www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

PAN ARCHIVES OF ACOUSTICS
Vol. **45**, No. 1, pp. 129–140 (2020)
DOI: 10.24425/aoa.2020.132489

Research Paper

Analysis of Features and Classifiers in Emotion Recognition Systems: Case Study of Slavic Languages

Željko NEDELJKOVIĆ*, Milana MILOŠEVIĆ, Željko ĐUROVIĆ

School of Electrical Engineering
University of Belgrade
Bulevar Kralja Aleksandra 73, Belgrade, Serbia
*Corresponding Author e-mail: nz135003p@student.etf.bg.ac.rs

(*received January 26, 2019; accepted December 3, 2019*)

Today's human-computer interaction systems have a broad variety of applications in which automatic human emotion recognition is of great interest. Literature contains many different, more or less successful forms of these systems. This work emerged as an attempt to clarify which speech features are the most informative, which classification structure is the most convenient for this type of tasks, and the degree to which the results are influenced by database size, quality and cultural characteristic of a language. The research is presented as the case study on Slavic languages.

Keywords: emotion recognition; speech processing; classification algorithms.

1. Introduction

Automatic emotion recognition from speech gains in popularity as the number of opportunities for real world application increases. Emotional awareness enables software to adapt behaviour to its current user's emotional state. For example, safety is enhanced when a car detects that the driver is angry or tired and can then react to it. It also furthers the learning efficiency of students using e-learning systems, by detecting boredom or disinterest. In the case of call centres, emotionally aware call management software can improve customer satisfaction (Cowie *et al.*, 2001; Kołakowska *et al.*, 2014; Vinola, Vimaladevi, 2015). Besides, it is necessary that the software interface is adapted to the user's language and culture so as to achieve the effect of natural interaction. Furthermore, since voice parameters are influenced by the nature of the language (Pell *et al.*, 2009b) and people within the same culture recognise emotions better (Pell *et al.*, 2009a) it is important to investigate how emotions are displayed in different languages.

Automatic speech emotion recognition experiments in up-to-date research have been rarely performed using Slavic languages, and comparative analysis has not been done so far. This may be due to a lack of research data in these languages and the current state of database availability. So far, Slavic language databases of acted emotional speech in Serbian (GEES) (Jovičić *et al.*, 2004), Polish (PES) (Cichosz, 2008; Igras, Ziółko, 2013; Staroniewicz, Majewski, 2009) and Russian (RUSLANA) (Makarova, Petrushin, 2002) have been compiled. Databases of spontaneous speech have recently been developed in Croatian (CrES) (Dropuljić *et al.*, 2011), Czech (CzED) (Uhrin *et al.*, 2014) and Slovenian (EmoLUKS) (Justin *et al.*, 2015).

On the Serbian database several classification methods have been used for the five emotions classification task. The use of discrete Hidden Markov Models (HMMs) with Linear Prediction Cepstral Coefficients (LPCC), Log Frequency Power Coefficients (LFPC), total signal energy (E), teager energy (TE), fundamental frequency (F0) and values of the formants (FF) has reached 72% of recognition rate (Nedeljković, Đurović, 2015). In the case of Support Vector Machine (SVM) approach, the results oscillated between 62.78% and 91.3% depending on which test setup was used (Hassan, Damper, 2010; Milošević *et al.*, 2016). The results obtained so far on the Polish database PES have been contrasted: 50.73% using k Nearest Neighbours (kNN) and Mel Frequency Cepstral Coefficients (MFCC) (Kamińska *et al.*, 2013), whereas phoneme level formant features combined with Binary Decision Trees (BDT) give 81.9% (Ślot *et al.*, 2009). MFCC-SVM combination has provided 40.5%

when solely MFCC have been used and 33.75% when deltas and double deltas have been added (Kamińska *et al.*, 2017). The CrES database contains utterances which are linguistically very diverse and a large number of speakers. On this database, recognition rate for five emotions classification task has been 65.4%, using fused feature set and Random Forest (RF) classification method (Dropuljić *et al.*, 2016b). Experiments on the RUSLANA database have been dedicated to analyzing acoustic features of different phonemes in Russian emotional speech (Makarova, Petrushin, 2012). To our knowledge, automatic emotion recognition results have not yet been published for the Russian, Czech and Slovenian emotional speech databases. A more detailed overview of systems for recognising basic emotions in speech, tested on Slavic databases, is given in Table 1.

Numerous recent researches have been based on the general set of features defined in the INTERSPEECH 2009 Challenge (Schuller *et al.*, 2009a), extracting it usually using openSMILE:) toolkit (Eyben, Schuller, 2014). Kamińska *et al.* (2017) and Delić *et al.* (2012) have suggested in their researches on the Polish and Serbian databases that spectral features perform similarly or even better than prosodic features. As far as classification methods are concerned, Ayadi *et al.* (2011) has reviewed that a variety of methods have been used so far, each entailing both benefits and limitations.

In the research, we implemented and examined typical spectral features (Bitouk *et al.*, 2010): LPCC, LFPC and MFCC in combination with the most used classification methods based on HMM, SVM and Deep Neural Network (DNN). We used GEES, PES and RUSLANA databases. The tests were also conducted on the most used database in emotional speech recognition, Berlin (Burkhardt *et al.*, 2005), to obtain results for comparison. All features were tested separately combined with each classification method in order to examine the efficiency of their combinations for each of the three Slavic languages – Serbian, Polish and Russian. For the purpose of measuring the effi-

Table 1. Basic emotion recognition systems tested on databases of Slavic languages.

Language	Feature set	Classifiers	Emotions	Reference
Serbian	LPCC, LFPC, E, TE, F0, FF	HMM	anger, fear, joy, neutral, sadness	(Nedeljković, Đurović, 2015)
	MFCC, E, F0, FF, harmonicity, duration, loudness, voice source	SVM	anger, fear, joy, neutral, sadness	(Shaukat, Chen, 2008; 2011)
	openEAR feature set	SVM	anger, fear, joy, neutral, sadness	(Hassan, Damper, 2010)
	MFCC	SVM	anger, fear, joy, neutral, sadness	(Milošević *et al.*, 2016)
	MFCC, Temporal Discrete Cosine Transform	Artificial Neural Network (ANN)	anger, fear, joy, sadness, threat	(Popović *et al.*, 2013)
	MFCC, E, F0	Linear Bayes (LB), kNN	anger, fear, joy, neutral, sadness	(Delić *et al.*, 2012; Bojanić *et al.*, 2014)
Croatian	MFCC, E, F0, FF, voice source, linguistic features	SVM, RF	anger, fear, joy, neutral, sadness	(Dropuljić *et al.*, 2016a; 2016b)
Polish	E, F0, FF, duration, zero crossing rate	BDT	anger, fear, joy, neutral, sadness, boredom	(Ślot *et al.*, 2009)
	MFCC, Human Factor Cepstral Coefficients (HFCC)	kNN	anger, fear, joy, neutral, sadness, boredom	(Kamińska *et al.*, 2013)
	E, F0, FF, spectral perceptual features (MFCC, HFCC, and others)	kNN, SVM	anger, fear, joy, neutral, sadness, anticipation, surprise, disgust	(Kamińska *et al.*, 2017)
	E, F0, FF, Linear Prediction Coefficients (LPC)	LB, SVM, ANN, kNN, BDT, Linear Discriminant Analysis	anger, fear, joy, neutral, sadness, surprise, disgust	(Staroniewicz, 2011)

ciency of a feature-classifier combination for a particular database, success coefficients (SCs) of the features, classifiers and feature-classifier pairs were introduced. The emotions, which were classified, were: joy, anger, sadness, fear and a neutral emotional state, since this was the broadest selection of emotional states that all databases have in common.

