DOI 10.1007/s10772-017-9396-2

Vocal-based emotion recognition using random forests and decision tree

Fatemeh Noroozi[1] · **Tomasz Sapiński**[2] · **Dorota Kamińska**[2] · **Gholamreza Anbarjafari**[3,4]

Received: 29 June 2016 / Accepted: 5 January 2017 / Published online: 9 February 2017

Abstract This paper proposes a new vocal-based emotion recognition method using random forests, where pairs of the features on the whole speech signal, namely, pitch, intensity, the first four formants, the first four formants bandwidths, mean autocorrelation, mean noise-to-harmonics ratio and standard deviation, are used in order to recognize the emotional state of a speaker. The proposed technique adopts random forests to represent the speech signals, along with the decision-trees approach, in order to classify them into different categories. The emotions are broadly categorised into the six groups, which are happiness, fear, sadness, neutral, surprise, and disgust. The Surrey Audio-Visual Expressed Emotion database is used. According to the experimental results using leave-one-out cross-validation, by means of combining the most significant prosodic features, the proposed method has an average recognition rate of 66.28%, and at the highest level, the recognition rate of 78% has been obtained, which belongs to the happiness voice signals. The proposed method has 13.78% higher average recognition rate and 28.1% higher best recognition rate compared to the linear discriminant analysis as well as 6.58% higher average recognition rate than the deep neural networks results, both of which have been implemented on the same database.

Keywords Vocal emotion recognition · Human–computer interaction · Random forests · Decision tree classifier

1 Introduction

In the existing literature, audiovisual human emotion recognition has been dealt with by means of numerous combinations of perceptions and features (Zeng et al. 2007; Anbarjafari and Aabloo 2014; Burget et al. 2011). Computer-based vocal emotion recognition has been under investigation and research for decades, and tends to attract more attention from scientists and engineers, being influenced by the development of artificial intelligence (Scherer 2013; Scherer et al. 2015; Kamińska and Pelikant 2012; El Ayadi et al. 2011; Cowie et al. 2001; Lüsi et al. 2016; Koolagudi and Rao 2012). It is aimed at inferring the emotional state of the speaker on the basis of their speech signal.

The capability of recognizing human emotions plays an essential role in many human-robot interaction scenarios (Zhang et al. 2013; Vogt et al. 2008; Anbarjafari and Aabloo 2014). An effective vocal emotion recognition system will help rendering the interaction between human and robot more naturalistic and user-friendly (Esposito et al. 2015; Rabiei and Gasparetto 2014; Bellantonio et al. 2016).

✉ Gholamreza Anbarjafari
shb@icv.tuit.ut.ee

Fatemeh Noroozi
fatemeh.noroozi@ut.ee

Tomasz Sapiński
tomasz.sapinski@p.lodz.pl

Dorota Kamińska
dorota.kaminska@p.lodz.pl

[1] Institute of Technology, University of Tartu, Nooruse 1, 50411 Tartu, Estonia

[2] Institute of Mechatronics and Information Systems, Łodz University of Technology, Lodz, Poland

[3] iCV Research Group, Institute of Technology, University of Tartu, Nooruse 1, 50411 Tartu, Estonia

[4] Department of Electrical and Electronic Engineering, Hasan Kalyoncu University, Gazinatep, Turkey

It has numerous applications in many fields such as business, entertainment (El Ayadi et al. 2011), call centre environments (Petrushin 2000), games (Bahreini et al. 2013) and education (Cowie et al. 2001). Besides, the correct perception of the emotions can encourage the timely detection of diseases that are known to affect the expressions of the feelings (Yoon and Park 2007; Petrushin 2000; Zeng et al. 2007).

Previous works on vocal emotion recognition have suggested and implemented various algorithms, in order to figure out the best possible solution to this problem (Palm and Glodek 2013), including the case approaching it from a multi-class perspective. Multi-class or multinomial classifiers are algorithms with the ability of separating more than two classes. Multi-class classification approaches are of two sorts, namely, single- and multi-labelled, which are distinguished from each other based on whether they employ a binary logic or not. Single-labelled classifiers make use of the induction rule (Liu and Motoda 2007).

