A Bimodal Approach for Speech Emotion Recognition using Audio and Text

Oxana Verkholyak[1,2,3]*, Anastasia Dvoynikova[1,2], and Alexey Karpov[1,2]
[1]St. Petersburg Federal Research Center of the Russian Academy of Sciences, St. Petersburg, Russia
overkholyak@gmail.com, {karpov, dvoynikova.a}@iias.spb.su
[2]ITMO University, Kronverkskiy Prospekt, 49, St Petersburg, Russia
[3]Ulm University, Helmholtzstraße 16, 89081 Ulm, Germany

Received: November 19, 2020; Accepted: February 9, 2021; Published: February 28, 2021

Abstract

This paper presents a novel bimodal speech emotion recognition system based on analysis of acoustic and linguistic information. We propose a novel decision-level fusion strategy that leverages both emotions and sentiments extracted from audio and text transcriptions of extemporaneous speech utterances. We perform experimental study to prove the effectiveness of the proposed methods using emotional speech database RAMAS, revealing classification results of 7 emotional states (happy, surprised, angry, sad, scared, disgusted, neutral) and 3 sentiment categories (positive, negative, neutral). We compare relative performance of unimodal vs. bimodal systems, analyze their effectiveness on different levels of annotation agreement, and discuss the effect of reduction of training data size on the overall performance of the systems. We also provide important insights about contribution of each modality for the best optimal performance for emotions classification, which reaches UAR=72.01% on the highest 5-th level of annotation agreement.

Keywords: Computational paralinguistics, Speech emotion recognition, Sentiment analysis, Bimodal fusion, Annotation agreement

1 Introduction

Emotions are psycho-physiological reactions that arise as a response to important events. They are widely recognized as key factors in decision-making processes, especially for impromptu decisions [18]. Because emotions are highly subjective in nature they are very difficult to analyze. Emotion recognition is an active research field that aims at identifying ways to quantify and measure emotional expressions in a wide variety of applications. Speech emotion recognition is a field of studying emotional expressions by analyzing acoustical properties of speech signals. Sometimes linguistic (lexical) characteristics are also considered, however emotions and other affective states are usually considered as paralinguistic phenomena, meaning the elements of communication that do not involve words. Nevertheless, linguistic cues may be important indicators of certain emotional states, and many bimodal systems (i.e. combining both acoustic and linguistic characteristics of speech) were proposed to take advantage of both channels of communication [29].

Sentiments are different from emotions in that they convey people's opinions rather than feelings. They can be of 2 types: rational and emotional [17]. Rational sentiment is based on rational reasoning and does not convey emotions, e.g. "This car is worth the price." Emotional sentiment in turn is based

Journal of Internet Services and Information Security (JISIS), volume: 11, number: 1 (February 2021), pp. 80-96
DOI:*10.22667/JISIS.2021.02.28.080*

*Corresponding author: St. Petersburg Federal Research Center of the Russian Academy of Sciences, Tel: +7-(812)-328-70-81, Web: http://hci.nw.ru/en/employees/5

on emotional attitudes and is usually used to directly convey the emotions, e.g. "I am scared," or "This makes me angry." When conducting sentiment analysis, usually three categories are considered: positive, negative and neutral.

In this paper we propose an effective bimodal speech emotion recognition system based on analysis of emotions and sentiment via acoustic and linguistic characteristics of speech. The following sections present proposed approach in detail. Section 2 provides an overview of the related methods for acoustic and linguistic modelling of emotions and sentiment, as well as multimodal fusion approaches. Section 3 reveals the proposed approach for audio-based, text-based and bimodal speech emotion recognition. Section 4 describes the data, experimental setup used in this study, as well as reveal the obtained results. Section 5 provides the discussions about the important experimental findings, and Section 6 draws the final conclusions.

2 Related work

Although a lot of acoustical features were recently proposed in the field of speech emotion recognition, for example Bag-of-Audio-Words [30], AuDeep [9] and DeepSpectrum [44], the functionals-based openSMILE features, which are hand-engineered statistical features, are still considered as de-facto standard and provide better reliability in terms of generalization and performance on new data since other more powerful methods show significant dependency on the data, especially when working on small datasets [31, 39]. These features are based on Low-Level Descriptors (LLDs) extracted at a frame level, which are further processed by aggregating statistics at an utterance level.

As contemporary research shows [5], neural network methods such as Word2Vec [21], FastText [2], BERT [4], and ELMO [24] are often used to extract linguistic features. The advantages of these methods are that they allow to vectorize words, taking into account the context of the text and the semantic proximity of words, in contrast to simple algorithms such as Bag-of-Words [40], One-hot Encoding [25], which take into account the occurrence of the word and its frequency, ignoring the grammar and word order in the sentence [16]. Also, the disadvantage of simple algorithms is the large size of the vector, as a rule, it is equal to the number of unique words in the text, which increases the training time of classifiers in contrast to the use of neural network methods. Another reason for using vectorization methods based on neural networks is the possibility of using pretrained models, which is relevant for solving Natural Language Processing (NLP) problems with a small amount of training data [10, 3]. If the pretrained models are used for text vectorization, then the dimension of the vector output is immutable, because it depends on the architecture of the neural network used. In this paper, we will use neural network methods of text vectorization, for which there are open access pretrained models for the Russian language, and one of the criteria for choosing methods is the small dimension of the vectors, this is necessary in order not to spend a lot of time resources for training classifiers for recognizing sentiment and emotions. These methods are Word2Vec, FastText with a vector dimension of 300, and BERT with a vector dimension of 768. The ELMO neural network also has pretrained models for the Russian language open access, but the dimension of its vector reaches 1024, which significantly affects the learning speed of classifiers, so ELMO will not be considered in experimental studies.

Acoustic and linguistic features are usually extracted and processed separately as their nature is extremely different [32]. The final contribution of each feature type for emotion recognition also varies greatly depending on the type of data used for training. Linguistic processing has almost no value for the scripted speech, whereas it gains importance relative to acoustic information in naturally occurring dialogues.

Most common approaches to combining multimodal data can be divided into 5 groups: 1) feature-level (early fusion), 2) feature-representation-level, 3) model-level, 4) decision-level (late fusion), 5)

hybrid approaches [43]. Feature-level fusion implies independent extraction and further concatenation of features from each modality. Then the resulting feature vector is passed on for classification step. This approach leads to a significant increase of dimensionality of the resulting feature space, and the feature vectors become very sparse, which may have a negative impact on classification performance on small datasets. Feature-representation-level fusion assumes independent feature extraction and processing, including feature selection and dimensionality reduction for each modality, which allows for partial solution of aforementioned problem, also known as the curse of dimensionality, however in practice this approach is often challenging as features from each modality have different nature and synchronization issues further complicate the analysis. Model-level fusion is possible via several classification models, such as Hidden Markov Models (HMM) and Neural Networks (NN), which have a capability to monitor the state of classification systems from other modalities. Such systems output a single final prediction, which takes into account the states of all classifiers from all modalities. Decision-level fusion allows to build independent, specific models for each modality. The predictions from all modalities are aggregated and transformed via a decision rule, which may vary according to the requirements.

