BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES, Vol. 67, No. 3, 2019
DOI: 10.24425/bpasts.2019.129647

Automatic speech based emotion recognition using paralinguistics features

J. HOOK[1], F. NOROOZI[1], O. TOYGAR[2], and G. ANBARJAFARI[1,3]*

[1] iCV Research Group, Institute of Technology, University of Tartu, Tartu 50411, Estonia
[2] Department of Computer Engineering, Eastern Mediterranean University, Famagusta, North Cyprus, via Mersin 10, Turkey
[3] Department of Electrical and Electronic Engineering, Hasan Kalyoncu University, Gaziantep, Turkey.

Abstract. Affective computing studies and develops systems capable of detecting humans affects. The search for universal well-performing features for speech-based emotion recognition is ongoing. In this paper, a small set of features with support vector machines as the classifier is evaluated on Surrey Audio-Visual Expressed Emotion database, Berlin Database of Emotional Speech, Polish Emotional Speech database and Serbian emotional speech database. It is shown that a set of 87 features can offer results on-par with state-of-the-art, yielding 80.21, 88.6, 75.42 and 93.41% average emotion recognition rate, respectively. In addition, an experiment is conducted to explore the significance of gender in emotion recognition using random forests. Two models, trained on the first and second database, respectively, and four speakers were used to determine the effects. It is seen that the feature set used in this work performs well for both male and female speakers, yielding approximately 27% average emotion recognition in both models. In addition, the emotions for female speakers were recognized 18% of the time in the first model and 29% in the second. A similar effect is seen with male speakers: the first model yields 36%, the second 28% a verage emotion recognition rate. This illustrates the relationship between the constitution of training data and emotion recognition accuracy.

Key words: random forests, speech emotion recognition, machine learning, support vector machines.

1. Introduction

The human voice carries information about the speaker's emotional state [1–3]. Emotion recognition can be used by many modalities [4, 5], among them speech-based emotion recognition (SER) is useful for speech-enabled human-machine interfaces (HMI) [6]. SER has applications in enhancing driver safety [7–9], online education [10–12], call centre environments [13–15], interactive games [16], health monitoring [17, 18], and virtual reality [19–21].

SER systems can be categorized by the types of features used as linguistic and paralinguistic. The first describes what is being said while the second describes how it is said. Linguistic SER systems face two main challenges: speech recognition and predicting the emotion from it while paralinguistic SER systems face the second. In addition, linguistic SER can be complicated in multi-language scenarios [22].

In the paralinguistic approach, the emotional state is thought to be represented by various characteristics of the speech signal, such as the tone and range, which can be analyzed in order to detect the emotion. These properties of speech signals can be extracted using signal processing techniques. The principal goal while designing paralinguistic SER systems is to maximize the recognition rate by finding a combination of features that can strongly discriminate between each emotion. At a higher level, such systems should preferably be robust and language-independent, which is challenging to achieve.

A universal set of paralinguistic SER features has not been discovered [23, 24]. The primary goal of this paper is finding a set of features that improve emotion recognition performance across multiple emotional speech databases (ESD), using support vector machines (SVM) as the classifier. The second goal is to keep the size of the feature set as small as possible. Reducing unnecessary computations translates directly to saved costs when an SER system is deployed on cloud-based pay-per-use services. Furthermore, a small feature set can enable SER in embedded applications, where computational resources are scarce.

In this paper, we propose a paralinguistic SER system. The system is tested on four different emotional speech databases to test the robustness and language independence of the features. In addition, an online SER system is created to measure the effect of gender in SER tasks. What is more, the online system serves as a benchmark for the features to measure SER in a plausible real-life scenario.

The paper is organized as follows. Section 2 describes related works and expands on what has been studied before on this topic. Section 3 describes the SER methodology proposed in this paper. In addition, the online system is described in detail. Section 4 describes the databases used and the experimental results. Section 5 discusses the results and their interpretation. Finally, Section 6 presents our conclusion.

*e-mail: shb@icv.tuit.ut.ee

Manuscript submitted 2018-02-19, revised 2018-04-13, initially accepted for publication 2018-05-08, published in June 2019.

2. Related works

A decade of developments in SER is reviewed in [24]. The authors expand upon different feature types, classifiers, emotional speech databases, common tools and more. Furthermore, the authors describe how the focus in SER has evolved over time with regards to feature types, features and classifiers.

During the early stages of SER research, pitch, duration and intensity were the main features studied [25–29]. Later, the attention shifted to voice quality low-level descriptors (LLDs) like harmonics to noise ratio, jitter, shimmer and spectral/cepstral measurements [30–32]. Finally, rhythm and sentence duration [33, 34] and non-uniform perceptual linear predictive (UN-PLP) features [35] and linear predictive cepstral coefficients (LPCCs) [36, 37] are used in conjunction with mel-frequency cepstral coefficients (MFCCs).

In [38], Shaukat and Chen were the first to study SER on the Serbian emotional speech database. The authors developed a multistage strategy with SVMs for emotion recognition. The first stage is the classification of the input as either active or passive. Active emotions are further classified as angry or happy and passive emotions as fear and non-fear, with the latter consisting of sad and neutral. The adoption of a divide-and-conquer strategy enabled them to outperform their peers.

Hassan and Damper [39] use binary SVMs for SER. They note that differrent structures binary decision trees have been used before [38]. SVMs are organized in two standard schemes (one-versus-one and one-versus-rest) and two hierarchies (directed acyclic graph, unbalanced decision tree). The four different structures are tested and compared on multiple ESDs. The authors are able to achieve state-of-the-art performance on two databases.

Shaukat and Chen [40] refine their previously developed multistage strategy. Inspired by psychology, the classification structure is modified to include all of the supposed basic emotions (anger, disgust, fear, happiness, sadness, surprise). This structure can be adopted to any database that contains a subset of these emotions. The new strategy helps the authors improve on their previous result.

Kobayashi and Calag [41] use ensemble methods like RF and kernel factories to improve SER accuracy. Segmental features are used instead of utterance, pointing to the complexity of word and syllable boundary identification. The segmentation strategy involves splitting the sample at fixed relative positions. This approach is argued to be more suitable for real-time processing and adaptable to stream analysis.

