameter	Parameter value
1	Audio signal sampling frequency	16000 Hz
2	Audio signal duration	3 seconds
3	Total number of signal samples	48000
4	Number of Fourier points	1024
5	Window size analysis	1024
6	Window offset analysis	256
7	Number of mel filters	80

Figure 6. Construction of a mel-spectrogram based on a speech signal

Рисунок 6. Построение мел-спектрограммы по речевому сигналу

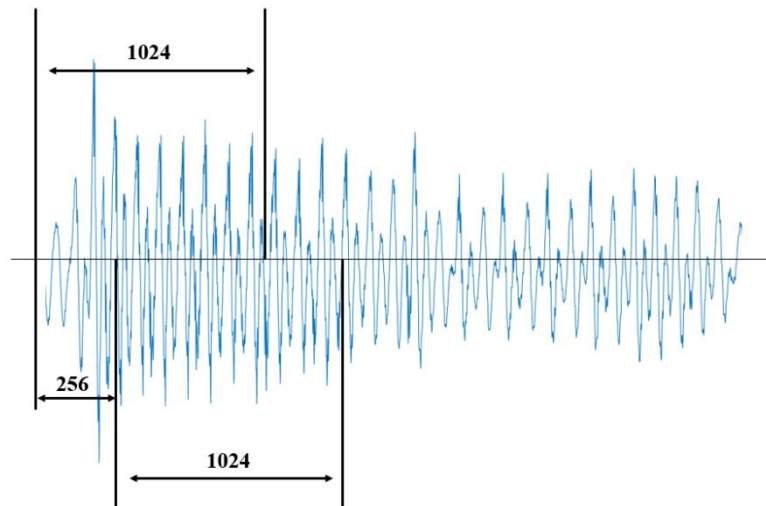
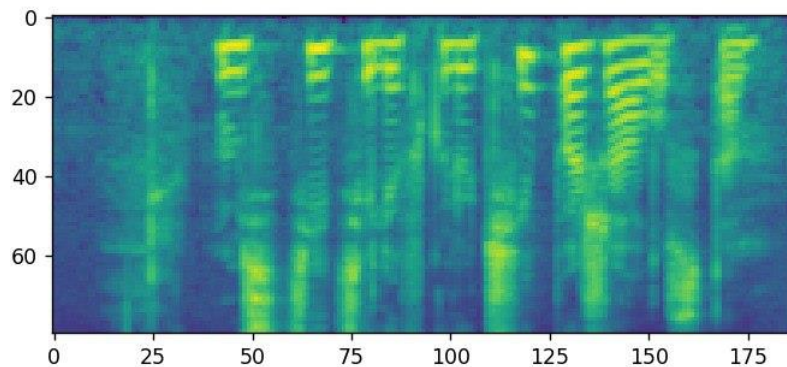


Figure 7. Spectrogram of a speech signal fragment

Рисунок 7. Спектрограмма фрагмента речевого сигнала



As a result, for each speech signal from the dataset a Mel-spectrogram was obtained, the size of which throw time was 188 (the number of 1024-size windows on 48000 samples with a shift of 256 samples) by 80 (the number of Mel filters used).

Developing and testing the neural network.

The convolutional architecture was chosen as the basic architecture of the neural network developed for recognizing the emotional state of a person by one's speech. This choice was attributed to several factors. Firstly, neural networks of the convolutional architecture require less resources for training and operation in general, which is important when using it with a limitation of computing resources. Secondly, it is planned to use a spectrogram of the speech signal as the initial data, and convolutional neural networks have

historically been created to work with images and show excellent results in solving image classification problems (Raudys, 2003).

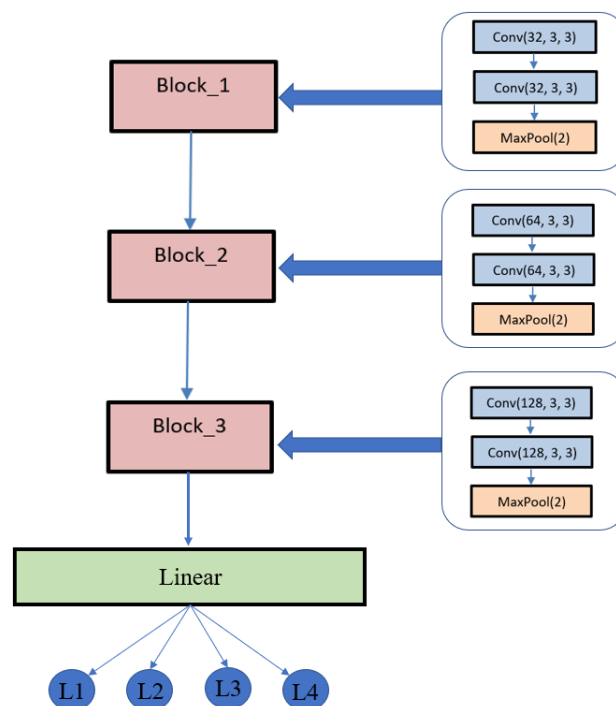
At the next stage of the experiment, it was decided to build a neural network based on the idea of VGG (Visual Geometry Group). This architecture was proposed by K. Simonyan and A. Zisserman from Oxford University in the article "Very Deep Convolutional Networks for Large-Scale Image Recognition".