Fusion on the decision level has several advantages compared to early fusion. First, it implies independent processing pipeline for each modality, which makes it possible to take into account peculiarities of each signal type, necessary for modelling corresponding phenomena, and better fit the models [1]. Second, it does not impose any restrictions on the methods used for modelling and classification, providing models with more flexibility [41]. And third, it allows to analyze how much contribution is necessary from each modality to achieve best possible performance [11]. This in turn makes it possible to determine the leading modality, or vice versa, provided a priory knowledge (for example, if some type of equipment is more prone to failure, or some type of signal is more noisy than others), to give more weight to the more reliable communication channel. In many applications, the performance of decision-level fusion of acoustic and linguistic information for emotion classification was shown to overcome the one obtained by early fusion techniques on the feature level [38]. Therefore, decision-level fusion remains a popular approach for information fusion in emotion classification research, particularly due to relative ease of implementation and numerous advantages outlined above.

One of the earliest decision-level approaches to combining acoustic and linguistic information in speech signals performed fusion at the decision level assuming statistical independence of each modality. Despite being an over-simplistic approach, it was proven effective [15] for recognition of negative vs. positive emotions, where the decision rule was formulated as a logical function "OR," i.e. the final emotion prediction was declared negative if either acoustic or linguistic model output a negative label, otherwise the final emotion prediction was declared positive. This approach does not take into account inter-dependency of the modalities, nor does it provide insights about their relative importance. Another decision level strategy was later proposed to account for confidence scores of each modality. By ranking modalities according to the confidence scores obtained from respective classifiers, authors selected the output with higher normalized confidence score to be the final output. In addition, to compensate the inherent difference in the performance of each modality (for example, noisy channel or faulty equipment), a constant weighting factor was applied to the confidence scores [26]. The performance of this approach was shown superior to any single-modality-based system proposed by the authors in the framework of the first Interspeech Computational Paralinguistics Challenge, however its performance relative to other participants, who did not use the fusion of acoustic and linguistic information, remained lower [35].

The decision-level fusion strategies discussed above have several limitations. Although they allow to improve classification accuracy over single-modality-based systems, they still do not consider relative performance of each modality and do not allow to account for the predictions of different modalities at once. To overcome these limitations, several soft decision fusion rules have been proposed. Soft decision is based on computing probabilities for each emotional class instead of outputting a single most probable label (hard decision). First and most simple approach uses couple-wise mean scores for each

emotion based on the acoustic and linguistic information followed by an adjacent maximum likelihood decision. This approach assigns an equal weight to each modality. Another more advanced decision-level strategy is based on weighting the probabilities obtained from each modality and computing the sum, which is further used to decide the most probable classification outcome. The fusion weights are generally learnt on a separate held-out dataset. The weights can also be tuned for each emotion class [13]. Another option is to build a meta-classifier, which receives probabilities from each modality as an input and outputs the final predictions [33]. This is most flexible approach, however it requires additional computational complexity.

3 Proposed method

We propose a novel approach towards classification of emotions into 7 categories (happy, surprised, angry, sad, scared, disgusted, neutral) using audio (acoustic) and text (linguistic) information fused at the decision-level. First, we fine-tune the performance on each modality separately, using specific methods to obtain best possible unimodal performance. Then, we define an effective rule for combining the information from both audio and text to get the final result. When making the final prediction we analyze both emotions and sentiment to complement their performance. The details of the proposed acoustic modelling, linguistic modelling and bimodal decision-level fusion are outlined in the following sections.

3.1 Audio modelling

The audio modelling is performed on the utterance level using labeled speech segments from the emotional speech corpus. First, acoustical features are extracted for each sample in the database. The resulting features undergo normalization and dimensionality reduction stages, after which they are input into a machine learning classifier to obtain the final predictions.

In this study we propose to use 2 different configurations of openSMILE features: INTERSPEECH 2010 Paralinguistics Challenge Feature Set (IS10 paraling [34]), and INTERSPEECH 2013 Paralinguistics Challenge Feature Set (IS13 ComParE [36]). The IS10 paraling feature set is based on 38 low-level descriptors extracted at 100 frames per second and their first order regression coefficients. 21 statistical functionals are applied to all frame-level features within a given utterance. Additionally F0 (Fundamental frequency) number of onsets and turn duration are added to the resulting utterance-level feature set. The low-level descriptors and functionals that were used in this study are shown in Table 1. Total number of utterance level features extracted for each training sample is 1582. The The IS13 ComParE feature set extends the Is10 paraling feature set by adding more LLDs and functionals and improving the numerical computation. Total number of utterance level features extracted for each training sample is 6373.

Support Vector Machine (SVM) and Logistic Regression (LR) are two most popularly used traditional machine learning classifiers in the field of speech emotion recognition [12]. Since SVM only outputs hard labels (classes), we opt for using LR to output soft labels (probabilities).

3.2 Text modelling

Linguistic information of speech utterances is contained in a textual modality, the study of which allows you to analyze speech audio data in full. To extract linguistic features from text data, it is necessary to use vectorization methods [6]. In this paper, the following methods are used for text vectorization: Word2Vec, FastText and BERT. The advantages of these methods are that there are pretrained models for the Russian language in the open access, as well as a small dimension of the vectors obtained using these methods, so that a large amount of time resources is not spent on training classifiers. Word2Vec [21] and FastText [2] are neural networks developed by Google and Facebook, respectively, they allow you to

Table 1: Low-Level Descriptors and Functionals of the openSMILE features [8]

Descriptors	Functionals
PCM Loudness	position max/min
MFCC [0-14]	arith. mean, std. deviation
log Mel Freq. Band [0-7]	skewness, kurtosis
LSP Frequency [0-7]	lin. regression coeff. 1/2
F0 by Sub-Harmonic Sum.	lin. regression error Q/A
F0 Envelop	quartile 1/2/3
Voicing Probability	quartile range 2-1/3-2/3-1
Jitter Local	percentile 1/99
Jitter DDP	percentile range 99-1
Shimmer Local	up-level time 75/90

study vector representations of words in natural language, considering the semantic proximity of words. The pretrained Word2Vec and FastText models for Russian language are available on the RusVectores website[1]. Based on previous studies [7], to vectorize the speech transcriptions of the RAMAS database, it is necessary to use pretrained models with the vector dimension 300 tayga_upos_skipgram_300_2_2019 and tayga_none_fasttextcbow_300_10_2019 for Word2Vec and FastText, respectively. Both models were trained on the 5 billion word Taiga corpus [37]. BERT (Bidirectional Encoder Representations from Transformers) [4] – a neural network developed by Google, allows you to extract vector representations of words from text, considering the context. BERT is a state-of-the-art method for many Nature Language Processing tasks. The pretrained BERT-Base Multilingual model[2], developed for 102 languages, including Russian, was also used to vectorize text transcriptions. The dimension of the vector obtained using this model is 768.

