DOI 10.15622/sp.18.1.30-56

O.V. VERKHOLYAK, H. KAYA, A.A. KARPOV

MODELING SHORT-TERM AND LONG-TERM DEPENDENCIES OF THE SPEECH SIGNAL FOR PARALINGUISTIC EMOTION CLASSIFICATION

Verkholyak O.V., Kaya H., Karpov A.A. **Modeling Short-Term and Long-Term Dependencies of the Speech Signal for Paralinguistic Emotion Classification.**

Abstract. Recently, Speech Emotion Recognition (SER) has become an important research topic of affective computing. It is a difficult problem, where some of the greatest challenges lie in the feature selection and representation tasks. A good feature representation should be able to reflect global trends as well as temporal structure of the signal, since emotions naturally evolve in time; it has become possible with the advent of Recurrent Neural Networks (RNN), which are actively used today for various sequence modeling tasks. This paper proposes a hybrid approach to feature representation, which combines traditionally engineered statistical features with Long Short-Term Memory (LSTM) sequence representation in order to take advantage of both short-term and long-term acoustic characteristics of the signal, therefore capturing not only the general trends but also temporal structure of the signal. The evaluation of the proposed method is done on three publicly available acted emotional speech corpora in three different languages, namely RUSLANA (Russian speech), BUEMODB (Turkish speech) and EMODB (German speech). Compared to the traditional approach, the results of our experiments show an absolute improvement of 2.3% and 2.8% for two out of three databases, and a comparative performance on the third. Therefore, provided enough training data, the proposed method proves effective in modelling emotional content of speech utterances.

Keywords: Speech emotion recognition, computational paralinguistics, affective computing, feature representation, context modelling, artificial neural networks, long short-term memory.

1. Introduction. Automatic emotion recognition has emerged as one of the most important and challenging research topics of affective computing [1, 2], a modern study concerned with recognizing and processing human feelings. Lying at the crossroads of computer and cognitive sciences, this rapidly growing field has gained its popularity due to advent of new trends and technologies that require monitoring of human's psychophysical state with higher level of personalization and adaptation, as well as the ability to simulate empathy for more natural human-computer interaction. The capacity to adapt to user's current emotional state is important because emotions greatly influence people's behavior: they affect communication, health and personal well-being, decision making processes and other important aspects of everyday life. Hence, developing systems that are aware of current user's state will help to incorporate emotional content into human-machine interaction and improve overall user experience.

Some essential issues inherent to the study field of vocal emotions are difficulties of defining emotions, specifying number of existing

emotions, and distinguishing between different emotional states. These questions are important because they define the way investigators approach the study — what emotions to model, what to measure, how to interpret results, etc. In a typical speech research, emotions are defined as "brief and intense reactions to goal-relevant changes in the environment" [3]. Currently, there are two major approaches to modeling the emotional states adopted by researchers — continuous and categorical [1].

Continuous approach assumes that every emotion can be represented as a point in a 2- or 3-dimensional space, where the dimensions represent essential emotion characteristics, such as valence (positive of negative), activation (calm or excited), and even dominance (active or passive). The two- and three- dimensional emotion spaces with some emotion interpretations are shown in Figures 1 and 2, respectively. Categorical approach defines a list of basic emotions, usually from 4 to 7, which can be considered universal: anger, happiness, sadness, surprise, fear and neutral state [4]. Both approaches are actively being exploited in the field; the choice is usually determined by the database of the interest. In this study, we will be using the categorical approach.

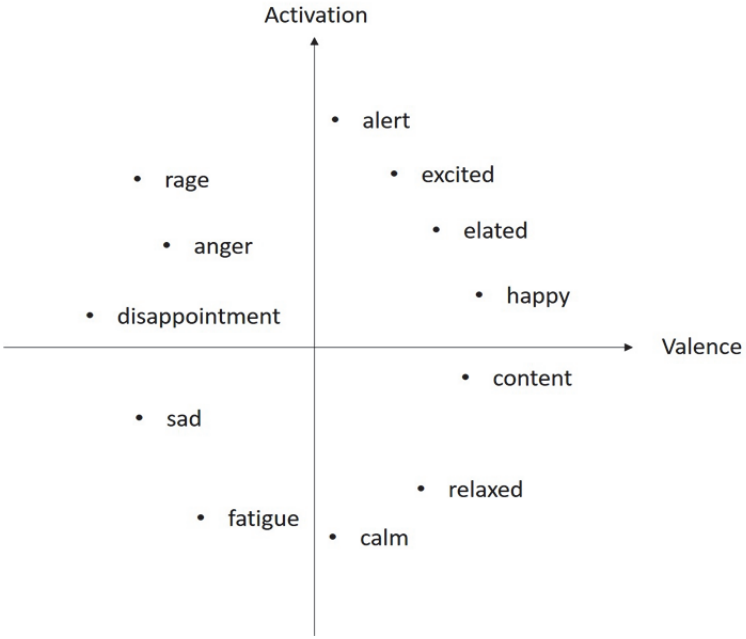


Fig. 1. Two-dimensional approach for emotion modelling by activation and valence dimensions

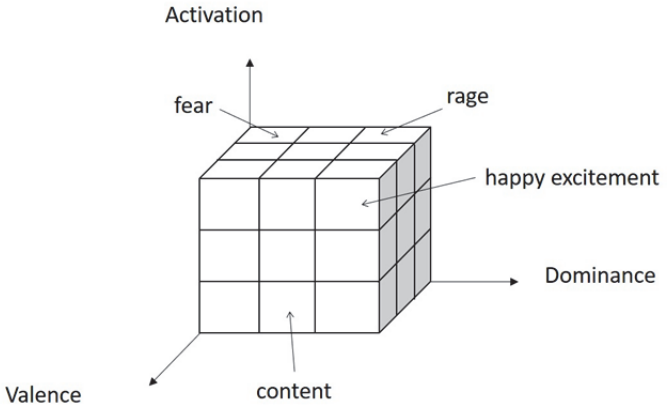


Fig. 2. Three-dimensional space for emotion modelling with some interpretations using activation, valence and dominance dimensions

Practical usage scenarios of emotionally-aware systems are numerous and range from small personalized applications to incorporation into big industries. Personal gadgets, such as smartphones and game consoles, will benefit from collecting data about user's emotional state to adapt their behavior and provide a better user experience; car electronics will monitor driver's stress level and sleepiness to prevent dangerous situations on the roads; banks and call-centers will be able to provide a better quality service; law enforcement agencies may escalate their security measures by automatically spotting suspicious activities. A good example of implementing such technologies on a bigger scale is medical treatment, where there is a need of continuous monitoring of a patient's emotional and psychophysical state, aiding an expert in health examination procedures, as well as better interacting and assisting people struggling to perceive emotions of others — such as children with autism. Involvement of automatic technologies proves beneficial insofar it allows capturing subtle characteristics that may escape from human's eyes and/or ears [2, 5].

Analysis of human emotions is possible from different information channels (modalities). To name a few — voice, face, gestures, postures, eye gaze and other physical measurements, such as electrocardiogram and skin conductance. Integrating several modalities in a single task is the main idea behind multimodal approaches towards emotion recognition. Multiple modalities are complimentary in nature and often bear redundant information. This allows to effectively battle the problem of missing values, where observations from one or several modalities may not be available at a particular time span. However, not all of them are equal in the amount of

information they provide. Voice, among other modalities, provides essential paralinguistic cues that are indicative of speaker's feelings. There are studies, which show that it is easier to read emotions from voice rather than from face [5], probably because people are better at hiding emotions from face, but not from voice [6]. Also, it has been established that face is more reliable conveying the valence of the portrayed emotion, while voice provides a better estimate of its intensity [7]. It means that speech is rather indicative of how calm or excited a person is, although it is not always easy to tell in which way — positive or negative. Thus, considering acoustic features of the signal, angry and happy voices sound similar in that they both carry a high amount of energy, higher fundamental frequency (F0), as well as wider F0 range, upward intonation contours, faster speaking rate and steeper spectral slope [6]. For the reasons described above this work is focused on automatic emotion recognition from speech.

2. Related work. There are several issues that impede the development of speech emotion recognition systems, one of them being absence of understanding what features should be used for classification [8]. Many current state-of-the-art systems use a large amount (on the order of thousands) of presumably relevant features with subsequent dimensionality reduction technique such as Principal Component Analysis (PCA) [9]. Another popular approach is to abandon predefined feature extraction and feed raw signals into a deep neural network, which finds a new feature representation without any hand-crafted engineering [10]. Another big issue, which renders the task of classifying emotions difficult, is the inherent inter-speaker and intra-speaker variability. Inter-speaker variability arises from different people having different age, gender, individual voice quality and characteristics. Intra-speaker variability adds uncertainty even more, for the voice quality of the same person changes with health condition, general mood, social environment, etc.

