УДК 81`3; 159.9.072 http://doi.org/10.32603/2412-8562-2019-5-5-136-152
Оригинальная статья / Original Article

Speech Emotion Recognition: Humans vs Machines

Stefan Werner[1], Georgii K. Petrenko[2]✉
[1]University of Eastern Finland, Joensuu, Kuopio, Finland
[2]Saint Petersburg State Electrotechnical University, St Petersburg, Russia
✉komrad-georgy2010@yandex.ru

Introduction. The study focuses on emotional speech perception and speech emotion recognition using prosodic clues alone. Theoretical problems of defining prosody, intonation and emotion along with the challenges of emotion classification are discussed. An overview of acoustic and perceptional correlates of emotions found in speech is provided. Technical approaches to speech emotion recognition are also considered in the light of the latest emotional speech automatic classification experiments.
Methodology and sources. The typical "big six" classification commonly used in technical applications is chosen and modified to include such emotions as disgust and shame. A database of emotional speech in Russian is created under sound laboratory conditions. A perception experiment is run using Praat software's experimental environment.
Results and discussion. Cross-cultural emotion recognition possibilities are revealed, as the Finnish and international participants recognised about a half of samples correctly. Nonetheless, native speakers of Russian appear to distinguish a larger proportion of emotions correctly. The effects of foreign languages knowledge, musical training and gender on the performance in the experiment were insufficiently prominent. The most commonly confused pairs of emotions, such as shame and sadness, surprise and fear, anger and disgust as well as confusions with neutral emotion were also given due attention.
Conclusion. The work can contribute to psychological studies, clarifying emotion classification and gender aspect of emotionality, linguistic research, providing new evidence for prosodic and comparative language studies, and language technology, deepening the understanding of possible challenges for SER systems.

Key words: emotional speech, speech emotion perception, speech emotion recognition, Russian emotional speech database, emotional speech corpora, emotion classification.

For citation: Werner S., Petrenko G. K. A Speech Emotion Recognition: Humans vs Machines. DISCOURSE. 2019, vol. 5, no. 5, pp. 136–152. DOI: 10.32603/2412-8562-2019-5-5-136-152

Conflict of interest. No conflicts of interest related to this publication were reported.

Received 25.09.2019; adopted after review 10.10.2019; published online 25.11.2019

Распознавание эмоций по речи: человек против компьютера

Ш. Вернер[1], Г. К. Петренко[2]✉
[1]Университет Восточной Финляндии, Йоенсуу, Куопио, Финляндия
[2]Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В. И. Ульянова (Ленина), Санкт-Петербург, Россия
✉komrad-georgy2010@yandex.ru

Введение. В исследовании рассмотрены восприятие эмоций в речи и распознавание эмоций по речи на основании одних только интонационных свойств. Обсуждаются

теоретические проблемы определения просодии, интонации и эмоции, а также классификации эмоций. Приводится обзор акустических и перцептивных характеристик, обнаруживающихся в речи в различных эмоциональных состояниях. Также рассматриваются технические подходы к распознаванию эмоций по речи в свете последних экспериментов по автоматической классификации эмоциональной речи.

Методология и источники. Нами выбрана распространенная классификация "большая шестерка", типичная для решения технических задач, и дополнена такими эмоциями, как отвращение и стыд. В условиях акустической лаборатории была создана база данных эмоциональной русской речи. Далее мы провели эксперимент по восприятию эмоциональной речи, используя экспериментальную среду ПО Praat.

Результаты и обсуждение. Выявлены возможности кросс-культурного распознавания эмоций, так как участники эксперимента из финской и международной групп распознали около половины образцов правильно. Тем не менее, носители русского языка, судя по всему, безошибочно различают больший процент эмоций. Влияние знания иностранных языков, музыкального образования и пола участников на результаты эксперимента недостаточно ярко выражены. Нами проведен анализ наиболее часто путаемых эмоций, таких как стыд и печаль, удивление и страх, злоба и отвращение, а также случаев, когда эмоционально окрашенная речь принималась за нейтральную.

Заключение. Данная работа может внести свой вклад в психологические исследования, проясняя некоторые вопросы классификации эмоций и гендерный аспект эмоциональности; лингвистику, предоставляя новые данные для просодических и сравнительных языковых исследований; языковые технологии, углубляя понимание возможных трудностей при построении систем распознования эмоций.

Ключевые слова: эмоциональная речь, восприятие эмоций в речи, распознавание эмоций по речи, база данных эмоциональной русской речи, корпусы эмоциональной речи, классификация эмоций.

Для цитирования: Вернер Ш., Петренко Г. К. Распознавание эмоций по речи: человек против компьютера // ДИСКУРС. 2019. Т. 5, № 5. С. 136–152. DOI: 10.32603/2412-8562-2019-5-5-136-152

Конфликт интересов. О конфликте интересов, связанном с данной публикацией, не сообщалось.

Поступила 25.09.2019; принята после рецензирования 10.10.2019; опубликована онлайн 25.11.2019

Introduction. Emotional speech has deservedly attracted quite a considerable amount of attention in the last few decades. At first, the interest in the subject was mainly shown by the researchers working in the field of experimental psychology. A bit later, scholars engaged in speech technology development joined the discussion.

In the present article we are going to try and give an overview the theoretical concepts, the results of work of both groups of scientists, as well as provide some of our own observations and experimental results. The study aims to bring the existing variety of approaches together and try to answer some of the most topical questions, such as what features are informative when differentiating between emotions, whether humans or machines are better at recognising emotions in speech, and if the emotions in speech are cross-culturally recognisable.

However, first of all, we will have to answer the following question: how are speech and emotion connected? Obviously, one of the most important means of our emotional expression is speech. Lexemes are capable of conveying emotional meaning by themselves, but as it is often noted in the theory of communication, segmental information accounts for only a small fraction of meaning. A much bigger portion of meaning comes non-verbally and paraverbally, the later term comprising prosodic (intonational) phenomena.

By just looking at the basic concepts involved in the discussion of emotional speech we can already see how complex the subject, in fact, is. Both 'emotion' and 'intonation' are, unfortunately, concepts that are far from being clear. Nevertheless, a sufficient amount of work has been done in the second half of the XX century in order to better define and classify the phenomena, as well as distinguish among the existing scientific approaches to both concepts. This makes it easier for contemporary researchers, as they can just make a choice of a paradigm.

The concepts of emotion and intonation. Classification of emotions.

According to the Russian Sociological Encyclopedia : "Emotion (from the Latin 'emovere' – stir up, excite) – a specific class of psychic processes and conditions characterising the attitude of a person to the world" [1, p. 639]. The authors go on to explain that "emotion's main function is the regulation of subject's activation through evaluation of internal and external signals' importance for their living activity". According to the article, the 3 basic components of every emotion are: "1) subjective experiencing as pleasure or displeasure etc.; 2) physiological changes in blood circulation, breathing and so on; 3) observable behavioural characteristics". However, it has been noted by various scholars that emotions neither have a commonly agreed theoretical definition nor a single comprehensive classification [2–4].

The numerous existing classifications can be grouped as *discrete* and *dimensional* emotion theories [3]. The latter normally put separate emotional phenomena as points in a multidimensional space along such axes as positive – negative, active – passive, strong – weak etc. [5, 4]. The former view emotions as a number of basic psychic phenomena. The number of such emotions varies from classification to classification, but what unites all these approaches is an attempt to reveal the most primary emotional phenomena. These emotions, in turn, can combine forming more complex ones.