The main feature of the VGG idea is the presence of a block structure in the architecture of the convolutional neural network. That is, in each block of the network there are several consecutive conv layers and one Pooling.

The network architecture is shown in Figure 8.

Figure 8. Convolutional neural network

Рисунок 8. Нейронная сеть сверточной архитектуры



Neural networks based on the VGG idea, according to the reference sources, have proven themselves well in solving the problem of classifying images from images.

Since the input data are spectrograms, this architecture was chosen.

The main elements and characteristics of the presented neural network are shown in Table 2.

Table 2. Main elements and characteristics of the neural network

Таблица 2. Основные элементы и характеристики нейронной сети

№	Element	Accepted values
1	Layers	Conv – convolutional MaxPooling – dimension reduction with preservation of the most significant features
2	Size of filters in Conv layers (kernels)	3×3
3	Padding	1 (inner indentation from the borders of the element to its contents)
4	Step	1
5	Number of feature maps at the output of the convolutional layer (feature map)	Conv 1 – 32 Conv 2 – 64 Conv 3 – 128
6	Dimensionality MaxPooling	2
7	Activation function	ReLU
8	Dropout	0,25
9	Optimizer	Adam (lr = 0,001)
10	Loss Functions	MSE
11	Duration of training	50 epochs
12	Hyperparameter batch size	64

The metrics used to evaluate the quality of the presented neural network were Precision, Recall and F_1 score. The Recall metric shows whether the algorithm can detect a given class at all. That was the reasoning used when choosing that particular metric. The Precision metric shows the proportion of objects marked positive by the neural network which are actually positive. F_1 score provides a balance between accuracy and completeness. These metrics are usually used in pairs to achieve better model evaluation during training. The calculation of the Precision, Recall and F_1 score metrics for each class was carried out using the expressions:

$$Precision = \frac{TR}{TR+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

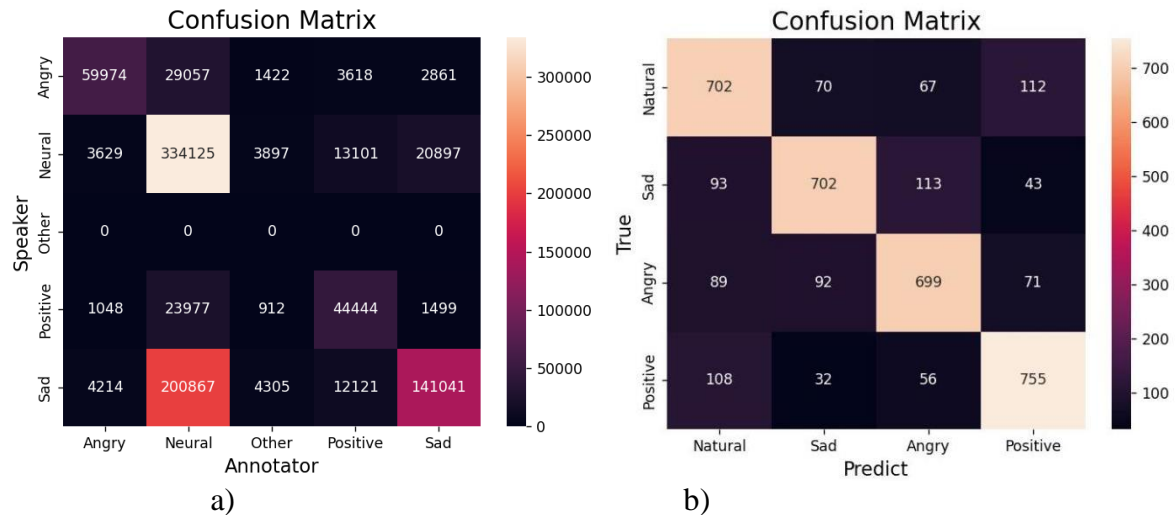
$$F_1 \text{ score} = \frac{2 \times (\text{precision} \times \text{recall})}{(\text{precision} + \text{recall})} \quad (3)$$

where TP is the proportion of positive objects correctly predicted by positive ones, FN is the proportion of positive objects incorrectly predicted as negative ones (type II error, false rejection), FP is the proportion of negative objects incorrectly predicted as positive ones (type I error, false acceptance).