To recognize emotions based on a text modality, you need to choose a classifier that meets the following requirements: it can be trained well on a small data set, and the ability to produce probabilistic predictions for each class of the test set. These requirements are met by the following machine classifiers: Logistic Regression (LR), Random Forest (RF) and Naive Bayes (NB).

3.3 Bimodal fusion

To fuse information from 2 modalities, audio and text, we propose to obtain 3 sets of predictions: probabilities of emotions from modelling acoustic parameters, probabilities of emotions from modelling linguistic parameters, and probabilities of sentiment from modelling linguistic characteristics.

As a rule, the text modality contains information about the polarity (valence) of the speaker's expressed emotion during a conversation, whereas the acoustic modality conveys the intensity of emotions [17]. Therefore, with the help of text transcriptions, we can perform not only the recognition of emotions, but also the recognition of sentiment. Sentiment analysis classes are obtained from the original annotation of emotions (6 emotional classes + neutral state) by grouping the following categories: Happy and Surprised (Positive sentiment); Angry, Sad, Disgusted, Scared (Negative sentiment); Neutral state (Neutral sentiment). This grouping corresponds to the circumplex model of emotions proposed by James Russel [28]. Some of the relevant emotional categories placed on a 2-dimensional space are depicted in Figure 1. Emotional categories Astonished and Excited, which roughly correspond to Surprise in the current study, are shown in yellow. Such emotions as Annoyed, Distressed and Frustrated, which are

[1] https://rusvectores.org/ru/

[2] https://github.com/google-research/bert/blob/master/multilingual.md

close to Disgust, are shown in orange. Positive emotional categories Happy, Pleased and Glad are shown in red. Other relevant to the present study categories such as Angry, Scared (Afraid) and Sad have direct correspondence with the diagram.

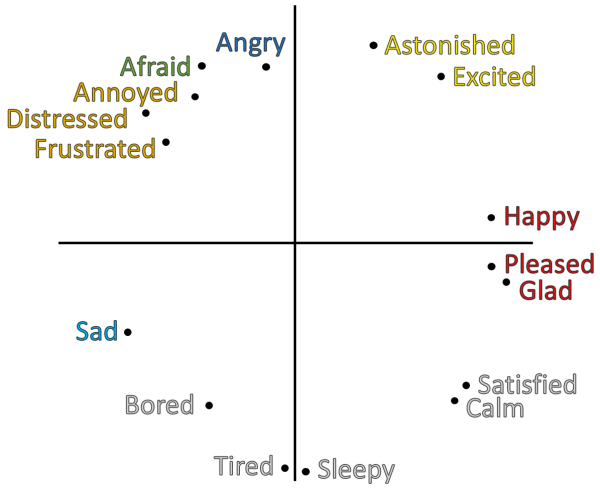

Figure 1: Circumplex model of emotions proposed by James Russel [28]. Some irrelevant to the study emotional categories are omitted for convenience.

Since the prediction of emotions from audio are made on the utterance level, and the predictions of sentiment and emotions from text are made on the whole dialogue level, we propose the following approach to merge the probabilities. Emotion predictions from text are interpolated to all the emotional intervals within the given dialogue, i.e. all the annotated intervals are given the same prediction as the overall prediction of the dialogue. The same procedure is repeated for the sentiment, however, because the sentiment predictions only have 3 classes, and the emotion predictions have 7 classes, the sentiment predictions need to be further interpolated to the emotional categories. With the help of the Russell's circumplex model, we perform interpolation by repeating the sentiment prediction for all corresponding emotions: positive sentiment prediction is interpolated to the happy and surprised classes, negative sentiment prediction is interpolated to the Angry, Sad, Scared, and Disgusted classes, and the Neutral sentiment prediction corresponds to the Neutral state class (see Figure 2).

4 Experiments

All experiments are conducted on the RAMAS [22] [23] dataset with a predefined train/test split. The classification performance is measured in terms of Unweighted Average Recall (UAR), which is average of recalls from each emotional category. This performance measure is preferred over the simple accuracy when dealing with unbalanced data, since it provides better estimation of the performance across all classes. Following is the description of data and experimental setup used in the current study.

4.1 RAMAS dataset

RAMAS [22] [23] is a multimodal corpus of dyadic interactions intended for modelling affective phenomena such as emotions. It was collected in 2016-2017 by Neurodata Lab company and is free and open-source for research purposes. 10 semi-professional actors (5 males and 5 females) in the age of 18-28 years old recorded 580 video clips 7 hours in total. Each video clip is approximately 30 seconds

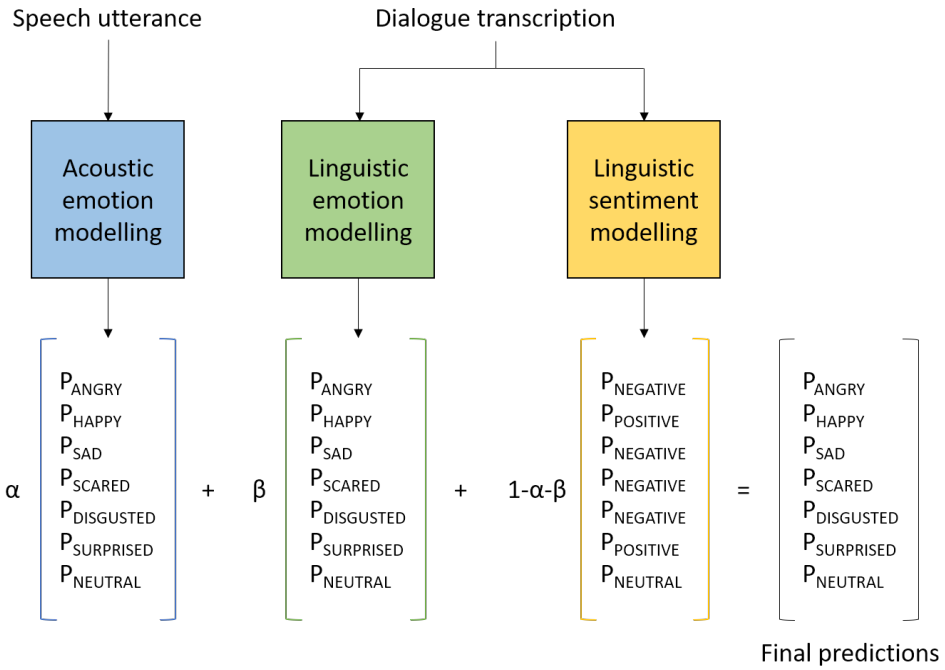

Figure 2: Proposed strategy for fusion of acoustic and linguistic information

long and contains speech utterances from 2 actors, a male and a female. Each actor had a dedicated lav microphone attached to the neckline and connected to a wireless recording system. Actors performed according to predefined scenarios, however they were free to improvise and choose words and phrases to make the conversation more natural. In total, there were 13 different scenarios including interactions between friends and coworkers on different topics: travel, work, health etc. Each scenario implied presence of 2 different emotions (1 per each actor) from the pool of 6 basic emotions (happiness, surprise, anger, sadness, fear, disgust) and neutral state. Furthermore, each actor was assigned a predefined social role (dominating or submissive). All emotions and social roles were equally distributed between actors and scenarios. The performance of actors was coordinated by a professional teacher from Russian State University of Cinematography to make sure that the portrayed emotions correspond to the intended roles and scenarios.