The term *'big six'* is commonly used in speech technology in regard to *fear, anger, happiness, sadness, surprise and disgust*, though, as Seppänen et al. point out, there is no general agreement even here. Nevertheless, as we are describing emotions to a large extent from the point of view of speech recognition, we will stick to the given classification. Another reason to do so is that the further we move away from primary emotions, the more complex, culture-bound or even individual they become.

Considering the phenomenon of intonation, we should note that in Western linguistics the term usually refers to 'speech melody' only or, in stricter acoustic terminology, F0 movement. Prosody, in contrast, is used as an 'umbrella-term' for such components of paraverbal communication as pace, rhythm, loudness and pitch. We would rather adhere to the broader understanding of intonation as in the classic definition of a prominent Russian scholar E. Bryzgunova: "Intonation is various proportion of quantitative pitch, timbre, intensity and duration changes used to express differences in meaning and emotion of utterances" [6, p. 99].

As we have seen, there is a direct link between emotional and intonational phenomena, namely the former often employ the latter to express themselves. We should now take a closer look at what particular changes in speech can we notice when a person is in one or another emotional state. This will give us a better understanding of what tasks both a human brain and a computer classifier, e. g. an artificial neural network, are confronted with while trying to decipher a speaker's emotion from speech signal.

Speech in different emotional states, Speech Emotion Perception.

It was already in the second half of the 19th century when Charles Darwin made his famous observations about speech in various emotional conditions [7, 5]. It was not until the middle of the 20th century, however, that psychologists systematised their vision of the parameters of emotional speech. P. F. Ostwald (1964) pointed out that sadness is characterised by slow, weak melody and

sighs, happiness is reflected in vivid melody, fear is connected with interrupted, downward moving melody, fury is manifested in rough articulation [8]. A slightly more formalised approach was used by C. Williams and K. Stevens [9]. According to them, fury leads to F0 rise, broader range of F0, pace and the frequency of the first formant rise; fear results in F0 somewhat lower than normal, peaks on the melodic curve and pace that is slower than normal; sadness is manifested by F0 lower than normal, small F0 range change, a rise in durations of sounds and pauses. Figure 1 provides an example of a pitch contour of Russian surprised speech plotted with Praat [10].

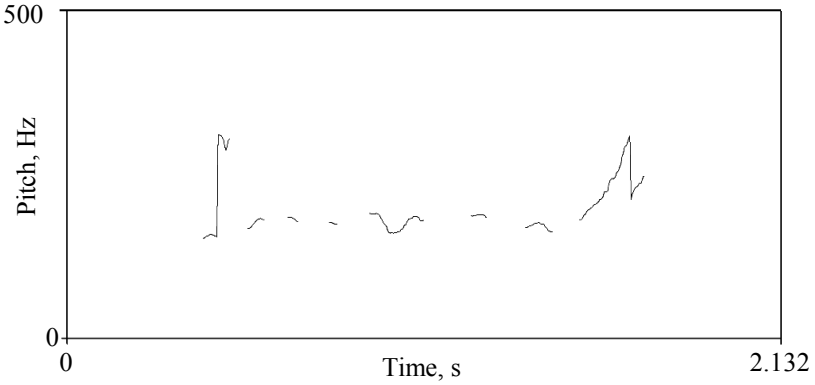

Fig. 1. Typical Russian surprised speech pitch contour

An important point to make here is that even though intonation in broader sense (prosody) appears to be an absolute language universal [11–12], intonation in narrower sense, i. e. expressive F0 movement, does not occur in all the languages of the world. For example, one of the Mexican Otomanguean languages "Itunyoso Triqui possesses a complex tonal system and does not possess either pitch accents or boundary tones" [13]. Logically enough, the expressive use of intonation in most tonal languages where tone is a means of inter-lexeme and/or inter-morpheme distinction is narrower compared to non-tonal languages.

Nonetheless, emotional expression through prosody seems to be almost an absolute universal. At the same time, many emotional speech phenomena are culture-bound as well. Otherwise, speakers of various languages would understand each other's emotions when listening to utterances making very few mistakes, which has proved wrong. It has been demonstrated, for example, that people are not as good at understanding the emotional component of speech in foreign languages as in native ones [14]. Not only are similar emotions leading to confusion, but even quite distant ones can be perplexing for both humans and machines [15, 4, 14]. This brings us to the issue of *speech emotion perception*.

A logical question arises whether native speakers are infallible at recognizing emotions expressed through speech in their language. A number of experiments carried out in the late 1980s early 1990s help to shed some light on the issue.

Manerov (1993) has shown that the main feature used by listeners in identifying speech emotion was the degree of speech-motor arousal [16]. Listeners have bigger difficulty determining the kind of emotion experienced by the speaker, than the degree of emotional arousal. The researcher also concluded from the results that the basic emotions are the easiest to distinguish, surprise and uncertainty are more difficult, whereas contempt and disgust are the hardest. Besides that, he mentioned that the accuracy of recognition is affected by the ability of speaker to express emotional states through speech and the experience of the listener [2, 16].

In Manerov's experiment listeners had to select an emotion from a set of 15 options. Some of these options, e. g. contempt, despair or spleen are often referred to as feelings. The latter are understood by some authors, such as a classic of Russian psychology Leontiev, as more sustainable psychic states that are of object-oriented nature, unlike emotions that are defined as situational states expressing a person's evaluation of current or possible future situation and of his or her activity in the situation [2, 17]. Still, the 25 participants of the experiment managed to guess 54.30% of the 'emotions' in the experiment above correctly [16].

Similar research was carried out a few years earlier by Galunov and his colleagues [18]. They claimed that in their experiments of dichotic listening of small vocal speech (singing) extracts grief, fear and fury were recognized better. The researchers set forward a theory, according to which these negative emotions have been evolutionarily more important and earlier ones, unlike joy, for example, that is why the most reliable bioacoustics means of communication were attached to them.

Vartanyan also claims that emotional modulation of sounds in humans precedes verbal communication and references a work by Klix (1983) who has shown that both the production and the perception of affective sounds are inborn. Most modern scientists acknowledge the ability of primates (and even other animal species) to communicate their emotional states to the members of their groups [19]. Darwin suggested that there is a similarity between human and primate emotions back in 1872 [7] and the guess seems to be confirmed experimentally [19]. Rozaliev (2007) states that intonation as a whole is now believed to be an earlier evolutionary feature of the humankind than the language itself [20].

As to the accuracy of the 8 participants' answers, in this case it reached a mean of 76.85 %. However, the difference to the study by Manerov is hardly surprising, as this time the subjects had to only choose one out of 5 options: joy, sorrow, neutral state, fear and fury.

Though providing some curious findings, both studies provoke a few methodological questions:

1) The size of the sample in both cases is most probably insufficient.

2) The choice of emotions by Manerov seems to be quite disputable (see above).

3) Although the collection of speech samples gathered by Manerov is impressive, as it comprised 300 speakers, all of the speakers were men.

Of course, many principles of modern speech data collection were not yet formulated at that time. Luckily for us, we could use the principles in our work to make the perception experiment we conducted a more objective and less error-prone one. Our goal was to find out about the perception of emotional speech in Russian by native and non-native speakers. This added a few more interesting dimensions to the issue of emotional speech perception (see below).