Cultural differences are another source of ambiguity encountered when modeling emotional states. A lot of debates were spawn around the question whether emotions are universal among people of all ages, genders, cultures and languages [11]. Do these factors define the way emotions are expressed and perceived and if so, how do emotions differ? There are few cross-corpus studies showing results on how well a given model trained on a particular database generalizes to other language databases [12]. Nevertheless, currently there is not enough data to prove or disprove the argument.

Feature extraction is an important step in the overall emotion recognition pipeline. The features can be extracted on two different levels: frame level and utterance level.

On the frame level, a certain amount of features are being extracted from analysis windows of the signal with a predefined frequency, usually

about 100 frames per second. This gives localized characteristics of the signal. To obtain the utterance level features, summarizing functionals are popularly applied to low level descriptors to form global utterance characteristics. Mere extraction of the features from the raw signal may not be enough; to gain a good performance, feature representation learning can be used to build high-level features from low-level features [13, 14]. Utterance-level representation of features was also explored by Kim [15]. Combination of different feature representations has also been explored; for example, Li Yang and Yunxin Zhao proposed to apply a shifting short-time window to extract short-term features and then applied functionals to the resulting sequences to obtain long-term feature representation, following with PCA dimensionality reduction and classification [16]. Dan-Ning Jiang and Lian-Hong Cai made use of temporal features alongside statistical features with GMM and HMM to benefit from both representations [17]. Some authors used combined frame and turn level analysis via HMM and statistical functionals [18].

Various types of classifiers have been used for the task of emotion classification from speech. Some of the desirable characteristics involve ability to work with small sets of data, handling missing values and outliers. As a result of Deep Neural Networks (DNN) becoming more and more powerful [19], various deep architectures, such as convolutional and recurrent NNs are actively being exploited [20-22]. A special type of recurrent neural network (RNN) called Long Short-Term Memory (LSTM) is particularly popular due to its ability to model arbitrarily large temporal sequences [7, 23]. It is an important property for emotion classification since emotions naturally evolve in time and therefore emotionally colored speech signals preserve a temporal structure.

3. Proposed method. General pipeline of a machine learning setup consists of feature extraction, preprocessing and training/testing stage. The baseline methods against which we compare the results consists of a single branch feature representation using predefined INTERSPEECH 2010 feature set (utterance-level functionals extracted via openSMILE toolkit), and a single LSTM Neural Network. The overall baseline pipelines are depicted in Figures 3 and 4. Preprocessing included feature normalization. The classification method was chosen to be logistic regression for the reasons discussed below.

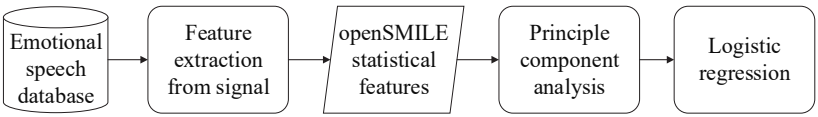


Fig. 3. Baseline method 1: PCA + Logistic Regression

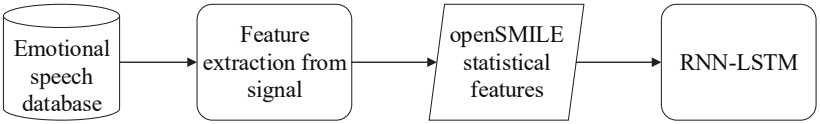


Fig. 4. Baselie method 2: LSTM-RNN

In our work, we propose a novel approach to implement a hybrid feature representation combining two different feature levels to take advantage of both short-time and long-time signal characteristics (Figure 5). The two levels of feature extraction are frame-level and utterance-level. The former corresponds to low level descriptors (LLDs), which are being extracted from every audio frame and form a sequence of feature vectors for every utterance. The length of the sequence depends on the duration of the audio signal and may vary for every utterance. To account for the temporal changes in these features as well as to match the resulting feature vectors of every utterance to have the same size, we let the sequences of LLDs to pass through an LSTM network and set the output of the network from the last frame to be the resulting feature vector describing the given utterance. Because the LSTM network has memory cells that allow accumulating information, the output from the last frame will have accumulated information from all the previous frames.

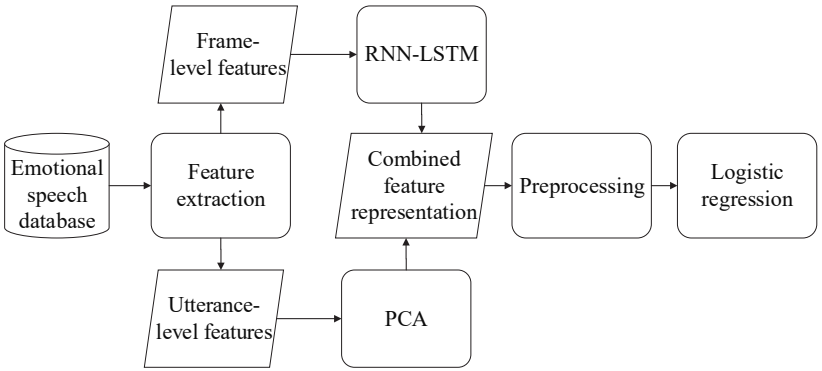


Fig. 5. General scheme of the proposed method

On the other hand, the utterance-level features represent statistical functionals applied to all the LLDs within one utterance and form a feature vector of a fixed size. Because the dimensionality of the resulting feature vector is very high, principal component analysis is used to reduce the number of features as well as to decorrelate them. The resulting two feature representations are concatenated to form a single feature vector, which is further fed into a logistic regression classifier that makes predictions about class labels.

The reason for combining the two different representations is that short-time characteristics (frame level features) together with appropriate modelling techniques allow capturing the temporal structure of the signal, while long-time characteristics (utterance-level features) are capable of expressing general trends [24]. Therefore, by combining the two approaches it is possible to benefit from both temporal dynamics as well as the big picture provided by statistical functionals.

4. Feature extraction and selection. The de-facto standard for feature extraction in the field of affective computing is openSMILE toolkit provided by German company audEERING [25]. There are predefined feature sets that were released during the series of annual INTERSPEECH Computational Paralinguistics Challenge (ComParE) [26-29]. The number of LLD features, as well as the number of applied statistical functionals and the total number of resulting utterance-level features are summarized in Table 1. After experimenting with these feature sets it turned out that the feature set, released in the year 2010, showed a better performance in comparison with other feature sets. The details about the LLDs and functionals in this set are presented in Table 2, the results of the comparison can be seen in Figure 6. Only the most significant number of principal components is shown.

Table 1. openSMILE configuration sets

Configuration set	LLD	Functionals	Total
INTERSPEECH-2009	16	12	384
INTERSPEECH-2010	38	21	1582
INTERSPEECH-2011	60	33	4368
INTERSPEECH-2013	65	54	6373

Table 2. INTERSPEECH 2010 paralinguistics challenge feature set

Low-Level Descriptors	Functionals
PCM loudness	Position maximum/minimum
MFCC [0-14]	Arithmetic mean, Standard deviation
Log Mel Freq. Band [0-7]	Skewness, Kurtosis,
LSP Frequency [0-7]	Linear regression coefficients 1/2
F0 by sub-Harmonic sum	Linear 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

Four different feature sets were compared by the accuracy, which was possible to obtain on RUSLANA corpus using particular feature set with the rest classification scheme being equal. The INTERSPEECH 2009 (IS_09) feature set consisted of 384 features, which was considered small enough not to apply any dimensionality reduction techniques. For the rest of the feature sets Principal Component Analysis (PCA) was applied in order to reduce the size of the feature vectors and decorrelate the features. The number of principal components was ranging from 10 to 1000. The INTERSPEECH 2010 feature set comprised 1582 features, INTERSPEECH 2011 feature set — 4368 features and the INTERSPEECH 2013 feature set included 6373 features in total. These features are obtained by applying certain statistical functionals to LLDs extracted on the frame level and represent utterance-level features.