The database was labeled on the frame level using categorical labels by 21 annotators. Each video clip was annotated by at least 5 different annotators. All the annotators passed an emotional intelligence test and scored average and above. The annotation was performed with the ELAN tool [42] from Max Planck Institute for Psycholinguistics (the Netherlands). Annotators were asked to mark the beginning and the end of each emotional expression that seemed natural. Due to high subjectivity of the task, the resulting labeled intervals vary significantly, and some intervals have multiple heterogeneous labels assigned by different annotators. The authors of the database reported an average agreement score between annotators (Krippendorff's alpha) as 0.44, which is considered as moderate agreement. The incongruities in annotations provided by different annotators give raise to a substantial problem of finding the ground truth of the provided annotations, and further complicate the analysis and classification of speech utterances.

The annotation of RAMAS dataset is unconventional in that it uses frame-level categorical annotations, whereas all other emotional speech corpora annotated on the frame level, for example RECOLA [27], SEMAINE [19], CreativeIT [20], SEWA [14] etc. use annotation of emotional pa-

rameters, such as activation and valence, instead of the emotional categories. Moreover, due to the extemporaneous nature of dyadic interactions, a significant amount of speech between the actors within one dialogue overlaps, which brings certain challenges in terms of extracting acoustical features and using speech recognition software to obtain the transcriptions.

4.2 Experimental setup

The experimental setup was divided into 3 stages. In the first and second stages, which were conducted in parallel, we performed individual analysis of the acoustic and linguistic modalities. In the third stage, we fused the audio and text analysis to obtain the final bimodal result.

To be able to fairly compare the performance of audio-based, text-based and bimodal systems we chose only those samples from the database for which it was possible to obtain the transcriptions using Automatic Speech Recognition (ASR) software. We used 2 online cloud-based platforms from Google[3] and Yandex[4]. This resulted in 263 audio recordings with corresponding transcriptions. The train/test split of the data for each experiment was kept constant at a ratio 70/30 for each emotion. The distribution of number of samples for each emotion and for each sentiment category in train and test sets is shown in Figure 3 a) and b), respectively.

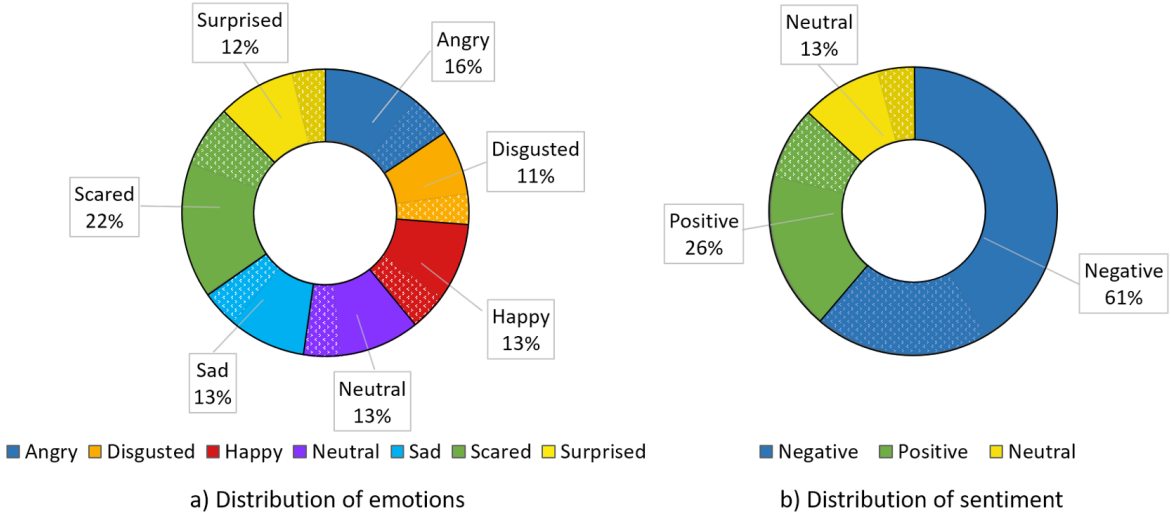

Figure 3: Distribution of samples between annotated classes: (a) emotional categories, (b) sentiment categories. Solid color indicates train partitioning, patterned color - test partitioning

In each stage we repeated all the experiments for 2 levels of annotation confidence: using the agreement of 4 and 5 annotators. For each set of the experiments, we filtered the training data according to the required level of agreement, which means that for each annotation confidence level we used a separate subset of data obtained from the original RAMAS corpus. This resulted in 223 and 207 training samples (audio + transcription) for the 4-th and 5-th level of annotation agreement, respectively. The details of the number of samples available for training and testing at each level of annotation confidence are summarized in the Figure 4. The first level of agreement corresponds to the total number of samples available in the dataset.

[3]https://pypi.org/project/SpeechRecognition
[4]https://cloud.yandex.ru/services/speechkit

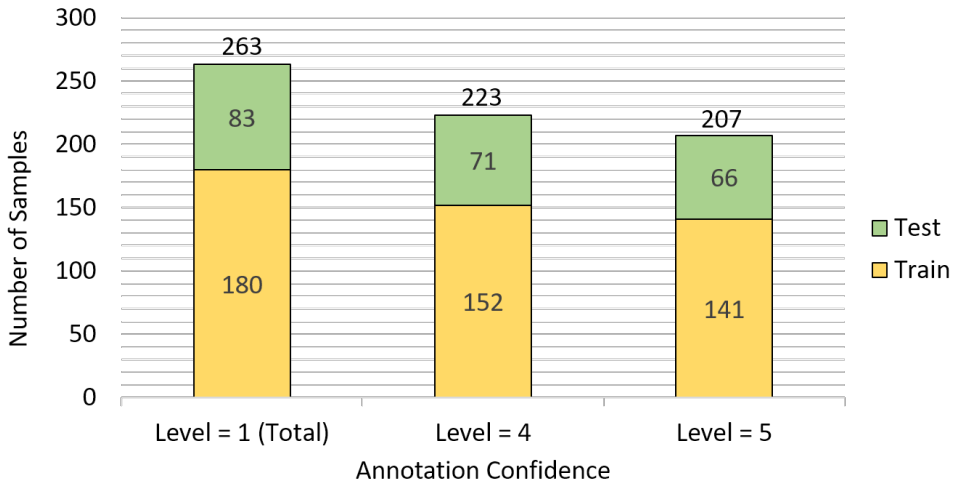

Figure 4: Distribution of samples in train and test set according to the level of annotator agreement

4.2.1 Audio-based experiments and results

For each level of annotation agreement, the audio samples were obtained by creating audio chunks that correspond to the intersection of annotation intervals of required confidence level. For example, the confidence level of 4 requires at least 4 out of 5 annotators to label the same speech utterance with the same emotional category. The beginning and the end of the resulting annotation (and corresponding audio chunk) were set to match the intersection of all 4 annotations from different annotators. The original audio recordings were segmented according to the provided annotation. All the frames that were not labeled by the annotators as emotional were discarded. The neutral class was trained only on speech segments that were specifically marked as neutral.