Speech Emotion Recognition.

SER, standing for speech emotion recognition, is a generally accepted term for technical approaches to recognising emotions. At first, most speech technology researchers were interested in speech recognition, but now that significant success has been reached in that field, emotion recognition has become an important additional focus for them. As Ververidis et al. [21, p. 1162) have noted, "The first investigations (in SER) were conducted around the mid-1980s using statistical properties of certain acoustic features (Van Bezooijen 1984; Tolkmitt and Scherer 1986)".

The first big achievements happened in the 1990s: "in environments like aircraft cockpits, speech recognition systems were trained by employing stressed speech instead of neutral (Hansen and Cairns 1995)" [21, p. 1162). The later studies were aimed at expanding the area of application

of SER systems to ticket reservation systems, call centre quality assurance and even diagnostic tools in medicine and psychology [21, 4]. Today SER is successfully applied in emotion monitoring and for automatic speech recognition improvement [22].

Let us now compare the systems' performance to the humans' one. Commercially available systems of today claim to have an average emotion recognition rate of about 70 %. The highest results are demonstrated by SER employing sophisticated varieties of neural networks. Khitrov et al. have managed to design and train a Support Vector Machine classifier to recognise 84 % of emotions correctly [15]. Strikingly, the distribution of results by Khitrov et al. are in line with those by Morozov, Vartanyan, Galunov et al. cited in the section above: sadness, fury and fear all overtake joy significantly [18].

It is important to mention though, that task the SVM classifier faced was rather easy, because it had to recognise only 5 emotional states: neutral, sad, scared, happy and furious. This task did not require the system to tell more similar emotions, such as shame and sadness, or disgust and anger from each other. It also did not include surprise, which is a distinct emotion that, however, seems quite dubious when recognized (see "Results and discussion").

Most probably, the same objective acoustic features are 'extracted' by both us and the machines to discriminate between emotions. Let us provide an overview of the characteristics that are commonly given attention in speech research and language technology.

Parameters commonly used for SER are:

1) amplitude and intensity (in acoustic terms – corresponding to loudness in perceptional terms);

2) changes in integral spectrum, though researchers differ on the importance of the factor;

3) F0 (fundamental frequency – or pitch) changes – commonly recognised as a crucial component of speech;

4) pace changes;

5) jitter and shimmer – voice fundamental frequency and amplitude modulation that results in a different voice quality, e. g. adding creakiness [5, 15].

There is no common agreement among speech researchers from the field of language technology on the temporal scope of the phenomena though. Some of them believe that processing the whole utterance ('turn-based processing) yields more information on the emotion. At the same time, others find the 'frame-based' approach analysing duration, height and other properties of short speech sample segments, called frames, more efficient for the SER task [23].

So far, we could see that the array of parameters used to build SER systems resembles quite strongly the set humans employ (or at least can notice and conceptualize – see above). However, more recent works show that in many cases the characteristics that bring a large amount of information are not explicit and clear to humans, for instance, cepstral characteristics (cepstrum is the energy spectrum logarithmical function) [23] and mel-cepstral characteristics [15].

All of the above-mentioned relatively simple characteristics are further processed statistically, and such values as normalised mean of spectrum, normalised time of the signal's being in the spectrum's band, linearly predicted spectrum and cepstrum, energy Teager operator and the like are chosen depending on their informativeness [23, 15, 21].

As many speech researchers agree, feature selection for automatic classification of speech samples is of paramount importance for the SER task [24]. Already in the early 2000s the number of speech features that could be derived from the relatively simple and clear parameters of signal could reach thousands. Some algorithms to limit this overwhelming variety were suggested, e. g.

by Fewzee and Karray [25] or by Semenkin et al. [26]. Eyben et al. (2010) even created an open source audio signal feature extractor – 'OpenSMILE' [27].

Historically, scholars from them field of speech technology first tried to apply high-dimensional feature sets that include a lot of acoustic parameters to try and capture all variances [22]. However, it made the learning process in most machine learning algorithms too complex and increased the likelihood of overfitting (raising one's algorithm performance greatly on a given dataset without being able to keep the level on a new sample). It also made the calculations extremely computationally expensive (Eyben, Huber, Marchi et al. 2015). As a consequence, Fayek and his colleagues offered "to investigate the application of deep learning to SER", as the method can help overcome the problem of feature selection and because "SER is an excellent test bed for exploring various deep learning architectures since the task itself can be formulated in multiple ways" [22, p. 60–61].

Standard Methods of Automatic Speech Emotion Classification.

Humans seem to be able to demonstrate emotions and perceive them since their infancy (see above). Of course, their sensibility to emotion develops through nurture and social communication too, but the machines, obviously, have to be taught to classify samples of speech by emotion from point 0. Let us briefly summarise the main stages of this learning process.

In any SER task one has to begin with speech data. Either an existing emotional speech corpus is selected, or a new one is created specifically to meet the research purposes. On the one hand, opting for an existing corpus sounds like an easy way out, however, the corpora of emotional speech are mostly only commercially available or even closed, remaining corporate secrets of speech technology companies or research groups. For instance, English emotional speech corpora on the website of the Linguistic Data Consortium, such as the ones created by M. Lieberman et al. [28] and B. Ramabhadran et al. [29], require quite a considerable license fee to get access to. The corpus of Russian emotional speech employed by Khitrov et al. [15] can serve as another example: being created for commercial applications development, it is a closed corporate tool. Yet another issue with the type of corpora in question is that as soon as a research project is finished, the corpora are often left unattended, like the Database of German Emotional Speech "EmoDB" [30] or the "RUSLANA" emotional speech databases of Russian [31], and eventually become unavailable for financial or technical reasons.

Nonetheless, when a researcher gets hold of an emotional speech corpus (either finding a free one, or buying a licence, or creating their own one), there are further steps to take. The standard procedure is described very well by E. Shriberg et al. for the similar task of dialogue acts recognition [32].

First, a training data set must be prepared, which implies labelling of the original corpus. A label is attached to every speech sample specifying the emotion type present in it. Then the training set is ready, the system is presented with it and learns what unique features accompany each label. This can involve probabilistic models, a perceptron classifier, Hidden Markov Models, or a different statistical learning mechanism [33, 34]. Recently all the other classifiers have been overtaken by deep neural networks (DNNs), which can be either programmed from scratch or downloaded freely as libraries in different programming languages.

After a system has been trained, the next step is testing it on the experimental data, a speech dataset it has never come across before. On this stage the scholars can check if the training has been successful and the system is robust enough to deal with unfamiliar data without a drastic drop in performance compared to the one on the training set.

Once the test set has been classified, it is checked against 'a key' for performance evaluation. In case researchers are not satisfied with the performance they usually change the properties of the system: the number of hidden layers, the number of neurons etc. – and reiterate. If the performance was satisfactory, an analysis of features can be conducted to try and understand, which features (F0, pausing, duration, energy or something more abstract and implicit – see above) actually enabled the system to make correct conclusions about specific speech exemplars.

Methodology and sources. As we announced above, we decided to conduct our own perception experiment, bearing both the technological and the intercultural context in mind. For the reasons given above, we chose to record our own speech database. We recorded *simulated emotional phrases* uttered by non-professional native speakers under laboratory conditions. Four native speakers of Russian (two men and two women) aged 28–37 were employed. The choice of non-professionals is motivated by the 'overacting' phenomenon that is often present in actors' speech which we wanted to escape [4].