In order to assess the quality of emotion recognition in speech by a neural network vs the same task done by a human, confusion matrices of the accuracy of recognition by a human and a neural network were constructed based on the Dusha emotional speech dataset (Figure 9).

Figure 9. Confusion matrix for recognizing emotions from speech by a) a human, b) a neural network

Рисунок 9. Confusion matrix по распознаванию эмоций по речи а) человеком, б) нейронной сетью



The results of precision and recall metrics for emotion recognition in speech

signal by a human and neural network are shown in Table 3.

Table 3. Comparison of the quality of emotion recognition by a human and a neural network
Таблица 3. Сравнение качества распознавания эмоций человеком и нейронной сетью

№	Recognition Method	Metric	Neutral	Sadness	Anger	Positive	Average
1	Human Recognition	Precision	0,5682	0,8481	0,8709	0,6065	0,7234
		Recall	0,8895	0,3890	0,6187	0,6183	0,6289
		F_1 score	0,69344	0,533362	0,72345	0,612343	0,672848
2	CNN (VGG based)	Precision	0,7077	0,7835	0,7476	0,7696	0,7521
		Recall	0,7382	0,7382	0,7350	0,7939	0,7513
		F_1 score	0,722628	0,760176	0,741246	0,781561	0,7517

Analyzing the precision and recall metrics, as well as the error matrices, we can conclude that the developed neural network recognizes the considered emotions almost equally. The average precision and recall metrics indicate a higher quality of emotion recognition from oral speech by the developed neural network in comparison with human results.

Thus, the use of the developed neural network for recognizing emotions from oral speech can be used in various areas of human activity: in security systems, systems for analyzing conversations with clients in call

centers, in smart home systems, in analyzing human conditions, in production, in diagnosing the initial stages of depression and other diseases, etc.

However, considering the solution to the problem of recognizing emotions from oral speech in the framework of interlingual communication, neural networks may cause some problems due to the fact that a model trained on a Russian-language dataset will produce low quality emotion recognition in another language. To investigate this issue, an English-language dataset (RAVDESS) containing emotional statements of the four

classes was used. RAVDESS contains 1,440 files recorded by 24 professional actors (12 women and 12 men). Speech emotions include expressions of calmness, joy, sadness, anger, fear, surprise, and disgust. Each expression has two levels of emotional intensity (normal, strong) and an additional

neutral expression (Hozjan, Kačič, 2003; Vu et al., 2021).

Figure 10 and Table 4 show the results of emotion recognition from English-language speech by a neural network trained on a Russian-language dataset.

Figure 10. Confusion matrix for recognizing emotions in English speech

Рисунок 10. Confusion matrix по распознаванию эмоций по англоязычной речи

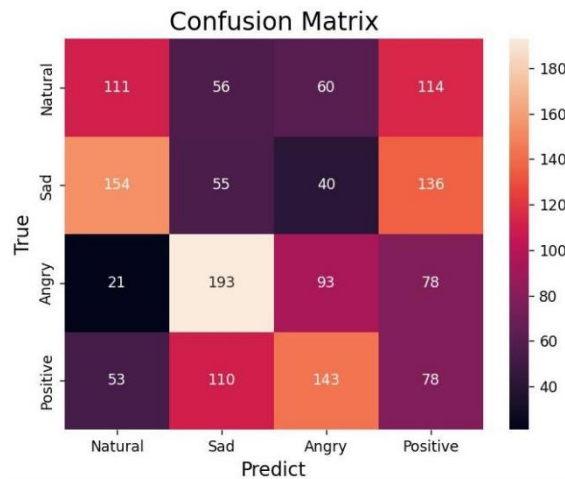


Table 4. Quality of recognizing emotions in English speech when training a neural network on a Russian-language dataset

Таблица 4. Качество распознавания эмоций англоязычной речи при тренировке нейронной сети на русскоязычном датасете

№	Metric	Neutral	Sadness	Anger	Positive	Average
1	Precision	0,3255	0,1429	0,2416	0,2031	0,2283
2	Recall	0,3274	0,1329	0,2768	0,1921	0,2323
3	<i>F1 score</i>	0,326447	0,137719	0,258005	0,197447	0,230283

The results of the experiment show that the use of a neural network trained on a monolingual dataset in interlingual communication is impractical, since the result of classifying a statement into a particular class does not exceed random distribution.

However, having trained the proposed model on the English-language dataset, the obtained emotion recognition results were close to the results of the Russian-language dataset.

The size of the total sample of the English-language dataset was 7472 statements, of which 5977 statements were used as a training sample, 1495 – a test sample. Preprocessing was carried out in the same way as with the Russian-language dataset.

The results are presented in the confusion matrix in Figure 11 and in Table 5.