The aim of our research was to:

1) evaluate robustly and reliably features, classifiers and their combinations;
2) investigate which emotion recognition system is the most convenient for being used in a particular language or a group of languages, or if the choice is language independent;
3) explore in which way database properties influence the final classification result.

The remainder of the paper is organised as follows. Section 2 provides information about the databases used. Section 3 describes the emotion recognition system – feature extraction algorithms and classification methods, and defines a success measure. Section 4 presents and discusses the results. Section 5 concludes the paper.

2. Databases

Technical information about the databases GEES, PES, RUSLANA and Berlin, which are used in the research, is summarised in Table 2. All four databases are of acted emotional speech. The amount of data recorded per speaker is the largest in the GEES database and the smallest in the PES database, whereas RUSLANA contains the largest number of speakers and the most training data per emotion. GEES and Berlin have the highest validation rate from human judges, while no precise data for PES and RUSLANA exists. In the case of PES, RUSLANA and Berlin databases all sentences containing joy, anger, sadness, fear and a neutral emotional state were used. In the case of the GEES database, only long sentences were used for these emotional states.

3. Emotion recognition system

The task of an emotion recognition system is to assess the emotional state of the test utterance, based on previously trained models of emotions.

A model of emotion is created in an appropriate form for each classification method, using feature vectors extracted from training utterances. The feature vectors were constructed from 36 values which are 12 feature coefficients with their delta and double delta coefficients on the frame level. The feature statistics were calculated over all feature vectors for one utterance. The calculated statistics were: minimum, maximum, mean, variance and range. As far as the input format is concerned, one HMM input was an array of feature vectors which represented a single utterance. One input in SVM and DNN was a vector of feature statistics calculated over a single utterance. Feature extraction algorithms, classification methods, and formulas for SCs are presented below.

3.1. Feature extraction

First, the signal was divided into frames of 16 ms length, with an overlap of 9 ms. Next, the Hamming window (1) was applied to each frame, to minimise spectral leakage:

$$w(n) = 0.54 - 0.46 \cdot \cos\left(2\pi \frac{n}{N-1}\right), \quad n = 0, ..., N-1, \tag{1}$$

where N is the number of samples in the frame/window. Then specific calculations were performed. A block diagram of feature extraction is shown in

Table 2. Overview of database information.

Attribute	GEES	PES	RUSLANA	Berlin
Emotions available	anger, happiness, fear, sadness, neutral	anger, happiness, fear, sadness, neutral, boredom	anger, happiness, fear, sadness, neutral, surprise	anger, happiness, fear, sadness, neutral, boredom, disgust
Language	Serbian	Polish	Russian	German
Speakers	3 male, 3 female	4 male, 4 female	12 male, 49 female	5 male, 5 female
Utterances	32 words, 30 short sentences, 30 long sentences, 1 passage	5 short sentences	10 long and short sentences	10 long and short sentences
Human validation	93.33–96.06%, depending on emotion	60–84%, depending on speaker	no validation data	79.6–96.9%, depending on emotion
Sampling frequency	22.05 kHz	44.1 kHz	32 kHz	16 kHz

Fig. 1. The details of each feature calculation algorithm are given in the following sub-sections.

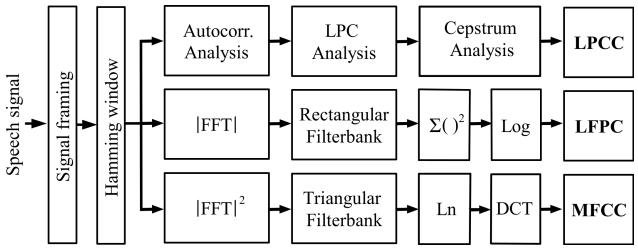

Fig. 1. Feature extraction from speech signal.

3.1.1. Linear Prediction Cepstral Coefficients

LPCC (Farsi, Saleh, 2014; Rabiner, Juang, 1993) are based on the assumption that one speech sample at the present time can be predicted as a linear combination of past speech samples. The following procedure was used to calculate LPCC from a framed and windowed speech signal $x(n)$, $n = 0, ..., N-1$:

1) The autocorrelation function was estimated using modified covariance method:

$$R_{i,j} = \frac{1}{2(N-p)}\left(\sum_{n=p}^{N-1} x(n-i)x(n-j) + \sum_{n=0}^{N-1-p} x(n+i)x(n+j)\right), \quad 0 \le i,j \le p, \tag{2}$$

where $p = 8$ is the number of autocorrelation coefficients.

2) Next, Linear Prediction Coefficients (LPC) a_m, $m = 1, ..., p$ were obtained by solving the matrix equation:

$$-\begin{bmatrix} R_{1,1} & R_{1,2} & \dots & R_{1,p} \\ R_{2,1} & R_{2,2} & \dots & R_{2,1} \\ \vdots & & \ddots & \vdots \\ R_{p,1} & & & R_{p,p} \end{bmatrix}\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} R_{1,0} \\ R_{2,0} \\ \vdots \\ R_{p,0} \end{bmatrix}. \tag{3}$$

3) Cepstral coefficients were derived using the following recursive formulas:

$$c_0 = R_{0,0}, \tag{4}$$

$$c_m = a_m + \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k a_{m-k}, \quad 1 \le m \le p, \tag{5}$$

$$c_m = \sum_{k=m-p}^{m-1}\left(\frac{k}{m}\right)c_k a_{m-k}, \quad p < m \le M-1, \tag{6}$$

where M is the number of LPCC coefficients. In the present research we set $M = 13$ and coefficients from c_1 to c_{12} were used.