Alternatively, we could have chosen to try and gather a sample of natural emotional speech, but this would have implied a few issues, such as:

1) very long recording and selection times, as emotional speech is not so common in everyday contexts;

2) as Seppänen [3, p. 2469] notes: 'the "uncontrolled" speech situation may cause the speech signal to be too weak or distorted for acoustic analysis' - and even more so for further application in a perception experiment;

3) secretly recording people raises ethical issues, whereas informing them of their being recorded inevitably changes the way people speak [35].

A seemingly easy solution could be using some recorded speech from TV shows, radio programmes and the like. However, in this case there is "no way of ascertaining the intended emotion" and "copyright problems are unavoidable" [3, p. 2469].

In case of simulated speech, an unvarying lexical context is normally used, and the same limited set of words or phrases is pronounced with different emotions [3]. This seemed more reliable to us when asking the speakers to read the text, because while reading, from our point of view, we quickly move away from our natural prosody and switch to 'reader's' or 'narrator's' tone. It can lead to either 'becoming an actor' and exaggerating or becoming more monotonous than usual.

Consequently, we decided to record only two utterances (a statement and a question) eight times for each speaker, each time with a varying emotional message, as the "big six" classification (see above) was chosen and modified to include neutral emotion and shame. The neutral emotion can serve as a useful benchmark against which all the other speech manifestations should theoretically be judged by the participants. As to shame, it is more complex, than other emotions (it seems to be a certain mixture of fear and sadness), but we believe it has a very distinct prosodic realisation in Russian, which should enable the subjects to distinguish it from all the other emotions. The tiny amount of phrases allowed for quick learning of the lexical content by the speakers, which enabled them to say the phrases rather than read them.

As a result, we recorded *a database of Russian emotional speech* consisting of 64 samples (32 statements and 32 questions). This collection, according to Gut & Voormann [35], cannot be called a phonological corpus, as it contains only a small amount of controlled speech samples, it was recorded in a sound lab and its potential application is probably limited to speech emotion recognition or perception experiments. Table 1 below summarises the main parameters of the database.

Table. Main features of the Russian emotional speech database.

Language	Speakers	Emotions	Utterance types	Utterances N	Recording conditions
Russian	4 (2 male + 2 female) non-professional natives	Neutral, sadness, joy, anger, fear, surprise, disgust, shame	Affirmation, question	64	Sound lab

In the course of preparation, we excluded all the 'emotionally charged' lexemes from the sentences so that neither the speakers, nor the listeners in the later part of the experiment would distort the emotion of the utterance or its perception according to this charge, as explained by El Ayadi et al. [4]. It was also crucial to make the sentence easy to pronounce phonetically and natural to say with different kinds of emotion. The resulting statement was, "Я буду приходить каждый день" [ja 'budu prjixə'ditj 'kaʐdʲij djenj], – 'I will come every day'. The question was formulated as, "У Вас завтра есть время?" [u vas 'zavtrə jestj 'vrjemjə], – 'Do you have time tomorrow?'

The speech samples were recorded in a sound-proof laboratory using a professional recorder Marantz PMD670 that does not produce any noise itself. The only noticeable noise in the recordings is created by the ventilation system which, unfortunately, was irremovable. However, this low frequency noise is relatively quiet, thus, it does not hinder listening to the stimuli. The noise can even make the samples sound more natural to the listeners, as opposed to absolutely 'clean' recordings.

In the next part of our work a perception experiment was conducted. Praat software's experimental interface (ExperimentMFC) was employed for the purpose [10]. A script was written that started speech database samples as soon as a subject has listened to the previous sample and given his/her reply. The database was randomised and each stimulus was repeated twice to ensure accuracy and intrasubject consistency. A training part was added to acquaint the participants with the layout of screen buttons and the nature of the task itself, as well as check the sound and mouse settings of the computer.

A total of 61 person participated in the experiment. The three groups of participants compared for some of the hypotheses below were native speakers of Russian (20), native speakers of Finnish (20) and a mixed international group (21). Part of the data was collected at the University of Eastern Finland, another part was gathered at Saint Petersburg State Electrotechnical University, and yet another one was received via email.

The participants were appropriately instructed and provided with Praat software, the sound and experiment files, the questionnaire and the consent form. The questionnaire included questions regarding subjects' educational, linguistic and musical background, age and gender. Questionnaires and consent forms were not in any way associated with each other, so the study remained anonymous. Questionnaires were marked with a number, so that they could be matched with the Praat experiment result files, which were saved with the same numbers.

The experiment was made in quiet computer or apartment rooms where no serious disturbances occurred. When data were gathered in distance mode, such conditions were highly recommended.

The experiment consisted of 2 repetitions of each of the 64 samples, therefore, 128 decisions had to be made by the subjects. The design of the experiment allowed for 2 breaks to be taken after every 50 stimuli. 'Replay' and 'change answer' buttons were not provided, as we were mainly interested in the spontaneous reaction of the subjects. Below, in Figure 2, you can see the layout of our experiment's window.

We wanted to see what percentage of samples was going to be recognized correctly across all three groups, as well as prove or challenge the seven hypotheses below:

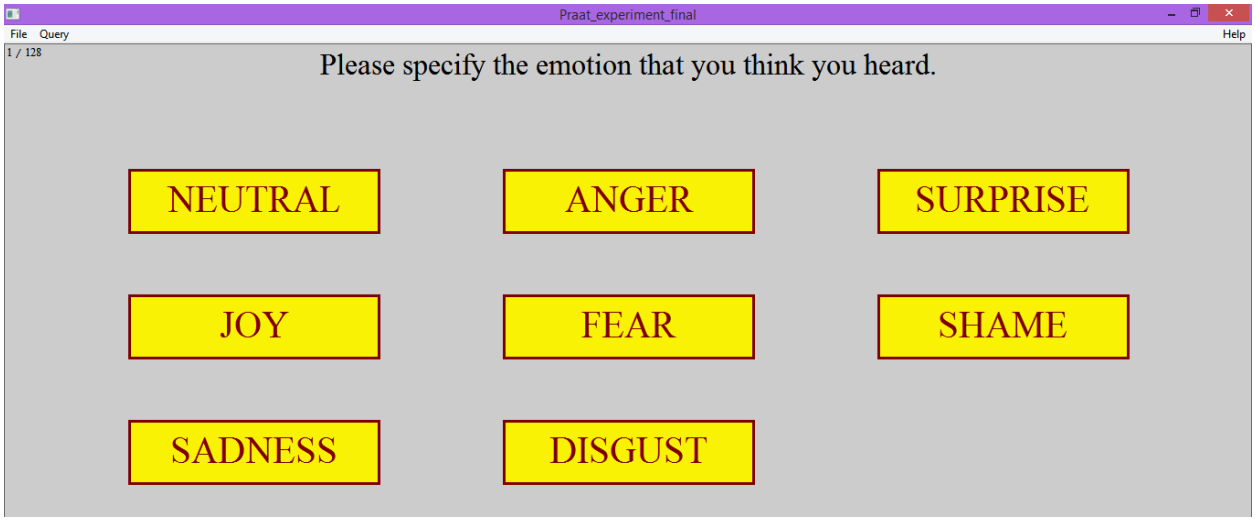

Fig. 2. Praat ExperimentMFC window as seen by experiment paticipants

1) *Speech emotions are cross-culturally recognisable. Even 'intonationally distant"cultures, such as Finnish and Russian, can recognise each other's emotions in speech using prosodic clues.* If the hypothesis proves true, this would be an argument for the universality of prosodic emotion expression. If it does not, this would imply that speech emotions may be more culturally specific than it has been previously thought.

2) *Native speakers and foreigners who have studied some Russian will be better at recognising speech emotions in Russian than those who have had little exposure to the language or none before.* Normally, foreign language acquisition and contact with speakers of a language lead to better production, among other things, on the level of suprasegmental phonology. However, as it is commonly noted (see 'Results and Discussion') prosodic transfer is one of the biggest difficulties for a language learner to overcome.

3) *Emotions in interrogative Russian sentences will be more poorly recognised by Finnish speakers than the ones in statements, as interrogative intonation is expressed in the two languages in very different ways* (Finnish intonation tending to demonstrate no rising pitch at the end of a question (Intonation Systems 1998) [36].

4) *Aurally and acoustically more similar emotions are more prone to be confused, e. g. shame vs sadness, anger vs disgust*. It appears logical that emotions sharing both valence and arousal should be more difficult to distinguish. However, as some studies have demonstrated (see above), some seemingly dissimilar emotions might often be confusing as well.

5) *Subjects speaking a larger number of foreign languages at higher levels will recognise speech emotions better*. Our idea behind this hypothesis is that extensive foreign language education may have an effect on perceptivity of subjects to intonation in the broad sense, as well as the emotional component therein.

6) *Female subjects might be more sensitive to emotion in speech than male ones recognising a larger amount of stimuli correctly*. This hypothesis aims to check whether the wide-spread idea of women being more sensitive to emotions of others and more emphatic is going to be confirmed or not. "While many studies have supported such a gender difference, some have found no difference, or that the female advantage occurs only under limited conditions" [37, p. 228].

7) *Subjects who have had musical training and/or practice music regularly will recognise speech emotions better*. Music is often referred to as the language of emotions, which makes us

wonder whether regular musical activity or boader music education of some subjects will be reflected in their performance in our experiment.

Results and discussion. The mean accuracy of recognition by the participants in our experiment reached 52.30 %. The relatively low recognition rate might partly be explained by higher complexity of the task we asked our participants to perform: both the number of emotions and the necessity to recognise them in a foreign language (for 2/3 of the subjects) might have contributed therein.

Let us also briefly summarise the results obtained for each hypothesis:

1) The 1st hypothesis has been proved: emotions seem to be cross-culturally recognisable. The average correct emotion recognition level was 49.38 % for the Finnish group and 47.43 % for the international group. Here the variation is statistically insignificant (two sample $t(39) = 1.01$, $p = 0.31$).

However, the probability of chance guessing is only 12.50 % in each trial for each subject. That is why we can make a positive conclusion regrading the possibility of cross-cultural speech emotion recognition. This once again proves that emotions, though not completely international, are to a large extent a universal phenomenon in human speech.

2) This hypothesis has partly proven correct. It has come as no surprise that native speakers of Russian have shown a higher average rate of correct recognition: 60.07 %. The difference with the levels above is statistically significant ($t(38) = 6.91$, p $p < 0.001$ for the Russian and Finnish groups, and $t(39) = 6.80$, $p < 0{,}001$ for the Russian and international groups). However, quite unexpectedly, there seems to be no correlation in the Finnish and international groups between the recognition level and the knowledge of Russian or exposure to it.

3) At first sight, the hypothesis has proved true. Finnish speakers, indeed, recognized a larger number of emotions correctly in statements compared to questions (52.19 % vs 46.56 %). Nevertheless, the same tendency was observable in the international group (50.45 % vs 44.42 %). Furthermore, in the native Russian group emotions in questions were recognized more precisely than in statements (61.33 % vs 58.83 %).

The results also raise the question of whether better recognition of questions by Russian native speakers is a chance difference owning to the size of sample, or is it an overall tendency for speakers of all languages to pay more attention to emotions in questions, as they imply more interaction on the part of the listener.

In any case, bearing the first two groups' performances in mind, we can hardly conclude that the better recognition of the affirmatives' underlying emotion is owing to the interrogative intonation differences between Russian and Finnish, as we originally expected. However, what indirectly points at the role of interrogative intonation pattern differences for this part of the experiment is the fact that 80.84 % of all the misattributions of samples to surprised speech happened in interrogative sentences (481 out of 595) and out of those only 21.62 % (or 104) of mistakes were made by native speakers of Russian (who amounted to almost a third of all the subjects). Thus, it looks like for native speakers interrogative intonation is more clearly distinct from surprised one, while for foreigners these phenomena remain dubious, as the interrogation in their languages can be expressed by means of different intonational patterns.

The situation described correlates with common observations from the field of foreign language teaching, where it is noted that correct intonation patterns are normally mastered among the last of all other skills. It is also stated that intonation interference or transfer (the phenomenon of using native tonal patterns in a foreign language) is one of the hardest to overcome for language learners. H. Palmer already back in 1924 noted that even if a non-native speaker is perfectly fluent

in all the other components of a language, the use of alien intonation patterns will immediately uncover his/her being a foreigner, and often even lead to poor understanding by the natives [38].

4) This seemed to be valid too. Our experimental data have shown that the most commonly confused pairs of emotions have been: shame – sadness, surprise – neutral, shame – neutral, surprise – fear, anger – disgust. We calculated the confusions in both directions, i. e. for the first pair, for example, it was both the cases when shame was mistaken for sadness and sadness was mistaken for shame.

It appears that the participants commonly employed two different strategies in situations when an emotional speech sample was confounding: to either choose a neutral emotion, or to choose the closest similar one. The only big surprise here is that fear and surprise turned out to be a common confusion pair. It does not seem illogical from the aural perception point of view though, as both emotions imply raising of the pitch and loudness (i. e. F0 and intensity in acoustic terms). However, the emotions should seem quite dissimilar if one analyses the amplitude and voice quality. Still, as Manerov has noted (see above), in perception experiments it is much easier to attribute a speech sample to a high or low degree of arousal than to figure out the valence of emotion (positive or negative).

As to the other pairs, we have pointed out earlier that shame and sadness are related, so it was quite predictable for them to be mixed up, even though there is a notable specificity in the way shame is demonstrated in Russian. Our guess regarding the latter is challenged by almost the same proportion of such confusions in the responses of our native Russian participants (29,41 % out of all mistaken samples for this pair of emotions).

Finally, anger and disgust are of similar emotional nature as well, sometimes implying each other, but disgust is characterised by creakier voice quality, which probably becomes explicit at least partly through the disgusted facial expression. The voice quality and the disgusted grimace are quite likely to be evolutionarily related too.