Figure 11. Confusion matrix for recognizing emotions in English speech

Рисунок 11. Confusion matrix по распознаванию эмоций по англоязычной речи

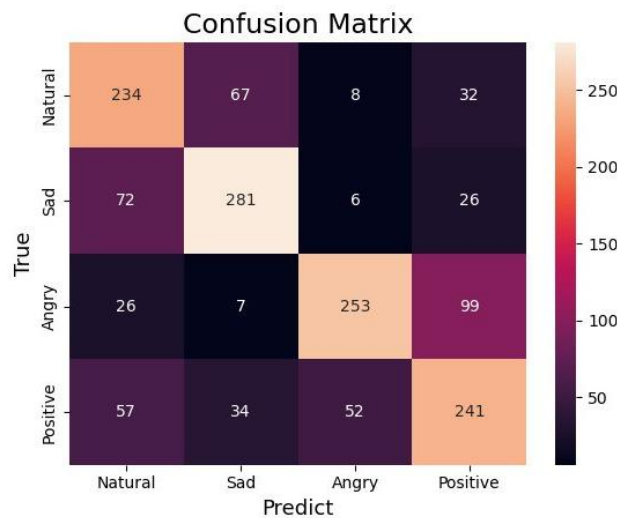


Table 5. Quality of emotion recognition in English speech when training a neural network on an English-language dataset

Таблица 5. Качество распознавания эмоций англоязычной речи при тренировке нейронной сети на англоязычном датасете

№	Metric	Neutral	Sadness	Anger	Positive	Average
1	Precision	0,6015	0,7224	0,7931	0,6055	0,6806
2	Recall	0,6862	0,7299	0,6571	0,6276	0,6752
3	F_1 score	0,641064	0,726131	0,718723	0,616352	0,677889

Thus, when using algorithms based on the neural network approach to solve paralinguistic problems in general and to recognize emotions from oral speech in particular, the issue of interlingual communication should be considered.

At the first stage, it is possible to use a neural network that recognizes the speaker's language (e.g., from Meta) as one of the solutions in interlingual communication, and then apply a version of the proposed neural network trained on the required language.

Conclusions

The article presents a neural network of convolutional architecture that allows to recognize the emotional state of a speaker directly from a speech signal without taking into account the semantic component. Particular attention is paid to the formation of a dataset for training and testing the model, since the quality of the developed SER

directly depends on the quality of the processing of this stage. Mel-spectrograms of the speech signal being used as features for recognizing emotions made it possible to increase the recognition accuracy and the speed of the neural network compared to the low-level descriptors. The current paper presents a neural network of convolutional architecture that helps to recognize four human emotions (sadness, joy, anger, neutral) from speech. Particular attention is paid to the formation of a dataset for training and testing the model, since there are currently practically no open speech datasets for dealing with paralinguistic phenomena (especially in Russian). The results of experiments on the test dataset indicate that the proposed neural network successfully copes with the task of recognizing four emotions, demonstrating high performance (about 75%) in metrics such as perception, recall and F_1 score. It is also

shown that the developed neural network can be used to recognize speaker emotions in interlingual communication.

However, it should be noted that while the proposed solution allows to identify only four emotions (Neutral, Sadness, Anger, Positive), other emotions can be recognized in a person's speech at the same time. Another obstacle to the correct recognition of emotion by speech may be a situation in which the emotion manifests itself very rapidly and lasts no more than a second. Also, the use of the presented SER is limited to two languages: Russian and English. Another factor negatively affecting the operation of the presented SER may be the presence of a noise component in the speech signal, for example, in the form of the white Gaussian noise. These aspects are the subject for further research.