3.1.2. Log Frequency Power Coefficients

LFPC (Nwe *et al.*, 2003) provide information about spectral energy distribution, which matches critical perceptual bands of the human ear. These coefficients were calculated as follows:

1) Framed and windowed speech was transformed into the frequency domain using the Fast Fourier Transformation (FFT) algorithm.
2) The spectral content was separated into $M = 12$ bands by a set of rectangular filters. Central frequencies and bandwidths of filters were obtained as follows:

$$b_1 = C, \tag{7}$$

$$b_m = \alpha b_{m-1}, \quad 2 \le m \le M, \tag{8}$$

$$f_m = f_1 + \sum_{j=1}^{m-1} b_j + \frac{b_m - b_1}{2}, \tag{9}$$

where b_m and f_m are the bandwidth and central frequency of the m-th filter. The adopted values of the constants are $C = 54$ Hz, $f_1 = 127$ Hz and $\alpha = 1.4$. The rectangular window W_m was defined as:

$$W_m(f) = \begin{cases} 1, & f_m - \frac{b_m}{2} \le f \le f_m + \frac{b_m}{2}, \\ 0, & f < f_m - \frac{b_m}{2} \ \vee \ f > f_m + \frac{b_m}{2}, \end{cases} \tag{10}$$

where $m = 1, ..., M$, $f \in \{nF_s/N,\ n = 0, ..., N/2\}$ and F_s is sampling frequency.

3) The energy was obtained as the square sum of each filter output:

$$S(m) = \sum_{f=f_m-\frac{b_m}{2}}^{f_m+\frac{b_m}{2}} \left(X(f)W_m(f)\right)^2, \quad m = 1, ..., M, \tag{11}$$

where $X(f)$ is the FFT spectral component at frequency f.

4) The final energy measure of the frequency band was calculated by taking the logarithm and scaling by the filter length:

$$\mathrm{SE}(m) = \frac{10\log_{10}\left(S(m)\right)}{N_m}, \tag{12}$$

where N_m is the number of spectral components in the m-th filter. The result is 12 LFPC.

3.1.3. Mel Frequency Cepstral Coefficients

MFCC (Davis, Mermelstein, 1980) is the most widely used speech feature. These coefficients represent an audio signal based on human perception. MFCC was calculated according to the following procedure:

1) Framed and windowed speech was transformed into the frequency domain using the FFT algorithm.

2) A power spectrum was obtained as the square of FFT magnitude.
3) Then a triangular filter bank with $M = 12$ filters was constructed. These filters were equidistant on the mel-scale:

$$H_m(\phi) = \begin{cases} \dfrac{\phi - \phi_{b_{m-1}}}{\phi_{b_m} - \phi_{b_{m-1}}}, & \phi_{b_{m-1}} \le \phi \le \phi_{b_m}, \\ \dfrac{\phi_{b_{m+1}} - \phi}{\phi_{b_{m+1}} - \phi_{b_m}}, & \phi_{b_m} \le \phi \le \phi_{b_{m+1}}, \\ 0, & \phi < \phi_{b_{m-1}} \ \vee \ \phi > \phi_{b_{m+1}}, \end{cases} \tag{13}$$

where $m = 1, ..., M$ is the index of a filter, and ϕ represents a discrete frequency on the mel-scale. The boundary frequencies $\phi_{b_0}, ..., \phi_{b_{M+1}}$ divided the mel scale into $M + 1$ equal frequency bands. The maximum mel-scale frequency corresponded to $F_s/2$ on the linear (Hz) scale.
4) Filters were transformed to linear scale by using the relation between the linear (f) scale and the mel (ϕ) scale:

$$\phi = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right). \tag{14}$$

5) The filter bank was normalised in such a way that the sum of coefficients for every filter equalled one. With this step the filter bank got its final shape.
6) Applying this filter bank on the power-spectrum resulted in the mel power spectrum.
7) In the end, mel cepstral coefficients were generated by discrete cosine transformation to the logarithm of the mel power spectrum.

The result was a mel frequency cepstar of 12 MFCC.

3.2. Classifiers

HMM was selected as the traditional method in speech processing, SVM turned out to be superior in many pattern recognition tasks, whereas DNN was added as representative of the deep learning approach, which is an emerging method in speech processing tasks. The following is an overview of every classification method used.

3.2.1. Hidden Markov Model

HMM (Rabiner, 1989; Rabiner, Juang, 1993) is a structure much used in speech recognition problems. Although there is no strict physical interpretation of hidden states, HMM is often used as a classifier for emotion recognition tasks (Ayadi *et al.*, 2011; Lin, Wei, 2005; Nwe *et al.*, 2003; Schuller *et al.*, 2009b). The main advantage of HMM is the possibility to model the dynamic of changes in speech features, which can be useful for emotion classification. There are different ways of implementing HMM for the task of emotion recognition (Ayadi *et al.*, 2011), and we used discrete ergodic HMM with four hidden states as it showed superior performance levels to left-right structure (Nwe *et al.*, 2003). One model was trained for each emotional state. The discrete HMM model takes the sequence of scalar values as an input, so it is necessary to do vector quantisation of feature vectors. K-means clusterisation with 64 clusters was used for this purpose.

HMM is a doubly embedded stochastic process (Rabiner, Juang, 1993). The first, underlying stochastic process describes the transition between hidden states and it can be observed only through the second stochastic process. The second process generates different observations depending on the current state of the first process. The hidden (underlying) process changes states with a given probability of change. The future state of the process depends only on its current state (it is independent of past changes). Figure 2 illustrates a possible sequence of states and observations.

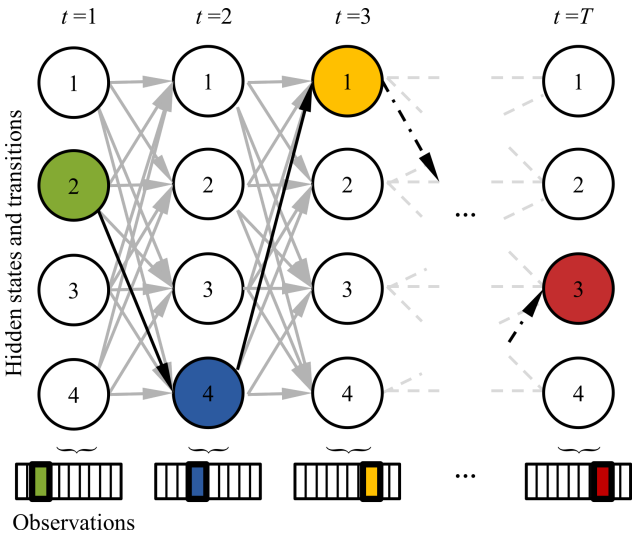

Fig. 2. Ergodic Hidden Markov Model – hidden state transition.

The Hidden Markov Model is described using three parameters: state transition probability matrix $\mathbf{A}$, observation probability matrix $\mathbf{B}$, and starting state probability vector $\boldsymbol{\pi}$:

$$\lambda = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}). \tag{15}$$

The possible states are from the state alphabet set S, and the observations are from the observation alphabet set V:

$$S = (s_1, s_2, ..., s_N), \tag{16}$$

$$V = (v_1, v_2, ..., v_M), \tag{17}$$

where N is the number of hidden states and M is the dimension of the observation alphabet.