5) Apparently, there is no correlation between the amount of foreign languages spoken by our subjects and the performance in the experiment. At the same time, perhaps, there is a rather weak correlation between the level at which a person speaks his/her foreign language(-s) and the amount of samples recognised correctly. In the group with Upper-Intermediate or Advanced general level the rate of recognition was 53.54 %; in the group speaking a foreign language(-s) at the Intermediate level the rate amounted to 52.39 %; finally, in the group possessing Pre-Intermediate or Elementary level of fluency the rate went down to 49.36 %. Native and non-native speakers have spread among these groups quite evenly. However, as the t-test demonstrates, statistically the rates above do not differ (the comparison between the first two groups resulted in $t(45) = 0.47$, $p = 0.64$, while the last group and the first one differ slightly stronger, but still insufficiently: $t(31) = 1.78$, $p = 0.09$).

It thus becomes obvious that a larger sample has to be gathered in order to make the picture clear. The composition of the questionnaire has to be better formalized when it comes to education (language and musical) and the level of subjects' command of foreign languages. It is quite probable that subjects might have flattered themselves regarding the number of foreign languages and the level they speak them at. They might have also underestimated themselves. As a result, the data obtained from the questionnaire are not very dependable.

6) Again, as with the previous hypothesis, we face the problem of insufficient sample size. In the female group the mean rate of recognition was, indeed, higher (53.34 % vs 51.13 % in the male group). However, the two groups are not statistically dissimilar: $t(59) = 1.09$, $p = 0.28$. To be able to contribute to the emotionality and empathy discussion (see above) we would need to collect a larger sample.

7) The analysis of data, as in the case of two previous hypotheses, also did not let us make definite conclusions. As well as before, there seems to be a minor variation between the groups consisting of subjects who: a) have musical education and practice regularly; b) have musical education and practice irregularly or do not practice; c) have no musical education and do not practice. In groups (a) and (b) the rate of recognition was 52.96 % and 53.61 % respectively, whereas in group (c) the rate was slightly lower: 50.69 %. And once more, we cannot be convinced by these data for statistical reasons, even comparing the more distinct groups (b) and (c): $t(45) = 1.19$, $p = 0.24$.

Conclusion. As we have seen, emotional speech is a complex and somewhat perplexing phenomenon. The research in this area comprises a wide range of issues from theoretical definition of the basic concepts through difficult methodological choices to the interpretation of experimental results. Nevertheless, the promising prospects this interdisciplinary field offers have become a solid basis for ever growing interest in the emotional speech nature, perception and recognition. Scientists from the areas of biology, psychology, linguistics, speech technology and even music studies and neuroscience are all eager to unveil the truth behind emotional speech.

Thus, emotion recognition and perception studies can be of practical value for all of these sciences, as well as for the sphere of language education. We hope that our theoretical review will be helpful for those going to immerse themselves into the field of emotional speech studies, whereas the experimental part can be insightful for researchers looking for the implicit features of the speech, the peculiarities of its perception by representatives of different cultures, and, finally, the possible teaching techniques to overcome the obstacles in the way of foreign language learners.

We would like the studies to be continued. The perception experiment could be repeated or expanded using a stricter version of the questionnaire (see 'Results and Discussion'). The next step could be running a neural network classifier on the samples we have gathered in order to see if it is going to be more accurate than human participants. It would be interesting to try and see whether machines face the same challenges as people do recognizing certain types of emotions, as the results we have obtained and the studies by Manerov, Galunov et al. and Khitrov et al. referenced in the 'Introduction' section suggest. Astonishingly, the results also demonstrate that if a discrete approach to emotions (like the 'big six' or similar ones) is employed for speech perception and recognition experiments, humans are, actually, left behind the 'heartless' machines.

So far, this leadership of the computers is only in a narrow domain and a specific type of experiments. Most probably, people would still be better at recognising more complex emotions and integrating the results of this recognition into the communication discourse. However, with the huge leap in most spheres of AI application we are experiencing now, it would not be a great surprise to see machines working in the spheres of teaching and psychotherapy soon or, at least, replacing the call centre operators, as they are already replacing taxi drivers and cashiers around the globe.

REFERENCES

1. Osipov, G.V. (ed.) (1999), *Rossiiskaya sotsiologicheskaya entsiklopediya* [Russian Sociological Encyclopedia], NORMA-INFRA-M, Moscow, available at: http://sociologicheskaya.academic.ru/1401/ (accessed 03.11.2015).

2. Ilyin, E.P. (2013), *Emotions and Feelings*. 2nd ed., Piter, SPb, Russia.

3. Seppnen, T., Toivanen, J. and Vyrynen E. (2003), "Mediateam speech corpus: a first large finnish emotional speech database", *Proceed. of XV International Conf. of Phonetic Science*, vol. 3, Barcelona, Spain, 3–9 aug. 2003, pp. 2469–2472.

4. El Ayadi, M., Kamel, M.S. and Karray, F. (2011), "Survey on speech emotion recognition: Features, classification schemes, and databases", *Pattern Recognition*, vol. 44, iss. 3, pp. 572–587. DOI: https://doi.org/10.1016/j.patcog.2010.09.020.

5. Galunov, V.I. (2008), "On the possibility of speaker's emotional state recognition from speech", *Speech Technology*, vol. 1, pp. 60–66.

6. Bryzgunova, E.A. (1980), "Intonation", in Shvedova, N.Yu. (ed.) *Russian Grammar*, vol. 1, Nauka, Moscow, USSR, pp. 96–122.

7. Darwin, C. (1897), *The Expression of the Emotions in Man and Animals.* D. Appleton & Company, NY, USA.

8. Ostwald, P.F. (1964), "Acoustic Manifestations of Emotional Disturbance", *Disorders of Communication*, XLII, pp. 450–465.

9. Williams, C.E. and Stevens, K.N. (1972), "Emotions and speech: Some acoustical correlates", *The Journal of the Acoustical Society of America*, vol. 52, no. 4, pp. 1238–1250.

10. Boersma, P. (2002) "Praat, a system for doing phonetics by computer", *Glot International*, vol. 5, iss. 9/10, pp. 341–345.

11. Nash, R. (1968), *Intonational Interference in the Speech of Puerto Rican Bilinguals, an Instrumental Study Based on Oral Readings of a Juan Bobo Story*, Inter American Univ., San Juan, PR.

12. Svetozarova, N.D. (1982), *Intonatsionnaya sistema russkogo yazyka* [*The Intonation System of the Russian Language*], Leningrad Univ. Publishing House, Leningrad, USSR.

13. DiCanio C. and Hatcher, R. (2018), "On the non-universality of intonation: Evidence from Triqui", *The Journal of the Acoustical Society of America*, vol. 144, iss. 3, DOI: https://doi.org/10.1121/1.5068494 (accessed 15.09.2019).

14. Petrenko, G.K. and Shumkov, A.A. (2014), *Speech and Music: Points of Contact*, ETU Publishing House, SPb, Russia.

15. Khitrov, M.V., Davydov, A.G., Tkachenya, A.V., Kiselev, V.V. and Romashkin, Yu.N. (2012), "Automatic Speech Emotion Recognition Using the Support Vector Method and Gini Coefficient", *Speech Technology*, vol. 4, pp. 34–43.

16. Manerov, V.H. (1993), "*Experimental and Theoretical Foundations of Social Identification of Speaker Interpretation",* Abstract of Dr. Sci. (psychology) dissertation, The Herzen State Pedagogical Univ. of Russia, SPb, Russia.