Emotion recognition based on speech is reported to be rather simple for human beings, to some extend, since they possess a natural talent to analyse such signals (Ingale and Chaudhari 2012). However, it may be significantly challenging to a machine. For example, Devillers and Vidrascu (2006) deals with processing real-world speech data collected at a call centre. In Devillers and Vidrascu (2006), it is stated that one of the fundamental problems is to distinguish the components of the speech signal related to a certain emotion from the ones naturally produced as a part of the conversation. It is also stated that various “linguistic” and “paralinguistic” features are present in any speech signal. Paralinguistic features can be extracted by signal processing techniques. They are not dependent on the content of the words, but they contain emotions. Amongst the paralinguistic features, “prosodic” ones are the most common (Devillers et al. 2005).

Many classification techniques such as support vector machine (SVM) (Schuller et al. 2007; Vlasenko et al. 2007), Gaussian mixture model (GMM) (Neiberg et al. 2006; Pribil and Pribilova 2013), and meta decision tree (MTD) (Borchert and Dusterhoft 2005; Anagnostopoulos et al. 2000) have been adopted for vocal emotion recognition. Nwe et al. proposed a text-independent method for speech-based emotion recognition, which benefits from short-time log frequency power coefficients (LFPC) for representing the speech signals, along with a discrete hidden markov model (HMM) as the classifier (Nwe et al. 2003). Atassi et al., on the other hand, performed an analysis on high-level features for vocal emotion recognition (Atassi et al. 2011). Besides, Wu and Liang employed GMM, SVM and multilayer perceptron (MLP), to model the acoustic-prosodic information, based on speech features (Wu and Liang 2011).

One of the prevalent classification methods used for vocal emotion recognition in the literature is decision tree (Anagnostopoulos et al. 2000). By definition, this algorithm randomly creates a new decision-making structure at every iteration (Liu and Motoda 2007). Decision trees work based on splitting the data into two subsets. Several algorithms exist that are constructed from more than one decision tree, and are referred to as ensemble methods. Examples of the foregoing approaches include Bootstrap Aggregating (bagging) decision tree (Sebe et al. 2007), boosting tree (Sun et al. 2006), rotation forests (Rodriguez et al. 2006) and the Random Forest (RF) algorithm (Zhou 2012). RF is an ensemble method with a composite structure (Breiman 2001). It is an extension of bagging techniques, and is characterized by random selection of the features.

The fact that the RF utilizes multiple randomly generated decision trees enables it to take advantage of all the virtues of decision trees, ensemble methods and bagging approaches. The RF classifier predicts the class label of an input data by majority voting on the predictions made by a set of tree classifiers. A portion of the data is kept for the test. The rest is used for training. This is known as leave-one-out cross-validation (LOOCV) method. Multiple decision trees are constructed. Each of the decision trees is created by randomly choosing a subset of the training set, and contains a certain number of samples (Breiman 2001; Liaw and Wiener 2002). Each tree classifier divides the feature space into a number of partitions, and its output for an input data is determined according to the partition the data lies in.

While bagging methods use deterministic decision trees, where the evaluation is based on all the features, RF only evaluates a subset of them. From technical point of view, the RF decision making structure is similar to bagging algorithms. In other words, RF is a learning ensemble consisting of a bagging of un-pruned decision tree learners. However, despite bagging processes, it utilizes a randomized selection of the features at each split. Due to this property, in terms of training, RF is generally more efficient than bagging (Nordhausen 2013).

In contrast with the aforementioned previous works in the field of vocal emotion recognition, this paper, as its main contribution, makes use of the RF method, which enables the algorithm to avoid basing the decision-making process on a single decision tree. In other words, the RF brings about the capability of interconnecting the emotion recognition results. Besides, as the last module, the overall decision is made on the basis of majority voting, i.e. corresponds to the choice demonstrating the highest frequency of occurrences throughout the iterations. This will help ensure that the best possible classification choice is opted for, via implementing multi-mode, rather than single-mode, analysis. To be more clear, it will further increase the chances of accomplishing a higher recognition rate (Stiefelhagen et al. 2004).

The rest of the paper is structured as follows. First, the methodology of the proposed emotion recognition system is introduced and discussed in Sect. 2. Next, in Sect. 3, the proposed method is applied to the Surrey Audio-Visual Expressed Emotion (SAVEE) Database and the results are presented and discussed. Finally, in Sect. 4, the paper concludes through providing a summary, followed by hints to possible subjects of future studies.

2 The proposed method

The proposed vocal-based emotion recognition method applies the RF decision making algorithm to the speech signals comprising six emotion categories, namely, happiness, sad, disgust, fear, neutral and surprise. The corresponding emotion labels are then assigned to each voice signal by means of multi-class classification.

The general structure of the vocal emotion recognition method proposed in this paper is illustrated in Fig. 1. As shown in the diagram, first, the speech signal including the set of emotions to be analysed is produced. Taking into account speech signals from both genders is vital for verifying the applicability of the system in terms of reliable performance, since there are fundamental differences in their structural speech features. The foregoing features are, mainly, represented through a set of parameters, including frequency range is and tone intensity (Gorham-Rowan and Laures-Gore 2006; Puts et al. 2007).

Afterwards, the features are extracted from speech signal, using signal processing techniques. The quantitative features describing the variations of the speech signal are extracted, holding a non-linguistic feature selection viewpoint. In accordance with the non-linguistic approach, the variations of pitch and intensity are analysed while ignoring the linguistic information. The reliability of the latter quantities is affected by the voice quality, which, independently from the word identity, is related to the spectral properties (Anagnostopoulos et al. 2000). As the last stage, classification is performed. As aforementioned, for the goal of this paper, RF decision trees are adopted for the purpose of classifying and recognising the emotions.

The RF and bagging algorithms have similar structures. One of the principal functionalities of bagging is to average noisy and unbiased models, in order to reduce the variance of the whole model. It randomly generates N subsets of the original set S through replacement.

According to Breiman (2001), and using a similar notation, assuming that the dataset, $\boldsymbol{S}$, contains m signals with n features each, it could be represented as a block matrix, consisting of a block, namely, $\boldsymbol{F}$, containing the features, and an m-dimensional vector $\boldsymbol{l}$ listing the class labels, as follows:

$$\boldsymbol{S}_{(m\times n+1)} = \left[\boldsymbol{F}_{(m\times n)} \middle| \boldsymbol{l}_{(m\times 1)}\right], \tag{1}$$

where for $i = 1, \dots, m$, $j = 1, \dots, n$, $[\boldsymbol{F}]_{(i,j)}$ stands for the jth feature of the ith signal, and $[\boldsymbol{l}]_{(i)}$ denotes its class label. It could be inferred that the ith signal is tantamount to the ith row of the matrix $\boldsymbol{F}$, i.e. $\left[\boldsymbol{v}_i\right]_{(j)} = [\boldsymbol{F}]_{(i,j)}$.

In order to utilize a prescribed dataset for training the classifier, first, the value of N, i.e. the number of the subsets, should be determined arbitrarily. Every subsequent subset of the dataset will have the same structure. In other words, every voice signal will contain the same number of features as the global set, $\boldsymbol{S}$. Thus N arbitrarily structured decision trees will be developed, with randomly chosen data and variables. Since the subsets might repeate the voice signal rows, the trees might overlap with each other. Noticing that N subsets are created from the whole dataset, which, as aforementioned, contains m voice signal in total, the number of the voice signals allocated for each subset will be $\beta = \frac{m}{N}$.

At the test level, each input voice signal is searched for throughout the forest containing the decision trees exhaustively. In other words, all the N trees are considered when searching for the training voice signals possibly matching the test voice signal. Afterwards, every decision tree, from its own perspective, reports the most probable class the input voice signal belongs to, as a single vote. Finally, a class label is assigned to the input vector, based on majority voting from the forest of the trees, which will be a basis for

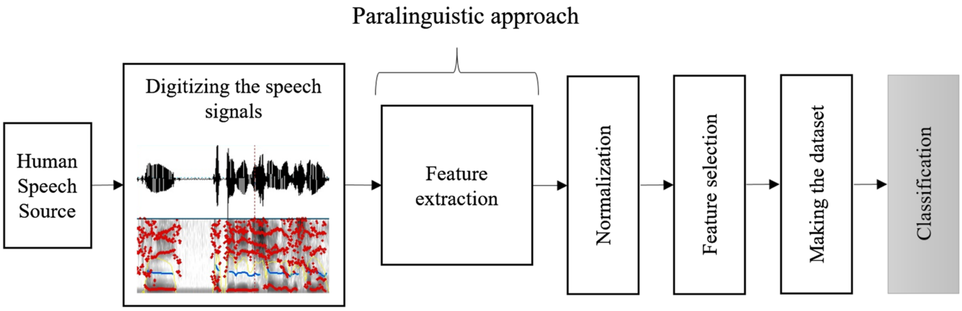

Fig. 1 Schematic presentation of the general structure of the proposed vocal emotion recognition system

predicting the class that the test voice signal is associated with. The RF classification algorithm is summarized by a pseudo-code in Algorithm 1.

Algorithm 1 A pseudo-code representing the RF classification method.

```
Require: $N$: The number of training sets, which is equal to the number of trees
Require: $\boldsymbol{S}$: The voice signals dataset
Require: $m$: The number of features in each feature vector
Require: $\boldsymbol{v}_i$, $i = 1, \ldots, m$: the $i^{\text{th}}$ voice signal
Require: $Build\ tree$: a function to construct of the trees
Ensure: $output\_label$
  for $i$=1 to $N$ do
    Pick up the voice signals from $\boldsymbol{S}$ to make the $i^{\text{th}}$ training set, $\boldsymbol{S}_i$, randomly and by replacement
    Create a root for the $\boldsymbol{S}_i$ to compare the feature values
    Make a decision tree based on $\boldsymbol{S}_i$ and the determined root nodes
    Select one of the feature vectors for the $i^{\text{th}}$ decision tree by splitting
    Choose the feature $f_i$ with the highest information gain
    while exists a test voice signal do
      Create the child nodes of the $i^{\text{th}}$ decision tree for the feature vectors
      for $i$=1 to $m$ do
        Compare the content of the nodes of the $i^{\text{th}}$ decision tree with the contents of the test feature vector
        Call "build tree" to make the rest of the tree
      end for
    end while
    Extract the emotion label from every decision tree
    Perform majority voting between the extracted emotion labels to determine the $output\_label$
  end for
```

The *k*-fold cross validation is also used in order to choose different partitions of the training and validation sets (Refaeilzadeh et al. 2009). Every time, a voice signal is considered for the test, and the rest for training. This is implemented exhaustively on every voice signal.

3 The experimental results and discussion

In this section, the proposed vocal-based emotion recognition method is implemented on the SAVEE database (Jackson and Haq 2014). In this work all the 360 voice signals of this dataset with the basic emotions are used. This case-study is aimed at verifying the efficiency and applicability of the proposed method. Six basic emotions, namely, happiness, sadness, disgust, fear, surprise and neutral, are studied. The SAVEE samples consist of audio–video (AV) clips. In this research work, for the experimental results, only the audio data is used.

The audio files have been released in "wav" format. As mentioned earlier, it is possible to use linguistic or paralinguistic feature categories. In this paper, the features have been chosen from the paralinguistic family. Overall, 95 features are extracted from every speech signal in the mentioned category. However, only 14 features from this list have rational distances within and between classes, which are used in our experiments. They are listed as pitch, which is the fundamental frequency of the speech signal, intensity, first through fourth formants and their bandwidths, mean harmonics-to-noise ratio,

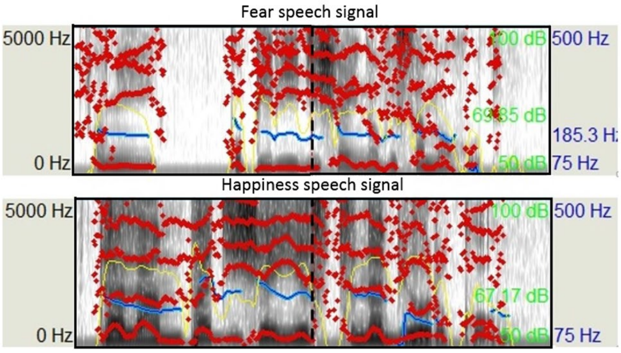

Fig. 2 Different features extracted from the speech, fear and happiness signals, using PRAAT (Boersma and Weenink 2013)

mean noise-to-harmonics ratio,standard deviation and auto-correlation.

In general, the first and second formants with the lowest frequencies are the most informative ones. However, the phonetic vowels that are used in different languages are not the same. In addition, their acoustic properties, which include formants, are different. Therefore, in order to describe the front vowels precisely, the third formant is needed as well. It is necessary for distinguishing between the vowels with similar sounds. Moreover, according to Hunter and Kebede (2012) and Deterding (1997), the third and fourth formants are important for analyzing the spectral properties of the voice in the higher magnitudes, because they are stronger in shouting than in speaking normally (Millhouse et al. 2002). Besides, the necessity of using the first four formants for precisely analyzing the vocal features can be understood by noticing their usage in the PRAAT software (Boersma and Weenink 2013).

The PRAAT software is used to extract these acoustic features from the voice signal. The mean values of the features have been computed throughout the whole duration of the speech signals, in the time domain, directly from the acoustic waveforms, for the voiced parts only. Figure 2 shows the spectrograms of fear and happiness emotion speech signals. In this figure, local minimum and maximum values have been highlighted with red points. Their frequencies range is from 0 to 5000 Hz. The speech signal might not be continuous during the whole time interval, because of the interruptions of the voice. As apparent from the latter figure, in the most cases, the speech signals representing the fear emotion demonstrate more smoothness, with more distance between the frequencies corresponding to the local minimums and maximums, till the middle of the spectrum, compared to that of the happiness. This gets reversed in the second half. This phenomenon and the variations of the duration of the speeches make differences between the emotions, and affect the recognition rate directly.

After extracting the features from every voice signal, all of them are kept in a two-by-two array. The settings for conducting the tests in the context of this study are such that for the RF classification, following the notation introduced in Sect. 2, $m = 360$ and $n = 14$.

According to the number of emotion classes taken into account, six different labels exist, each of which applies to 60 voice signals belonging to the associated class. Besides, N is set to 15. Afterwards, the following two methods are used with the RF classifier. The first approach with name of LOOCV method chooses one voice signal as the test data, and the remaining 359 are considered for training. These 359 samples are divided into 15 subsets. Then forests with 15 trees are created. Each tree has more than $\beta = \frac{m}{N} = \frac{359}{15} \simeq 24$ nodes, because of the overlaps between them. This procedure continues till all the voice signals have been used once in the test. The second approach uses the $k-$fold method. According to the Weka standard (Bouckaert et al. 2013), the best possible result is obtained when 10-fold cross validation is considered. In both of the cases,

Table 1 The results of implementing the proposed RF-based vocal emotion recognition method, considering 15 decision trees

	Happiness	Fear	Sadness	Neutral	Suprise	Disgust	Accuracy rate (%)
Happiness	**47**	5	0	0	7	1	78
Fear	8	**27**	2	1	9	13	45
Sadness	0	1	**45**	13	1	0	75
Neutral	1	0	5	**43**	0	11	73.66
Suprise	10	11	0	0	**38**	1	63
Disgust	3	3	6	15	1	**38**	63
					Average:		66.28

Table 2 Classification results using the 10-fold cross-validation algorithm

	Happiness	Fear	Sadness	Neutral	Suprise	Disgust	Accuracy rate (%)
Happiness	**40**	6	4	0	10	0	66.67
Fear	11	**14**	5	3	24	3	23.33
Sadness	3	2	**27**	15	1	12	45
Neutral	1	2	9	**40**	3	5	66.67
Suprise	19	16	1	1	**21**	2	35
Disgust	5	1	8	20	4	**22**	36.67
					Average:		45.51

the result is obtained trough majority voting on the decisions made by the trees.

The results of the recognition process using the proposed method, which is based on RF, and 10-fold classification are shown by means of confusion matrices (Townsend 1971) in Tables 1 and 2, respectively. Each row of the confusion matrix represents the recognition rate related to an emotion class , where each cell shows the number of the voice signals of that class being classified into the class specified by the column label. This means that the summation of the elements on every row should show the number of the voice signals included in the corresponding class, which, in the context of this study, is 60. More clearly, the numbers appearing on the diagonal of the confusion matrix represent the number of correctly recognised voice signals, while the rest denote the number of the misclassification instances. According to the results shown in Table 1, the average recognition rate using the proposed method with LOOCV is 66.28%. The results of using 10-fold classification are also shown in Table 2, which has an average recognition rate of 45.51%. The outcomes show that LOOCV outperforms the 10-fold classification, since both of the methods have been implemented on the same database and under the same conditions.Besides, the best accuracy rate resulted from 10-fold classification is 66.67%, which has been obtained when recognising the happiness and neutral emotions. However, a performance rate as high as 78% is accomplished for recognising the happiness emotion by LOOCV approach. The reason is that higher number of training samples results in a better training performance.

In the case of 10-fold classification, dividing the dataset to 10 parts and using one tenth of the samples as the test set reduces the number of the samples in the training set, which weakens the learning performance, and decreases the recognition rate, compared to the case where only one voice signal is considered for the test.

Moreover, the overall performance rate reported by Haq et al. (2008), which has made use of features such as pitch, intensity, formants, MFCC, jitter and Shimmer, has been 52.5% on average, with linear discriminate analysis (LDA), and in the best state, LDA has earned a recognition rate of 59.7%. These values, which have been obtained on the same dataset, i.e. SAVEE, are smaller than that of the proposed method by 13.78 and 28.1%, respectively. Fayek et al. have made use of the deep neural networks (DNN) method based on a features including pitch, energy, MFCC and teager energy operator (Fayek et al. 2015). The average recognition rate has been 59.7%, which is 6.58% less than the recognition rate achieved by the proposed method, as shown in Table 3.

Thus putting the results of implementing the proposed method in contrast with that of the foregoing studies, leads to the inference that it outperforms such state-of-the-art techniques.

According to Tables 1 and 2, feature values of happiness samples have the least within class distances and the most between class distances. Fear signals have the least recognition rate with both of the methods. They have the most similar feature values with the surprised emotion. Many surprise voice signals have been misclassified as fear. Nine fear voice signals have been misclassified as surprise, where also 24 fear voice signals are misclassified as surprise, with 10-fold classification. As a relevant conclusion to be made based on the results, structural similarities between the emotion categories, which cause recognition conflicts, could be drawn.

Similar to the relevant studies, in this work, we are finding a suitable feature vector length, and selecting an appropriate combination of the paralinguistic features. But this can be a limitation of this category of works on vocal emotion recognition, because due to the high number of the available paralinguistic features, it is not guaranteed that the optimal choice is made.

4 Conclusion

In this paper, a vocal-based emotion recognition system using RF was proposed and implemented. To this end, features such as pitch, intensity, first through fourth formants and their bandwidths, mean autocorrelation, mean harmonics-to-noise ratio, mean noise-to-harmonics ratio and standard deviation were used. Each pair of the features was considered as a criterion for classifying the voice signals using RF, which adopts the decision tress approach for recognition. The performance of the proposed technique was evaluated through applying it to the SAVEE Database. The

Table 3 Comparison of the average and best recognition rates resulted from the proposed method with LDA and DNN classification approaches reported in Haq et al. (2008) and Fayek et al. (2015), respectively

	RF with LOOCV	10-fold RF	LDA reported in Haq et al. (2008)	DNN reported in Fayek et al. (2015)
Average recognition rate	66.28	45.51	52.50	59.7
Best recognition rate	78.00	66.67	59.90	89.00

experimental results for the recognition of six well-known emotions, namely, happiness, fear, sadness, neutral, surprise and disgust, were reported. The average recognition accuracy rate was 66.28% by keeping a sample for the test and using the rest for training. The highest recognition performance, i.e. 78%, was accomplished in case of the happiness emotion. It was inferred that the feature values of the happiness emotion have the least within class distances and the most between class distances, compared to other labels, which leads to high recognition rate. Paralinguistic features such as pitch and intensity were able to distinguish between the emotions, thus they are decisive in the recognition of vocal emotions through signal processing techniques.

Since we used the powerful algorithm of Random Forest, which is accurate but time-consuming, we tried to select the feature vectors with an affordable length, i.e. 14, in order to reduce the running time, and achieve a balance between time-consumption and accuracy. The performance of the proposed method was also compared with the existing techniques, which verified its efficiency and reliability. The average recognition rate by using the proposed RF based method was higher by 13.78% than the LDA used by Haq et al., as a statistical classification method, on the same database, i.e. SAVEE, including the feature categories pitch, intensity, formants, MFCC, jitter and Shimmer. The average recognition rate obtained by the proposed method was also compared with another work using the DNN approach, by 6.58%, where the same dataset was utilized while taking features such as pitch, energy, MFCC and teager energy operator into account.

A large number of paralinguistic features exist, which are useful for describing the emotions that are included in the human voice. In the future works, one can investigate whether it is possible to improve the result of applying the method by changing the selection of the features, and including new features, such as MFCCs and filter bank energies (FBEs).

Acknowledgements This work has been partially supported by Estonian Research Grant (PUT638), the Estonian Centre of Excellence in IT (EXCITE) funded by the European Regional Development Fund, Estonian-Polish Joint Research Project and the European Network on Integrating Vision and Language (iV&L Net) ICT COST Action IC1307.

References

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

Anbarjafari, G., & Aabloo, A. (2014). Expression recognition by using facial and vocal expressions. *V&L Net*, *2014*, 103–105.

Atassi, H., Esposito, A., Smekal, Z. (2011). Analysis of high-level features for vocal emotion recognition. In 2011 34th international conference on telecommunications and signal processing (TSP) (pp. 361–366). IEEE

Bahreini, K., Nadolski, R., Westera, W. (2013). Filtwam and voice emotion recognition. In Games and learning alliance (vol. 8605, pp. 116–129). Springer.

Bellantonio, M., Haque, M. A., Rodriguez, P., Nasrollahi, K., Telve, T., Escarela, S., Gonzalez, J., Moeslund, T. B., Rasti, P., Anbarjafari, G. (2016). Spatio-temporal pain recognition in cnn-based super-resolved facial images. In International conference on pattern recognition (ICPR). Springer.

Boersma, P., & Weenink, D. (2013). *Praat software*. Amsterdam: University of Amsterdam.

Borchert, M., Dusterhoft, A. (2005). Emotions in speech-experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments. In Proceedings of 2005 IEEE international conference on natural language processing and knowledge engineering, 2005. IEEE NLP-KE'05 (pp. 147–151). IEEE.

Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., Scuse, D. (2013). Weka manual for version 3-7-8.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Burget, R., Karasek, J., & Smekal, Z. (2011). Recognition of emotions in czech newspaper headlines. *Radioengineering*, *20*(1), 39–47.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., et al. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, *18*(1), 32–80.

Deterding, D. (1997). The formants of monophthong vowels in standard southern british english pronunciation. *Journal of the International Phonetic Association*, *27*(1–2), 47–55.

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

Devillers, L., