Int. J. Appl. Math. Comput. Sci., 2013, Vol. 23, No. 4, 797–808
DOI: 10.2478/amcs-2013-0060

COMPARISON OF SPEAKER DEPENDENT AND SPEAKER INDEPENDENT EMOTION RECOGNITION

JAN RYBKA *, ARTUR JANICKI **

* Institute of Computer Science
Warsaw University of Technology, ul. Nowowiejska 15/19, 00-665 Warsaw, Poland
e-mail: Jan.Rybka@r4system.net

**Institute of Telecommunications
Warsaw University of Technology, ul. Nowowiejska 15/19, 00-665 Warsaw, Poland
e-mail: A.Janicki@tele.pw.edu.pl

This paper describes a study of emotion recognition based on speech analysis. The introduction to the theory contains a review of emotion inventories used in various studies of emotion recognition as well as the speech corpora applied, methods of speech parametrization, and the most commonly employed classification algorithms. In the current study the EMO-DB speech corpus and three selected classifiers, the k-Nearest Neighbor (k-NN), the Artificial Neural Network (ANN) and Support Vector Machines (SVMs), were used in experiments. SVMs turned out to provide the best classification accuracy of 75.44% in the speaker dependent mode, that is, when speech samples from the same speaker were included in the training corpus. Various speaker dependent and speaker independent configurations were analyzed and compared. Emotion recognition in speaker dependent conditions usually yielded higher accuracy results than a similar but speaker independent configuration. The improvement was especially well observed if the base recognition ratio of a given speaker was low. Happiness and anger, as well as boredom and neutrality, proved to be the pairs of emotions most often confused.

Keywords: speech processing, emotion recognition, EMO-DB, support vector machines, artificial neural networks.

1. Introduction

It has been known for years that speech is more than just words. It usually comes with an underlying emotion. As was shown by Mehrabian and Wiener (1967), when communicating feelings and attitudes, the spoken words (verbal attitude) constitute only 7% of the message. The rest of the message comprises non-verbal vocal attitude (38%) and visual one, that is, facial expression (55%). When communicating messages other than feelings or attitudes, the impact of non-verbal aspects is lower, but still it carries important information which scientists have tried to explore.

Several studies have been conducted on multimodal emotion recognition, for example, involving face analysis and gesture recognition. However, a remarkable part of communication is carried out remotely using audio transmission only, for example, over the phone, in which the visual part is not available. Therefore speech-based emotion recognition has become a fast-developing sub-domain of speech processing, and for this reason this work focuses on speech-based recognition only.

Speech-based emotion recognition algorithms can be employed in various applications. They can help in evaluating the operation of a call center by detecting conversations with customers who become angry during service (Erden and Arslan, 2011). Other systems can detect an increase in the stress of a person performing a responsible task, for example, a pilot or surgeon (He *et al.*, 2008). Emotion recognition algorithms can be helpful in educational software, for example, by detecting whether the user gets bored during training (Schuller *et al.*, 2006). Last but not least, they can improve human–machine interfaces by detecting the emotional state of the user and, for example, simplifying the menu if user annoyance is detected. What is more, trials are being carried out to apply emotions to robots (Kowalczuk and Czubenko, 2011), so the "expressive" layer of communication between humans and machines becomes even more important.

Emotion recognition is a difficult task even for

humans. Scherer (2003) claims that based on speech a human achieves a recognition accuracy of only 60% when recognizing an emotion of an unknown person, that is, when acting in speaker independent mode. When a speaker is known to a listener, then the recognition can be treated as speaker dependent, that is, the prior knowledge about the speaker is taken into account during the recognition process. This speaker dependence obviously improves humans' recognition performance. The objective of this project was to compare speaker dependent and speaker independent conditions for emotion recognition algorithms.

This paper begins with a description of the theoretical background of speech based emotion recognition, and then defines the aim of the project. Next, the experimental setup will be described and the obtained results will be presented. Finally, the discussion of the results and final conclusions will be given.

2. Theoretical background

Generally speaking, most emotion recognition systems employ learning algorithms, so prior to use they must undergo proper training. Typical training of an emotion recognition system consists in extracting parameters from the training recordings and making a classifier learn the emotional classes based on the labels of the recordings. The testing of such a system consists in classifying the tested recordings and comparing the results with the expected emotions. In order to use the available speech data efficiently, testing and training often follow a cross-validation scheme.

With regards to the details, a huge variety of studies dealing with automatic emotion recognition based on speech analysis exist. They differ in terms of the range of emotions considered for recognition, and they employ various sets of speech parameters and classifiers. Last but not least, in their experiments researchers use various speech corpora to train and test their systems. All of this makes comparison between studies a difficult task. The aim of this section is to highlight the main tendencies in speech-based emotion recognition.

2.1. Emotions being recognized. The most basic recognition which is sometimes analyzed is just a differentiation between neutral and emotionally flavored speech, that is, detection of whether or not a speaker expresses some emotion. A more advanced approach tries to evaluate the polarity of the emotional state, that is, whether the emotion is positive or negative.

Many studies follow Paul Ekman's early theory of emotions (Ekman, 1972), which postulates that all emotional states are combined out of basic six emotions: anger, happiness, sorrow, surprise, fear and disgust. These studies usually recognize seven classes: the above mentioned six emotional states and a neutral emotion. Sometimes single emotions from this list are replaced by others (for example, boredom, anxiety) or simply dropped. Some researchers follow Plutchik's model of emotions (Kaminska and Pelikant, 2012).

There is also a group of studies which try to detect a selected basic emotion, such as fear (Clavel *et al.*, 2007) or anger (Erden and Arslan, 2011). Other studies research the possibility of recognizing complex emotional states, such as stress (He *et al.*, 2008), certainty (Liscombe *et al.*, 2005), interest (Schuller *et al.*, 2006), speaker engagement (Yu *et al.*, 2004) or deception (Hirschberg et al., 2005).

A different approach which is sometimes researched consists in abandoning recognition of discrete emotion classes and moving towards a continuous emotional space (Lugger and Yang, 2007). Such a space usually consists of the following dimensions:

- *valence* (or evaluation): this dimension describes whether the emotion is positive or negative and by how much;
- *activation* (or arousal): this dimension describes whether a human is aroused and by how much; for example, a sad or bored man will be described as having low activation, but an angry, happy or anxious man will have high activation (Lugger and Yang, 2007);
- optional—*dominance* (or potency): this dimension describes whether and how strongly the emotion empowers a human to undertake further actions; for example, an anxious man will have low dominance, but an angry or happy one will show high dominance.