17. Leont'ev, A.N. (1971), *Needs, Motives and Emotions*, Moscow State Univ., Moscow, USSR.

18. Vartanyan, I.A., Galunov, V.I., Dmitrieva, E.S., Zaitseva, K.A., Koroleva, I.V., Kuzmin, Yu.I., Morozov, V.P. and Shurgaya, G.G. (1988), *Vospriyatie rechi. Voprosy funktsional'noi asimmetrii mozga* [Speech perception. Functional brain asymmetry issues], Nauka, Leningrad, USSR.

19. Vartanov, A.V., Kosareva, Yu.I. (2015), "Emotions of a person and a monkey: subjective scaling of vocalizations", *Moscow University Psychology Bulletin*, vol. 2, pp. 93–109. DOI: 10.11621/vsp.2015.02.93.

20. Rozaliev, V.L. (2007), "Construction the model of emotions on speech of the person", *Izvestiya VolgGTU*, iss. 3, no. 9 (35), pp. 65-68.

21. Ververidis, D. and Kotropoulos, C. (2006), "Emotional Speech Recognition: Resources, Features, and Methods", *Speech Communication*, vol. 48, iss. 9, pp. 1162–1181. DOI: 10.1016/j.specom.2006.04.003.

22. Fayek, H.M., Lech, M. and Cavedon, L. (2017), "Evaluating deep learning architectures for Speech Emotion Recognition", *Neural Networks*, vol. 92, pp. 60–68. DOI: 10.1016/j.neunet.2017.02.013.

23. Sidorov, K.V. and Filatova, N.N. (2012), "Analysis of Signs of Emotive Speech", *Vestnik TvGTU*, no. 20, pp. 26–31.

24. Xiao, Z., Dellandrea, E., Dou, W. and Chen, L. (2005), "Features extraction and selection for emotional speech classification", IEEE Conference on Advanced Video and Signal Based Surveillance, Como, Italy, 5–16 Sept. 2005, pp. 411–416. DOI: 10.1109/AVSS.2005.1577304.

25. Fewzee, P. and Karray, F. (2012), "Dimensionality Reduction for Emotional Speech Recognition", *International Conference on Privacy, Security, Risk and Trust (PASSAT), International Conference on SocialCom, IEEE*, 03–05 Sept., 2012, Amsterdam, Netherlands. pp. 532–537. DOI: 10.1109/SocialCom-PASSAT.2012.83.

26. Brester, K.Yu., Semenkin, E.S. and Sidorov, M.Yu. (2014), "Automatic Feature Selection System for Human Emotion Recognition in Speech Communication", *Software and Systems*, no. 4 (108), available at: http://cyberleninka.ru/article/n/sistema-avtomaticheskogo-izvlecheniya-informativnyh-priznakov-dlya-raspoznavaniya-emotsiy-cheloveka-v-rechevoy-kommunikatsii (accessed 15.07.2019).

27. Eyben, F., Wöllmer, M. and Schuller, B. (2010) "OpenSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor", *Proceedings of the 18th ACM international conference on Multimedia, o*ct. 25–29, 2010, Firenze, Italy, pp. 1459–1462. DOI: 10.1145/1873951.1874246.

28. Liberman, M., Davis, K., Grossman, M., Martey, N.and Bell, J. (2002), *Emotional Prosody Speech and Transcripts LDC2002S28*. *Web Download*. Philadelphia: Linguistic Data Consortium.

29. Ramabhadran, B., Gustman, S. Byrne, W., Hajič J., Oard D., J. Scott Olsson, Picheny M. and Psutka J. (2012), *USC-SFI MALACH Interviews and Transcripts English LDC2012S05*, DVD, Linguistic Data Consortium, Philadelphia, USA.

30. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. and Weiss, B. (2005) "A Database of German Emotional Speech", *9th European Conference on Speech Communication and Technology*, Lisboa, Portugal, sept. 4–8, 2005, pp. 1–4.

31. Makarova, V. and Petrushin, V. (2002), "RUSLANA: a database of Russian emotional utterances", *7th International Conference on Spoken Language Processing, ICSLP2002 – INTERSPEECH 2002*, available at: https://www.researchgate.net/publication/221491469_RUSLANA_a_database_of_Russian_emotional_utterances/ (accessed 23.06.2018).

32. Shriberg, E., Bates, R., Stolcke, A., Taylor, P., Jurafsky D. et al. (1998), "Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?", *Language and Speech*, vol. 41 (3–4), pp. 443–492.

33. Coleman, J. (2005), *Introducing Speech and Language Processing,* Cambridge Univ. Press. Cambridge, UK.

34. Dickinson, M., Brew, C. and Meurers, D. (2012), *Language and Computers*, John Wiley & Sons Hoboken, NJ, USA.

35. Durand, J., Gut, U. and Kristoffersen, G. (2014), *The Oxford handbook of corpus phonology*, Oxford Univ. Press, Oxford, UK.

36. Hirst, D. and Di Cristo, A. (ed.) (1998), *Intonation Systems: A Survey of Twenty Languages*, Cambridge Univ. Press, Cambridge, UK.

37. Rueckert, L. (2011), "Gender Differences in Empathy", in Scapaletti, D.J. (ed.) *Psychology of Empathy*, Nova Science Publishers, NY, USA, pp. 221–234.

38. Palmer, H.E. (1924), *English Intonation with Systematic Exercises*, Heffer, Cambridge, UK.

Information about the authors.

Stefan Werner – PhD (Linguistics) (2000), Professor, University of Eastern Finland, FI-80100 Joensuu, Finland; FI-70210 Kuopio, Finland. The author of 40 scientific publications. Areas of expertise: speech technology, pathological speech, prosody. ORCID: http://orcid.org/0000-0001-5176-8114. E-mail: stefan.werner@uef.fi

Georgii N. Petrenko – Assistant Lecturer at the Department of Foreign Languages, Saint Petersburg Electrotechnical University, 5 Professora Popova str., St Petersburg 197376, Russia. The author of 4 scientific publications. Areas of expertise: language technology, prosody, emotional speech. ORCID: https://orcid.org/0000-0003-3616-427X. E-mail: komrad-georgy2010@yandex.ru

СПИСОК ЛИТЕРАТУРЫ

1. Российская социологическая энциклопедия / под ред. Г. В. Осипова. М.: НОРМА-ИНФРА-М, 1999. URL: http://sociologicheskaya.academic.ru/1401/ (дата обращения: 03.11.2015).

2. Ильин Е. П. Эмоции и чувства. 2-е изд., перераб. и доп. СПб.: Питер, 2013.

3. Seppnen, T., Toivanen, J. and Vyrynen E. Mediateam speech corpus: a first large finnish emotional speech database // Proceed. of XV International Conf. of Phonetic Science, vol. 3, Barcelona, Spain, 3–9 aug. 2003, pp. 2469–2472.

4. El Ayadi M., Kamel M. S., Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases // Pattern Recognition. 2011. Vol. 44. Iss. 3. P. 572–587. DOI: https://doi.org/10.1016/j.patcog.2010.09.020.

5. Галунов В. И. О возможности определения эмоционального состояния говорящего по речи // Речевые технологии. 2008. № 1. С. 60–66.

6. Брызгунова Е. А. Интонация // Русская грамматика / гл. ред. Н. Ю. Шведова. М.: Наука, 1980. Т. 1. С. 96–122.