As can be seen from Figure 6, the INTERSPEECH 2010 feature set has shown the best performance in comparison to other available feature sets. Therefore, further research was focused to include only features contained in this set. 38 LLDs were extracted from the audio signal at the frame rate of 100 fps, with the windows of various types and lengths. Hamming window of 25 msec was used for all the features, except the fundamental frequency (F0). The window applied to extract F0 was Gaussian with the length of 60 msec. Moving average filter was applied to all of the characteristics. 21 functionals as well as first order regression coefficients were applied to LLDs. 16 zero features were removed from the set (such as minimum F0 value — always zero). Two other additional features were included — the number of raises and the length of the F0 curve.

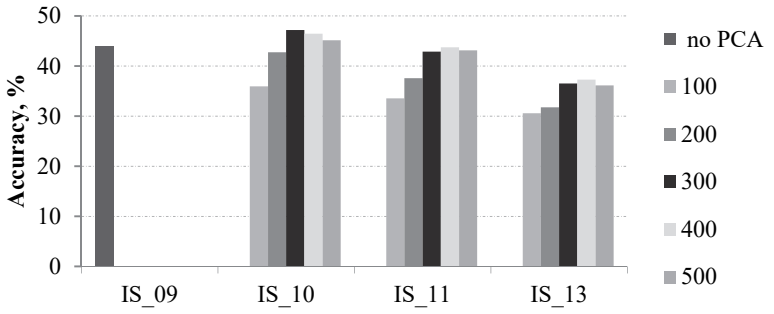
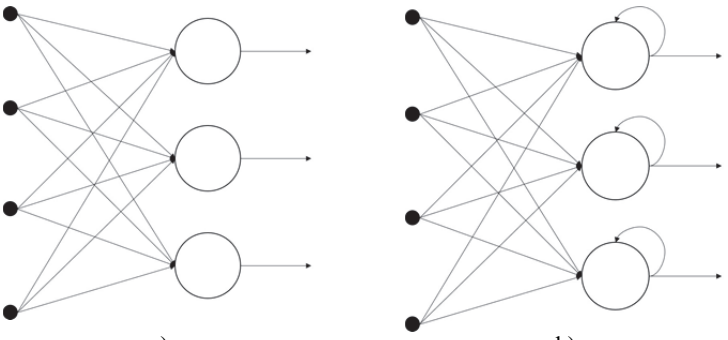


Fig. 6. Comparison of 4 openSMILE feature sets on RUSLANA database using Logistic Regression

PCM loudness is the normalized intensity raised to a power of 0.3. MFCC [0-14] refers to Mel-frequency cepstral coefficients from 0 to 14. Log

Mel Frequency Band [0-7] is the logarithmic power of Mel-frequency bands from 0 to 7 (distributed over a range from 0 to 8 kHz). LSP Frequency [0-7] is the 8-line spectral pair frequencies computed from 8 LPC (Linear Prediction Coding) coefficients. F0 envelop is the envelope of the smoothed fundamental frequency contour. The vioicing probability reflects the likelihood that the frame is pitched. Jitter and shimmer are characteristics of voice quality that reflect frequency and amplitude instability, respectively. Jitter local is the frame-to-frame pitch period length deviations. Shimmer local is the frame-to-frame amplitude deviations between pitch periods. Jitter DDP is the differential frame-to-frame jitter (the jitter of the jitter). Some functionals are not applied to every low-level descriptor. For example, minimum, maximum, mean and standard deviation are not applied to voice related LLDs except for fundamental frequency F0.

5. LSTM Recurrent Neural Networks. Recurrent Neural Networks [30] are a special type of Neural Networks that have feedback connections, i.e. the output of an RNN unit is connected to the input of the same unit in order to provide the network ability to the previous activations. This allows to store memory about previous elements of a sequence and make better decisions in future. The general idea of an RNN network with feedback connections is shown below in Figure 7. On the left (a), the figure shows connections of a hidden layer from a Feedforward Neural Network, where the information propagates strictly forward. On the right (b), the same piece of network is shown with added feedback connections to each of the hidden neurons that turns the architecture into a Recurrent Neural Network. The feedback connections may also connect the neurons on the current layer with the previous layers, which allows to build more complex models.



a) b)

Fig. 7. Recurrent Neural Network structure in comparison with a simple Feedforward Neural Network: a) – feedforward Neural Network; b) – recurrent Neural Network

However, this kind of architecture has several drawbacks, one of which is inability to store sufficiently long context information. Some estimates suggest that the network forgets information after ten steps of iteration. For these reason Recurrent Neural Networks are said to possess short term memory. Vanishing gradients also pose difficulty for applying such networks in practice.

In order to cope with these drawbacks, a special type of cell was invented that can memorize and store an arbitrary amount information in time. It was called Long Short-Term Memory to emphasize that it overcomes the loss of information in RNNs. A typical LSTM cell has the structure depicted in Figure 8. It has an Input Gate i_t, Forget Gate f_t, and Output Gate o_t that regulate the amount of information which is being stored in the cell. The flow of data within the cell also allows to pass through the cell unchanged so that the problem of vanishing gradients is no longer a problem for LSTM. The gates are regulated by equations 1-3, where σ denotes sigmoid function. Each gate has its own weight coefficients W_f, W_i, W_o and biases b_f, b_i, b_o, which are optimized during training. The cell state C_t and output hidden state h_t are given in equations 4-5. The cell state is updated in accordance with the previous cell state and new candidates at the input gate. Output h_t represents a filtered cell state.

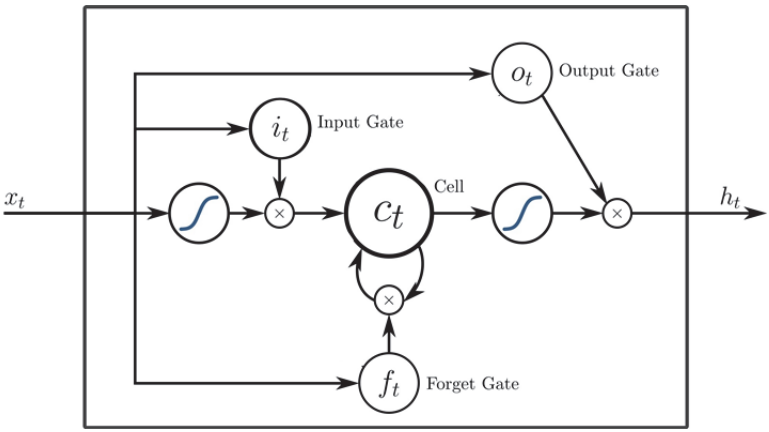


Fig. 8. Structure of a basic LSTM cell [30]

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right), \tag{1}$$

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right), \tag{2}$$

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right), \tag{3}$$

$$C_t = f_t * C_{t-1} + i_t * tanh\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_c\right), \tag{4}$$

$$h_t = o_t * tanh\left(C_t\right). \tag{5}$$

Another improvement in LSTM cell structure is the additional peephole connections, which are depicted in Figure 9. Peephole connections are allowing for information to flow directly to the Input, Output and Forget Gates, letting them access to the current cell state.

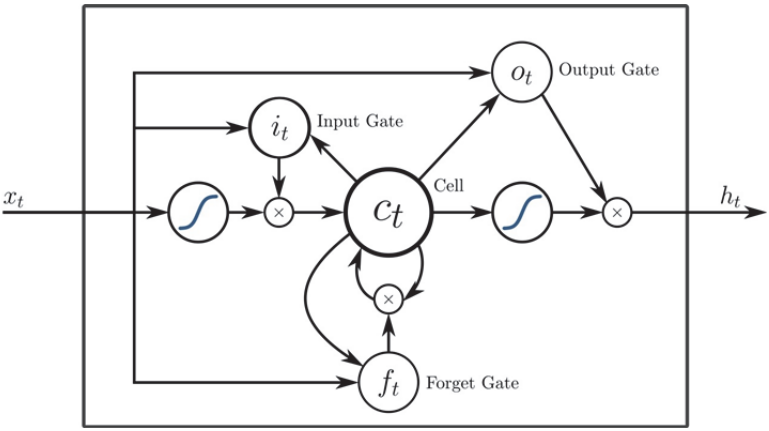


Fig. 9. Structure of an LSTM cell with peephole connections [30]

Different implementations will use some peepholes and not the others. One more variation is to couple forget and input gates. The implication is that the decision on what to store and what to forget are no longer independent and are made jointly, i.e. forgetting happens only when there is new data to put in the memory. Another popular variation on the LSTM cell is the Gated Recurrent Unit (GRU), introduced by Cho et al. [31]. In this particular implementation, the input gate and the forget gate are combined in a single update gate. Cell state and the output hidden state are also merged together. The structure of such unit is simpler than a standard LSTM cell and requires less computation, while providing similar performance.