We compared the performance of two acoustical feature sets, namely 1582 features from the INTERSPEECH 2010 Paralinguistics Challenge Feature Set (IS10 paraling [34]), and 6373 features from the INTERSPEECH 2013 Paralinguistics Challenge Feature Set (IS13 ComParE [36]).

We also tested two different strategies for normalization of features: Min-Max normalization and Z-score normalization. Min-Max normalization results in feature values that fall in the range [0, 1]. Z-score normalization results in feature vectors where the distribution of each feature has zero mean and unit standard deviation.

In order to reduce the size of high-dimensional feature representations and de-correlate the features we applied Principle Component Analysis (PCA) that projects the data into lower-dimensional space, preserving maximum variance. In each experiment, the optimal number of principle components was determined empirically. Logistic Regression (LR) was used as classification method to output emotion probabilities.

Classification results in terms of UAR (%) obtained on a test set for various levels of annotation agreement are shown in Tables 2 and 3. Bold script indicates highest maximum performance in the given experiment. Baseline system does not use any normalization and dimensionality reduction strategies and performs classification directly on the extracted features.

4.2.2 Text-based experiments and results

In the RAMAS database, the dialogues between the two speakers were played out according to dyadic scenarios, and the speakers could think through their speech in advance before expressing emotions, re-

Table 2: Audio Modality Test Set Classification Result (UAR, %) for 7 Emotional Categories on the 4-th Level of Annotation Agreement

System	IS10 paraling	IS13 ComParE
Baseline	23.02	14.23
Min-Max	39.27	40.90
Z-normalization	43.39	48.09
Z-normalization + PCA	**44.17**	**49.48**
Min-Max + PCA	41.45	44.29

Table 3: Audio Modality Test Set Classification Result (UAR, %) for 7 Emotional Categories on the 5-th Level of Annotation Agreement

System	IS10 paraling	IS13 ComParE
Baseline	26.62	13.23
Min-Max	41.79	32.34
Z-normalization	44.13	36.95
Z-normalization + PCA	**44.44**	**40.01**
Min-Max + PCA	43.74	35.76

flecting the role from the scenarios in terms of semantic content. Based on this, for text transcriptions, the markup corresponding to the emotions prescribed in the scenarios was used. Before you build classifiers for recognizing sentiment and emotions, you need to select relevant information from the text. To do this, you need to remove punctuation marks, stop words (words without semantic content, such as prepositions, conjunctions, etc.), lower the case of all words, and normalize words using lemmatization. Then linguistic information was extracted from the preprocessed text transcriptions, and the text was vectorized using Word2Vec, FastText, and BERT methods. The following methods were used as classifiers: Logistic Regression (LR), Random Forest (RF), and naive Bayes (NB). The training was performed using 3-fold validation with the selection of the best parameters for each classifier. The results (UAR,%) of experimental studies of sentiment recognition by text modality for 4-th and 5-th levels of annotation agreement are presented in Tables 4 and 5, respectively. To recognize 7 emotional categories, similar experiments were performed with the methods of vectorization and classification, the results of which are presented in Tables 6 and 7.

Table 4: Text Modality Test Set Classification Result (UAR, %) for 3 Sentiment Categories on the 4-th Level of Annotation Agreement, UAR,%

	RF	LR	NB
Word2Vec	74.91	**90.11**	76.13
FastText	83.05	82.54	74.89
BERT	68.98	80.61	49.21

4.2.3 Bimodal fusion experiments and results

The fusion of acoustic and linguistic information was performed via weighted sum of the probabilities of each modality. The linguistic information was represented by 2 sets of probabilities: emotion predictions and sentiment predictions. Therefore, in total 3 terms contributed to the final result: emotion probabilities using audio, emotion probabilities using text, and sentiment probabilities using text. 2 weighing

Table 5: Text Modality Test Set Classification Result (UAR, %) for 3 Sentiment Categories on the 5-th Level of Annotation Agreement, UAR,%

	RF	LR	NB
Word2Vec	76.09	**87.58**	69.56
FastText	85.78	82.22	83.93
BERT	60.78	85.78	62.90

Table 6: Text Modality Test Set Classification Result (UAR, %) for 7 Emotional Categories on the 4-th Level of Annotation Agreement

	RF	LR	NB
Word2Vec	42.83	**69.13**	40.87
FastText	54.97	66.47	51.28
BERT	47.98	55.96	26.36

coefficients were introduced for weighting the sum:α and β. α controls the contribution of audio modality, while β controls the contribution of text modality. The third coefficient is computed based on the values of α and β to make sure that all three coefficients sum to 1.

$$P_b = \alpha * P_A + \beta * P_T + (1 - \alpha - \beta) * P_S, \quad \beta \leq 1 - \alpha$$

where P_b is the final set of emotion probabilities from bimodal system, P_A is a set of emotion probabilities from the acoustic model (using audio), P_T is a set of emotion probabilities from the linguistic model (using text), and P_S is a set of sentiment probabilities from the linguistic model (using text); α and β are weighing coefficients that control the contribution (importance) of each set of probabilities. The constraint of β being less than or equal to α is necessary to make sure that all 3 coefficients sum to 1.

The results of the proposed fusion approach are shown in Figures 5 (for the 4-th level of annotation confidence) and 6 (for the 5-th level of annotation confidence). In bold are highlighted the numbers indicating highest maximum performance for the given level of annotation agreement. The tables are color-coded, with the green color indicating better performance and red color indicating poor performance. The weights of α and β are incremented by 0.1, the third coefficient (1-α-β) is implied and not shown in the tables. As a result, the higher-left corner represents classification accuracy of the sentiment analysis without considering emotion modelling ($\alpha = \beta = 0$), the higher-right corner indicates classification accuracy of the acoustic emotion modelling without considering linguistic modality ($\beta = 0$), and the lower-left corner shows classification accuracy of the linguistic emotion modelling without considering acoustic parameters and sentiments ($\alpha = 0$). All other values in the tables indicate mixed performance of fusing several sets of probability predictions.