References

- Abramov, K. V., Balabanova, T. N. and Gaivoronskaya, D. I. (2024). Ispolzovanie nejronnyh setej dlja raspoznavanija agressii po rechevomu signal [Using neural networks to recognize aggression by speech signal], *Information Systems and Technologies*, № 2 (142), 28–36. (In Russian)
- Albornoz, E. M., Milone, D. H. and Rufiner, H. L. (2011). Spoken emotion recognition using hierarchical classifiers, *Computer Speech & Language*, 25 (3), 556–570. (In English)
- Ayadi, M. El., Kamel, M. S. and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases, *Pattern Recognition*, 44 (3), 572–587. (In English)
- Balabanova, T. N., Abramov, K. V. (2023). Paralingvisticheskiy analiz dlja raspoznavanija agressii po rechi cheloveka [Paralinguistic analysis for recognizing aggression from human speech], *Naukoemkie tehnologii i innovacii (XXV nauchnye chtenija): Sbornik dokladov Mezhdunarodnoj nauchno-prakticheskoy konferencii, Belgorod, Belgorodskij gosudarstvennyj tehnologicheskij universitet im. V. G. Shuhova*, 697–700. (In Russian)
- Balabanova, T. N., Abramov, K. V., Boldyshev, A. V. and Dolbin, D. M. (2023). Automatic Detection of Anger and Aggression in Speech Signals, *Economics. Information technologies*, 50 (4), 944–954. DOI: 10.52575/2687-0932-2023-50-4-944-954 (In Russian)
- Chen, L., Mao, X., Xue, Y. and Cheng, L. L. (2012). Speech emotion recognition: Features and classification models, *Digital Signal Processing*, 22 (6), 1154–1160. (In English)
- Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. and Taylor, J. G. (2001). Emotion recognition in human-computer interaction, *IEEE Signal Processing Magazine*. 18 (1), 32–80. (In English)
- Dellaert, F., Polzin, T. and Waibel, A. (1996). *Recognizing emotion in speech, Proceeding of Fourth International Conference on Spoken Language Processing (ICSLP)*, 1970–1973. (In English)
- Dvoynikova, A. A., Karpov, A. A. (2020). Analiticheskij obzor podhodov k raspoznavaniju tonal'nosti russkojazychnyh tekstovyh dannyh [An analytical review of approaches to recognizing the tonality of Russian-language text data], *Informacionno-upravljajushhie sistemy*, 4 (107), 20–30. DOI: 10.31799/1684-8853-2020-4-20-30 (In Russian)
- Fedotov, D., Kaya, H. and Karpov A. (2018). Context Modeling for Cross-Corpus Dimensional Acoustic Emotion Recognition: Challenges and Mixup, *Proceedings of 20th International Conference on Speech and Computer (SPECOM-2018)*, 155–165. DOI: 10.1007/978-3-319-99579-3_17 (In English)
- Gorshkov, Yu. G., Dorofeev, A. V. (2003). Rechevye detektory lzhi kommercheskogo primeneniya [Speech lie detectors for commercial use], *Informacionnyj most (INFORMOST). Radioelektronika i Telekommunikacija*, 6, 13–15. (In Russian)
- Grimm, M., Kroschel, K., Mower, E. and Narayanan, S. (2007). Primitives-based evaluation and estimation of emotions in speech, *Speech Communication*, 49 (10–11), 787–800. (In English)
- Holden, K. T. and Hogan, J. T. (1993). The emotive impact of foreign intonation: An experiment in switching English and Russian intonation, *Language and Speech*, 36 (1), 67–88. (In English)
- Hozjan, V. and Kačić, Z. (2003). Context-Independent Multilingual Emotion Recognition

from Speech Signals, *International Journal of Speech Technology*, 6, 311–320. (In English)

Hsu, W. N., Bolte, B., Tsai, Y.-H. H., Lakhota, K., Salakhutdinov, R., Mohamed, A.-r. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 3451–3460. (In English)

Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Ali Mahjoub, M. and Cleder, C. (2020). *Automatic Speech Emotion Recognition Using Machine Learning*, Virginia Commonwealth University, United States of America. (In English)

Kim, J., Truong, K. P., Englebienne, G., Evers, V. (2017). Learning spectro-temporal features with 3D CNNs for speech emotion recognition, *Proceedings of the 7th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 383–388. DOI: 10.1109/ACII.2017.8273628 (In English)

Lemaev, V. I., Lukashevich, N. V. (2024). Avtomaticheskaja klassifikacija jemocij v rechi: metody i dannye [Automatic classification of emotions in speech: methods and data], *Litera*, 4, 159–173. DOI: 10.25136/2409-8698.2024.4.70472 (In Russian)

Makarova, V. (2000). Acoustic cues of surprise in Russian questions, *Journal of the Acoustical Society of Japan (E)*, 21 (5), 243–250. DOI: 10.1250/ast.21.243 (In English)

Maysak, N. V. (2010). Matrica social'nyh deviacij: klassifikacija tipov i vidov deviantnogo povedenija [The matrix of social deviations: classification of types and types of deviant behavior], *Sovremennye problemy nauki i obrazovanija*, 4, 78–86. (In Russian)

Neiberg, D., Elenius, K. and Laskowski, K. (2006). Emotion recognition in spontaneous speech using GMMs, *INTERSPEECH 2006 – ICSLP, Ninth International Conference on Spoken Language Processing*, 809–812. (In English)

New, T. L., Foo, S. W. and De Silva, L. C. (2003). Speech emotion recognition using hidden Markov models, *Speech Communication*, 41 (4), 603–623. (In English)

Nogueiras, A., Moreno, A., Bonafonte, A., Mariño, J.B. (2001) Speech emotion recognition using hidden Markov models, *Proceedings of EUROSPEECH 2001, 7th European conference on speech communication and technology*, 746–749. (In English)

Perepelkina, O., Kazimirova, E., Konstantinova, M. (2018). RAMAS: Russian Multimodal Corpus of Dyadic Interaction for studying emotion recognition, *PeerJ Preprints*, 6:e26688v1.