In an LSTM network, LSTM units replace feedforward neurons. Typically, a hidden layer of LSTM cells would be followed by a fully connected layer, that connects the LSTM layer with the output. The input is represented by a 3-dimensional array, where samples are sequences of

features. Assuming that the number of samples is n, the sequences are of length s and feature vector contains f features, the final topology of the network is depicted in Figure 10.

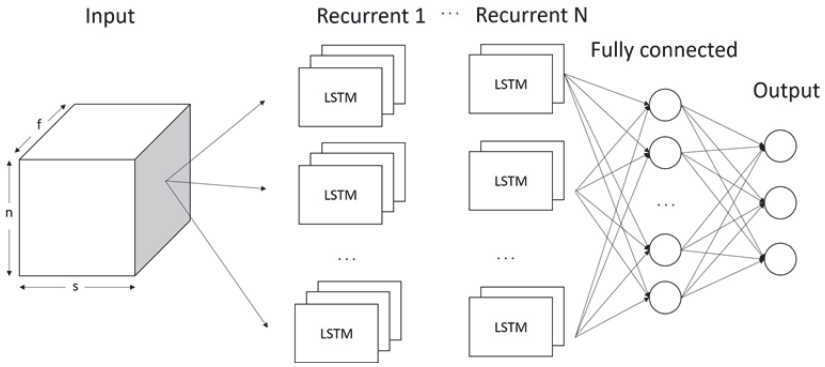


Fig. 10. Topology of an LSTM network: n – number of samples, s – length of the sequence, f – number of features in each observation in the sequence, N – number of recurrent layers used in the network

6. Feature representation using LSTM. In the second branch, frame-level LLD sequences are extracted via openSMILE toolkit in accordance with the feature selection methods described above. These sequences are passed through a unidirectional one-layer RNN-LSTM network with 300 hidden units, one frame at a time. After the last frame has propagated through the network, the output of the network is saved as a feature vector, containing information about temporal structure of the signal. The process of representing an utterance with LSTM unrolled in time is depicted in the Figure 11. The LLD sequence of one particular utterance is denoted as $f_0, f_1, \ldots, f_d$, where d –the number of frames in the utterance. The LSTM network depicted in the diagram is the same network, unrolled in time, to show that only the last output of the LSTM is considered as the feature vector representation for that one utterance.

The training of LSTM network is done as if it performs the function of a classifier — that is, a pair of (feature vector, label) is presented to the network, the weights get updated and the procedure repeats until the network has seen all the samples and the training process converges. Next, the last layer of the network, the softmax layer with the number of units equal to the number of prediction classes, is removed and the output of the network is no longer a probability of belonging to a particular class, but rather a complex 300-dimensional non-linear representation of the input features. The two feature representations are concatenated together to form a single feature vector that is passed to the classification step.

The architecture of the used RNN-LSTM network contains one hidden layer of 128 LSTM units. The initial learning rate is set to 0.001 and is decreased every 100 epochs by the factor of 0.1. The learning process was stopped when the learning curve is not improved for 15 consecutive epochs. Dropout was not applied out of the data size concern: most of the datasets are small in size. L2-regularization was implemented alongside cross-entropy loss function. In all the experiments the optimizer of choice was Adam. Mini-batches of size 250 were used. In all the experiments LSTM implementation was carried out with TensorFlow framework [32].

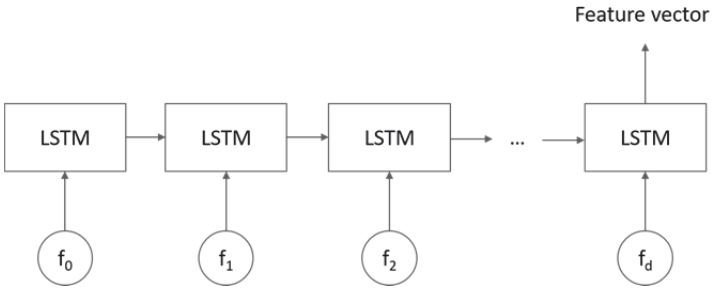


Fig. 11. Feature vector representation via LSTM network. The LLD sequence of one particular utterance is denoted as f_0, f_1, …, f_d, where d is the number of frames

7. Feature representation using PCA. The proposed method consists of two different feature representation methods, implemented in parallel, as can be seen in Figure 8. In the first branch, utterance-level functional are computed from LLDs. Statistical functional are known to have more expressive power than LLDs, considering the suprasegmental nature of emotions [24], however they fail to take into account the temporal changes in the signal, which are also of great importance since all emotions evolve in time. The number of possible functionals is very big and it is not clear which ones are more informative than others. They also often have a high correlation. For these reasons and a common problem known as curse of dimensionality [33], PCA is performed to reduce the dimensionality of resulting feature vectors and de-correlate the features.

Mathematically, PCA [34] is defined as an orthogonal linear projection of data to a new coordinate system of the form $\mathbf{T} = \mathbf{XW}$, where $\boldsymbol{X} = \left(x_1,\ldots,x_n\right)^T$ is the data matrix, $\mathbf{W} = \left(w_1,\ldots w_p\right)$ is a $p \times p$ weight matrix, whose columns are eigenvectors of $\mathbf{X}^T\mathbf{X}$, and $\mathbf{T}$ is the new PCA representation of the data. The transformation maps each row vector of data $x_{(i)}$ to a new vector of principal component scores $t_{(i)} = \left(t_1,\ldots,t_m\right)_{(i)}$ given

by $t_{k(i)} = x_{(i)} \cdot \boldsymbol{w}_{(k)}$ for $i = 1 \ldots n, k = 1 \ldots m$ with each weight vector $\boldsymbol{w}$ being a unit vector. The greatest variance happens to lie on the first coordinate, which is also called the first principal component, the second greatest variance – on the second principal component and so on. Principal components are visualized in the directions of greatest variance in Figure 12.

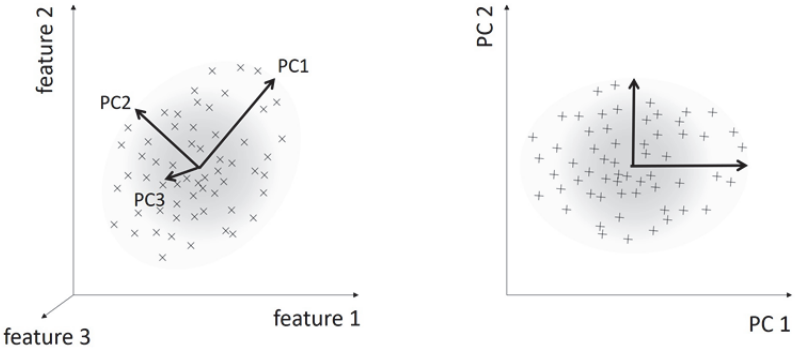


Fig. 12. Illustrations of the Principal Component Analysis: principal components point in the direction of greatest variance

To maximize variance, the first weight vector $\boldsymbol{w}_{(1)}$ must satisfy the following condition:

$$\mathbf{w}_{(1)} = \arg\max_{w=1} \left\{ \mathbf{X}\mathbf{w}^2 \right\} = \arg\max_{w=1} \left\{ \mathbf{w}^T \mathbf{X}^T \mathbf{X}\mathbf{w}^2 \right\} = \arg\max \left\{ \frac{\mathbf{w}^T \mathbf{X}^T \mathbf{X}\mathbf{w}}{\mathbf{w}^T \mathbf{w}} \right\}. \quad (6)$$

After the $\boldsymbol{w}_{(1)}$ is found, the first principal component of a data vector $x_{(i)}$ can be found in the transformed coordinates as $t_{1(i)} = x_{(i)} \cdot \boldsymbol{w}_{(1)}$. The k-th principal component can be found by subtracting from $\mathbf{X}$ the first k-1 principal components and then finding the weight vector corresponding to the maximum variance from the new data $\hat{\mathbf{X}}_k$.

$$\hat{\mathbf{X}}_k = \mathbf{X} - \sum_{s=1}^{k-1} \mathbf{X}\mathbf{w}_{(s)}\mathbf{w}_{(s)}^T. \quad (7)$$

8. Classification. A comparative study [35] was conducted on the RUSLANA database to find out the best classification scheme suited for the task. The results of the study are summarized in Figure 13. As can be seen

from the figure, the best prediction accuracy was achieved via Logistic Regression (Log-R) classifier. The second-best approach turned out to be Support Vector Machine (SVM), and the third best — a simple artificial Neural Network (NN) with one hidden layer. Other classifiers, that were used in the comparative study, are: Linear Regression (Lin-R), Naïve Bayes (NB), k-Nearest Neighbours (kNN), and Random Forest (RF).

Linear kernel gave the best result for SVM classification, which proves that with a high number of features non-linear projections don't improve the efficiency of the system. The optimal value for C parameter was chosen empirically. In kNN approach, the best k parameter was also found empirically and was equal to 10 neighboring points. The number of hidden layers in NN varied from 1 to 2, number of neurons in each layer — from 50 to 500. The number of epochs and learning rate ranged from 100 to 500, from 0.1 to 0.0001 respectively. RF parameter was optimized experimentally.

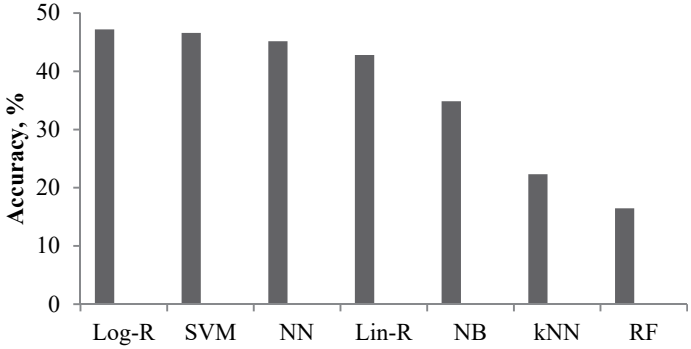


Fig. 13. Classification results of various classifiers applied to RUSLANA

In order to achieve speaker-independent properties of the system we would like to know whether a model trained on a particular set of persons generalizes well to the group of previously unseen voices. To implement this, we need to ensure that all the samples in the validation set are not represented at all in the training set. This is achieved via leave-one-subject-out cross validation strategy, which ensures that the same subject is not represented in both testing and training sets. Hence, the overall experiments are handled as 61-fold subject-independent cross-validation, i.e. the training was implemented on 60 speakers and the testing on the remaining speaker to achieve speaker-independent properties of the system.

9. Logistic Regression. Regression models are used for defining the relationship between one dependent variable and several independent variables. When the dependent variable is in discrete form, the regression turns into a binary classification task. Let us assume that output y_i is a realization of

a random variable Y_i, which takes on one of the two values: 0 and 1 with probabilities π_i and $1-\pi_i$. Such distribution is called a Bernoulli distribution and can be written in the form $\Pr\{Y_i = y_i\} = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$. When building a linear regression an assumption is being made about probabilities π_i having linear dependency on observations: $\pi_i = x_i'\beta = \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n$, where β is a vector of regression coefficients.

In order to limit the prediction values to the range from 0 to 1, probabilities π_i are transformed to $\frac{\pi_i}{1-\pi_i}$, which reflects the chance of a particular event happening, which can take on any positive value. After that the transformed values undergo logistic transformation $\eta_i = \text{logit}(\pi_i) = \log\frac{\pi_i}{1-\pi_i}$. With such representation, when the chance is close to 0, logit is approaching $-\infty$. From the other side, when the probability is close to 1, both the chance and the logit approach $+\infty$. Therefore, the logit transformation projects probability from the range {0, 1} on to the whole rational number space. When the probability is 0.5, the logistic function is 0. Negative numbers correspond to probabilities < 0.5, and positive numbers correspond to probabilities > 0.5. This logit transformation is unique and therefore reversable: $\pi_i = \text{logit}^{-1}(\eta_i) = \frac{e^{\eta_i}}{1+e^{\eta_i}}$. Logistic regression assumes that probability logit is a linear model: $\text{logit}(\pi_i) = x_i'\beta$. Interestingly enough, logistic regression provided a better performance in comparison to simple feedforward Neural Network, despite having a simpler structure. This phenomenon can be explained by regularization effects.

Regularization can be used in all regression models by adding another term to the loss function, which does not allow for coefficients to grow arbitrarily large, causing model overfitting. Hence, regularization helps to generalize the model and abstract away from the particularities.

There are several underlying assumptions on which the logistic regression models are build. First, it is assumed that there is no linear dependency between regression model factors; second, the variance is assumed to be constant. In practice, it is not always possible to comply with such conditions. Despite of that, logistic regression models are commonly used in practice.

10. Emotional speech corpora. We used 3 open source speech databases to evaluate our proposed method. These corpora are:

RUSLANA [36] (Russian speech), EMODB [37] (German speech), and BUEMO [38] (Turkish speech). The databases were chosen so that they have a similar structure and different languages to see how well the model can be generalized. All of them are well-balanced, i.e. having the same number of samples per each class, acted by non-professional actors, with a predefined set of categorical emotions.

RUSsian LANguage Affective corpus RUSLANA is a Russian language emotional speech database collected at the Department of Phonetics, St. Petersburg State University, Russia, in a sound proof recording studio. It contains audio recordings of 61 subjects (12 male and 49 female), in the age of 16-28 years old, all of whom were university students and native speakers of the standard Russian language. Each speaker pronounced 10 phonetically representative decontextualized sentences, which included all the Russian phonemes and most commonly encountered consonant clusters. Each sentence represented one of the 10 syntactic types, corresponding to distinct intonation contours, which coincide with different intonation contours in Bryzgunova's classification [39] inherent to the Russian language. Every sentence was pronounced with the following six basic emotional states: Neutral (N), Surprise (S), Happiness (H), Anger (A), Sadness (D) and Fear (F). Therefore, there are 61 speakers × 10 sentences × 6 emotion classes = 3660 audio files in total, each ranging in length from 2 to 5 seconds.

RUSLANA was chosen for the training purposes for several reasons. First, it is a well-balanced corpus containing an equal number of emotional utterances in every class; second, the actors are not professional and thus they do not reflect exaggerated intonation and other prosodic characteristics, common to professional performers [40]; third, the corpus construction strategy takes into account peculiarities of the Russian language and therefore provides a useful framework for developing a speech emotion recognition system from Russian speech, research on which until now has been sparse and inconsistent.

A database of German emotional speech (EMODB) was collected at the Department of Technical Acoustics of the Technical University of Berlin, in an anechoic chamber with high-quality recording equipment. Ten non-professional actors (5 male and 5 female) whose age ranged from 21 to 35 years old repeated 10 utterances with the following 7 different emotions: Neutral, Anger, Fear, Joy, Sadness, Disgust and Boredom. The total number of utterances is 10 speakers × 10 sentences × 7 emotion classes = 700 with some additional second versions reaching up to 800 recordings. The utterances of two types (a short sentence and a two phrase sentence) were constructed from everyday life usage.

The speech recordings of the BUEMODB dataset were collected in Bogazici University in Istanbul. There are 121 utterances for each of four emotional classes (Angry, Happy, Neutral and Sad) from 11 amateur theatre

actors (4 male and 7 female) by means of Stanislavskiy effect for generating emotional utterances [41]. The total number of audio files is 11 speakers × 11 sentences × 4 emotion classes = 484. The information about all three databases are summarized in Tables 3 and 4. Table 3 shows general information of how many speakers participated in database collection, what language was used for recording and number of emotion classes. Table 4 shows distribution of number of utterances among the classes.

Table 3. Corpora summary

Corpus	Lang.	Speakers	Females	Classes
EMODB	German	10	5	7
RUSLANA	Russian	61	49	6
BUEMODB	Turkish	11	7	4

Table 4. Corpora class distribution. Classes are abbreviated as follows: N – Neutral, A – Anger, H – happiness, S – Sadness, D – Disgust, F – Fear, B – boredom

Corpus	N	A	H	S	D	F	B	All
EMODB	79	127	71	62	51	69	81	535
RUSLANA	610	610	610	610	610	610	-	3600
BUEMODB	121	121	121	121	-	-	-	484

11. Experimental results and discussion. After experimenting with the number of components used in the PCA analysis we can conclude that for different datasets the optimal number of principal components differs, which may be explained by different data distributions due to varying recording conditions and audio signal quality. This is in consent with the recent study of [42], which showed evidence that the nature of dataset gives an idea to select relevant classifiers. The results are presented in the Table 5.

Table 5. Classification accuracy (%) of logistic regression with various numbers of principal components

	Classification accuracy (%)		
PCA #	EMODB	RUSLANA	BUEMO
10	52.1	31.7	49.4
50	*56.3*	34.2	52.5
100	48.3	36.9	53.4
200	38.4	43.6	52.8
300	51.1	*47.2*	55.1
400	52.3	44.9	*56.0*
500	52.2	44.8	56.2

EMODB showed the best performance at the minimum number of principal components being equal 50, which allowed obtaining classification accuracy of 56.3%. Both RUSLANA and BUEMO datasets required more number of principal components in order to achieve optimal performance. In case of RUSLANA, 300 principal components obtained via PCA resulted in 47.2% classification accuracy. Among other datasets optimal performance on BUEMO was achieved with the highest number of principal components being equal 400. In all the cases the original number of features was reduced more than three times. This indicates that high number of features indeed suffers from curse of dimensionality. The apparent conclusion is that feature decorrelation and dimensionality reductions techniques are an important preprocessing step in emotion classification task.

The baseline classification method, which consisted of single branch openSMILE utterance-level feature extraction, showed maximum classification accuracy of 47.2% on the RUSLANA dataset, 56.3 on the EMODB dataset, and 56.0 on the BUEMO dataset. The second baseline using a single LSTM network gave the worst results, as expected: 40.5% on the RUSLANA dataset, 45.3% on the EMODB dataset, and 39.9% on the BUEMO dataset. The implementation of the single LSTM network alone showed a severe overfitting problem, presumably due to a small size of the dataset.

Implementing the newly proposed method allowed us to obtain a relative improvement over the first baseline (PCA + Logistic regression) 2.3 % in classification accuracy on the RUSLANA corpus and 2.8% on the BUEMO corpus. EMODB, on the contrary, did not show a relative improvement but a comparative performance with the combination of the proposed techniques. The relative improvement over the second baseline was much more prominent 9.0%, 10.8%, 18.9%. The details of the obtained results are summarized in Table 6.

Table 6. Classification accuracy (%) of the baseline and the proposed methods

Modeling method	RUSLANA (6 classes)	EMODB (7 classes)	BUEMO (4 classes)
PCA + Logistic regression	47.2	56.3	56.0
LSTM	40.5	45.3	39.9
Proposed Combination	49.5	56.1	58.8

The number of classes differed from corpus to corpus and therefore the interpretation of the results should be made accordingly. In case of 4

classes (BUEMO), the chance level is 25%, and so the proposed method worked more than twice better than the chance level. In the case of 6 classes (RUSLANA) the chance level is 16.(6)% and so the proposed implementation achieved performance almost 3 times better than the chance level. In case of 7 classes (EMODB), the chance level is at 14.29% giving relative improvement of the proposed approach almost 4 times better than the chance level.

Training an LSTM network on EMODB and BUEMO corpora showed significant overfitting results despite implementing various regularization techniques, such as early stopping and dropout. This can be explained by the small dataset sizes (114 samples per class on average for EMODB, and 121 samples per class for BUEMO). RUSLANA, on the contrary, had 610 samples per class and therefore showed more consistent results on the training. The reader is referred to [43] for more experimental results on the BUEMO corpus.

A sentiment analysis based on automatic text processing can help additionally improving the quality of bimodal speech emotion recognition [44]; automatic speech recognition techniques (e.g. [45, 46]) should be applied for speech-to-text transformation in this case.

12. Conclusions. We have proposed a new method for combining two feature representations for emotion classification from speech: a frame-level representation of low-level descriptors and an utterance-level representation of LLD functionals. The proposed approach is motivated by the need to account for dynamic nature of emotion evolution in time, as well as the trade-off between local LLD features, which give an insight on the temporal changes in signal, and global statistical functionals, which are known to better capture the general trends. Our method was built on traditional application of statistical openSMILE features with PCA dimensionality reduction combined with the recent state-of-the-art LSTM RNN technologies. The optimal number of principal components lies in the range 50 to 400. Principal components less than 50 do not allow adequately modeling the underlying nature of emotions and therefore do not render optimal performance. Principal components more than 400 turn out to possess redundancy and high correlation, which also hinders the effectiveness of classification and shows worse performance in terms of classification accuracy. More features turn out to give worse performance probably due to bigger correlation, redundancy, noise and curse of dimensionality problem.

The experimental results were compared to two different baseline methods. One consisted of a single branch openSMILE utterance-level feature extraction, Principal Component Analysis dimensionality reduction and Logistic Regression classification. The other was based on a single LSTM-

RNN Neural Network trained on utterance-level openSMILE features. The proposed method showed increased classification accuracy compared to the baseline methods, however the results were not consistent across all the datasets. The explanation lies in the different corpus sizes: more samples per class guarantee better performance of the system and less overfitting issue. One of the drawbacks of the proposed method is that it requires a lot of data to be trained. However, with recent advances in cross-corpus analysis it is possible to combine different corpora in order to have more training data and more robust and stable learning process. The method proved effective combining the temporal dynamic changes in frame level features and general trends of utterance-level functionals. Therefore, the direction of future research will be to investigate possible ways of post-processing of the obtained feature representations, scaling and normalization techniques, as well as possibility of conducting cross-corpus analysis in order to upsample the training data.

References

1. Swain M., Routray A., Kabisatpathy P. Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*. 2018. vol. 21. no. 1. pp. 93–120.
2. Mustafa M.B., Yusoof M.A., Don Z.M., Malekzadeh M. Speech emotion recognition research: an analysis of research focus. *International Journal of Speech Technology*. 2018. vol. 21 no. 1. pp. 137–156.
3. Laukka P., Juslin P., Bresin R. A dimensional approach to vocal expression of emotion. *Cognition and Emotion*. 2005. vol. 19 no. 5. pp. 633–653.
4. Nogueira P.A., Rodrigues R., Oliveira E., Nacke L.E. Modelling human emotion in interactive environments: Physiological ensemble and grounded approaches for synthetic agents. *Web Intelligence*. 2015. vol. 13. no. 3. pp. 195–214.
5. Kraus M.W. Voice-only communication enhances empathic accuracy. *American Psychologist*. 2017. vol. 72. no. 7. pp. 644–654.
6. Kreiman J., Sidtis D. Foundations of voice studies: an interdisciplinary approach to voice production and perception. John Wiley & Sons. 2013. 512 p.
7. Wöllmer M. et al. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling. Proc. INTERSPEECH. Japan. 2010. pp. 2362–2365.
8. Anagnostopoulos C.-N., Iliou T., Giannoukos I. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*. 2015. vol. 43 no. 2. pp. 155–177.
9. Shabani S., Norouzi Y. Speech recognition using Principal Components Analysis and Neural Networks. 8th International Conference on Intelligent Systems. 2016. pp. 90–95.
10. Trigeorgis G. et al. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2016. pp. 5200–5204.
11. Kaya H. et al. Emotion, age, and gender classification in children's speech by humans and machines. *Computer Speech & Language*. 2017. vol. 46. pp. 268–283.
12. Kaya H., Karpov A.A. Efficient and effective strategies for cross-corpus acoustic emotion recognition. *Neurocomputing*. 2018. vol. 275. pp. 1028–1034.
13. Bengio Y., Courville A., Vincent P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2013. vol. 35. no. 8. pp. 1798–1828.

14. Ghosh S., Laksana E., Morency L.-P., Scherer S. Representation Learning for Speech Emotion Recognition. Proc. INTERSPEECH. 2016. pp. 3603–3607.
15. Kim Y., Provost E.M. Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2013. pp. 3677–3681.
16. Li Y., Zhao Y. Recognizing emotions in speech using short-term and long-term features. Fifth International Conference on Spoken Language Processing. 1998. vol. 6. pp. 2255.
17. Jiang D.-N., Cai L.-H. Speech emotion classification with the combination of statistic features and temporal features. ICME. 2004. pp. 1967–1970.
18. Vlasenko B., Schuller B., Wendemuth A., Rigoll G. Combining frame and turn-level information for robust recognition of emotions within speech. Proc. INTERSPEECH. Belgium. 2007. pp. 2249–2252.
19. Sainath T.N., Vinyals O., Senior A., Sak H. Convolutional, long short-term memory, fully connected deep neural networks. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2015. pp. 4580–4584.
20. Mao Q., Dong M., Huang Z., Zhan Y. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Transactions on Multimedia*. 2014. vol. 16. no. 8. pp. 2203–2213.
21. Tran D. et al. Learning spatiotemporal features with 3d convolutional networks. Proceedings of the IEEE international conference on computer vision. 2015. pp. 4489–4497.
22. Kim J., Truong K.P., Englebienne G., Evers V. Learning spectro-temporal features with 3D CNNs for speech emotion recognition. Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). 2017. pp. 383–388.
23. Chao L. et al. Long short term memory recurrent neural network based multimodal dimensional emotion recognition. Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge. 2015. pp. 65–72.
24. Schuller B. et al. The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals. Proc. INTERSPEECH. 2007. pp. 2253–2256.
25. 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. 2010. pp. 1459–1462.
26. Schuller B., Steidl S., Batliner A. The INTERSPEECH 2009 emotion challenge. Tenth Annual Conference on the International Speech Communication Association INTERSPEECH. 2009. pp. 312–314.
27. Schuller B. et al. The INTERSPEECH 2010 paralinguistic challenge. Proc. INTERSPEECH. 2010. pp. 2794–2797.
28. Schuller B. et al. The INTERSPEECH 2011 speaker state challenge. Proc. INTERSPEECH. 2011. pp. 3201–3204.
29. Schuller B. et al. The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. Proc. INTERSPEECH. 2013. pp. 148–152.
30. Greff K. et al. LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*. 2017. vol. 28. no. 10. pp. 2222–2232.
31. Cho K. et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078. 2014.
32. Abadi M. et al. Tensorflow: a system for large-scale machine learning. OSDI. 2016. vol. 16. pp. 265–283.
33. Keogh E., Mueen A. Curse of dimensionality. Encyclopedia of machine learning. Springer. 2011. pp. 257–258.
34. Jolliffe I. Principal component analysis. International encyclopedia of statistical science. Springer. 2011. pp. 1094–1096.

35. Verkholyak O., Karpov A. Combined Feature Representation for Emotion Classification from Russian Speech. International Conference on Artificial Intelligence and Natural Language AINL-2017. Springer CCIS. vol. 789. 2018. pp. 68–73.
36. Makarova V., Petrushin V.A. RUSLANA: A database of Russian emotional utterances. Seventh International Conference on Spoken Language Processing. 2002. pp. 2041–2044.
37. Burkhardt F. et al. A database of German emotional speech. Ninth European Conference on Speech Communication and Technology. 2005. pp. 1517–1521.
38. Meral H.M., Ekenel H.K., Ozsoy A. Analysis of emotion in Turkish. XVII National Conference on Turkish Linguistics. 2003.
39. Bryzgunova, E.A. *Zvuki i intinaciya v russkoy rechi* [Sounds and intonation of Russian Speech]. Moscow: Russkiy yazyk. 1977. 281 p. (In Russ.).
40. Anikin A., Lima C.F. Perceptual and acoustic differences between authentic and acted nonverbal emotional vocalizations. *The Quarterly Journal of Experimental Psychology*. 2017. pp. 1–21.
41. Kaya H., Salah A.A., Gurgen S.F., Ekenel H. Protocol and baseline for experiments on Bogazici University Turkish emotional speech corpus. 22nd Signal Processing and Communications Applications Conference (SIU). 2014. pp. 1698–1701.
42. Koolagudi S.G., Murthy Y.S., Bhaskar S.P. Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition. *International Journal of Speech Technology*. 2018. vol. 21. no. 1. pp. 167–183.
43. Kaya H., Karpov A.A., Salah A.A. Robust acoustic emotion recognition based on cascaded normalization and extreme learning machines. Proc. International Symposium on Neural Networks ISNN-2016. LNCS 9719. 2016. pp. 115–123.
44. Montacié C., Caraty M.J. Vocalic, lexical and prosodic cues for the INTERSPEECH 2018 self-assessed affect challenge. Proc. INTERSPEECH. 2018. pp. 541–545.
45. Besacier L., Barnard E., Karpov A., Schultz T. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication.* 2014. vol. 56. pp. 85–100.
46. Kipyatkova I., Karpov A. A Study of Neural Network Russian Language Models for Automatic Continuous Speech Recognition Systems. *Automation and Remote Control*. 2017. vol. 78. no. 5. pp. 858–867.

Verkholyak Oxana Vladimirovna — Junior Researcher of Speech and Multimodal Interfaces Laboratory, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS). Research interests: automatic emotion recognition is speech, affective computing, computational paralinguistics. The number of publications — 9. overkholyak@gmail.com; 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone: +7(812)328-0421, fax: +7(812)328-0421.

Kaya Heysem — Ph.D., Assistant Professor of Computer Engineering Department of Çorlu Faculty of Engineering, Namık Kemal University. Research interests: machine learning, pattern recognition, speech signal processing, emotions in speech. The number of publications — 40. hkaya@nku.edu.tr; Silahtarağa Mahallesi, Üniversite, 1, 59860 Çorlu/Tekirdağ, Turkey, office phone: +902822502346.

Karpov Alexey Anatolyevich — Ph.D., Dr. Sci., Associate Professor, Head of Speech and Multimodal Interfaces Laboratory, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS). Research interests: automatic speech recognition, multimodal interfaces, computational paralinguistics. The number of publications — 250. karpov@iias.spb.su; 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone: +7(812)328-0421, fax: +7(812)328-0421.

Acknowledgments. This research is supported by the Russian Science Foundation (project № 18-11-00145).

УДК 004.89 **DOI** 10.15622/sp.18.1.30-56

О.В. Верхоляк, Х. Кайя, А.А. Карпов

МОДЕЛИРОВАНИЕ КРАТКО- И ДОЛГОВРЕМЕННЫХ ЗАВИСИМОСТЕЙ РЕЧЕВОГО СИГНАЛА ДЛЯ ПАРАЛИНГВИСТИЧЕСКОЙ КЛАССИФИКАЦИИ ЭМОЦИЙ

Верхоляк О.В., Кайя Х., Карпов А.А. **Моделирование кратко- и долговременных зависимостей речевого сигнала для паралингвистической классификации эмоций.**

Аннотация. Распознавание эмоций в речи стало одним из важных направлений в области аффективных вычислений. Это комплексная задача, трудности которой во многом определяются необходимостью выбора признаков и их оптимального представления. Оптимальное представление признаков должно отражать глобальные характеристики, а также локальную структуру сигнала, поскольку эмоции естественным образом длятся во времени. Подобное представление возможно моделировать с помощью рекуррентных нейронных сетей (РНС — RNN), которые активно используются для различных задач распознавания, предполагающих работу с последовательностями. Предлагается смешанный подход к представлению признаков, который объединяет традиционные статистические признаки с последовательностью значений, полученных на выходе РНС с длинной кратковременной памятью (ДКП — LSTM) и хорошо моделирующих временную структуру сигнала. Таким образом, удается получить одновременное представление как кратковременных, так и долгосрочных характеристик, позволяющих использовать преимущества обоих подходов к моделированию признаков речевого сигнала. Для экспериментальной проверки предложенного метода была произведена оценка его эффективности на трех различных базах данных эмоционально окрашенной речи, находящихся в свободном доступе: RUSLANA (русская речь), BUEMODB (турецкая речь) и EMODB (немецкая речь). В сравнении с традиционным подходом результаты наших экспериментов показывают абсолютный прирост в точности распознавания эмоций в 2.3% и 2.8% для двух из вышеупомянутых корпусов, в то время как для третьего корпуса предложенный метод не уступает базовой системе. Следовательно, данный подход можно признать эффективным для моделирования эмоциональной окраски речевых высказываний при условии достаточного количества обучающих данных.

Ключевые слова: распознавание эмоций в речи, компьютерная паралингвистика, аффективные вычисления, представление признаков, моделирование контекста, искусственные нейронные сети, длинная кратковременная память.

Верхоляк Оксана Владимировна — младший научный сотрудник лаборатории речевых и многомодальных интерфейсов, Федеральное государственное бюджетное учреждение науки Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН). Область научных интересов: автоматическое распознавание эмоций в речи, аффективные вычисления, компьютерная паралингвистика. Число научных публикаций — 9. overkholyak@gmail.com; 14-я линия В.О., 39, Санкт-Петербург, 199178, РФ; р.т.: +7(812)328-7081, факс: +7(812)328-7081.

Кайя Хейсем — к-т техн. наук, доцент, Университет Намыка Кемаля (Турция), инженерный факультет Чорлу, Отдел вычислительной техники, Старший преподаватель. Область научных интересов: машинное обучение, распознавание образов, обработка речевых сигналов, эмоции в речи. Число научных публикаций — 40. hkaya@nku.edu.tr; р-н Силахтарага, ул. Университетская 1., 59860 Чорлу, Текирдаг, Турция; р.т.: +902822502346

Карпов Алексей Анатольевич — д-р техн. наук, доцент, заведующий лабораторией речевых и многомодальных интерфейсов, Федеральное государственное бюджетное учреждение науки Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН). Область научных интересов: автоматическое распознавание речи, многомодальные интерфейсы, компьютерная паралингвистика. Число научных публикаций — 250. karpov@iias.spb.su. 14-я линия В.О., 39, Санкт-Петербург, 199178, РФ; р. т. +7(812)328-0421, факс +7(812)328-0421.