5 Discussion

The experiments with acoustic modality show that the base system works better with IS10 paraling features. The performance of the base system on the IS13 ComParE features remained on the chance level (UAR=14.29%, 13.23%), which can be explained by high dimensionality of the feature space and sparseness of feature vectors. Normalization greatly improves the performance on both feature sets for both 4-th and 5-th level of annotation agreement, and Z-score normalization gives better results than Min-Max. PCA further improves the performance reaching UAR=49.48% on the 4-th level and UAR=44.44% on the 5-th level of annotation agreement. The influence of normalization procedure before applying

Table 7: Text Modality Test Set Classification Result (UAR, %) for 7 Emotional Categories on the 5-th Level of Annotation Agreement

	RF	LR	NB
Word2Vec	47.96	**62.74**	39.64
FastText	52.02	62.66	48.98
BERT	43.19	57.21	29.72

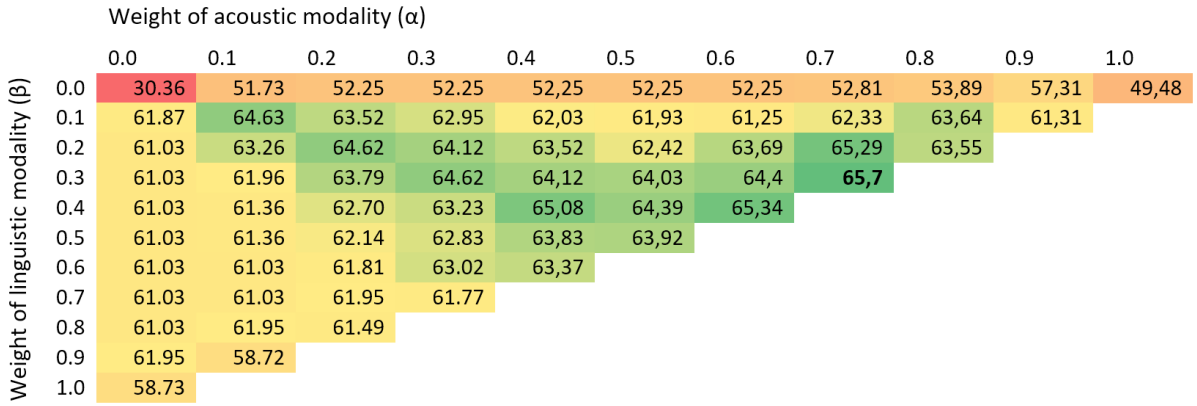

Figure 5: Bimodal Fusion Test Set Classification Results (UAR, %) for 7 Emotional Categories on the 4-th Level of Annotation Agreement

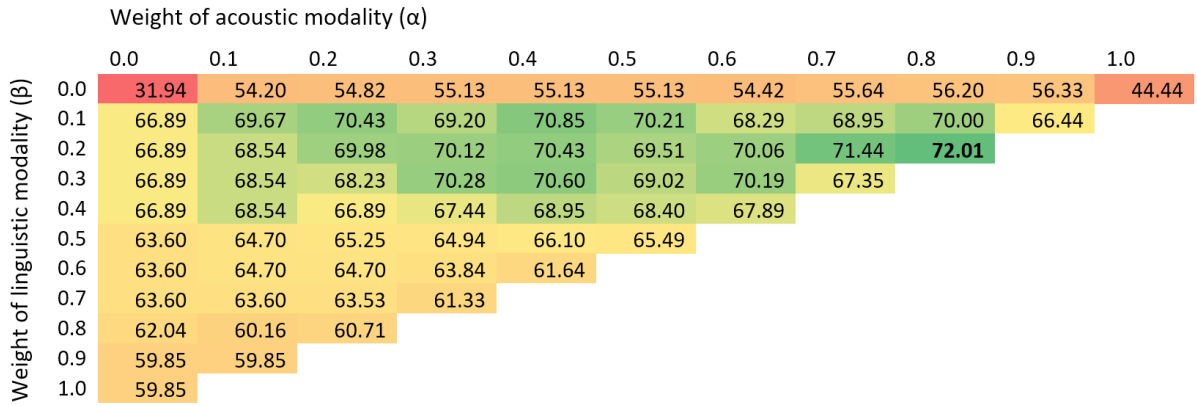

Figure 6: Bimodal Fusion Test Set Classification Results (UAR, %) for 7 Emotional Categories on the 5-th Level of Annotation Agreement

PCA is important and in out experiments, PCA produced better results with the Z-score normalization. It is worth to mention that on the 4-th level of annotation agreement, the best result was obtained using PCA on the IS13 ComParE features, however on the 5-th level of annotation agreement the best result was achieved by using PCA on the IS10 paraling feature set. This can be explained by the fact that on the 5-th level the training data is not enough to effectively train the classifiers in high dimensional feature space. However, on the 4-th level the training data size is enough to optimally train the model and show best performance. The optimal number of PCA components varied from 50 to 300.

From the results of experimental studies of the recognition of sentiment and emotions by linguistic

modality, presented in Tables 4, 5, 6, 7, the following conclusions can be drawn. The best UAR in all experiments is achieved by combining the Word2Vec vectorization method and Logistic Regression. When recognizing 3 categories of sentiment (negative, neutral, and positive) UAR reaches 90.11% and 87.58%, when recognizing 7 categories of emotions (anger, sadness, joy, neutral, disgust, fear, surprise) UAR = 69.13% and 62.74%, for level 4 and 5 of annotator agreement, respectively. The results of text data classification at level 5 of annotator agreement are lower than at level 4, this is because there was less text data consistency at level 5, hence the classifier was trained on a smaller data set that was less representative. We can also notice that when using the state-of-the-art BERT vectorization method, the classification results are lower than with other vectorization methods. The best classification results when using BERT are achieved in combination with logistic regression, but nevertheless the results lag behind the final ones by 10% and 2% for the recognition of 3 categories of sentiment, and by 14% and 5% for the recognition of 7 categories of emotions for the 4 and 5 levels of annotator agreement, respectively.

The experimental results of the proposed bimodal fusion of acoustic and linguistic parameters of speech show similar trends for both 4-th level and 5-th level of annotation agreement. At the 4-th level, the best performance UAR=65.7% was obtained by setting α=0.7 and β=0.3, which sum to 1. At the 5-th level, the best performance UAR=72.01% was obtained by setting α=0.8 and β=0.2, which also sum to 1. This means that using the probabilities of sentiment categories (positive, negative, neutral) did not play any role in final decision. However, it does not mean that the sentiment indicators are useless. As can be seen from the color coding in Table 6, when acoustic modality is assigned a weight α=0.7, the best performance is achieved by considering both emotions and sentiments from the text data. Therefore, depending on the exploitation requirements (for example, when the weight of a certain modality is a priory given according to exploitation conditions), sentiment analysis may play an important role when making the final decision. Moreover, as seen from the tables, when acoustic modality weight is set to 0, the best performance of the linguistic modality is achieved when both emotions and sentiments are considered.