<https://doi.org/10.7287/peerj.preprints.26688v1> (In English)

Russell, J. A., Posner, J., Peterson, B. S. (2005). The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology, *Dev Psychopathol.* 17 (3), 715–34. DOI: 10.1017/S0954579405050340 (In English)

Raudys, S. (2003). On the universality of the single-layer perceptron model, *Neural Networks and Soft Computing. Physica, Heidelberg*, 79–86. (In English)

Sadiq, S., Mehmood, A., Ullah, S., Ahmad, M., Sang Choi, G., On, Byung-Won. (2021). Aggression detection through deep neural model on twitter, *Future Generation Computer Systems*, 114, 120–129. (In English)

Sahoo, S., Routray, A. (2016). Detecting aggression in voice using inverse filtered speech features, *IEEE Transactions on Affective Computing*, 9 (2), 217–226. DOI: 10.1109/TAFFC.2016.2615607 (In English)

Santos, F., Durães, D., Marcondes, F. M., Hammerschmidt, N., Lange, S., Machado, J., Novais, P. (2021). In-car violence detection based on the audio signal, *Proceedings of the International Conference on Intelligent Data Engineering and Automated Learning. Springer*, 437–445. https://doi.org/10.1007/978-3-030-91608-4_43 (In English)

Shakhovskiy, V. I. (2009). Jemocii kak obekt issledovanija v lingvistike [Emotions as an object of research in linguistics], *Voprosy psiholingvistiki*, 9, 29–43. (In Russian)

Siging, W. (2009). Recognition of human emotion in speech using modulation spectral features and support vector machines: master of science thesis, Department of Electrical and Computer Engineering Queen's University, Kingston, Ontario. (In English)

Surabhi, V., Saurabh, M. (2016). Speech emotion recognition. A review, *International Research Journal of Engineering and Technology (IRJET)*, 03, 313–316. (In English)

Svetozarova, N. D. (1982). Intonacionnaja sistema russkogo jazyka [Intonation system of the Russian language], *Izd-vo Len. un-ta*, Leningrad. (In Russian)

Uzdyaev, M. Yu. (2020). Nejrosetevaja model' mnogomodal'nogo raspoznavanija chelovecheskoj agresii [Neural network model of multimodal recognition of human aggression], *Vestnik KRAUNC. Fiziko-matematicheskie nauki*, 33 (4), 132–149. DOI: 10.26117/2079-6641-2020-33-4-132-149 (In Russian)

Velichko, A., Markitantov, M., Kaya, H., Karpov, A. (2022). Complex Paralinguistic Analysis of Speech: Predicting Gender, Emotions and Deception in a Hierarchical Framework, *Proceedings of Interspeech*, 4735–4739. DOI: 10.21437/Interspeech.2022-11294 (In English)

Vu, M. T., Beurton-Aimar, M. and Marchand, S. (2021). Multitask multi-database emotion recognition, *Proceedings of IEEE/CVF International Conference on Computer Vision*, 3637–3644. DOI: 10.1109/ICCVW54120.2021.00406 (In English)

Wang, J., Xue, M., Culhane, R., Diao, E., Ding, J., Tarokh, V. (2020). Speech emotion recognition with dual-sequence LSTM architecture, *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6474–6478. DOI: 10.1109/ICASSP40776.2020.9054629 (In English)

Wu, S., Falk, T. H. and Chan, W. Y. (2011). Automatic speech emotion recognition using modulation spectral features, *Speech Communication*, 53, 768–785. (In English)

Zeiler, M. D., Fergus, R. (2013). Visualizing and understanding convolutional networks, *Computer Vision and Pattern Recognition (ECCV 2014)*, 818–833. DOI: 10.48550/arXiv.1311.2901 (In English)

Список литературы

Абрамов К. В., Балабанова Т. Н., Гайворонская Д. И. Использование нейронных сетей для распознавания агрессии по речевому сигналу // Информационные системы и технологии. 2024. № 2(142). С. 28–36.

Albornoz E. M., Milone D. H., Rufiner H. L. Spoken emotion recognition using hierarchical classifiers // *Computer Speech & Language*. 2011. №25 (3). Pp. 556–570.

Ayadi M. El., Kamel M. S., Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases // *Pattern Recognition*. 2011. №44 (3). Pp. 572–587.

Балабанова Т. Н., Абрамов К. В. Паралингвистический анализ для

распознавания агрессии по речи человека // Научные чтения: Сборник докладов Международной научно-практической конференции, Белгород, 23 ноября 2023 года. Белгород: Белгородский государственный технологический университет им. В.Г. Шухова. 2023. С. 697–700.

Балабанова Т. Н., Абрамов К. В., Болдышев А. В., Долбин Д. М. Автоматическое обнаружение гнева и агрессии в речевых сигналах // *Экономика. Информатика*. 2023. №50 (4). С. 944–954. DOI: 10.52575/2687-0932-2023-50-4-944-954

Chen L., Mao X., Xue Y., Cheng L. L. Speech emotion recognition: Features and classification models // *Digital Signal Processing*. 2012. №22 (6). Pp. 1154–1160.

Cowie R., Douglas-Cowie E., Tsapatsoulis N., Votsis G., Kollias S., Fellenz W., Taylor J. G. Emotion recognition in human-computer interaction // *IEEE Signal Processing Magazine*. 2001. №18 (1). Pp. 32–80.

Dellaert F., Polzin T., Waibel A. Recognizing emotion in speech // *Recognizing emotion in speech, Proceeding of Fourth International Conference on Spoken Language Processing (ICSLP)*. 1996. Pp. 1970–1973.

Двойникова А. А., Карпов А. А. Аналитический обзор подходов к распознаванию тональности русскоязычных текстовых данных // Информационно-управляющие системы. 2020. № 4 (107). С. 20–30. DOI: 10.31799/1684-8853-2020-4-20-30

Fedotov, D., Kaya, H., Karpov A. Context Modeling for Cross-Corpus Dimensional Acoustic Emotion Recognition: Challenges and Mixup // *Proceedings of 20th International Conference on Speech and Computer (SPECOM-2018)*. 2018. С. 155–165. DOI: 10.1007/978-3-319-99579-3_17

Горшков Ю. Г., Дорофеев А. В. Речевые детекторы лжи коммерческого применения // Информационный мост (ИНФОРМОСТ). Радиоэлектроника и Телекоммуникация. 2003. №6. С. 13–15.

Grimm M., Kroschel K., Mower E., Narayanan S. Primitives-based evaluation and estimation of emotions in speech // *Speech Communication*. 2007. №49 (10–11). Pp. 787–800.

Holden K. T., Hogan J. T. The emotive impact of foreign intonation: An experiment in

switching English and Russian intonation // *Language and Speech*. 1993. №36 (1). Pp. 67–88.

Hozjan V., Kačič, Z. Context-Independent Multilingual Emotion Recognition from Speech Signals // *International Journal of Speech Technology*. 2003. №6. Pp. 311–320.

Hsu W. N., Bolte B., Tsai Y.-H. H., Lakhota K., Salakhutdinov R., Mohamed A.-r. Hubert: Self-supervised speech representation learning by masked prediction of hidden units // *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2021. №29. Pp. 3451–3460.

Kerkeni L., Serrestou Y., Mbarki M., Raouf K., Ali Mahjoub M., Cleder C. *Social Media and Machine Learning*. Virginia Commonwealth University, United States of America: IntechOpen, 2020. С. 96 с. DOI: 10.5772/intechopen.78089

Kim J., Truong K. P., Englebienne G., Evers V. Learning spectro-temporal features with 3D CNNs for speech emotion recognition // *Proceedings of the 7th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017. Pp. 383–388. DOI:10.1109/ACII.2017.8273628

Лемаев В. И., Лукашевич Н. В. Автоматическая классификация эмоций в речи: методы и данные // *Litera*. 2024. № 4. С. 159–173. DOI: 10.25136/2409-8698.2024.4.70472

Makarova, V. Acoustic cues of surprise in Russian questions // *Journal of the Acoustical Society of Japan (E)*. 2000. №21 (5). Pp. 243–250. DOI: 10.1250/ast.21.243

Майсак Н. В. Матрица социальных девиаций: классификация типов и видов девиантного поведения // *Современные проблемы науки и образования*. 2010. № 4. С. 78–86.

Neiberg D., Elenius K., Laskowski K. Emotion recognition in spontaneous speech using GMMs // *INTERSPEECH 2006 – ICSLP*, Ninth International Conference on Spoken Language Processing. 2006. Pp. 809–812.

New T. L., Foo S. W., De Silva L. C. Speech emotion recognition using hidden Markov models // *Speech Communication*. 2003. №41 (4). Pp. 603–623.

Nogueiras A., Moreno A., Bonafonte A., Mariño J.B. Speech emotion recognition using hidden Markov models // *Proceedings of EUROSPEECH 2001*, 7th European conference on speech communication and technology. 2001. Pp. 746–749.

Perepelkina O., Kazimirova E., Konstantinova M. RAMAS: Russian Multimodal Corpus of Dyadic Interaction for studying emotion recognition // *PeerJ Preprints*. 6:e26688v1. 2018. <https://doi.org/10.7287/peerj.preprints.26688v1>

Russell, J. A., Posner, J., Peterson, B. S. The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology // *Dev Psychopathol*. 2005. 17 (3), Pp. 715–734. DOI: 10.1017/S0954579405050340.

Raudys S. On the universality of the single-layer perceptron model // *Neural Networks and Soft Computing*. Physica, Heidelberg. 2003. Pp. 79–86.

Sadiq S., Mehmood A., Ullah S., Ahmad M., Sang Choi G., On B.-W. Aggression detection through deep neural model on twitter // *Future Generation Computer Systems*. 2021. №114. Pp. 120–129.

Sahoo S., Routray A. Detecting aggression in voice using inverse filtered speech features // *IEEE Transactions on Affective Computing*. 2016. №9 (2). Pp. 217–226. DOI: 10.1109/TAFFC.2016.2615607

Santos F., Durães D., Marcondes F. M., Hammerschmidt N., Lange S., Machado J., Novais P. In-car violence detection based on the audio signal // *Proceedings of the International Conference on Intelligent Data Engineering and Automated Learning*. Springer. 2021. Pp. 437–445. https://doi.org/10.1007/978-3-030-91608-4_43

Шаховский В. И. Эмоции как объект исследования в лингвистике // *Вопросы психолингвистики*. 2009. № 9. С. 29–43.

Siging W. Recognition of human emotion in speech using modulation spectral features and support vector machines: магистерская диссертация / Siqing Wu ; Department of Electrical and Computer Engineering Queen's University. Kingston, Ontario, Canada. 2009. С. 126

Surabhi V., Saurabh M. Speech emotion recognition. A review // *International Research Journal of Engineering and Technology (IRJET)*. 2016. №03. Pp. 313–316.

Светозарова Н. Д. Интонационная система русского языка. Л.: Изд-во Лен. ун-та. 1982. 176 с.

Уздяев М. Ю. Нейросетевая модель многомодального распознавания человеческой

агрессии // Вестник КРАУНЦ. Физико-математические науки. 2020. Т. 33. №. 4. С. 132–149.

Velichko A., Markitantov M., Kaya H., Karpov A. Complex Paralinguistic Analysis of Speech: Predicting Gender, Emotions and Deception in a Hierarchical Framework // Proceedings of Interspeech. 2022. Pp. 4735–4739. DOI:10.21437/Interspeech.2022-11294.

Vu M.T., Beurton-Aimar M., Marchand S. Multitask multi-database emotion recognition // Proceedings of IEEE/CVF International Conference on Computer Vision. 2021. Pp. 3637–3644. DOI:10.1109/ICCVW54120.2021.00406

Wang J., Xue M., Culhane R., Diao E., Ding J., Tarokh V. Speech emotion recognition with dual-sequence LSTM architecture // Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020. Pp. 6474–6478. DOI:10.1109/ICASSP40776.2020.9054629

Wu S, Falk T. H., Chan W. Y. Automatic speech emotion recognition using modulation spectral features // Speech Communication. 2011. № 53. Pp. 768–785.

Zeiler M. D., Fergus R. Visualizing and understanding convolutional networks // Computer Vision and Pattern Recognition (ECCV 2014). 2013. Pp. 818–833. DOI: 10.48550/arXiv.1311.2901

The authors have read and approved the final manuscript.

Авторы прочитали и одобрили окончательный вариант рукописи.

Конфликты интересов: у авторов нет конфликтов интересов для декларации.

Conflicts of interests: the authors have no conflicts of interest to declare.

Tatiana N. Balabanova, PhD in Technical Sciences, Associate Professor, Associate Professor of the Department of Information and Telecommunication Systems and Technologies, Belgorod State National Research University, Belgorod, Russia.

Балабанова Татьяна Николаевна, кандидат технических наук, доцент, доцент кафедры информационно-телекоммуникационных систем и технологий института инженерных и цифровых технологий НИУ «БелГУ», Белгород, Россия.

Diana I. Gaivoronskaya, PhD in Technical Sciences, Associate Professor of the Department of Information and Telecommunication Systems and Technologies, Belgorod State National Research University, Belgorod, Russia.

Гайворонская Диана Игоревна, кандидат технических наук, доцент кафедры информационно-телекоммуникационных систем и технологий института инженерных и цифровых технологий НИУ «БелГУ», Белгород, Россия.

Anna N. Doborovich, PhD in Philology, Associate Professor, Associate Professor of the Department of English Philology and Cross-cultural Communication, Belgorod State National Research University, Belgorod, Russia.

Доборович Анна Николаевна, кандидат филологических наук, доцент, доцент кафедры английской филологии и межкультурной коммуникации института межкультурной коммуникации и международных отношений НИУ «БелГУ», Белгород, Россия.