Let us define Q as a fixed state sequence of length T, and corresponding observation sequence O:

$$Q = q_1, q_2, ..., q_T, \tag{18}$$

$$O = o_1, o_2, ..., o_T. \tag{19}$$

The transition probability matrix $\mathbf{A}$ defines the probabilities that state j follows state i, independent of time t:

$$\mathbf{A} = [a_{ij}], \qquad a_{ij} = P(q_t = s_j | q_{t-1} = s_i). \tag{20}$$

The observation probability matrix $\mathbf{B}$ defines the probabilities that observation k is produced from a hidden state i, independent of time t:

$$\mathbf{B} = [b_i(k)], \qquad b_i(k) = P(o_t = v_k | q_t = s_i). \tag{21}$$

The starting state probability vector $\boldsymbol{\pi}$ defines the probability of each state being the starting state:

$$\boldsymbol{\pi} = [\pi_i], \qquad \pi_i = P(q_1 = s_i). \tag{22}$$

Two assumptions are made by the model. The first is that the current state is dependent only on the previous state. This represents the memory of the model:

$$P(q_t | q_1, ..., q_{t-1}) = P(q_t | q_{t-1}). \tag{23}$$

The second assumption is that the output observation at time t is dependent only on the current state of the model, and that it is independent of previous observations and states:

$$P(o_t | o_1, ..., o_{t-1}, q_1, ..., q_{t-1}) = P(o_t | q_t). \tag{24}$$

The Baum-Welch algorithm (Rabiner, 1989; Rabiner, Juang, 1993) was used to obtain the set of HMM model parameters. The model probability matrixes and the probability vector were initialised by random values. The basic implementation of the Baum-Welch procedure was extended for parameter estimation using a multiple training sequence. A scaling procedure was added to parameter estimation to prevent potential underflows caused by multiplication of plural low probability values. Additionally, in order to reduce the impact of insufficient training data, an extra threshold constraint was applied to the model parameters to ensure that the estimated probabilities do not fall below a specified value. The forward procedure was used to evaluate the probability of the test observation sequence.

3.2.2. Support Vector Machine

The main idea behind this method is to separate space into two subspaces by finding the hyperplane which maximises the gap between two classes. There are various ways to apply this method to the multiple class separation task. Based on previous research (Hassan, Damper, 2010) and our own preliminary tests, we implemented SVM classification as 10 binary classifiers for each pair of emotions (Fig. 3). The final decision was based upon the majority vote.

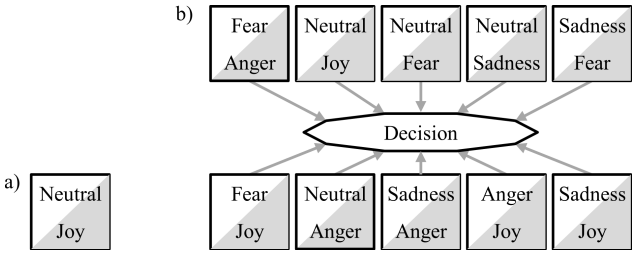

Fig. 3. a) SVM which decides between two classes; b) decision scheme for multi class SVM.

The binary classification mechanism is constructed using the following steps:

1) We describe input data as a set of pairs:

$$X = \{(x_1, y_1), ..., (x_n, y_n)\}, \tag{25}$$

$$x_i \in \mathbb{R}^d, \quad y_i \in \{-1, +1\}, \quad i = 1, ..., n, \tag{26}$$

where n is the number of input vectors.

2) Kernel function selected *a priori* performs nonlinear mapping of original values to a high dimensional space where these values become linearly separable:

$$\Phi : \mathbb{R}^d \to \mathbb{R}^D \ (D \gg d), \tag{27}$$

$$x \to \Phi(x). \tag{28}$$

A polynomial kernel function of the third order was used in this research.

3) In the high-dimensional space, decision function coefficients should be selected in such a way that the margin between the classes is maximised. The decision function is defined as:

$$f(x) = \text{sgn}\left(\langle w \cdot \Phi(x) \rangle + b\right), \tag{29}$$

where b and w are the hyperplane parameters and $\langle \cdot \rangle$ represents the inner product.

4) Finally, under the conditions described in (Pierna *et al.*, 2004), the decision function is presented as:

$$f(x) = \text{sgn}\left(\sum_{i=1}^{n_{SV}} \alpha_i y_i \langle \Phi(x_i) \cdot \Phi(x) \rangle + b\right), \quad 0 \le \alpha_i \le C, \tag{30}$$

where n_{SV} is the number of support vectors, α_i are parameters learned from the data, and $C = 1$ is the regularisation parameter for the trade-off between error minimization and margin maximization.

3.2.3. Deep Neural Network

DNN (Hendy, Farag, 2013; Lange, Riedmiller, 2010) is an artificial neural network with multiple layers hidden between input and output layers. Multiple-layers neural networks emerge as very useful for complex data classification due to the fact that each layer can learn with a different abstraction (generalisation) level (Stuhlsatz *et al.*, 2011). It can be configured in different ways (Stuhlsatz *et al.*, 2011). The present research used a single DNN with 180 nodes in the input layer, four hidden layers with 160, 120, 80, and 40 nodes, and five nodes corresponding to emotion labels in the output layer.

The problem here lies in the actual training of the neural network. One efficient way of multiple-layer neural network training is performed by separate training of the network layers (Fig. 4). DNN layers were formed as follows:

1) A feed-forward neural network was structured with one hidden layer and input and output layers with the same number of nodes (the dimension of an input vector), having less nodes in the hidden layer than in the input/output one.
2) Training was conducted by reflecting the input vector to the output of the network, using a back-propagation algorithm.
3) After training, the output layer was removed and the hidden layer became a new output layer. In that way a compressed image of the input vector was created.

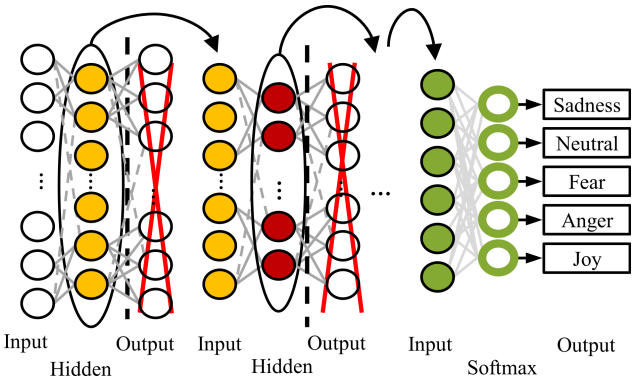

Fig. 4. DNN training process.

The next hidden layers were formed in a similar way, using the previous hidden layer output as their input. The final layer of the network was created using supervised training, so that the input for this layer was the output of the last hidden layer in the network and class labels were the output of this layer. The decision function in the last layer was softmax.

In the final step, the layers of the network trained separately were connected and the network was fine-tuned using the back-propagation algorithm.

3.3. Success measure

Comparisons of different systems are based on a large number of results and should be simplified. In order to achieve this goal, a success measure was introduced. Let us first define the recognition rate matrix $\mathbf{R}(f, c, d)$, which is a three dimensional matrix of test results obtained using feature f, classifier c, applied on database d, where:

$$f \in \{\text{LPCC}, \text{LFPC}, \text{MFCC}\},$$

$$c \in \{\text{HMM}, \text{SVM}, \text{DNN}\},$$

$$d \in \{\text{GEES}, \text{PES}, \text{RUSLANA}, \text{Berlin}\}.$$

We defined success measure through a set of success coefficients of features, classifiers and their combinations by formulas (31)–(35).

1) Feature success coefficient:

$$Q_1(f, d) = \frac{a^*}{3}, \tag{31}$$

where

$$a^* = \mathbf{R}(f, \text{HMM}, d) + \mathbf{R}(f, \text{SVM}, d) + \mathbf{R}(f, \text{DNN}, d).$$

2) Overall feature success coefficient:

$$Q_2(f) = \frac{b^*}{4}, \tag{32}$$

where

$$b^* = Q_1(f, \text{GEES}) + Q_1(f, \text{PES}) + Q_1(f, \text{RUSLANA}) + Q_1(f, \text{Berlin}).$$

3) Classifier success coefficient:

$$Q_3(c, d) = \frac{c^*}{3}, \tag{33}$$

where

$$c^* = \mathbf{R}(\text{LPCC}, c, d) + \mathbf{R}(\text{LFPC}, c, d) + \mathbf{R}(\text{MFCC}, c, d).$$

4) Overall classifier success coefficient:

$$Q_4(c) = \frac{d^*}{4}, \tag{34}$$

where

$$d^* = Q_3(c, \text{GEES}) + Q_3(c, \text{PES}) + Q_3(c, \text{RUSLANA} + Q_3(c, \text{Berlin}).$$

5) Feature-classifier success coefficient:

$$Q_5(f, c) = \frac{e^*}{4}, \tag{35}$$

where

$$e^* = \mathbf{R}(f, c, \text{GEES}) + \mathbf{R}(f, c, \text{PES}) + \mathbf{R}(f, c, \text{RUSLANA}) + \mathbf{R}(f, c, \text{Berlin}).$$

4. Results and discussion

4.1. Experimental setup

The tests were performed in speaker dependent (SD) and speaker independent (SI) setups. The SD setup used the first 83% of each speaker's sentences for classifier training. The remaining 17% was used for testing. This way the classifier learned from the voices of all available speakers and no new voice was introduced in the test. The lexical content of the training sentences was different from that of the test sentences. In the SI setup, all sentences from one speaker were used for testing and all sentences from the other speakers were used for training. This means that the lexical content of the test sentences was the same as the lexical content of the training sentences. It also means that the voice of the test speaker was unfamiliar to the classification system. Training and testing of the classification system in the SI setup were repeated several times – each time with a different speaker left out for testing. The final results were based on all test runs.

4.2. Recognition rate

The recognition rates obtained from all classification tests are shown in Table 3. These results are comparable to previous researches conducted in similar setups. In the case of GEES database, MFCC-SVM in the SI setup, Shaukat and Chen (2011; 2008) reported 63.90% using hierarchal SVM, whereas our system yielded 66.78%. A relevant MFCC-SVM approach provided 33.75% on the PES database (Kamińska *et al.*, 2017), in contrast to our 70.00% in the SD and 55.50% in the SI setup. Ultimately, it is interesting to comment on DNN as a classification method on the Berlin database, although the majority of previous researchers have classified seven emotions. The results we obtained are comparable to those of Albornoz *et al.* (2014). The features they used were MFCC and prosodic features. They reported 69.14% using the Deep Belief Network and 68.10% using Multilayer Perceptron in the SD setup, and 60.32% and 51.65%, respectively, in the SI setup. These results can be compared to our MFCC-DNN results: 66.44% in SD and 66.32% in SI setup.

Table 3. Classification rate of all experiments.

Setup	Classifier	Feature	GEES	PES	RUSL	Berlin
SD	HMM	LPCC	72.41	52.75	32.56	54.41
		LFPC	72.41	54.75	32.89	63.24
		MFCC	60.69	52.75	31.40	**72.06**
	SVM	LPCC	79.33	57.50	34.43	65.75
		LFPC	85.33	67.50	**46.07**	69.86
		MFCC	83.33	**70.00**	43.11	63.01
	DNN	LPCC	76.87	57.25	35.26	66.30
		LFPC	**87.87**	67.50	43.69	69.72
		MFCC	85.07	64.25	42.41	66.44
SI	HMM	LPCC	57.24	52.50	30.31	55.22
		LFPC	65.18	49.75	31.73	62.50
		MFCC	53.45	44.50	30.27	59.73
	SVM	LPCC	65.11	55.00	34.00	59.80
		LFPC	66.67	55.50	**43.31**	65.93
		MFCC	66.78	55.50	42.92	62.25
	DNN	LPCC	63.46	50.50	30.58	59.83
		LFPC	65.39	**56.00**	42.16	**67.72**
		MFCC	**67.85**	54.00	41.23	66.32

In Table 3 the best results per database are in bold, but it is not clear what combination performs the best. It is interesting to note, from Table 3, that the MFCC-HMM pair in the SD setup from all tests performed on the Berlin database yielded the best classification results, but the poorest if the same classification setup was applied on the GEES database. The differences in results indicate that the database used and cultural background of emotion display might influence the choice of optimal system in a certain way. Also, this illustrates why it was not possible to arrive at a straightforward conclusion regarding which feature-classifier combination provided the best results. For that reason, SCs as defined in Subsec. 3.3, were calculated based on raw recognition rates. The results are presented in Figs 5–10 and discussed in the following sub-sections.

4.3. Feature success coefficient

The feature SC Q_1, and overall feature SC Q_2, are displayed in Fig. 5 for SD setup and Fig. 6 for SI setup. In the SD setup, LFPC features were the most informative in the case of all databases, although in the case of the PES and Berlin databases, LFPC outperformed MFCC just slightly (Fig. 5). In the SI setup, LFPC features were the most useful for all databases again (Fig. 6). Overall, LFPC turned out to be the most informative in the SD setup (Fig. 5), and SI setup (Fig. 6).

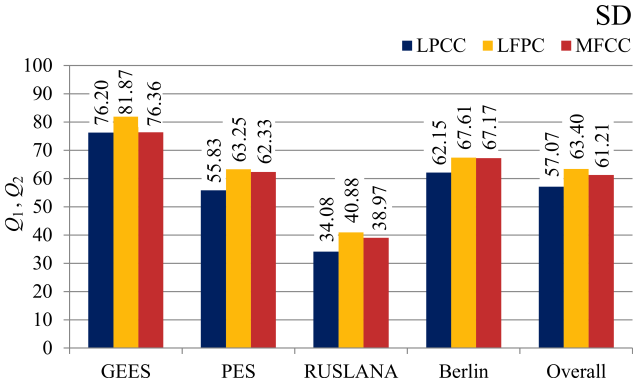

Fig. 5. Feature SCs in the SD setup (individual databases and all databases together).