At the 4-th level of annotation agreement, when α=1, which corresponds to emotion recognition using only acoustical parameters (upper right corner of Table 2), the best UAR=49.48%. When β=1, which corresponds to emotion recognition using only linguistic features (lower left corner of the Table 2), the best UAR=58.73%. This result has an absolute improvement of 9.25% (relative improvement of 15.75%) as compared to the performance of the acoustical features. This indicates that emotion recognition using linguistic information is more reliable, probably due to the fact that audio modality in RAMAS dataset is quite noisy. However, to reach the optimal performance, the fusion weight of the acoustic modality is higher (α=0.7) than the linguistic (β=0.3) more than twice. On the 5-th level of agreement the optimal distribution of fusion weights is even more imbalanced: α=0.8 and β=0.2. An important conclusion here is that emotional states are more likely to be expressed via acoustic rather linguistic cues. Therefore, when expressing emotions, it is more important how a person pronounces the utterance, rather than what he says. However we should not ignore the linguistic content completely, as shown from the experiments, there is a huge improvement in performance when using bimodal emotion recognition relative to acoustic modelling: 16.22% absolute (24.69% relative) increase in UAR. On the 5-th level of annotation agreement, the increase in performance is even more drastic: 27.57% absolute (38.29% relative), reaching a maximum value of UAR=72.01%.

The comparison of performance of the proposed system on different levels of annotation agreement reveals interesting results. In unimodal experimental setups, both acoustic and linguistic, there is a performance drop on the 5-th level of annotation agreement, which can be attributed to a decrease in available training data. However, the bimodal experiments show that increasing the level of annotation agreement actually leads to an increased performance. This may indicate that the proposed bimodal setup is more robust against training data size.

6 Conclusions

In this study we proposed an effective bimodal emotion recognition system based on acoustic and linguistic information from speech signals. The experimental findings presented in this study reveal several important conclusions. First, the performance of the bimodal system greatly overcomes the performance of any single modality. Second, the contribution of acoustic modality is far greater than linguistic, gaining 2-4 times more weight. However, linguistic modality is still important to consider since fusing the linguistic information with acoustic provides 24.69%-38.29% relative improvement. When processing linguistic modality, it is imperative to consider both emotions and sentiments since their mutual performance provides the best outcome in a variety of application scenarios. Finally, the comparison of the performance of the proposed system on different levels of annotation agreement shows that the proposed system is more robust against the training data size reduction as compared to any single modality system. It is also worth noting that the level of annotation agreement affects the result of bimodal recognition of emotions, since level 5 of agreement exceeds level 4 in performance by 6.31%. The best performance was achieved on the 5-th level of annotation agreement, reaching UAR=72.01%.

Acknowledgements

This research was financially supported by the Russian Science Foundation (project No. 18-11-00145, research and development of the audio-based speech emotion recognition), by the Russian Foundation for Basic Research (project No. 18-07-01407, research and development of the text-based and bimodal Russian speech emotion recognition), as well as by the state research No. 0073-2019-0005.

References

[1] H. Al Osman and T. Falk. Multimodal affect recognition: Current approaches and challenges. *Emotion and Attention Recognition Based on Biological Signals and Images*, pages 59–86, February 2017.

[2] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, June 2017.

[3] P. Conde-Cespedes, J. Chavando, and E. Deberry. Detection of suspicious accounts on twitter using word2vec and sentiment analysis. In *Proc. of the 11th International Conference on Multimedia and Network Information System (MISSI'18), Wroclaw, Poland*, volume 833 of *Advances in Intelligent Systems and Computing*, pages 362–371. Springer, Cham, September 2018.

[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'19), Minneapolis, Minnesota, USA*, pages 4171–4186, June 2019.

[5] A. Dvoynikova and A. Karpov. Analytical review of approaches to russian text sentiment recognition. *Information and Control Systems*, 4:20–30, November 2020.

[6] A. Dvoynikova, O. Verkholyak, and A. Karpov. Analytical review of methods for identifying emotions in text data. In *Proc. of the 3rd International Conference on R. Piotrowski's Readings in Language Engineering and Applied Linguistics (PRLEAL'19), Saint Petersburg, Russia*, pages 8–21. CEUR-WS, November 2019.

[7] A. Dvoynikova, O. Verkholyak, and A. Karpov. Emotion recognition and sentiment analysis of extemporaneous speech transcriptions in russian. In *Proc. of the 22nd International Conference on Speech and Computer (SPECOM'20), St. Petersburg, Russia*, volume 12335 of *Lecture Notes in Computer Science*, pages 136–144. Springer, Cham, October 2020.

[8] F. Eyben and B. Schuller. opensmile:) the munich open-source large-scale multimedia feature extractor. *ACM SIGMultimedia Records*, 6(4):4–13, January 2015.

[9] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller. audeep: Unsupervised learning of representations from audio with deep recurrent neural networks. *The Journal of Machine Learning Research*, 18(1):6340–6344, January 2017.

[10] A. Golubev and N. Loukachevitch. Improving results on russian sentiment datasets. In *Proc. of the 9th Conference on Artificial Intelligence and Natural Language (AINL'20), Helsinki, Finland*, volume 1292 of *Communications in Computer and Information Science*, pages 109–121. Springer, Cham, October 2020.

[11] Z. He, Z. Li, F. Yang, L. Wang, J. Li, C. Zhou, and J. Pan. Advances in multimodal emotion recognition based on brain–computer interfaces. *Brain Sciences*, 10(10):687, October 2020.

[12] W. Jiang, Z. Wang, J. Jin, X. Han, and C. Li. Speech emotion recognition with heterogeneous feature unification of deep neural network. *Sensors*, 19(12):2730, June 2019.

[13] Q. Jin, C. Li, S. Chen, and H. Wu. Speech emotion recognition with acoustic and lexical features. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'15), South Brisbane, Queensland, Australia*, pages 4749–4753. IEEE, April 2015.

[14] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. Schuller, K. Star, E. Hajiyev, and M. Pantic. SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):1022–1040, October 2019.

[15] C. Lee, S. Narayanan, and R. Pieraccini. Combining acoustic and language information for emotion recognition. In *Proc. of the 7th International Conference on Spoken Language Processing (ICSLP'02), Denver, Colorado, USA*. International Speech Communication Association, September 2002.

[16] S. Lee. Document vectorization method using network information of words. *PLOS ONE*, 14(7):e0219389, July 2019.

[17] B. Liu. Many facets of sentiment analysis. In *A practical guide to sentiment analysis*, volume 5 of *Socio-Affective Computing*, pages 11–39. Springer, Cham, 2017.

[18] J. Luo and R. Yu. Follow the heart or the head? the interactive influence model of emotion and cognition. *Frontiers in psychology*, 6:573, May 2015.

[19] G. McKeown, M. Valstar, R. Cowie, and M. Pantic. The semaine corpus of emotionally coloured character interactions. In *Proc of. the 2010 IEEE International Conference on Multimedia and Expo (ICME'10), Singapore*, pages 1079–1084. IEEE, July 2010.