In this approach the discrete emotions are often mapped onto a 2D or 3D emotional space. When recognition is performed and values of the two (or three) dimensions are determined, then these values can be mapped back to the discrete categories (Janicki and Turkot, 2008).

2.2. Speech corpora used in emotion recognition studies. In general, research is conducted using two types of speech corpora: corpora with *acted speech* and those with *spontaneous speech*. Spontaneous speech can be acquired in a natural environment, for example, in a call center as in the work of Erden and Arslan (2011), or it can be recorded in a Wizard of Oz scenario, where certain emotions can be provoked in a special scenario, as was done by Batliner *et al.* (2005), when speech was acquired from children playing with Sony's AIBO pet robot.

Studies with spontaneous speech seem more realistic, as these corpora contain real-world recordings with real emotions. On the other hand, when recording acted speech, we can ask actors to repeat the same sentences

with different emotions, which makes it possible to conduct numerous comparative studies. The following speech corpora are mostly used in speech-based emotion recognition:

- *Linguistic Data Consortium (LDC) Emotional Prosody Speech and Transcripts* (Liberman *et al.*, 2002): a commercial database containing recordings of seven actors, each saying 10 sentences in 15 emotions;
- *EMO-DB*: *Berlin Emotional Database* (Burkhardt *et al.*, 2005), with almost 800 recordings from 10 professional German actors, each saying 10 sentences in seven emotional states, with some repetitions;
- *DES*: *Danish Emotional Database* (Engberg *et al.*, 1997), with recordings from four non-professional actors, each saying two words, nine sentences and two passages in five emotional states;
- *SUSAS*: *Speech Under Simulated and Actual Stress*, a database used for research on stress detection and analysis, containing recordings of 32 actors in various stressful situations, investigated, for example, by He *et al.* (2008);
- numerous *private corpora*: databases with recordings dedicated to a given project, usually not open to the public, for example, the recordings used by Erden and Arslan (2011), or ITSPOKE, which was employed by Liscombe *et al.* (2005).

2.3. Speech parameterization. Emotion recognition from speech is based on parameters extracted from the speech signal. Dozens, hundreds, or even thousands of parameters are calculated, in order to capture the features containing information of an underlying emotion. The following are the most commonly met sets of parameters:

- $F0$*-based parameters*, that is, parameters related with pitch—the fundamental frequency of speech, e.g., the maximum, minimum, mean, median and variance of pitch for an utterance being tested, the first and third quartiles of $F0$, and the first and ninth deciles of $F0$, such as in the work of Ayadi *et al.* (2011);
- *energy-based parameters*, e.g., the maximum, mean, median, and variance of energy, such as in the work of Yu *et al.* (2004);
- *MFCC (Mel-Frequency Cepstral Coefficients)*, *LPC (Linear Prediction Coding) parameters*, *LFPCs (Log Frequency Power Coefficients)*, their maximum, minimum and mean values, and the range, standard deviation, mean, and standard deviation of their derivatives (Iliou and Anagnostopoulos, 2010);
- *formant-based parameters*, such as the maximum, mean, median, and variance of formant frequencies and formant bandwidth (Janicki and Turkot, 2008);
- other spectral parameters, such as spectral moments, and spectral flatness;
- parameters related to *speech rate*, such as the maximum, mean, median and variance of speech rate, voicing change rate, and voicing ratio, that is, the ratio of voiced speech parts to total length of speech;
- *voice quality parameters*, such as incompleteness of glottal closure (IC), spectral gradients, skewness coefficients (Lugger and Yang, 2007), $F0$ jitter, and amplitude shimmer;
- *Teager Energy Operator (TEO)-derived measures* (Ayadi *et al.*, 2011), such as normalized TEO autocorrelation envelope (He *et al.*, 2008) or TEO-decomposed FM variation;
- *linguistic features*: higher-level features related to, for example, the vocabulary used, part-of-speech distribution, and so on (Seppi *et al.*, 2008).

Since often hundreds (or even thousands) of parameters are calculated out of a single utterance, typically these features undergo a selection process. One of the most popular techniques, which gives the best results, is the Sequential Forward Selection (SFS) technique (used, for example, by Grimm *et al.* (2007)). In some projects, however, features are not selected, but they all feed a classifier, even if there are thousands of them (Hassan and Damper, 2010).

2.4. Classification algorithms. Actual emotion recognition is performed by a classifier, previously trained on the training corpus. Researchers use a couple of classification algorithms, known to perform well in various classification tasks. The most promising classifiers are briefly described below.

Among the classifiers that yield the best results and are therefore the most popular are *Support Vector Machines* (SVMs). This algorithm was proposed and described by Vapnik (1982). It is well known for solving various classification problems in different areas of science, for example, in medicine (Jeleń *et al.*, 2008), where it outperformed neural networks in breast cancer detection, or in psychology (Gorska and Janicki, 2012), where it helped with classifying personality traits based on handwriting. In its basic form an SVM consists in dividing, based on training data, a feature space into two parts by an optimal hyperplane defined by the so-called support vectors. It is often used in emotion recognition (e.g., Devillers and Vidrascu, 2006; Erden and

Arslan, 2011; Schuller *et al.*, 2006; Seppi *et al.*, 2008). When there are more than two classes (which is often the case in emotion recognition), a set of SVMs is used with a selected voting algorithm, thus forming a multi-class SVM. A variant of SVM, called the Support Vector Regression (SVR) model, which aims at finding a regression model within emotional speech data, was proposed by Grimm *et al.* (2007).

Another classifier is *Hidden Markov Models* (HMMs), an algorithm used, for example, by Kang *et al.* (2000), which allows a temporal sequence of observations (in this case, speech features) to be modeled, while the actual sequence of states remains unknown ("hidden"). Separate HMM models are trained for different emotions. During recognition the likelihood of generating the tested sequence of observations by a given HMM is calculated, and the emotion with the highest likelihood is selected. Monophone-based HMMs were also proposed for modeling frame-level acoustic features by Gajsek *et al.* (2013), who achieved classification improvement both in emotion recognition and the alcohol detection, compared with the other state-of-the-art methods. Usually HMMs with Gaussian outputs are used.

Gaussian Mixture Models (GMMs) are a classifier similar to HMMs but use only one state and Gaussian outputs (unlike in HMMs, Gaussian outputs are compulsory in GMMs). In speech processing, GMMs are used intensively in speaker recognition, for example, by Janicki (2012), as well as in emotion recognition (e.g., Clavel *et al.*, 2007; Erden and Arslan, 2011; He *et al.*, 2008). Gaussian Mixture Vector AutoRegressive (GMVAR) models are variants of GMMs, which are quite successfully employed for emotion recognition by Ayadi *et al.* (2007).

A *k-Nearest Neighbor* (k-NN) classifier is a very simple one that makes decisions based on k training vectors which are closest to the tested vector(s). Despite its simplicity, it is also often used in emotion recognition (e.g., Kang *et al.*, 2000; Grimm *et al.*, 2007; Kaminska and Pelikant, 2012), often as a baseline classifier, to show the complexity of the classification task and as a reference to other, more complex classifiers.

Artificial Neural Networks (ANNs) have been used for decades in machine learning. They consist of interconnected artificial neurons. The output of a neuron is usually activated by a non-linear activation function (for example, sigmoid) based on a weighted sum of the inputs of the neuron. ANNs can have various architectures, of which the feed-forward one is mostly met in emotion recognition (e.g., Iliou and Anagnostopoulos, 2010; Yacoub *et al.*, 2003). ANNs are usually trained in supervised mode, using the back-propagation algorithm. Another type of ANN architecture is represented by recurrent networks, used, e.g., in automatic control systems described by Patan and Korbicz (2012).

Yet another classification method uses *Decision Trees* (DTs). DTs are decision-support algorithms, which have a graphical form of a tree in which each node is a decision point. A properly constructed decision tree can provide good results in emotion recognition, as described by Cichosz and Slot (2007), where in each decision node one emotion was identified, based on frequency-related, energy-related, and duration-related features.

Other classifiers that are sometimes used are naive Bayes classifiers, the Maximum Likelihood Bayes (MLB) classifier (used, for instance, by Kang *et al.*, (2000) or Lugger and Yang (2007)), Linear Discriminant Classifiers (LDCs) employed by Batliner *et al.* (2005), Quadratic Discriminant Analysis (QDA), and Fisher's linear classifier. Combinations of the above-described classification algorithms are also met.

Recognition rates achieved with these classifiers are difficult to compare, because of the variety of factors which influence the results (speech corpora used, number of emotion classes considered, employed testing methodology, etc.). According to Ayadi *et al.* (2011) they range from 51.19% to 81.29%, but some researchers report even higher results, especially for binary emotion classification.

3. Aim of this study

Many researchers describe emotion recognition in either speaker independent or speaker dependent modes. There are only a few studies which describe these two approaches in one study (e.g., Cichosz and Slot, 2007; Iliou and Anagnostopoulos, 2010); however, precise comparisons detailing the amount of speaker data used in training were not described.

The aim of this study is to fill this gap. We wanted to compare the performance of emotion classification between speaker independent and speaker dependent configurations, using the same environment, that is, the same speech features and the same classifiers, at the same time as controlling the number of speaker recordings present in the training set (that is, controlling the "depth" of speaker dependence). The use of a widely investigated speech corpus containing multiple emotions of the same speakers and selected classification algorithms chosen from among the best and most commonly used ones was planned. Therefore we decided to use the EMO-DB corpus (Burkhardt *et al.*, 2005) and the following classifiers: the k-NN, the ANN and the SVM.

4. Experimental setup

4.1. Speaker independent and speaker dependent configurations. To perform experiments on both speaker dependent and speaker independent classification

and allow comparison between the results, three experimental configurations were applied:

- *SIF*: Speaker Independent, Full. In this configuration evaluation was performed by cross-validation in the leave-one-speaker-out strategy, for all available utterances of the classified speaker;
- *SI-i*: Speaker Independent, limited to i utterances. In this case the experiments were run similarly to SIF, but the number of utterances in the testing set for each emotion for each speaker was limited to $i = L_{\min}$, where $L_{\min}$ is the number of emotions in the least numerous emotion class for a given speaker;
- *SD-i*: Speaker Dependent, limited to i utterances, similarly to SI-i. However, in this configuration n utterances of the speaker that are recognized are moved to the training set, with the aim of adapting the classifier to the given speaker. This procedure is repeated using the cross-validation scheme (in each iteration n utterances are moved to the training set and $k - n$ utterances remain to be tested, where k denotes the initial number of the utterances in the testing set). At the same time, n other randomly chosen utterances are removed from the training data set to keep its size constant.

The SIF configuration is mostly used in other studies on the EMO-DB corpus, so it was added here to allow a comparison with other researchers. Having the same number of utterances with each emotion in the training sets in the SI-i and SD-i configurations allows the most accurate comparison to be made between speaker dependent and speaker independent classification.

4.2. Speech data and their parametrization. To observe the influence of the presence of speaker samples in the training set on classification performance, a two-step cross-validation was performed. It was also considered to be a way of overcoming the relatively small number of samples in the EMO-DB corpus. Unfortunately, this meant that speakers whose number of samples of emotion was lower than the number of samples moved to the training set incremented by one had to be removed from the test set (for proper testing, at least one of each emotion had to be left in the test set). The emotion label "disgust" had to be dropped from the SI-i and SD-i configurations because of the small sample representation and unbalanced distribution between the speakers, so these modes used six-class emotion recognition. The final summary of samples for each i value is shown in Table 1. The emotion recognition for the SIF configuration remained a seven-class one.

Speech data were parametrized and 431 parameters were calculated from each recording. They included MFCC, LPC and LFPC coefficients, and $F0$-based, energy-based, speech rate-based and voice quality parameters, as detailed in Table 2. $F0$ values and all $F0$-related parameters were calculated using the SWIPE' algorithm (Camacho and Harris, 2008) while voice quality parameters were obtained following the procedure described by Lugger *et al.* (2006), and the remaining parameters were extracted using the functions from the Voicebox toolbox for Matlab (Brooks, 2012).

Next, the SFS procedure was run for feature selection. In order to be able to compare speaker dependent and speaker independent configurations, as well as various classifiers, we wanted to have the same set of parameters in all configurations. To avoid the curse of dimensionality, we decided to keep the space dimension relatively low and to verify the results using the cross-validation procedure. We also observed the standard deviation of the results, since experiments were repeated several times and the results varied, e.g., due to the training nature of ANNs. High deviation would mean overfitting to some parts of the data and poor recognition of the other parts of the data. Since the achieved classification results were satisfactory and the standard deviation of the results was low, we agreed that the adopted dimensionality, i.e., 14 (ca. two dimensions per emotion), was not too excessive.

Unfortunately, the SFS mainly returned different parameters for each classifier, and only a few of them appeared regularly. Therefore they were selected

Table 1. Number of speakers and their recordings for various i, used both in the SI-i and SD-i configurations.

i	No. of speakers taking part in validation	Average No. of recordings per speaker in one validation run	Total No. of recordings in one validation run
1	9	28.7	258
2	7	29.1	204
3	7	24.0	168
4	5	30.0	150
5	3	36.0	108
6	2	42.0	84

i	Total no. of recordings	No. of cross validation folds in emotion CV	No. of recordings in a test set in emotion CV
1	454	from 3 to 7	from 2 to 6
2	384	from 2 to 3	2 or 4
3	384	4	1
4	285	5	1
5	172	6	1
6	114	7	1

for further experiments, together with a couple of heuristically added ones. In all, we selected the following 14 parameters: three $F0$-based, five formant-based, two MFCC-based, three spectral, and one speech-rate parameter.

4.3. Configuration of classification algorithms. Each of the selected classification methods was first evaluated in the SD-1 configuration to choose the best parameters. Then experiments with the configurations SIF, SI-i and SD-i for $i = 1, 2, 3, 4, 5$, and 6 were performed.

As for the k-NN algorithm, the number of neighbors k was chosen based on experiments. Various values of k between 1 and 100 were tried and classification accuracy was assessed. The results are shown in Fig. 1. Based on these experiments, we decided to use the k-NN algorithm with $k = 40$ neighbors as the optimal value, and with the Euclidean metric.

As for the ANN classifier, a feed-forward, two-layer back-propagation neural network was chosen, with a sigmoid function in the hidden layer and a linear output layer. It is a simple, easy to train and frequently used ANN configuration. The initial results showed that the optimal strategy is one-versus-one.

The number of neurons in the hidden layer was adjusted experimentally. Various numbers between 1 and 20 were tried. It was found that for more than 10 neurons the classifier exhibited overfitting tendency. The results for five neurons in the hidden layer proved to assure both the sufficient generalization capabilities of the classifier and the good stability of the classification results (below 1.5% for all the tested configurations). Therefore, in the subsequent experiments, a two-layer ANN with five neurons in the hidden layer was used.

The experiments with the SVM classifier were performed with polynomial and radial basis (RDF) kernels. The initial experiments showed that the latter was superior, so the further experiments focused on the RDF kernel. The parameters, γ and the cost C, were set up heuristically, based on the average and standard deviation of classification error, the share of support vectors in the total number of vectors and the training error. These values for various γ and C are presented in Table 3. Based on these results, the values $C = 32$ and $\gamma = 0.125$ were chosen.

Table 2. Parameters extracted from the speech signal.

Group	Description	Count
$F0$-based	$F0$, mean, median, min, max, stddev, range 95%(90%,80%,25%)-range delta, delta^2	38
energy-based	energy, mean, median, min, max, stddev, range 95%(90%,80%,25%)-range delta, delta^2	66
speech rate	voicing ratio voicing change per sec	2
formant-based	$F1-F4$, mean, median, min, max, stddev,range 95%(90%,80%,25%)-range delta, delta^2 $B1-B4$ formant width	208
spectral-based	MFCCs, LPCs, LFPCs mean, median, min, max delta, delta^2	72
voice quality	spectral flatness measure incompleteness of closure spectral center skewness gradient mean, median, min, max delta, delta^2	45
Total:		431

5. Results of experiments

Results for all three classifiers were evaluated based on the mean classification accuracy, both for the whole corpus and for each speaker independently. The best result in the SIF configuration was 68.7% and was achieved with the SVM classifier. The other two classifiers yielded worse results: 63.22% and 56% for the k-NN and ANN, respectively.

With regard to the SI-i and SD-i configurations, the results will be described separately for each of the classifiers. Confusion matrices for the classes being recognized will be presented and analyzed too. In

Table 3. Classification evaluation for various values of γ (in rows) and C (in columns) for the SVM classifier (values in percentages). Results for the configuration selected as optimal are printed in bold.

$\log_2(\gamma)$	Parameter	8	32	128	512
-7	Avg. error	62.19	31.30	28.45	28.95
-7	Std. dev.	1.25	1.47	1.28	1.40
-7	SV share	72.33	50.68	33.62	24.21
-7	Train error	22.89	15.24	9.46	6.01
-5	Avg. error	31.54	28.55	28.99	29.12
-5	Std. dev.	1.54	1.38	1.40	1.37
-5	SV share	50.82	33.70	24.30	19.20
-5	Train error	15.28	9.46	5.99	3.96
-3	Avg. error	27.95	**28.51**	28.94	30.20
-3	Std. dev.	1.38	**1.43**	1.23	1.26
-3	SV share	34.04	**24.55**	19.41	17.19
-3	Train error	9.51	**5.90**	3.72	2.36
-1	Avg. error	28.51	30.08	32.64	35.24
-1	Std. dev.	1.27	1.25	1.34	1.60
-1	SV share	25.56	20.59	18.82	18.27
-1	Train error	5.73	3.31	1.67	0.69

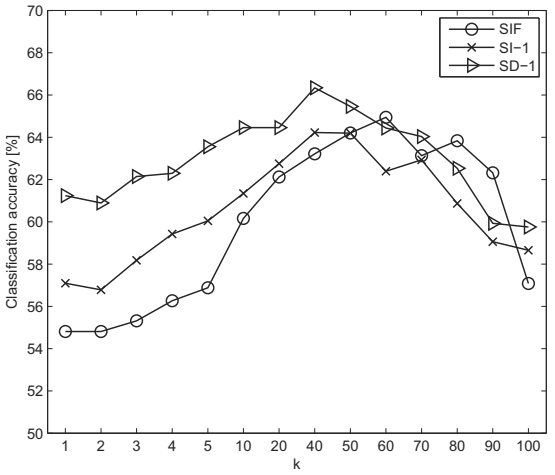

Fig. 1. Emotion classification error for various k for the k-NN classifier, for speaker dependent and speaker independent configurations.

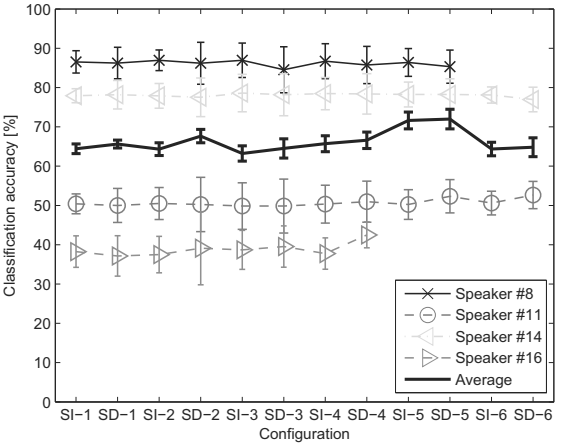

Fig. 2. Emotion classification error for various configurations for the k-NN classifier.

the following section the results will be summarized, discussed, and compared with other studies.

5.1. *k*-NN algorithm. The total results for the k-NN classifier are shown in Table 4. It can be observed that for $i < 5$ the classification accuracy for the SD-i configurations is 1–2 percentage points (p.p.) higher than for the speaker independent configuration, while the performances for $i = 5$ and $i = 6$ are almost equal (within the confidence level of the test).

The impact of adding speaker samples to the training set was highly dependent on the evaluated speaker. The general trend was that speakers who were classified with high results in SI-i (e.g., speakers $\#8$ and $\#14$; see Fig. 2) did not improve their results in the SD-i configuration, or even showed a slight drop. On the other hand, speakers with low results in SI-i (e.g., speakers $\#11$ and $\#16$) performed much better when their samples were present in the training set. The gain for SD-i reached over 9 p.p. in some cases.

In Fig. 2 it can be seen that the average classification performance for each speaker has only a small variation. The highest differences are for $i = 4, 5, 6$, where the variation of average classification accuracy is caused by the elimination of speakers from the test set. The increase for $i = 5$ is caused by the elimination of the poorly classified speaker $\#16$. The decrease for $i = 6$, on the other hand, is caused by the elimination of speaker $\#8$, whose classification error was very low.

The confusion matrix shown in Table 5 presents a high rate of misclassification between happiness and anger, equal to about 50%, for both SI and SD configurations. We believe that the reason for this is that both emotions are characterized by high arousal and are therefore difficult to distinguish. Further investigation of the results for each speaker in SI-1 revealed that five out of nine speakers did not have even one correctly classified sample of happiness (mostly misclassified as anger), and

Table 4. Classification results (in percentages) for various numbers of speaker samples (i) in the training set for the k-NN classifier.

i	1	2	3	4	5	6
SI-i	64.41	64.33	63.23	65.71	71.62	64.35
SD-i	65.63	67.64	64.51	66.59	71.97	64.81

Table 5. Confusion matrices for the SI-1 (upper) and SD-1 (lower) configurations for k-NN (in percentages of recognitions). The diagonals show the percentages of correctly recognized emotions.

	Neu	Ang	Hap	Sad	Bor	Fea
Neu	**74.80**	0.00	0.00	2.53	13.69	9.30
Ang	1.51	**94.56**	1.69	0.00	0.00	2.24
Hap	8.08	52.24	**19.42**	0.00	5.67	14.59
Sad	7.53	0.00	0.00	**84.83**	3.87	3.78
Bor	28.14	0.73	0.99	6.48	**57.70**	5.96
Fea	21.10	17.27	3.43	0.00	2.50	**55.70**

	Neu	Ang	Hap	Sad	Bor	Fea
Neu	**72.53**	0.01	0.00	2.89	15.63	8.94
Ang	2.02	**94.07**	1.67	0.00	0.00	2.25
Hap	7.36	49.44	**25.19**	0.00	4.05	13.97
Sad	4.87	0.00	0.00	**87.59**	3.81	3.74
Bor	26.24	0.97	0.62	6.53	**60.48**	5.16
Fea	22.90	14.68	3.84	0.02	2.82	**55.74**

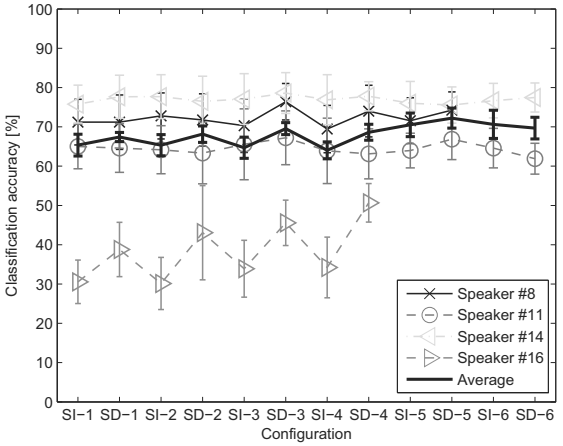

Fig. 3. Emotion classification error for various configurations for the ANN classifier.

the results of the other four showed that 20% of answers were correct. In contrast, the situation of classifying anger as happiness was very rare. It is supposed that in the k-NN classifier anger was represented by a set of vectors (potential "neighbors") not accompanied by representatives of happiness, whilst happiness was often accompanied by representatives of anger. It is noticeable, however, that the recognition of happiness increased from 19.42% to 25.19% when switching to the speaker dependent configuration, which probably caused an increase in the "distinct" representatives of happiness.

5.2. Artificial neural network. Table 6 shows that the average classification accuracy values for $i = 1, \ldots, 4$ are rather stable and variates at the level of 65% for SI-i and 68% for SD-i, so here the speaker dependent configuration improves the recognition accuracy by ca. 3 p.p. on the average. A performance increase for $i = 5$ can be also noticed, as for the k-NN classifier, due to the elimination of speaker #16. For $i = 6$, speaker dependent emotion recognition seems slightly inferior.

As shown in Fig. 3, the highest increase of emotion recognition accuracy is observed for speaker #16, reaching 16 p.p., when switching from SI-4 to SD-4. For other speakers the difference was not so distinct. Also, due to the low number of samples in the test set for low values of i (for example, $i = 2$), the confidence interval (at the confidence level of 0.95) shown as the error bars in Fig. 3 is fairly high.

The confusion matrix presented in Table 7 for $i = 1$ clearly demonstrates that in general the performance of speaker dependent recognition of individual emotions is significantly better than it was for the k-NN. It is mostly visible for fear, where the difference is greater than 9 p.p.. Happiness is still the least recognizable emotion, but misclassification between happiness and anger is much lower. Similar results were also observed for other values of i.

5.3. SVM algorithm. As shown in Table 8, for $i = 1, \ldots, 4$ the recognition accuracy of the SVM classifier variates at the level of 69% for SI-i and 72% for SD-i. These results are significantly better than those for the k-NN and ANN. Similarly to the previous classifiers, the speaker dependent configuration improves the recognition accuracy by ca. 3 p.p. For $i = 5, 6$ the increase is only minor.

Figure 4 shows the results of emotion recognition for the most characteristic speakers. The highest gain was again observed for speaker #16, for whom it reached 21 p.p. in the case of $i = 4$. Speaker #15 also yielded a remarkable improvement. For the other speakers the SD configuration did not improve the results much, and for

Table 6. Classification results (in percentages) for various numbers of speaker samples (i) in the train set for the ANN classifier.

i	1	2	3	4	5	6
SI-i	65.35	65.33	64.68	64.02	70.52	70.63
SD-i	67.40	68.14	69.56	68.62	72.22	69.68

Table 7. Confusion matrices for the SI-1 (upper) and SD-1 (lower) configurations for the ANN (in percentages of recognitions). The diagonals show the percentages of correctly recognized emotions.

	Neu	Ang	Hap	Sad	Bor	Fea
Neu	**54.17**	0.00	3.93	4.05	22.26	15.60
Ang	0.36	**78.45**	17.38	0.00	0.00	3.81
Hap	5.12	38.69	**46.07**	0.00	4.17	5.95
Sad	5.48	0.00	0.00	**84.17**	3.81	6.55
Bor	18.21	0.83	2.98	5.24	**65.83**	6.90
Fea	10.12	14.29	12.26	1.07	2.86	**59.40**

	Neu	Ang	Hap	Sad	Bor	Fea
Neu	**59.76**	0.00	3.57	4.05	17.62	15.00
Ang	0.24	**80.48**	16.90	0.00	0.00	2.38
Hap	3.81	35.24	**50.95**	0.00	4.48	4.52
Sad	3.10	0.00	0.00	**88.81**	4.29	3.81
Bor	18.10	0.71	1.90	4.76	**68.57**	5.95
Fea	10.71	10.00	8.33	0.48	1.67	**68.81**

Table 8. Classification results (in percentages) for various numbers of speaker samples (i) in the training set for the SVM classifier.

i	1	2	3	4	5	6
SI-i	69.41	69.82	68.37	69.13	75.31	74.54
SD-i	71.81	72.8	72.61	72.45	75.44	75.07

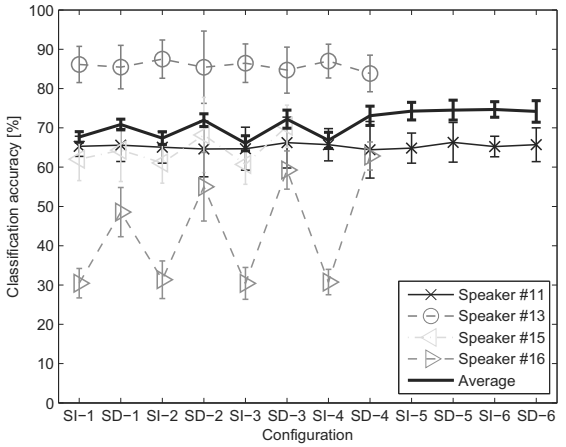

Fig. 4. Emotion classification error for various configurations for the SVM classifier.

speaker #13 the results even showed a small decrease.

Table 9 shows confusion matrices for the SVM in configurations SI-1 and SD-1. The high recognition rate of sadness compared to the other emotions is noticeable. This is the most well recognized emotion and the one that is least frequently misclassified as other emotions. The recognition of happiness improved further. Fear and boredom are sometimes confused with neutral emotion, in both configurations. For example, in the case of speaker #3, whose neutral emotion was recognized with a 100% success rate, 61% of samples of boredom were classified as neutral. On the other hand, for speaker #10, whose boredom was recognized with a 78% correct-classification rate, 51% of samples of neutral emotion were recognized as boredom.

Table 9. Confusion matrices for the SI-1 (upper) and SD-1 (lower) configurations for the SVM (in percentages of recognitions). The diagonals show the percentages of correctly recognized emotions.

	Neu	Ang	Hap	Sad	Bor	Fea
Neu	**63.17**	2.03	4.65	3.40	17.99	8.75
Ang	1.48	**83.34**	13.63	0.00	0.00	1.54
Hap	5.06	32.09	**52.47**	0.00	4.04	6.34
Sad	2.97	0.00	0.00	**87.15**	4.83	5.06
Bor	17.27	0.64	4.65	7.18	**65.32**	4.94
Fea	10.99	12.44	7.12	1.66	2.76	**65.03**

	Neu	Ang	Hap	Sad	Bor	Fea
Neu	**66.28**	0.71	4.30	3.85	17.91	6.96
Ang	2.06	**82.53**	14.56	0.00	0.00	0.85
Hap	3.14	31.44	**57.09**	0.00	3.40	4.93
Sad	2.51	0.00	0.00	**92.49**	2.24	2.76
Bor	16.55	0.81	5.19	7.88	**65.28**	4.29
Fea	11.98	9.31	7.02	1.30	3.17	**67.22**

6. Summary and discussion of results

All the tested classification methods yielded classification accuracies between 64% and 75%. It is worth remembering that in the case of six classes the choice level is $1/6 = 16.67\%$, so the accuracy results are far above this level. They are also higher than the estimated level of human performance in speaker independent conditions (60%, shown by Scherer (2003)).

With regard to the comparison between speaker dependent and speaker independent conditions, Fig. 5 shows that in almost all configurations the speaker dependent configuration improved recognition; however, this increase was minor for the k-NN classifier. The higher improvement was observed for speakers who were originally poorly recognized (e.g., speaker #16). The best recognition results were achieved by the SVM, followed by the ANN and k-NN classifiers.

The k-NN classifier had the worst happiness recognition, i.e., 0% for five out of nine speakers. Speaker dependent recognition had a small influence on the recognition rate mostly due to the specificity of the k-NN algorithm: for complex problems it requires a high value of k (the number of neighbors, in this case $k = 40$). The addition of one to six samples to the training set had only a slight influence on the decision border, compared with the number of 40. Only speakers #15 and #16 yielded better results in SD-i, which influenced the overall result.

The ANN classifier had a similar recognition rate, although with better emotion recognition distribution. The classification performance for speakers was more uniform—eight out of nine speakers had classification accuracies between 65% and 70%, and only speaker #16 yielded an accuracy of 30%. Another positive aspect is that the recognition rate grew in the SD-i configuration in the case of four speakers and increased when more samples were inserted into the training set.

The SVM classifier had the best overall recognition rate of 75.44% and the best performance in the SD-5 configuration, with a classification error below 25%. Emotion recognition was at a different level for each speaker, which places this classifier between the distributions achieved by the k-NN and ANN.

A somewhat strange behavior of the tested classifiers for $i > 4$ was caused by deficiencies of the corpus used: EMO-DB unfortunately did not contain enough samples to obtain a representative training and testing set of speakers and their emotional recordings. Therefore the results for $i < 5$ should be treated as more reliable.

All classifiers showed the presence of pairs of emotions which were often confused, for example, anger and happiness, boredom and neutrality (see Tables 5, 7 and 9). We believe that this was caused by high class infiltration, that is, there was no dimension that could distinguish between these emotions. This can

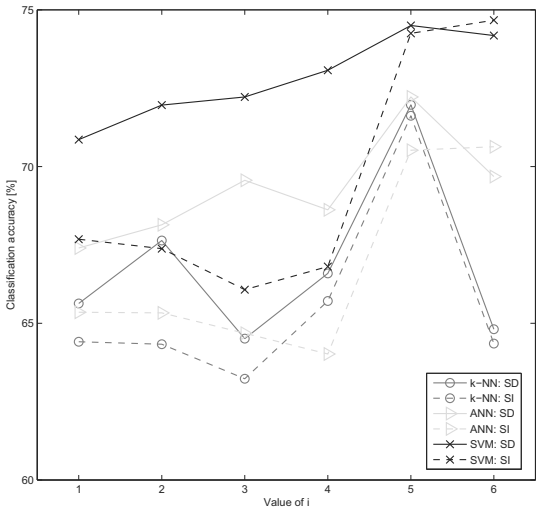

Fig. 5. Comparison of speaker independent and speaker dependent configurations for all three classifiers tested.

also be shown by analyzing the SVM training error: the SVM consisted of many one-versus-one classifiers, each distinguishing between a pair of emotions. For the happiness–anger pair the training error was the highest. Both these emotions show high arousal and are therefore often confused. The differ as for valence (positive vs. negative); however, this feature is much more difficult to be captured using speech signal parameters. We think that replacing some of the currently used speech parameters with novel ones (e.g., TEO based) could possibly improve it.

The SVM was chosen as the best classifier, as it reached the highest recognition rates in all configurations, both at the level of overall performance and in the emotion recognition of each speaker.

6.1. Comparison with other studies of EMO-DB. As for the six-class recognition, the best result obtained in the SI-1 configuration (69.41%) turned out to be slightly inferior compared to 71.2% described by Neiberg *et al.* (2010), who used time varying constant-Q cepstral coefficients and $F0$ normalization. As for the speaker dependent configurations, the best result (75.44%) obtained in the SD-5 configuration in this study is similar to that described by Cichosz and Slot (2007) (74.40%) as well as Ayadi *et al.* (2007) (76.00%). The confusion matrices also showed similar tendencies.

Iliou and Anagnostopoulos (2010) claim that they attained 84% using a speaker dependent configuration, SVMs and seven emotions. However, various details of this study remain unclear, such as how it was possible to achieve speaker dependent conditions with only 45 recordings of "disgust" for all 10 speakers and using 10-fold cross-validation.

With regard to the seven-class emotion recognition, we were able to compare the results achieved in the SIF configuration, that is, with all samples, in speaker independent mode. The best classification accuracy achieved in this study (68.7%) proved to be higher than 66.5% achieved by Janicki and Turkot (2008); however, it was lower than results achieved in other studies, for example, 78.32% by Xiao *et al.* (2006), 81.9% by Liu *et al.* (2010), 84.6% by Schuller *et al.* (2009), and as high as 92.3% by Hassan and Damper (2010) when using SVMs and Directed Acyclic Graphs (DAGs). It must be noted, however, that Xiao *et al.* (2006) used a different methodology, in which only female voices from EMO-DB were employed and half of them were employed for training while the whole set was used for testing. This in fact gave speaker dependent conditions, which clearly boosted the result. Also Liu *et al.* (2010) used another methodology; namely, they employed only one fifth of the data for testing, without carrying out a cross-validation procedure, so a direct comparison with our results is not possible. On the other hand, both Hassan and Damper (2010) as well as Schuller *et al.* (2009) used 6552 features per utterance, without any feature selection, which seems a weak point of their approach.

7. Conclusions

In our study we showed that adding even a few sentences spoken by a speaker improves the emotion recognition rate. This effect was especially highly visible for speakers with low recognition rates in the speaker independent scenario. Therefore it is recommended that some reference recordings of the examined speaker(s) be added to the system wherever feasible, that is, when any historical recordings are available.

Direct comparison with some of the other studies turned out to be difficult. Where such a comparison was possible, the obtained results showed similar tendencies. Some studies yielded better results thanks to novel parameters, among others things.

Comparison between the three tested classifiers showed superiority of the SVM classifier. Further studies could include experiments with other algorithms, such as HMMs, and possibly experiments with other features which proved successful in other studies, as well as improving the methods of attribute selection.

References

Ayadi, M.E., Kamel, M.S. and Karray, F. (2007). Speech emotion recognition using Gaussian mixture vector autoregressive models, *IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP 2007), Honolulu, HI, USA, Vol. 4, pp. IV–957–IV–960.

Ayadi, M.E., Kamel, M.S. and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases, *Pattern Recognition* **44**(3): 572–587.

Batliner, A., Steidl, S., Hacker, C., Noth, E. and Niemann, H. (2005). Tales of tuning—prototyping for automatic classification of emotional user states, *Interspeech 2005, Lisbon, Portugal*, pp. 489–492.

Brooks, M. (2012). Voicebox: Speech processing toolbox for Matlab, `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. and Weiss, B. (2005). A database of German emotional speech, *Interspeech 2005, Lisbon, Portugal,* pp. 1517–1520.

Camacho, A. and Harris, J.G. (2008). A sawtooth waveform inspired pitch estimator for speech and music, *Journal of the Acoustical Society of America* **124**: 1638–1652.

Cichosz, J. and Slot, K. (2007). Emotion recognition in speech signal using emotion-extracting binary decision trees, *ACII 2007, Lisbon, Portugal*.

Clavel, C., Devillers, L., Richard, G., Vasilexcu, I. and Ehrette, T. (2007). Detection and analysis of abnormal situations through fear-type acoustic manifestations, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007), Honolulu, HI, USA*, Vol. 4, pp. IV–21–IV–24.

Devillers, L. and Vidrascu, L. (2006). Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs, *Interspeech 2006, Pittsburgh, PA, USA*, pp. 801–804.

Ekman, P. (1972). Universals and cultural differences in facial expressions of emotions, *in* J. Cole (Ed.), *Nebraska Symposium on Motivation*, Vol. 19, University of Nebraska Press, Lincoln, NE, pp. 207–282.

Engberg, I.S., Hansen, A.V., Andersen, O. and Dalsgaard, P. (1997). Design, recording and verification of a Danish emotional speech database, *Eurospeech 1997, Rhodes, Greece*.

Erden, M. and Arslan, L.M. (2011). Automatic detection of anger in human–human call center dialogs, *Interspeech 2011, Florence, Italy*, pp. 81–84.

Gajsek, R., Mihelic, F. and Dobrisek, S. (2013). Speaker state recognition using an HMM-based feature extraction method, *Computer Speech and Language* **27**(1): 135–150.

Gorska, Z. and Janicki, A. (2012). Recognition of extraversion level based on handwriting and support vector machines, *Perceptual and Motor Skills* **114(3)**(0031–5125): 857–869.

Grimm, M., Kroschel, K. and Narayanan, S. (2007). Support vector regression for automatic recognition of spontaneous emotions in speech, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007), Honolulu, HI, USA*, Vol. 4, pp. IV–1085–IV–1088, ID: 1.

Hassan, A. and Damper, R.I. (2010). Multi-class and hierarchical SVMs for emotion recognition, *Interspeech 2010, Makuhari, Japan*, pp. 2354–2357.

He, L., Lech, M., Memon, S. and Allen, N. (2008). Recognition of stress in speech using wavelet analysis and teager energy operator, *Interspeech 2008, Brisbane, Australia*, pp. 605–608.

Hirschberg, J., Benus, S., Brenier, J.M., Enos, F., Friedman, S., Gilman, S., Gir, C., Graciarena, M., Kathol, A. and Michaelis, L. (2005). Distinguishing deceptive from non-deceptive speech, *Interspeech 2005, Lisbon, Portugal*, pp. 1833–1836.

Iliou, T. and Anagnostopoulos, C.-N. (2010). Classification on speech emotion recognition—a comparative study, *International Journal on Advances in Life Sciences* **2**(1–2): 18–28.

Janicki, A. (2012). *On the Impact of Non-speech Sounds on Speaker Recognition*, Text, Speech and Dialogue, Vol. 7499, Springer, Berlin/Heidelberg, pp. 566–572.

Janicki, A. and Turkot, M. (2008). Speaker emotion recognition with the use of support vector machines, *Telecommunication Review and Telecommunication News* (8–9): 994–1005, (in Polish).

Jeleń, Ł., Fevens, T. and Krzyżak, A. (2008). Classification of breast cancer malignancy using cytological images of fine needle aspiration biopsies, *International Journal of Applied Mathematics and Computer Science* **18**(1): 75–83, DOI: 10.2478/v10006-008-0007-x.

Kaminska, D. and Pelikant, A. (2012). Recognition of human emotion from a speech signal based on Plutchik's model, *International Journal of Electronics and Telecommunications* **58**(2): 165–170.

Kang, B.S., Han, C.H., Lee, S.T., Youn, D.H. and Lee, C. (2000). Speaker dependent emotion recognition using speech signals *ICSLP 2000, Beijing, China*.

Kowalczuk, Z. and Czubenko, M. (2011). Intelligent decision-making system for autonomous robots, *International Journal of Applied Mathematics and Computer Science* **21**(4): 671–684, DOI: 10.2478/v10006-011-0053-7.

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

Liscombe, J., Hirschberg, J. and Venditti, J.J. (2005). Detecting certainess in spoken tutorial dialogues, *Interspeech 2005, Lisbon, Portugal*.

Liu, G., Lei, Y. and Hansen, J.H.L. (2010). A novel feature extraction strategy for multi-stream robust emotion identification, *Interspeech 2010, Makuhari, Japan*, pp. 482–485.

Lugger, M. and Yang, B. (2007). The relevance of voice quality features in speaker independent emotion recognition, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007), Honolulu, HI, USA*, Vol. 4, pp. IV–17–IV–20.

Lugger, M., Yang, B. and Wokurek, W. (2006). Robust estimation of voice quality parameters under realworld disturbances, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006), Toulouse, France*, Vol. 1, p. I.

Mehrabian, A. and Wiener, M. (1967). Decoding of inconsistent communications, *Journal of Personality and Social Psychology* **6**(1): 109–114.

Neiberg, D., Laukka, P. and Ananthakrishnan, G. (2010). Classification of affective speech using normalized time-frequency cepstra, *5th International Conference on Speech Prosody (Speech Prosody 2010), Chicago, IL, USA*, pp. 1–4.

Patan, K. and Korbicz, J. (2012). Nonlinear model predictive control of a boiler unit: A fault tolerant control study, *International Journal of Applied Mathematics and Computer Science* **22**(1): 225–237, DOI: 10.2478/v10006-012-0017-6.

Scherer, K.R. (2003). Vocal communication of emotion: A review of research paradigms, *Speech Communication* **40**(1–2): 227–256.

Schuller, B., Koehler, N., Moeller, R. and Rigoll, G. (2006). Recognition of interest in human conversational speech, *Interspeech 2006, Pittsburgh, PA, USA*, pp. 793–796.

Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G. and Wendemuth, A. (2009). Acoustic emotion recognition: A benchmark comparison of performances, *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2009), Merano, Italy*, pp. 552–557.

Seppi, D., Batliner, A., Schuller, B., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N. and Aharonson, V. (2008). Patterns, prototypes, performance: Classifying emotional user states, *Interspeech 2008, Brisbane, Australia*, pp. 601–604.

Vapnik, V.N. (1982). *Estimation of Dependences Based on Empirical Data*, Springer-Verlag, New York, NY, (translation of *Vosstanovlenie zavisimostei po empiricheskim dannym* by Samuel Kotz).

Xiao, Z., Dellandrea, E., Dou, W. and Chen, L. (2006). Two-stage classification of emotional speech, *International Conference on Digital Telecommunications (ICDT'06), Cap Esterel, Côte d'Azur, France*, pp. 32–32.

Yacoub, S., Simske, S., Lin, X. and Burns, J. (2003). Recognition of emotions in interactive voice response systems, *Eurospeech 2003, Geneva, Switzerland*, pp. 1–4.

Yu, C., Aoki, P. M. and Woodruff, A. (2004). Detecting user engagement in everyday conversations, *8th International Conference on Spoken Language Processing (ICSLP 2004), Jeju, Korea*, pp. 1–6.

Jan Jakub Rybka received his M.Sc. in computer science in 2011 from the Faculty of Electronics and Information Technology, Warsaw University of Technology (WUT). His research interests focus on human-computer interaction, including emotion recognition, as well as classification algorithms and data-driven techniques.

Artur Janicki received his M.Sc. and Ph.D. ('97 and '04, respectively, both with honours) in telecommunications from the Faculty of Electronics and Information Technology, Warsaw University of Technology (WUT). He is an assistant professor at the Institute of Telecommunications, WUT. His research and teaching activities focus on speech processing, including speaker recognition, speech coding and synthesis, emotion recognition, with elements of data mining and information theory. He is a member of the International Speech Communication Association (ISCA) and the European Association for Signal Processing (EURASIP).

Received: 3 January 2013
Revised: 26 April 2013