7. Darwin C. The Expression of the Emotions in Man and Animals. NY: D. Appleton & Company, 1897.

8. Ostwald P. F. Acoustic Manifestations of Emotional Disturbance // Disorders of Communication. 1964. XLII. P. 450–465.

9. Williams C. E., Stevens K. N. Emotions and speech: Some acoustical correlates // The Journal of the Acoustical Society of America. 1972. Vol. 52. № 4. P. 1238–1250.

10. Boersma P. Praat, a system for doing phonetics by computer // Glot International. 2002. Vol. 5. Iss. 9/10. P. 341–345.

11. Nash R. Intonational Interference in the Speech of Puerto Rican Bilinguals, an Instrumental Study Based on Oral Readings of a Juan Bobo Story. San Juan: Inter American Univ., 1968.

12. Светозарова Н. Д. Интонационная система русского языка. Л.: Изд-во ЛГУ, 1982.

13. DiCanio C., Hatcher R. On the non-universality of intonation: Evidence from Triqui // The Journal of the Acoustical Society of America. 2018. Vol. 144. Iss. 3, DOI: https://doi.org/10.1121/1.5068494 (дата обращения: 15.09.2019).

14. Петренко Г. К., Шумков А. А. Речь и музыка: точки соприкосновения. СПб.: Изд-во СПбГЭТУ «ЛЭТИ», 2014.

15. Автоматическое распознавание эмоций по речи с использованием метода опорных векторов и критерия джина / М. В. Хитров, А. Г. Давыдов, А. В. и др. // Речевые технологии. 2012. № 4. С. 34–43.

16. Манеров В. Х. Экспериментально-теоретические основы социальной идентификации и интерпретации говорящего: автореф. дис. ... д-ра психол. наук / РГПУ. СПб.,1993.

17. Леонтьев А Н. Потребности, мотивы и эмоции. М.: МГУ, 1971.

18. Восприятие речи. Вопросы функциональной асимметрии мозга / И. А. Вартанян, В. И. Галунов, Е. С. Дмитриева и др. Л.: Наука, 1988.

19. Вартанов А. В., Косарева Ю. И. Эмоции человека и обезьян: субъективное шкалирование вокализаций // Вестн. Моск. ун-та. Сер. 14. Психология. 2015. № 2. С. 93–109. DOI: 10.11621/vsp.2015.02.93.

20. Розалиев В. Л. Построение модели эмоций по речи человека // Изв. ВолгГТУ. 2007. Вып. 3. № 9 (35). С. 65–68.

21. Ververidis D., Kotropoulos C. Emotional Speech Recognition: Resources, Features, and Methods // Speech Communication. Vol. 48. Iss. 9. P. 1162–1181. DOI: 10.1016/j.specom.2006.04.003.

22. Fayek H. M., Lech M., Cavedon L. Evaluating deep learning architectures for Speech Emotion Recognition // Neural Networks. 2017. Vol. 92. P. 60–68. DOI: 10.1016/j.neunet.2017.02.013.

23. Сидоров К. В., Филатова Н. Н. Анализ признаков эмоционально окрашенной речи // Вестн. ТвГТУ. 2012. № 20. С. 26–31.

24. Features extraction and selection for emotional speech classification / Z. Xiao, E. Dellandrea, W. Dou et al. // IEEE Conference on Advanced Video and Signal Based Surveillance, Como, Italy, 5–16 Sept. 2005. P. 411–416. DOI: 10.1109/AVSS.2005.1577304.

25. Fewzee P., Karray F. Dimensionality Reduction for Emotional Speech Recognition // International Conference on Privacy, Security, Risk and Trust (PASSAT), International Conference on SocialCom, IEEE, Sept. 03–05, 2012. Amsterdam, Netherlands. P. 532–537. DOI: 10.1109/SocialCom-PASSAT.2012.83.

26. Брестер К. Ю., Семенкин Е. С., Сидоров М. Ю. Система автоматического извлечения информативных признаков для распознавания эмоций человека в речевой коммуникации // Программные продукты и системы. 2014. № 4 (108). URL: http://cyberleninka.ru/article/n/sistema-

avtomaticheskogo-izvlecheniya-informativnyh-priznakov-dlya-raspoznavaniya-emotsiy-cheloveka-v-rechevoy-kommunikatsii (дата обращения: 15.07.2019).

27. Eyben F., Wöllmer M., Schuller B. OpenSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor // Proceedings of the 18th ACM international conference on Multimedia, oct. 25–29, 2010. Firenze, Italy. P. 1459–1462. DOI: 10.1145/1873951.1874246.

28. Emotional Prosody Speech and Transcripts LDC2002S28 / M. Liberman, K. Davis, M. Grossman end al. *Web Download*. Philadelphia: Linguistic Data Consortium. 2002.

29. *USC-SFI MALACH Interviews and Transcripts English LDC2012S05 /* Ramabhadran B., Gustman S., Byrne W. et al. (2012). Philadelphia: Linguistic Data Consortium. DVD.

30. A Database of German Emotional Speech / F. Burkhardt, A. Paeschke, M. Rolfes end al. // 9th European Conference on Speech Communication and Technology, Lisboa, Sept. 4–8. 2005. P. 1–4.

31. Makarova V., Petrushin V., RUSLANA: a database of Russian emotional utterances, *7th International Conference on Spoken Language Processing, ICSLP2002 – INTERSPEECH 2002*, URL: https://www.researchgate.net/publication/221491469_RUSLANA_a_database_of_Russian_emotional_utterances/ (дата обращения: 23.06.2018).

32. Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech? / E. Shriberg, R. Bates, A. Stolcke et al. language and speech. 1998. Vol. 41 (3–4). P. 443–492.

33. Coleman J. Introducing Speech and Language Processing. Cambridge: Cambridge Univ. Press, 2005.

34. Dickinson M., Brew C., Meurers D. Language and Computers. Hoboken, NJ: John Wiley & Sons, 2012.

35. Durand J., Gut U., Kristoffersen G. The Oxford handbook of corpus phonology. Oxford: Oxford Univ. Press, 2014.

36. Hirst D., Di Cristo A. (ed.), Intonation Systems: A Survey of Twenty Languages. Cambridge: Cambridge Univ. Press, 1998.

37. Rueckert L. Gender Differences in Empathy / in D. J. Scapaletti (ed.) // Psychology of Empathy. NY.: Nova Science Publishers, 2011. P. 221–234.

38. Palmer H. E. *English Intonation with Systematic Exercises*. Cambridge: Heffer, 1924.

Информация об авторах.

Штефан Вернер – доктор филологических наук (2000), профессор университета Восточной Финляндии, FI-80100 Йоэнсуу, Финляндия; FI-70210 Куопио, Финляндия. Автор 40 научных публикаций. Сфера научных интересов: речевые технологии, патологическая речь, просодия. ORCID: http://orcid.org/0000-0001-5176-8114. E-mail: stefan.werner@uef.fi

Петренко Георгий Кириллович – ассистент кафедры иностранных языков Санкт-Петербургского государственного электротехнического университета «ЛЭТИ» им. В. И. Ульянова (Ленина), ул. Профессора Попова, д. 5, Санкт-Петербург, 197376, Россия. Автор 4 научных публикаций. Сфера научных интересов: языковая технология, просодия, эмоциональная речь. ORCID: https://orcid.org/0000-0003-3616-427X. E-mail: komrad-georgy2010@yandex.ru