[20] A. Metallinou, C. Lee, C. Busso, S. Carnicke, and S. Narayanan. The USC CreativeIT database: A multimodal database of theatrical improvisation. *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, page 55, May 2010.

[21] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, pages 3111–3119, October 2013.

[22] O. Perepelkina, E. Kazimirova, and M. Konstantinova. Ramas: Russian multimodal corpus of dyadic interaction for affective computing. In *Proc. of the 20th International Conference on Speech and Computer (SPECOM'18), Leipzig, Germany*, volume 11096 of *Lecture Notes in Computer Science*, pages 501–510. Springer, Cham, August 2018.

[23] O. Perepelkina, E. Kazimirova, and M. Konstantinova. Ramas: Russian multimodal corpus of dyadic interaction for studying emotion recognition. *PeerJ Preprints*, 6:e26688v1, March 2018.

[24] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations, February 2018.

[25] D. Pham and A. Le. Exploiting multiple word embeddings and one-hot character vectors for aspect-based sentiment analysis. *International Journal of Approximate Reasoning*, 103:1–10, December 2018.

[26] T. Polzehl, S. Sundaram, H. Ketabdar, M. Wagner, and F. Metze. Emotion classification in children's speech using fusion of acoustic and linguistic features. In *Proc. of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH'09), Brighton, UK*. International Speech Communication Association, September 2009.

[27] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *Proc. of the 10th IEEE international conference and workshops*

on automatic face and gesture recognition (FG'13), Shanghai, China, pages 1–8. IEEE, April 2013.

[28] J. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178, 1980.

[29] K. Sailunaz, M. Dhaliwal, J. Rokne, and R. Alhajj. Emotion detection from text and speech: a survey. *Social Network Analysis and Mining*, 8(1):28, December 2018.

[30] M. Schmitt, F. Ringeval, and B. Schuller. At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech. In *Proc. of 17th Annual Conference of the International Speech Communication Association (INTERSPEECH'16), San Francisco, California, USA*, pages 495–499. International Speech Communication Association, September 2016.

[31] B. Schuller, A. Batliner, C. Bergler, E. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen, and et. al. The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks. In *Proc. of the 2020 Annual Conference of the International Speech Communication Association (INTERSPEECH'20), Shanghai, China*. International Speech Communication Association, September 2020.

[32] B. Schuller, A. Batliner, S. Steidl, and D. Seppi. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9-10):1062–1087, November–December 2011.

[33] B. Schuller, G. Rigoll, and M. Lang. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *Proc. of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04), Montreal, Quebec, Canada*, pages I–577. IEEE, May 2004.

[34] B. Schuller, S. Steid, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan. The INTERSPEECH 2010 paralinguistic challenge. In *Proc. of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH'10), Makuhari, Chiba, Japan*, pages 2794–2797. International Speech Communication Association, September 2010.

[35] B. Schuller, S. Steidl, and A. Batliner. The interspeech 2009 emotion challenge. In *Proc. of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH'09), Brighton, UK*, pages 312—-315. International Speech Communication Association, September 2009.

[36] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. March, and M. Mortillaro. The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proc. of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH'13), Lyon, France*, pages 148–152. International Speech Communication Association, August 2013.

[37] T. Shavrina and O. Shapovalova. To the methodology of corpus construction for machine learning: "Taiga" syntax tree corpus and parser. In *Proc. of the 2017 international conference "CORPUS LINGUISTICS - 2017", St. Petersburg, Russia*, pages 78–84. Publishing house of St. Petersburg State University, June 2017.

[38] M. Sidorov, E. Sopov, I. Ivanov, and W. Minker. Feature and decision level audio-visual data fusion in emotion recognition problem. In *Proc. of the 12th International Conference on Informatics in Control, Automation and Robotics (ICINCO'15), Colmar, France*, volume 2, pages 246–251. IEEE, July 2015.

[39] G. Sogancioglu, O. Verkholyak, H. Kaya, D. Fedotov, T. Cadée, A. Salah, and A. Karpov. Is everything fine, grandma? acoustic and linguistic modeling for robust elderly speech emotion recognition. In *Proc. of the 2020 Annual Conference of the International Speech Communication Association (INTERSPEECH'20), Shanghai, China*. International Speech Communication Association, September 2020.

[40] G. K. Soumya and S. Joseph. Text classification by augmenting bag of words (bow) representation with co-occurrence feature. *IOSR Journal of Computer Engineering*, 16(1):34–38, January 2014.

[41] C. Tan, G. Ceballos, N. Kasabov, and N. Puthanmadam Subramaniyam. Fusionsense: Emotion classification using feature fusion of multimodal data and deep learning in a brain-inspired spiking neural network. *Sensors*, 20(18):5328, 2020.

[42] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. Elan: a professional framework for multimodality research. In *Proc. of the 5th International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy*, pages 1556–1559. European Language Resources Association (ELRA), May 2006.

[43] C. Wu, J. Lin, and W. Wei. Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. *APSIPA transactions on signal and information processing*, 3, November 2014.

[44] Z. Zhao, Y. Zhao, Z. Bao, H. Wang, Z. Zhang, and C. Li. Deep spectrum feature representations for speech emotion recognition. In *Proc. of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data (ASMMC-MMAC'18), Seoul, Republic of Korea*, pages 27–33. ACM, October 2018.

Author Biography

Oxana Verkholyak received the B.S. degree (2015) in electrical engineering at Korea University, Seoul, South Korea, and the M.S. degree (2017) in Software and Information science from ITMO University in Saint-Petersburg Russia. Currently she is a PhD student at ITMO and Ulm universities (double diploma), and a junior researcher at the Federal Research Institute of Russian Academy of Sciences (SPC RAS), formerly known as the Saint Petersburg Institute of Informatics and Automation of Russian Academy of Sciences (SPIIRAS). Her research interests include machine learning, deep learning, speech paralinguistic analysis, sentiment analysis, speech emotion recognition. She is a member of ISCA, ACL, IEEE and IEEE young professionals.

Anastasia Dvoynikova received the B.S. degree in Information Security from ITMO University, Saint-Petersburg in 2018. Currently she is taking a master's course at Speech Information Systems, ITMO University. Her research interests include Sentiment-analysis, Emotion Recognition in speech transcription.

Alexey Karpov received the M.S. degree in 2002 from the St. Petersburg State University of Aerospace Instrumentation, the Ph.D. degree in 2007, and the D.Sc. degree in 2013 in Computer Science from the St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS). Currently he is the head of the Speech and Multimodal Interfaces Laboratory at SPIIRAS Institute of the St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). He has published more than 250 peer-reviewed papers in international journals and proceedings of international conferences. He is a general chair of a series of International Conferences on Speech and Computer (SPECOM). His current research interests include computational paralinguistics, audio-visual speech processing, automatic speech recognition, multimodal user interfaces, etc. He is a member of ISCA, ACM, IEEE SPS and EURASIP associations. .