Поддержка исследований. Работа выполнена при поддержке Российского научного фонда (проект № 18-11-00145).

Литература

1. *Swain M.*, *Routray A.*, *Kabisatpathy P.* Databases, features and classifiers for speech emotion recognition: a review // International Journal of Speech Technology. 2018. vol. 21. no. 1. pp. 93–120.
2. *Mustafa M.B.*, *Yusoof M.A.*, *Don Z.M.*, *Malekzadeh M.* Speech emotion recognition research: an analysis of research focus // International Journal of Speech Technology. 2018. vol. 21. no. 1. pp. 137–156.
3. *Laukka P.*, *Juslin P.*, *Bresin R.* A dimensional approach to vocal expression of emotion // Cognition and Emotion. 2005. vol. 19. no. 5. pp. 633–653.
4. *Nogueira P.A.*, *Rodrigues R.*, *Oliveira E.*, *Nacke L.E.* Modelling human emotion in interactive environments: Physiological ensemble and grounded approaches for synthetic agents // Web Intelligence. 2015. vol. 13. no. 3. pp. 195–214.
5. *Kraus M.W.* Voice-only communication enhances empathic accuracy // American Psychologist. 2017. vol. 72. no. 7. pp. 644–654.
6. *Kreiman J.*, *Sidtis D.* Foundations of voice studies: an interdiscipinary approach to voice production and perception // John Wiley & Sons. 2013. 512 p.
7. *Wöllmer M. et al.* Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling // Proc. INTERSPEECH. 2010. pp. 2362–2365.
8. *Anagnostopoulos C.-N.*, *Iliou T.*, *Giannoukos I.* Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011 // Artificial Intelligence Review. 2015. vol. 43. no. 2. pp. 155–177.
9. *Shabani S.*, *Norouzi Y.* Speech recognition using Principal Components Analysis and Neural Networks // IEEE 8th International Conference on Intelligent Systems. 2016. pp. 90–95.
10. *Trigeorgis G. et al.* Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2016. pp. 5200–5204.
11. *Kaya H. et al.* Emotion, age, and gender classification in children's speech by humans and machines // Computer Speech & Language. 2017. vol. 46. pp. 268–283.
12. *Kaya H.*, *Karpov A.A.* Efficient and effective strategies for cross-corpus acoustic emotion recognition // Neurocomputing. 2018. vol. 275. pp. 1028–1034.
13. *Bengio Y.*, *Courville A.*, *Vincent P.* Representation learning: A review and new perspectives // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2013. vol. 35. no. 8. pp. 1798–1828.
14. *Ghosh S.*, *Laksana E.*, *Morency L.-P.*, *Scherer S.* Representation Learning for Speech Emotion Recognition // Proc. INTERSPEECH. 2016. pp. 3603–3607.
15. *Kim Y.*, *Provost E.M.* Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2013. pp. 3677–3681.

16. *Li Y.*, *Zhao Y.* Recognizing emotions in speech using short-term and long-term features // Fifth International Conference on Spoken Language Processing. 1998. vol. 6. pp. 2255.
17. *Jiang D.-N.*, *Cai L.-H.* Speech emotion classification with the combination of statistic features and temporal features // ICME. 2004. pp. 1967–1970.
18. *Vlasenko B.*, *Schuller B.*, *Wendemuth A.*, *Rigoll G.* Combining frame and turn-level information for robust recognition of emotions within speech // Proc. INTERSPEECH. 2007. pp. 2249–2252.
19. *Sainath T.N.*, *Vinyals O.*, *Senior A.*, *Sak H.* Convolutional, long short-term memory, fully connected deep neural networks // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2015. pp. 4580–4584.
20. *Mao Q.*, *Dong M.*, *Huang Z.*, *Zhan Y.* Learning salient features for speech emotion recognition using convolutional neural networks // IEEE Transactions on Multimedia. 2014. vol. 16. no. 8. pp. 2203–2213.
21. *Tran D. et al.* Learning spatiotemporal features with 3d convolutional networks // Proceedings of the IEEE international conference on computer vision. 2015. pp. 4489–4497.
22. *Kim J.*, *Truong K.P.*, *Englebienne G.*, *Evers V.* Learning spectro-temporal features with 3D CNNs for speech emotion recognition // Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). 2017. pp. 383–388.
23. *Chao L. et al.* Long short term memory recurrent neural network based multimodal dimensional emotion recognition // Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge. 2015. pp. 65–72.
24. *Schuller B. et al.* The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals // Proc. INTERSPEECH. 2007. pp. 2253–2256.
25. *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. 2010. pp. 1459–1462.
26. *Schuller B.*, *Steidl S.*, *Batliner A.* The INTERSPEECH 2009 emotion challenge // Tenth Annual Conference on the International Speech Communication Association INTERSPEECH. 2009. pp. 312–314.
27. *Schuller B. et al.* The INTERSPEECH 2010 paralinguistic challenge // Proc. INTERSPEECH. 2010. pp. 2794–2797.
28. *Schuller B. et al.* The INTERSPEECH 2011 speaker state challenge // Proc. INTERSPEECH. 2011. pp. 3201–3204.
29. *Schuller B. et al.* The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism // Proc. INTERSPEECH. 2013. pp. 148–152.
30. *Greff, K. et al.* LSTM: A search space odyssey // IEEE Transactions on Neural Networks and Learning Systems. 2017. vol. 28. no. 10. pp. 2222–2232.
31. *Cho K. et al.* Learning phrase representations using RNN encoder-decoder for statistical machine translation // arXiv preprint arXiv:1406.1078. 2014.
32. *Abadi M. et al.* Tensorflow: a system for large-scale machine learning // OSDI. 2016. vol. 16. pp. 265–283.
33. *Keogh E.*, *Mueen A.* Curse of dimensionality. Encyclopedia of machine learning. Springer. 2011. pp. 257–258.
34. *Jolliffe I.* Principal component analysis. International encyclopedia of statistical science. Springer. 2011. pp. 1094–1096.
35. *Verkholyak O.*, *Karpov A.* Combined Feature Representation for Emotion Classification from Russian Speech // International Conference on Artificial Intelligence and Natural Language AINL-2017. Springer CCIS. vol. 789. 2018. pp. 68–73.

36. *Makarova V.*, *Petrushin V.A.* RUSLANA: A database of Russian emotional utterances // Seventh International Conference on Spoken Language Processing. 2002. pp. 2041–2044.
37. *Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., et al.* A database of German emotional speech // Ninth European Conference on Speech Communication and Technology. 2005. pp. 1517–152.
38. *Meral H.M.*, *Ekenel H.K.*, *Ozsoy A.* Analysis of emotion in Turkish // XVII National Conference on Turkish Linguistics. 2003.
39. *Брызгунова Е.А.* Звуки и интонация в русской речи // Москва: Русский язык. 1977. 281 с.
40. *Anikin A.*, *Lima C.F.* Perceptual and acoustic differences between authentic and acted nonverbal emotional vocalizations // The Quarterly Journal of Experimental Psychology. 2017. pp. 1–21.
41. *Kaya H.*, *Salah A.A.*, *Gurgen S.F.*, *Ekenel H.* Protocol and baseline for experiments on Bogazici University Turkish emotional speech corpus // 22nd Signal Processing and Communications Applications Conference (SIU). 2014. pp. 1698–1701.
42. *Koolagudi S.G.*, *Murthy Y.S.*, *Bhaskar S.P.* Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition // International Journal of Speech Technology. 2018. vol. 21. no. 1. pp. 167–183.
43. *Kaya H.*, *Karpov A.A.*, *Salah A.A.* Robust acoustic emotion recognition based on cascaded normalization and extreme learning machines // Proc. International Symposium on Neural Networks (ISNN-2016). LNCS 9719. 2016. pp. 115–123.
44. *Montacié C., Caraty M.J.* Vocalic, lexical and prosodic cues for the INTERSPEECH 2018 self-assessed affect challenge // Proc. INTERSPEECH. 2018. pp. 541–545.
45. *Besacier L., Barnard E., Karpov A., Schultz T.* Automatic speech recognition for under-resourced languages: A survey // Speech Communication. 2014. vol. 56. pp. 85–100.
46. *Kipyatkova I., Karpov A.* A Study of Neural Network Russian Language Models for Automatic Continuous Speech Recognition Systems // Automation and Remote Control. 2017. vol. 78. no. 5. pp. 858–867.