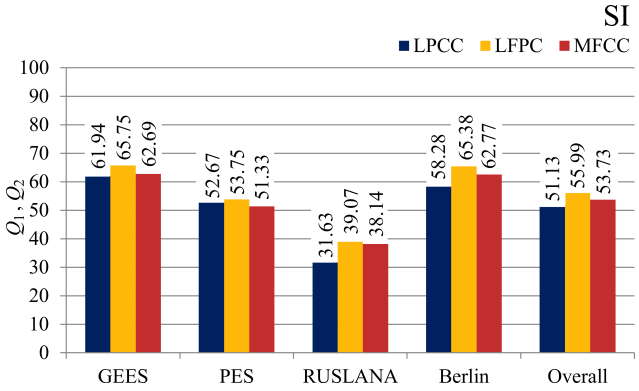

Fig. 6. Feature SCs in the SI setup (individual databases and all databases together).

4.4. Classifier success coefficient

The classifier SC Q_3, and overall classifier SC Q_4, are displayed in Fig. 7 for SD setup and Fig. 8 for SI setup. In the SD setup, SVM classifier was the most efficient for the PES and RUSLANA databases. In the case of the GEES and Berlin databases, the most efficient was DNN classifier, whereas in the case of the GEES database, scores of SVM and DNN classifiers were almost even (Fig. 7). Similar conclusions were derived in the SI setup. In the case of the PES and RUSLANA databases the best choice was SVM classifier, while in the case of the Berlin database it was DNN classifier. Again, in the case of the GEES database SVM and DNN classifiers were even, but this time SVM classifier performed slightly better (Fig. 8). In both setups, SVM and DNN had similar overall success coefficients, whereas SVM was slightly better (Figs 7 and 8). According to the calculated success coefficients, HMM classifier had the poorest performance, although raw results show it made the best result on the Berlin database in the SD setup.

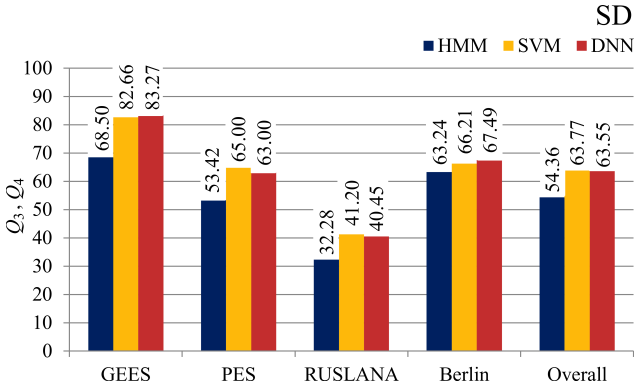

Fig. 7. Classifier SCs in the SD setup (individual databases and all databases together).

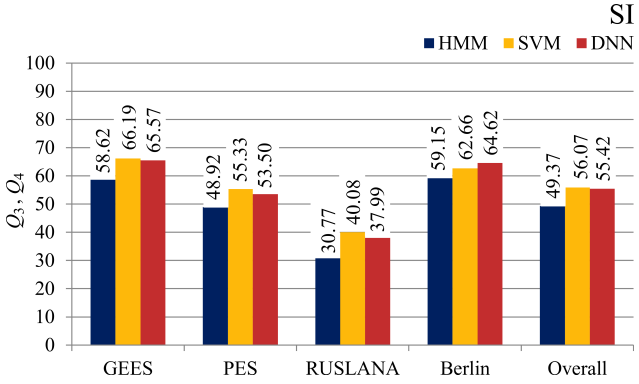

Fig. 8. Classifier SCs in the SI setup (individual databases and all databases together).

Upon analyzing the amplitudes of SCs, both for features and classifiers, in the SD and the SI setup (Figs 5–8) – it becomes apparent that there is a significant difference in performance only in the case of the GEES database. The GEES database had the highest SCs values and the steepest drop in performance, when SD and SI were compared. This database has a small number of speakers but the most data per speaker, when compared with the other databases. Besides, low recognition rates, low SCs and small difference in SD and SI setup results on the RUSLANA database (the largest of all databases with 61 speakers) may indicate that speaker characteristics influence the expression of emotions. This leaves room for future analysis of individual characteristics of speakers and their emotion expression, in order to separate data regarding emotional states from speaker-specific characteristics.

4.5. Feature-classifier success coefficient

The results presented so far indicate that features and classifiers, which show the best results in experiments on one database, do not necessarily remain the best when all databases are taken into consideration. The problem is obviously a very complex one because the success of the automatic emotion recognition system varies depending on the type of feature, structure and type of classifier, as well as on the characteristics of available databases and spoken language. Aiming to determine whether pairs feature-classifier that give the best results regardless of language i.e. database exist, or it might be the case that language particularities condition the applicability of certain features or classifiers, we calculated the success coefficient Q5 for all possible feature-classifier pairs, in both speaker dependent, and speaker independent setups. The results are presented in Figs 9 and 10.

Pairs that included LFPC exhibited the best performance per classifier, so in order to achieve the best performance LFPC should be chosen for feature regardless of the classifier choice. The second choice is MFCC feature in all cases except HMM classifier in the SI setup where the second best choice is LPCC feature.

SVM and DNN classifiers show very similar success coefficients, so it is not clear which one makes a bet-

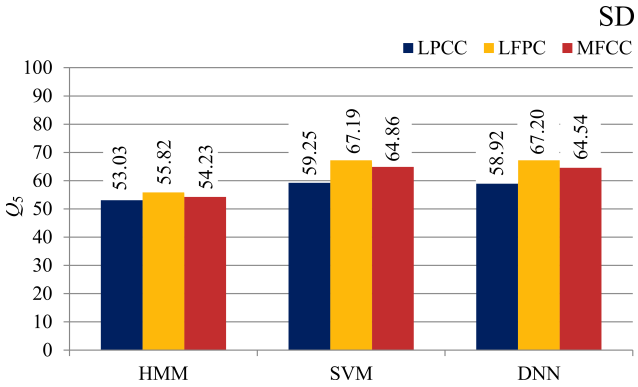

Fig. 9. Feature-classifier pair SCs in the SD setup.

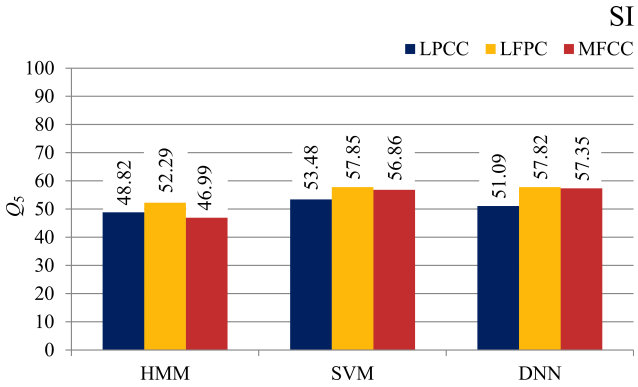

Fig. 10. Feature-classifier pair SCs in the SI setup.

ter choice. In case of the SD and SI setups LFPC-SVM or LFPC-DNN would be a good choice. In the SI setup MFCC-SVM and MFCC-DNN show a performance that is almost as good.

5. Conclusion

The results from our systematic evaluation of different automatic emotion recognition system configurations are presented in the paper. The test data were taken from four acted emotional speech databases. Three databases are of Slavic languages: GEES in Serbian, PES in Polish and RUSLANA in Russian, while the fourth database, Berlin, is in German and it was used for comparison. Two test setups have been used: speaker dependent (SD) and speaker independent (SI). For general evaluation purposes, we have introduced a success measure. The tests were performed with all possible combinations of one of the features: LPCC, LFPC and MFCC, with one of the classifiers: HMM, SVM and DNN, on all four databases in the two setups.

The following conclusions are drawn:

1) The proposed success coefficients, as a robust performance measure of automatic emotion recognition system, provided a very good quality estimate for evaluating various system configurations on different databases.
2) Based on the success measure, LFPC is the best choice among the tested spectral features. In the case of classifier, no unique conclusion can be made, and SVM and DNN both make a good choice for the classifying structure. The derived conclusions hold for the SD and SI setups.
3) The evaluated systems have shown sensitivity to the database construction in terms of quality, number of speakers, and speaker-specific characteristics rather than language.
4) Although HMM showed superiority in many speech recognition problems, we can conclude that there are better choices of classifiers when it comes to designing a single classifier emotion recognition system.

Conclusions that appeared as results of the presented research can serve as baseline for future research directed towards improving emotion classification systems. Another direction for future research could be aimed at constructing a complex classifying structure involving multiple baseline classifiers and a wide range of features with dimension reduction. In that case, the constructed structure should include all features and classifiers investigated in this work, each with the particular discriminatory capabilities pertaining to it. Also, it remains for future research to test constructed systems with spontaneous speech datasets in order to confirm the practical usability of the derived conclusions.

Acknowledgement

The authors would like to thank Professor Slobodan Jovičić from the School of Electrical Engineering, University of Belgrade, for providing access to the GEES database, and Dr. Valery A. Petrushin, Illuminated Numerati, Inc., Greater San Diego Area, USA, for providing access to the RUSLANA database and his useful comments and suggestions.

References

1. Albornoz E.M., Sánchez-Gutíerrez M., Martinez-Licona F., Rufiner H.L., Goddard J. (2014), Spoken emotion recognition using deep learning, [in:] *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, *Lecture Notes in Computer Science*, Bayro-Corrochano E., Hancock E. [Eds], Vol. 8827, pp. 104–111, Springer, Cham, doi: 10.1007/978-3-319-12568-8_13.
2. El Ayadi M., Kamel M.S., Karray F. (2011), Survey on speech emotion recognition: Features, classification schemes, and databases, *Pattern Recognition*, **44**(3): 572–587, doi: 10.1016/j.patcog.2010.09.020.
3. Bitouk D., Verma R., Nenkova A. (2010), Class-level spectral features for emotion recognition, *Speech Communication*, **52**(7–8): 613–625, doi: 10.1016/j.specom.2010.02.010.

4. Bojanić M., Delić V., Sečujski M. (2014), Relevance of the types and the statistical properties of features in the recognition of basic emotions in speech, *Facta Universitatis – Series: Electronics and Energetics*, **27**(3): 425–433, doi: 10.2298/FUEE1403425B.
5. Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W.F., Weiss B. (2005), A database of German emotional speech, *Proceedings of the 9th European Conference on Speech Communication and Technology*, pp. 1517–1520, Lisbon.
6. Cichosz J. (2008), *Database of polish emotional speech*, retrieved October 16th, 2015, from http://www.eletel.p.lodz.pl/med/eng.
7. Cowie R. *et al.* (2001), Emotion recognition in human-computer interaction, *IEEE Signal Processing Magazine*, **18**(1): 32–80, doi: 10.1109/79.911197.
8. Davis S., Mermelstein P. (1980), Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentence, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **28**(4): 357–366, doi: 10.1109/TASSP.1980.1163420.
9. Delić V., Bojanić M., Gnjatović M., Sečujski M., Jovičić S.T. (2012), Discrimination capability of prosodic and spectral features for emotional speech recognition, *Elektronika ir Elektrotechnika*, **18**(9): 51–54, doi: 10.5755/j01.eee.18.9.2806.
10. Dropuljić B., Chmura M.T., Kolak A., Petrinović D. (2011), Emotional speech corpus of Croatian language, *Proceedings of the 7th International Symposium on Image and Signal Processing and Analysis*, pp. 95–100, Dubrovnik.
11. Dropuljić B., Skansi S., Kopal R. (2016a), Analyzing affective states using acoustic and linguistic features, *Proceedings of Central European Conference on Information and Intelligent Systems*, pp. 201–206, Varaždin.
12. Dropuljić B., Skansi S., Kopal R. (2016b), Croatian emotional speech analyses on a basis of acoustic and linguistic features, *International Journal of Digital Technology & Economy*, **1**(2): 85–96.
13. Eyben F., Schuller B. (2014), openSMILE:): The Munich open-source large-scale multimedia feature extractor, *ACM SIGMultimedia Records*, **6**(4): 4–13, doi: 10.1145/2729095.2729097.
14. Farsi H., Saleh R. (2014), Implementation and optimization of a speech recognition system based on hidden Markov model using genetic algorithm, *2014 Iranian Conference on Intelligent Systems*, pp. 1–5, Bam, doi: 10.1109/IranianCIS.2014.6802533.
15. Hassan A., Damper R.I. (2010), Multi-class and hierarchical SVMs for emotion recognition, *Proceedings of the 11th Annual Conference of the International Speech Communication Association*, pp. 2354–2357, Makuhari.
16. Hendy N.A., Farag H. (2013), Emotion recognition using neural network: A comparative study, *International Journal of Computer and Information Engineering*, **7**(3): 433–439, doi: 10.5281/zenodo.1077145.
17. Igras M., Ziółko B. (2013), Database of emotional speech recordings [in Polish], *Studia Informatica*, **34**(2B): 67–77.
18. Jovičić S.T., Kašić Z., Đorđević M., Rajković M. (2004), Serbian emotional speech database: design, processing and evaluation, *Proceedings of the 9th International Conference Speech and Computer*, pp. 77–81, Saint-Petersburg.
19. Justin T., Štruc V., Žibert J., Mihelić F. (2015), Development and evaluation of the emotional Slovenian speech database – EmoLuks, [in:] *Text, Speech, and Dialogue, Lecture Notes in Computer Science*, Král P., Matoušek V. [Eds], Vol. 9302, pp. 351–359, Springer, Cham, doi: 10.1007/978-3-319-24033-6_40.
20. Kamińska D., Sapiński T., Anbarjafari G. (2017), Efficiency of chosen speech descriptors in relation to emotion recognition, *EURASIP Journal on Audio, Speech, and Music Processing*, **2017**: 3, doi: doi:10.1186/s13636-017-0100-x.
21. Kamińska D., Sapiński T., Niewiadomy D., Pelikant A. (2013), Comparison of perceptual features efficiency for automatic identification of emotional states from speech signal [in Polish], *Studia Informatica*, **34**(2B): 59–66, doi: 10.21936/si2013_v34.n2B.50.
22. Kołakowska A., Landowska A., Szwoch M., Szwoch W., Wrobel M.R. (2014), Emotion recognition and its applications, [in:] *Human-Computer Systems Interaction: Backgrounds and Applications 3*, *Advances in Intelligent Systems and Computing*, Hippe Z., Kulikowski J., Mroczek T., Wtorek J. [Eds], Vol. 300, pp. 51–62, Springer, Cham, doi: 10.1007/978-3-319-08491-6_5.
23. Lange S., Riedmiller M. (2010), Deep auto-encoder neural networks in reinforcement learning, *The 2010 International Joint Conference on Neural Networks*, pp. 1–8, Barcelona, doi: 10.1109/IJCNN.2010.5596468.
24. Lin Y.L., Wei G. (2005), Speech emotion recognition based on HMM and SVM, *Proceedings of 2005 International Conference on Machine Learning and Cybernetics*, pp. 4898–4901, Guangzhou, doi: 10.1109/ICMLC.2005.1527805.
25. Makarova V., Petrushin V.A. (2012), Phonetics: Tracing emotions in Russian vowels, [in:] *Russian Language Studies in North America: New Perspectives from Theoretical and Applied Linguistics*, Makarova V. [Ed.], pp. 3–42, Athem Press, London, New York, doi: 10.7135/UPO9780857286505.002.
26. Makarova V., Petrushin V.A. (2002), RUSLANA: A database of Russian emotional utterances, *Proceedings of the 7th International Conference on Spoken Language Processing*, pp. 2041–2044, Colorado.
27. Milošević M., Nedeljković Ž., Đurović Ž. (2016), SVM classifier for emotional speech recognition in software environment SEBAS, *Proceedings of 3rd International Conference on Electrical, Electronic and Computing Engineering*, pp. AUI4.1.1–4, Zlatibor.
28. Nedeljković Ž., Đurović Ž. (2015), Automatic emotion recognition from speech using hidden Markov

models [in Serbian], *Proceedings of 59th Conference on Electrical, Electronic and Computing Engineering*, pp. AU1.6.1–5, Silver Lake.

29. Nwe T.L., Foo S.W., Silva L.C.D. (2003), Speech emotion recognition using hidden Markov models, *Speech Communication*, **41**(4): 603–623, doi: 10.1016/S0167-6393(03)00099-2.
30. Pell M.D., Monetta L., Paulmann S., Kotz S.A. (2009a), Recognizing emotions in a foreign language, *Journal of Nonverbal Behavior*, **33**(2): 107–120, doi: 10.1007/s10919-008-0065-7.
31. Pell M.D., Paulmann S., Dara C., Alasseri A., Kotz S.A. (2009b), Factors in the recognition of vocally expressed emotions: A comparison of four languages, *Journal of Phonetics*, **37**(4): 417–435, doi: 10.1016/j.wocn.2009.07.005.
32. Pierna J.A., Baeten V., Renier A.M., Cogdill R.P., Dardenne P. (2004), Combination of support vector machines (SVM) and near-infrared (NIR) imaging spectroscopy for the detection of meat and bone meal (MBM) in compound feeds, *Journal of Chemometrics*, **18**(7–8): 341–349, doi: 10.1002/cem.877.
33. Popović B., Stanković I., Ostrogonac S. (2013), Temporal discrete cosine transform for speech emotion recognition, *Proceedings of IEEE 4th International Conference on Cognitive Infocommunications*, pp. 87–90, Budapest, doi: 10.1109/CogInfoCom.2013.6719219.
34. Rabiner L. (1989), A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE*, **77**(2): 257–286, doi: 10.1109/5.18626.
35. Rabiner L., Juang B.H. (1993), *Fundamentals of speech recognition*, Prentice Hall, New Jersey.
36. Schuller B., Steidl S., Batliner A. (2009a), The Interspeech 2009 Emotion Challenge, *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 312–315, Brighton.
37. Schuller B., Vlasenko B., Eyben F., Rigoll G., Wendemuth A. (2009b), Acoustic emotion recognition: A benchmark comparison of performances, *Proceedings of IEEE Workshop on Automatic Speech Recognition & Understanding*, pp. 552–557, Merano, doi: 10.1109/ASRU.2009.5372886.
38. Shaukat A., Chen K. (2011), Emotional state recognition from speech via soft-competition on different acoustic representations, *Proceedings of the International Joint Conference on Neural Networks*, pp. 1910–1917, San Jose, doi: 10.1109/IJCNN.2011.6033457.
39. Shaukat A., Chen K. (2008), Towards automatic emotional state categorization from speech signals, *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 2771–2774, Brisbane.
40. Ślot K., Bronakowski Ł., Cichosz J., Kim H. (2009), Application of Poincare-mapping of voiced-speech segments for emotion sensing, *Sensors*, **9**(12): 9858–9872, doi: 10.3390/s91209858.
41. Staroniewicz P. (2011), Automatic recognition of emotional state in Polish speech, [in:] *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues, Lecture Notes in Computer Science*, Esposito A., Esposito A.M., Martone R., Müller V.C., Scarpetta G. [Eds], Vol. 6456, pp. 347–353, Springer, Berlin-Heidelberg, doi: 10.1007/978-3-642-18184-9_30.
42. Staroniewicz P., Majewski W. (2009), Polish emotional speech database – recording and preliminary validation, [in:] *Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions, Lecture Notes in Computer Science*, Esposito A., Vích R. [Eds], Vol. 5641, pp. 42–49, Springer, Berlin-Heidelberg, doi: 10.1007/978-3-642-03320-9_5.
43. Stuhlsatz A., Meyer C., Eyben F., Zielke T., Meier G., Schuller B. (2011), Deep neural networks for acoustic emotion recognition: Raising the benchmarks, *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5688–5691, Prague, doi: 10.1109/ICASSP.2011.5947651.
44. Uhrin D., Partila P., Voznak M., Chmelikova Z., Hlozak M., Orcik L. (2014), Design and implementation of Czech database of speech emotions, *Proceedings of the 22nd Telecommunications Forum*, pp. 529–532, Belgrade, doi: 10.1109/TELFOR.2014.7034463.
45. Vinola C., Vimaladevi K. (2015), A survey on human emotion recognition approaches, databases and applications, *Electronic Letters on Computer Vision and Image Analysis*, **14**(2): 24–44, doi: 10.5565/rev/elcvia.795.