Chiou and Chen [42] set out to find the smallest set of features while trying to maximize emotion recognition accuracy. The motivation behind reducing the number of features is twofold. First, not all features contribute positively to emotion recognition accuracy. Second, more features require more resources to process. In their work, the authors start with 6552 baseline feature and an average accuracy of 85.2%. After reducing the baseline features down to 37, a 5% decrease in accuracy is reported.

In [43], Yüncü et al. devised a computational model mimicking the human auditory system. The authors note that the human auditory system has built-in adaptive mechanisms and performs frequency-dependent filtering. Instead of extracting features from the audio files, the files are used as input for the model mimicking the human auditory system. Features are extracted from the outputs of the auditory model. SVMs are used for classification and arranged in a binary decision tree structure. In addition, the authors performed a subjective listening test on one of the ESDs.

3. Methodology

3.1. Features. According to [24], SER features can be categorized as suprasegmental and segmental features. The first is calculated over the full duration of speech while the second is calculated over multiple short-duration segments. In this work, all features are calculated over the full duration of the sample. Although this strategy gives little insight to the detailed change of a feature over time, we can still see the amount of change reflected in standard deviation. Moreover, extracting the features over the full duration of the sample frees us from the choice of a partitioning strategy, adding to the simplicity of the system.

In [24], another way of categorizing features is as LLDs and functionals (applied to LLDs). Since we use suprasegmental features, all features related to LLDs (e.g. MFCCs) are functionals. To summarize, all features used in this work are suprasegmental functionals.

3.2. Feature extraction. Praat 6.0.36 [44] was used for feature extraction because of familiarity with the program and the scripting capabilities. A script extracting the necessary features was created. For each feature in the feature list, there exists a corresponding Praat object. To extract the features, the built-in Praat object functions were used. The parameters for the functions are left at default values presented by Praat, except for MFCCs: instead of the default 12 coefficients, we decided to calculate 24 coefficients. In total, 87 features were extracted from each sample.

3.3. Data preprocessing and kernel parameters. To keep the preliminary work short, the guidelines in [45] were followed. Data was scaled with *svm-scale* with no additional parameters. The kernel used for the SVMs was radial basis function (RBF). Kernel parameters for each database were found with *grid.py* provided by LIBSVM. The script searches for the optimal combination of parameters from a preset set of possible values in a coarse grid search fashion. Using the provided tools with default settings keeps the complexity of the system low and simplifies reproducibility.

3.4. Classification

3.4.1. Offline system. The offline system is the most common way of performing SER: features are extracted from the samples in the ESD, the extracted features are optionally further processed and are finally used for training and testing the machine learning classifier. Usually, the same database is used for

Table 1
Features used in the offline system. p_x is the x-th percentile

Feature	Functionals
Pitch	min, max, mean, p_{25}, p_{50}, p_{75}, stdev, mean absolute slope, slope without octave jumps
Intensity	min, max, mean, p_{25}, p_{50}, p_{75}, stdev
LTAS	min, fmin, max, fmax, mean, slope, stdev
Sound	min, max, mean, stdev, power, energy, RMS
Harmonicity	min, max, mean, stdev
Point process	periods, meanperiod, stdevperiod, jitterlocal, jitterppq5
MFCC(1-24)	mean, stdev

training, validation and testing. SVM [46] was chosen as the classifier for the offline system because there are many works that use it [38–43], enabling us to compare our results to others'. Moreover, the authors of LIBSVM have provided its users with excellent material [45] on how to use SVMs as classifiers. What is more, the provided tools are easy to use. Scikit-learn [47] is used to train and validate the models. The framework provides a way to calculate the confusion matrices for SVMs. Finally, their SVM classifier implementation uses LIBSVM.

In the offline system, 10-fold cross-validation is used for training and testing. This method is often used in SER research [3, 48–54]. With n-fold cross-validation, the data is divided to n disjoint sets of equal size. After the division, each one of the n sets will be used as the testing set while the other $n-1$ sets are used for training. This yields n results that can be averaged to get an idea of the model's accuracy. Choosing this method allows us to use each sample for testing exactly once, giving a more consistent estimate of the model's performance that is not affected by the constitution of the testing and training sets. The process is represented in Fig. 1.

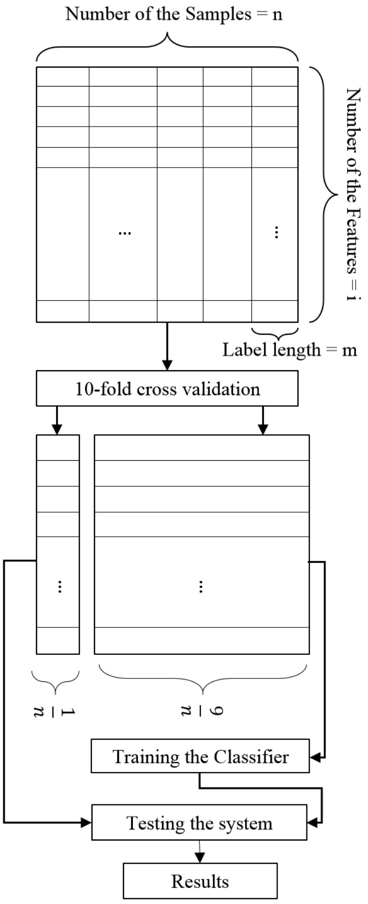

Fig. 1. Offline system representation

3.4.2. Online system.

The online system consists of a web application that acts as the user interface for the emotion recognition subsystem. The speaker will speak in different emotions and the speech will be recorded. The recording will be converted to an audio file, the features are extracted with Praat and the system will predict the emotion based on two trained models. The classification using trained models is done in Weka [55].

Table 2
Sentences spoken in the online system experiment. These sentences are a subset of the sentences used in SAVEE [61]

1	Who authorized the unlimited expense account?
2	Please take this dirty table cloth to the cleaners for me.
3	Call an ambulance for medical assistance.
4	Those musicians harmonize marvelously.
5	The prospect of cutting back spending is an unpleasant one for any governor.
6	The best way to learn is to solve extra problems.

The models will be trained on a subset of SAVEE and EMO-DB. We used SAVEE without surprise samples (60) and EMO-DB without boredom samples (81). This left us with two databases, SAVEE' (420) and EMO-DB' (454), respectively. The trained models will output the class probabilities of six emotions: anger, disgust, fear, happiness, neutral and sadness. The aforementioned databases were chosen because they share 6 common emotions out of 7, both of them are West Germanic languages, the resulting databases are of similar size and SAVEE' has only male samples, while EMO-DB' has both. This way we can observe the effects of how having no female samples in SAVEE' affects emotion prediction performance for female speakers.

In the online system, we use Random Forests (RF) as one of the strongest techniques from the category of ensemble decision trees [3]. RF was introduced in [56]. Decision trees mostly have low bias and high variance, and benefit from averaging processes [57, 58]. They can be excellent candidates for multi-label classification tasks [59]. RF is robust against noise and its accuracy is usually higher than boosting methods. It is also faster than bagging and boosting methods, and can be parallelized easily [60].

Four non-native speakers, two male and two female, participated in the online experiment. The speaker could record, listen and re-record the sample until the speaker felt the emotion was

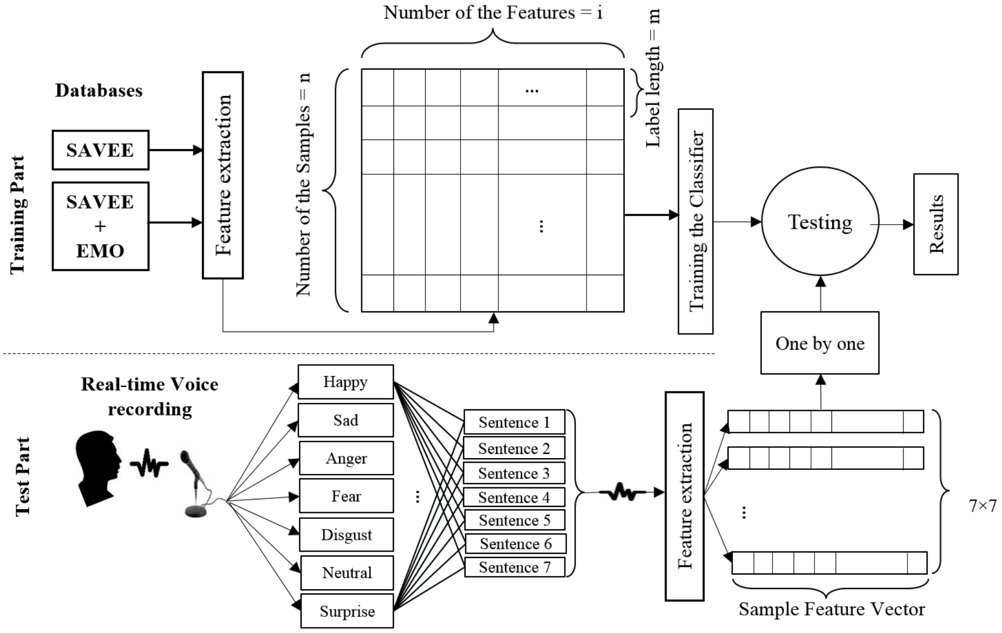

Fig. 2. Online system representation

acted out on a satisfactory level. Outputs of the models are then saved for further analysis. The online system's process is represented in Fig. 2.

4. Experimental results

4.1. Database descriptions.

4.1.1. Offline system. In this work, 4 different emotional speech databases were used:

- Surrey Audio-Visual Expressed Emotion database (SAVEE) [61]
- Berlin Database of Emotional Speech (EMO-DB) [62]
- Polish Emotional Speech Database (PESD) [63]
- Serbian emotional speech database (GEES) [64]

SAVEE contains a total of 480 samples from 4 male speakers. It consists of neutral (120), angry (60), disgust (60), fear (60), happiness (60), sadness (60) and surprise (60) samples. EMO-DB contains a total of 535 samples from 5 male and 5 female speakers. It consists of angry (127), boredom (81), disgust (46), fear (69), happiness (71), neutral (79) and sadness (62) samples. PESD contains a total of 240 samples. It consists of anger (40), boredom (40), fear (40), happiness (40), neutral (40) and sadness (40) samples. Each speaker (2 female, 2 male) voiced 10 samples in each emotional class.

From GEES we chose to use samples in the form of isolated words (32), short (30) and long (30) semantically neutral sentences from 3 male and 3 female speakers each [64]. This makes our results comparable to others and these parts of the database were also chosen to evaluate the database in [64], which makes our results comparable to the human listeners'.

The average lengths of the samples and the male to female ratio of the samples in each database are shown in Table 3.

Table 3
A detailed overview of the databases. M – samples by males, F – samples by females, T – total number of samples, M/T – part of male samples in a database, L – average length of the samples

Database	Labels	M	F	T	M/T	L
SAVEE	Anger	60	0	60	1	3.71
	Disgust	60	0	60	1	3.95
	Fear	60	0	60	1	3.75
	Happiness	60	0	60	1	3.8
	Neutral	120	0	120	1	3.61
	Sadness	60	0	60	1	4.48
	Surprise	60	0	60	1	3.8
Total	**7**	**480**	**0**	**480**	**1**	**3.84**
EMO-DB	Anger	60	67	127	0.472	2.64
	Boredom	35	46	81	0.432	2.78
	Disgust	11	35	46	0.239	3.35
	Fear	36	33	69	0.522	2.23
	Happiness	27	44	71	0.380	2.54
	Neutral	39	40	79	0.494	2.36
	Sadness	25	37	62	0.402	4.05
Total	**7**	**233**	**302**	**535**	**0.434**	**2.78**
PESD	Anger	20	20	40	0.5	2.06
	Boredom	20	20	40	0.5	2.86
	Fear	20	20	40	0.5	2.31
	Happiness	20	20	40	0.5	2.14
	Neutral	20	20	40	0.5	2.04
	Sadness	20	20	40	0.5	2.44
Total	**6**	**120**	**120**	**240**	**0.5**	**2.31**
GEES	Anger	276	276	552	0.5	2.61
	Fear	276	276	552	0.5	2.82
	Happiness	276	276	552	0.5	2.82
	Neutral	276	276	552	0.5	2.65
	Sadness	276	276	552	0.5	3.31
Total	**5**	**1380**	**1380**	**2760**	**0.5**	**2.84**

4.2. Offline system. As mentioned in section 3.4.1, the offline system represents the standard way of SER research. Looking at the results, patterns can be noticed across all databases. There is symmetrical confusion between anger and happiness in Tables 4, 5, 7, and 8. This also occurs in similar works [38,40,43]. The confusion is not limited to machine learning classifiers: humans in [64] also had trouble distinguishing between these two emotions.

Table 4
Row – the true label, column – the predicted label. Confusion matrix for SAVEE. Average accuracy: **80.21%**. Kernel parameters: $\gamma = 2^{-7}$, $C = 128$

	ANG	DIS	FEA	HAP	NEU	SAD	SUR
ANG	**85.00**	5.00	1.67	8.33	0.00	0.00	0.00
DIS	3.33	**81.67**	3.33	0.00	6.67	1.67	3.33
FEA	1.67	6.67	**71.67**	8.33	0.00	1.67	10.00
HAP	15.00	0.00	13.33	**61.67**	1.67	0.00	8.33
NEU	0.00	2.50	0.00	0.83	**94.17**	2.50	0.00
SAD	0.00	8.33	0.00	0.00	13.33	**78.33**	0.00
SUR	1.67	0.00	15.00	8.33	0.00	0.00	**75.00**

Another confusion, but less pronounced, occurs between neutral, sadness and boredom. If boredom is not present in the database, then neutral and sadness are mutually confusing as seen in Tables 4 and 8. This is also true for human listeners in [64]. The confusion changes when boredom is present in a database: in Table 5, discrimination between boredom and neutral is the hardest, but in Table 7, we can see that boredom is more often confused with sadness than with neutral, although the difference is small (2.5%).

Table 5
Row – the true label, column – the predicted label. Confusion matrix for EMO-DB. Average accuracy: **88.6%**. Kernel parameters: $\gamma = 2^{-7}$, $C = 32$

	ANG	BOR	DIS	FEA	HAP	NEU	SAD
ANG	**93.70**	0.00	0.00	0.79	4.72	0.79	0.00
BOR	0.00	**90.12**	1.23	0.00	1.23	6.17	1.23
DIS	0.00	2.17	**82.61**	4.35	4.35	4.35	2.17
FEA	4.35	0.00	1.45	**91.30**	1.45	1.45	0.00
HAP	21.13	0.00	1.41	9.86	**66.20**	1.41	0.00
NEU	0.00	6.33	0.00	1.27	0.00	**92.41**	0.00
SAD	0.00	0.00	0.00	0.00	0.00	1.61	**98.39**

In SAVEE, the most well-recognized emotion was neutral, however this emotion has twice the number of samples (120) compared to others. Comparing SAVEE to other databases, a few results stand out. SAVEE has the highest rate of happiness samples misclassified as fear (13.33%), followed by EMO-DB (9.86%). This was not present in PESD and GEES (Tables 7, 8). Mistaking sadness for neutral (13.33%) is comparable to PESD (10%). This also occurs with sadness: 13.33% of sadness samples were misclassified as neutral, compared to EMO-DB, where the rate was 1.61%.

Table 6
Comparing our results to human listeners' as described in [62]. Improvements are in **bold**

EMO-DB	This work	Humans	Difference
Anger	93.7	96.9	−3.2
Boredom	90.12	86.2	**3.92**
Disgust	82.61	79.6	**3.01**
Fear	91.3	87.3	**4**
Happiness	66.2	83.7	−17.5
Neutral	92.41	88.2	**4.21**
Sadness	98.39	80.7	**17.69**

Table 7
Row - the true label, column – the predicted label. Confusion matrix for PESD. Average accuracy: **75.42%**. Kernel parameters: $\gamma = 2^{-5}$, $C = 32$

	ANG	BOR	FEA	HAP	NEU	SAD
ANG	**77.50**	2.50	5.00	12.50	0.00	2.50
BOR	0.00	**67.50**	12.50	0.00	10.00	10.00
FEA	12.50	2.50	**70.00**	0.00	2.50	12.50
HAP	17.50	0.00	2.50	**80.00**	0.00	0.00
NEU	0.00	7.50	0.00	0.00	**87.50**	5.00
SAD	0.00	12.50	7.50	0.00	10.00	**70.00**

In EMO-DB, the most well-predicted emotion was sadness. Misclassification of happiness was the greatest, as 21.13% of the samples were classified as anger. This led to happiness having the worst recognition rate in the database. Although the misclassification rate is high, it is similar in Tables 4, 5 and 7. One of the more interesting results in EMO-DB is the classification of disgust. For both SAVEE and EMO-DB, the misclassifications are spread out almost evenly across other emotions. It is possible that disgust is hard to express through speech.

Since in [62] detailed data on human listeners' accuracy provided, we are able to compare our results to humans in Table 6. Looking at Table 6, happiness is poorly predicted compared to humans. For boredom, disgust, fear and neutral, the improvements are between 3.01 and 4.21%. In addition, sadness was very well predicted in our work, achieving a 17.69% increase compared to humans.

PESD is the smallest database (240) used in this work; it has two times less samples than SAVEE (480) and over 10 times less samples than GEES (2760). Average accuracy of 75.42% was reached despite the size of the database. Boredom was the worst performer, followed closely by fear and sadness. It seems that the small feature list is incapable of correctly classifying

Table 8
Row – the true label, column – the predicted label. Confusion matrix for GEES. Average accuracy: **93.41%**. Kernel parameters: $\gamma = 2^{-3}$, $C = 32$

	ANG	FEA	HAP	NEU	SAD
ANG	**91.85**	0.18	7.61	0.36	0.00
FEA	1.45	**95.47**	1.27	1.09	0.72
NEU	9.78	1.45	**87.86**	0.91	0.00
HAP	0.18	0.72	0.18	**96.56**	2.36
SAD	0.00	0.54	0.00	4.17	**95.29**

boredom, although a similar problem was not seen in EMO-DB, which is the only other database in this work that contains has boredom.

GEES is the largest of the databases and the one with the greatest average recognition rate: 93.41%. In addition, a large group of people (30) were used for the validation of the database. In [64], the confusion and performance of human listeners is described in detail, making it a good benchmark to compare machine learning classifiers to humans. What is more, on average 95% of the emotions were correctly classified in the database's validation, which is a testament to the good performance of the speakers and/or the listeners.

We compare our results to humans in Table 9. Some improvements can be seen compared to human listeners except for anger, where we underperform human listeners. According to Table 9, other results are similar to human listeners'.

Table 9
Comparing our results with human listeners' as described in [64]. Improvements are in **bold**

GEES	The proposed method	Humans	Difference
Anger	91.85	96.06	−4.21
Fear	95.47	93.33	**2.14**
Happiness	87.86	88.95	−1.09
Neutral	96.56	94.67	**1.89**
Sadness	95.29	96.04	−0.75

In this work, classification accuracy was improved while keeping the number of features low. The results are comparable to state-of-the-art results for two databases (EMO-DB and GEES). For GEES, we used approximately 75 times less features while achieving 98.7% of the state-of-the-art accuracy obtained in [39]. For EMO-DB, we achieved approximately 96% of the state-of-the-art accuracy with a similar reduction in the number of features used.

Results of similar works are shown in Table 10. These works used SVMs as the classifier and at least one database used in these works is also used in this paper. The results being compared are average emotion recognition rates over all labels and samples of the specified database. For SAVEE and PESD, approximately 1.7 times less features were used while improving average accuracy by 4.13% and 3.92%, respectively. For EMO-DB, almost 75 times less features were used to achieve an average accuracy of 88.6%, which is 3.7% less than state-of-the-art. However, the massive reduction in the number of features shortens training and kernel parameter search times. In addition, less computational resources are needed for feature extraction, scaling, training and testing. For the GEES database, the number of features is also approximately 75 times smaller, although the difference of our work compared to state-of-the-art is considerably smaller: only 0.89%.

Table 10
Comparison with similar works. The results of the current work are **bold** and have a dot in the reference column

Reference	Database	Labels	Features	Accuracy
.	**SAVEE**	**7**	**87**	**80.21**
[43]	SAVEE	7	566	73.81
[41]	SAVEE	7	153	76.08
.	**EMO-DB**	**7**	**87**	**88.6**
[43]	EMO-DB	7	566	82.9
[39]	EMO-DB	7	6553	92.3
[42]	EMO-DB	7	4368	86.1
[42]	EMO-DB	7	180	81.1
[41]	EMO-DB	7	153	85.13
.	**PESD**	**6**	**87**	**75.42**
[43]	PESD	6	566	71.3
.	**GEES**	**5**	**87**	**93.41**
[39]	GEES	5	6553	94.6
[40]	GEES	5	318	90.63
[40]	GEES	5	162	90.96
[38]	GEES	5	318	89.7

4.3. Online system. For males, anger (25% SAVEE', 33% EMO-DB') and neutral (92% SAVEE', 100% EMO-DB) were well recognized in both models. Fear was recognized much better in SAVEE' (50%) than in EMO-DB' (8%). Disgust, happiness and sadness were recognized almost twice as often in SAVEE' (17%) than in EMO-DB' (8%).

For females, EMO-DB' showed far better results than SAVEE'. The 100% fear recognition rate in SAVEE' is very likely a result of the system's misclassification, given the performance on all other emotions (only 8% for sadness, 0% for others). In EMO-DB', happiness is recognized 100% of the time, neutral 42% of the time.

The average emotion recognition rate for men (36%) is twice as high as for females (18%) in SAVEE'. In EMO-DB', the average recognition rate for men (28%) is very close to female recognition rate (29%). We can see that the average emotion recognition rate changes with the constitution of the training data: with female samples in EMO-DB', the performance for female speakers was significantly improved. Furthermore, with less male training samples, the performance for male speakers was lower. The average emotion recognition rate for both models is similar: 27% for SAVEE', 28% for EMO-DB' (Fig. 3).

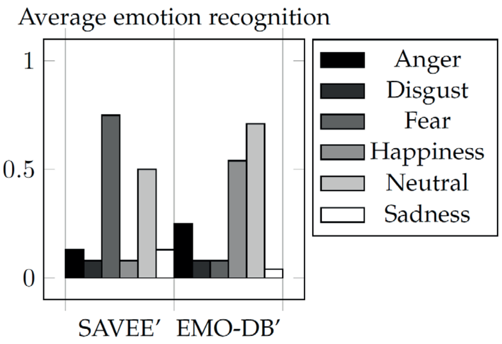

Fig. 3. Comparing emotion recognition rates between both models

If we observe how the predictions for both databases were distributed for both genders in Fig. 4, we can see that having less samples increases misclassification. SAVEE' male speaker distribution is the closest one to a uniform distribution, but the shape changes significantly for male speakers in EMO-DB'. The same can be observed for female speakers: having more female samples for training decreases misclassifications.

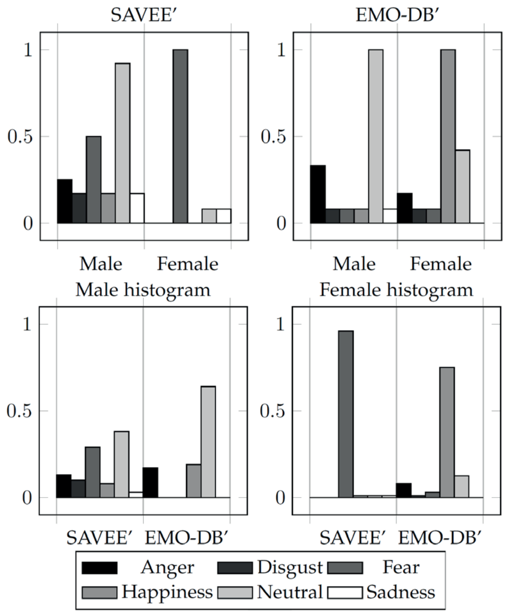

Fig. 4. Online system results. The top two charts show average emotion recognition accuracy in both models. The bottom charts show normalized distribution of predictions for males and females

These results show how having no samples from one gender affects the emotion recognition outcome. In addition, it seems that the features perform equally well for both genders (EMO-DB'). This suggests having female samples for training significantly improves SER rates for female speakers. The online system results are demonstrated in Table 11.

Table 11
The average emotion recognition rates for the online system

	SAVEE'			EMO-DB'		
	Male	Female	Both	Male	Female	Both
ANG	25.00	0.00	12.50	33.33	16.67	25.00
DIS	16.67	0.00	8.33	8.33	8.33	8.33
FEA	50.00	100.00	75.00	8.33	8.33	8.33
HAP	16.67	0.00	8.33	8.33	100.00	54.17
NEU	91.67	8.33	50.00	100.00	41.67	70.83
SAD	16.67	8.33	12.50	8.33	0.00	4.17
AVG	**36.11**	**19.44**	**27.78**	**27.78**	**29.17**	**28.47**

5. Discussion

5.1. Offline system. The main purpose of this work, as mentioned in the introduction, was to find universally well-performing features for SER. The small feature set provides good performance on all four databases, nearly matching or exceeding state-of-the-art performance. This suggests that the proposed feature set works well with any of the four different languages. Universal performance should be further tested by using this feature set on as many ESDs as possible, ideally on all the available ones, as this is the strongest possible way to show universality. However, almost matching or exceeding state-of-the-art performance on four databases, each in a different language, is a solid starting point and encourages further investigation. This also demonstrates that the features are not language- and gender-independent. The second claim is also confirmed and reflected in the online system's results. To improve performance, features capable of strong discrimination between anger and happiness can be appended to the presented features.

For GEES, the performance is comparable to human listeners as shown in Table 9. The same applies for EMO-DB in Table 6, with the exception of happiness, where we fall behind, and sadness, where we outperform human listeners. The average recognition rate is not overwhelmingly better for the machine learning model in EMO-DB, although the sadness recognition rate of 98.39% stands out. Given the recognition rate of humans (80.7%) it is hard to distinguish the thin line between true emotion detection and arbitrarily assigned label detection. One alternative to arbitrary labels would be per-sample distributions. These distributions can be acquired during database validation with little extra effort since the database creators already collect this data to show average recognition rates for emotions. Furthermore, distributions enable researchers to include more complex emotions in the databases (e.g. sad and angry). Finally, validation is often done by more people than were involved in database creation. The actors and database creators can also be involved in the voting process. This allows for a bigger voting pool which leads to more accurate descriptions of the samples and enables us to see emotions on a non-discrete, continuous scale.

5.2. Online system. The online system shows, first and foremost, that the features proposed in this method can be used for emotion recognition in real-world applications. Although the ESDs used for training the models were recorded in acoustically controlled environments, they were still capable of emotion recognition. This can suggest that the trained model actually learned something general, because the use of a laptop microphone in a regular room will certainly be reflected in the extracted features, making the recordings sound significantly different from the samples used for training. This, in addition to the offline system's good performance on multiple different ESDs, can be a strong indicator of universality of the proposed features. The online system also demonstrates the importance of having training samples from both genders. It is shown by the significant improvements in emotion recognition for women when female samples are available for training. If the system is trained only on male samples, the output for female speakers is almost always fear. This can be explained by the acoustic differences of male and female voices (e.g. female voices having a higher pitch) and differences between training and testing voices. In addition, it can be caused by the way male speakers expressed fear during the recording of the database. If, for example, the speakers spoke quietly and in a high-pitched voice, then the system has no way of discriminating between female and frightened male speakers. This is important to note when designing real-life SER systems. Usually ESDs are recorded with high quality audio equipment. What is more, the speakers were native speakers. Despite this, the online system was able to show signs of emotion detection even with commodity hardware (a common laptop microphone) and non-native speakers. Moreover, the fact that EMO-DB' was able to recognize emotions from sentences spoken in English instead of German demonstrates the systems capability for language independent emotion recognition. Finally, the online system results show that SER is not just for research in an isolated lab environment. While increasing raw emotion recognition rates for ESDs is desirable, it usually does not show how well the features work in the real world. Unless the recording environment is very controlled, the input to the system will be affected by noise, the acoustic properties of the environment etc. The online system demonstrated that the feature set proposed in this paper is capable of emotion recognition not only for ideal databases, but for the noisy real-life environment as well. It would benefit future SER research if similar real-life performance testing would be implemented. Standardization would be required to make these results comparable. The additional testing would add a new dimension to SER research and perhaps help the field evolve in a new direction.

6. Conclusion

In this paper, a small set of features showed competitive performance in speech-based emotion recognition (SER). Better than or close to state-of-the-art performance was seen in all tested emotional speech databases (ESD), demonstrating good performance in four different languages. The proposed features and the provided results look promising, but testing on other ESDs is required to better assess the universality of the features proposed in this method. However, the initial results look promising and warrant further investigation. To improve SER research, using class distributions instead of discrete labels was proposed.

The online experiment demonstrated the importance of gender in SER. Poor performance was seen when a gender was not represented in the training data. The opposite was also true: more samples for a particular gender increased SER for that gender. In addition, the online experiment demonstrated the viability of using these features in real-world environments, showing the robustness of the feature set.

In future works, we would like to do further testing on other ESDs to assess the quality of the suggested features. Second, we would like to improve our online system by incorporating more models and arranging them in a voting configuration in order to produce a robust SER system that can work on commodity hardware. Finally, we wish to study how to improve discrimination between anger and happiness due to the universal nature of the problem in the field of SER for machines and humans.

Acknowledgements. This work has been partially supported by Estonian Research Council Grants (PUT638), The Scientific and Technological Research Council of Turkey (TÜBİTAK) (Proje 1001 - 116E097) , the Estonian Centre of Excellence in IT (EXCITE) funded by the European Regional Development Fund and the European Network on Integrating Vision and Language (iV&L Net) ICT COST Action IC1307.

References

[1] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Fusion of classifier predictions for audio-visual emotion recognition", in *Pattern Recognition (ICPR), 2016 23rd International Conference on*. IEEE, 61–66, 2016.

[2] D. Kamińska, T. Sapiński, and G. Anbarjafari, "Efficiency of chosen speech descriptors in relation to emotion recognition", *EURASIP Journal on Audio, Speech, and Music Processing* 2017 (1), 3 (2017).

[3] F. Noroozi, T. Sapiński, D. Kamińska, and G. Anbarjafari, "Vocal-based emotion recognition using random forests and decision tree", *International Journal of Speech Technology* 20 (2), 239–246 (2017).

[4] J. Guo, Z. Lei, J. Wan, E. Avots, N. Hajarolasvadi, B. Knyazev, A. Kuharenko, J.C. Jacques, X. Baró, H. Demirel et al., "Dominant and complementary emotion recognition from still images of faces", *IEEE Access*, 2018.

[5] R.E. Haamer, K. Kulkarni, N. Imanpour, M.A. Haque, E. Avots, M. Breisch, K. Nasrollahi, S.E. Guerrero, C. Ozcinar, X. Baro et al., "Changes in facial expression as biometric: a database and benchmarks of identification", in *IEEE Conf. on Automatic Face and Gesture Recognition Workshops*. IEEE, 2018.

[6] A. Jaimes and N. Sebe, "Multimodal human–computer interaction: A survey", *Computer vision and image understanding*, 108 (1-2), 116–134 (2007).

[7] N. Kamaruddin and A. Wahab, "Heterogeneous driver behavior state recognition using speech signal", in *Proceedings of the 10th WSEAS international conference on System science and simulation in engineering*, 207–212, 2011.

[8] A. Tawari and M. Trivedi, "Speech based emotion classification framework for driver assistance system", in *Intelligent Vehicles Symposium (IV), 2010* IEEE, 174–178, 2010.

[9] M. Grimm, K. Kroschel, H. Harris, C. Nass, B. Schuller, G. Rigoll, and T. Moosmayr, "On the necessity and feasibility of detecting a driver's emotional state while driving", in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 126–138, 2007.

[10] A. Tickle, S. Raghu, and M. Elshaw, "Emotional recognition from the speech signal for a virtual education agent", in *Journal of Physics: Conference Series* 450 (1), 012053, IOP Publishing, 2013.

[11] M. Gong and Q. Luo, "Speech emotion recognition in web based education", in *Grey Systems and Intelligent Services, 2007. GSIS 2007. IEEE International Conference on* IEEE, 1082–1086, 2007.

[12] W. Li, Y. Zhang, and Y. Fu, "Speech emotion recognition in e-learning system based on affective computing", in *Natural Computation, 2007. ICNC 2007. Third International Conference on* IEEE, 5, 809–813, 2007.

[13] F.-M. Lee, L.-H. Li, and R.-Y. Huang, "Recognizing low/high anger in speech for call centers", in *International Conference on Signal Processing, Robotics and Automation*, 171–176, 2008.

[14] V. Petrushin, "Emotion in speech: Recognition and application to call centers", in *Proceedings of Artificial Neural Networks in Engineering* 710, 1999.

[15] D. Morrison, R. Wang, and L.C. De Silva, "Ensemble methods for spoken emotion recognition in call-centres", *Speech communication* 49 (2), 98–112 (2007).

[16] M. Szwoch and W. Szwoch, "Emotion recognition for affect aware video games", in *Image Processing & Communications Challenges* 6, 227–236, Springer, 2015.

[17] J. Torous, R. Friedman, and M. Keshavan, "Smartphone ownership and interest in mobile applications to monitor symptoms of mental health conditions", *JMIR mHealth and uHealth* 2 (1), e2 (2014).

[18] M.S. Hossain and G. Muhammad, "Cloud-assisted speech and face recognition framework for health monitoring", *Mobile Networks and Applications* 20 (3), 391–399 (2015).

[19] M.R. Kandalaft, N. Didehbani, D.C. Krawczyk, T.T. Allen, and S.B. Chapman, "Virtual reality social cognition training for young adults with high-functioning autism", *Journal of autism and developmental disorders* 43 (1), 34–44 (2013).

[20] I. Lüsi and G. Anbarjafari, "Mimicking speaker's lip movement on a 3d head model using cosine function fitting", *Bulletin of the Polish Academy of Sciences Technical Sciences* 65 (5), 733–739 (2017).

[21] G. Anbarjafari, R.E. Haamer, I. Lusi, T. Tikk, and L. Valgma, "3D face reconstruction with region based best fit blending using mobile phone for virtual reality based social media", *Bulletin of the Polish Academy of Sciences Technical Sciences* 67 (1), 125–132 (2019).

[22] J. Gorbova, I. Lüsi, A. Litvin, and G. Anbarjafari, "Automated screening of job candidate based on multimodal video processing", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 29–35, 2017.

[23] D. Kamińska and A. Pelikant, "Recognition of human emotion from a speech signal based on plutchik's model", *International Journal of Electronics and Telecommunications* 58 (2), 165–170 (2012).

[24] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011", *Artificial Intelligence Review* 43 (2), 155–177 (2015).

[25] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden markov models", *Speech Communication* 41 (4), 603–623 (2003).

[26] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition", in *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on* IEEE, 2, 401–401 (2003).

[27] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion recognition based on phoneme classes", in *Eighth International Conference on Spoken Language Processing*, 2004.

[28] K. Rychlicki-Kicior and B. Stasiak, "Multipitch estimation using judge-based model", *Bulletin of the Polish Academy of Sciences Technical Sciences* 62 (4), 751–757 (2014).

[29] F. Noroozi, D. Kaminska, T. Sapinski, and G. Anbarjafari, "Supervised vocal-based emotion recognition using multiclass support vector machine, random forests, and adaboost", *Journal of the Audio Engineering Society* 65 (7/8), 562–572 (2017).

[30] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous et al., "The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals", in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[31] M. Lugger and B. Yang, "An incremental analysis of different feature groups in speaker independent emotion recognition", in *16th Int. congress of phonetic sciences*, 2007.

[32] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing", in *International Conference on Affective Computing and Intelligent Interaction*, 139–147, Springer, 2007.

[33] C. Yang, L. Ji, and G. Liu, "Study to speech emotion recognition based on twinssvm", in *Natural Computation, 2009. ICNC'09. Fifth International Conference on* IEEE, 312–316, 2009.

[34] Y. Jin, Y. Zhao, C. Huang, and L. Zhao, "Study on the emotion recognition of whispered speech", in *Intelligent Systems, 2009. GCIS'09. WRI Global Congress on* IEEE, 3, 242–246, 2009.

[35] Y. Zhou, Y. Sun, L. Yang, and Y. Yan, "Applying articulatory features to speech emotion recognition", in *Research Challenges in Computer Science, 2009. ICRCCS'09. International Conference on* IEEE, 73–76, 2009.

[36] T.-L. Pao, W.-Y. Liao, Y.-T. Chen, J.-H. Yeh, Y.-M. Cheng, and C.S. Chien, "Comparison of several classifiers for emotion recognition from noisy mandarin speech", in *Intelligent Information Hiding and Multimedia Signal Processing, 2007. IIHMSP 2007. Third International Conference on* IEEE, 1, 23–26, 2007.

[37] X. Mao, L. Chen, and L. Fu, "Multi-level speech emotion recognition based on hmm and ann", in *Computer Science and Information Engineering, 2009 WRI World Congress on* IEEE, 7, 225–229, 2009.

[38] A. Shaukat and K. Chen, "Towards automatic emotional state categorization from speech signals", in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[39] A. Hassan and R.I. Damper, "Multi-class and hierarchical svms for emotion recognition", in *Proc. Interspeech*, 2010.

[40] A. Shaukat and K. Chen, *Emotional state categorization from speech: machine vs. human*, arXiv preprint arXiv:1009.0108, 2010.

[41] V. Kobayashi and V. Calag, “Detection of affective states from speech signals using ensembles of classifiers”, in *IET Intelligent Signal Processing Conference*, 2013.

[42] B.-C. Chiou and C.-P. Chen, “Feature space dimension reduction in speech emotion recognition using support vector machine”, in *Signal and information processing association annual summit and conference (APSIPA), 2013 Asia-Pacific* IEEE, 1–6, 2013.

[43] E. Yüncü, H. Hacihabiboglu, and C. Bozsahin, “Automatic speech emotion recognition using auditory models with binary decision tree and svm”, in *Pattern Recognition (ICPR), 2014 22nd International Conference on* IEEE, 773–778, 2014.

[44] P. Boersma and D. Weenink, “Praat: doing phonetics by computer (version 6.0.36) [computer program], retrieved january 1, 2018”, 2018.

[45] C.-W. Hsu, C.-C. Chang, C.-J. Lin *et al.*, *A practical guide to support vector classification*, 2003.

[46] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines”, *ACM Transactions on Intelligent Systems and Technology* 2, 27:1–27:27 (2011), software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[47] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python”, *Journal of Machine Learning Research* 12, 2825–2830 (2011).

[48] P. Patel, A. Chaudhari, R. Kale, and M. Pund, “Emotion recognition from speech with gaussian mixture models & via boosted gmm”, *International Journal of Research in Science & Engineering*, 3 (2017).

[49] S.R. Krothapalli and S.G. Koolagudi, “Speech emotion recognition: a review”, in *Emotion Recognition using Speech Features*, 15–34, Springer, 2013.

[50] D. Gharavian, M. Bejani, and M. Sheikhan, “Audio-visual emotion recognition using fcbf feature selection method and particle swarm optimization for fuzzy artmap neural networks”, *Multimedia Tools and Applications* 76 (2), 2331–2352 (2017).

[51] J. Deng, X. Xu, Z. Zhang, S. Frühholz, D. Grandjean, and B. Schuller, “Fisher kernels on phase-based features for speech emotion recognition”, in *Dialogues with Social Robots*, 195–203, Springer, 2017.

[52] N. Yang, J. Yuan, Y. Zhou, I. Demirkol, Z. Duan, W. Heinzelman, and M. Sturge-Apple, “Enhanced multiclass svm with thresholding fusion for speech-based emotion classification”, *International Journal of Speech Technology* 20 (1), 27–41 (2017).

[53] Z.-T. Liu, M. Wu, W.-H. Cao, J.-W. Mao, J.-P. Xu, and G.-Z. Tan, “Speech emotion recognition based on feature selection and extreme learning machine decision tree”, *Neurocomputing*, 2017.

[54] F. Noroozi, N. Akrami, and G. Anbarjafari, “Speech-based emotion recognition and next reaction prediction”, in *Signal Processing and Communications Applications Conference (SIU), 2017 25th*, 1–4, IEEE, 2017.

[55] F. Eibe, M. Hall, I. Witten, and J. Pal, “The weka workbench”, *Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques*, 4, 2016.

[56] L. Breiman, “Random forests”, *Machine learning* 45 (1), 5–32 (2001).

[57] J. Gama, R. Rocha, and P. Medas, “Accurate decision trees for mining high-speed data streams”, in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 523–528, ACM, 2003.

[58] L. Breiman, “Bias, variance, and arcing classifiers”, 1996.

[59] M.-L. Zhang and Z.-H. Zhou, “A review on multi-label learning algorithms”, *IEEE Transactions on Knowledge and Data Engineering* 26 (8), 1819–1837 (2014).

[60] G. Louppe, *Understanding random forests: From theory to practice*, arXiv preprint arXiv:1407.7502, 2014.

[61] P. Jackson and S. Haq, *Surrey audio-visual expressed emotion(savee) database*, University of Surrey: Guildford, UK, 2014.

[62] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, “A database of german emotional speech”, in *Interspeech*, 5, 1517–1520 (2005).

[63] P. Staroniewicz, “Polish emotional speech database–design”, in *Proc. of 55th Open Seminar on Acoustics, Wroclaw, Poland*, 373–378, 2008.

[64] S.T. Jovicic, Z. Kasic, M. Dordevic, and M. Rajkovic, “Serbian emotional speech database: design, processing and evaluation”, in *9th Conference Speech and Computer*, 2004.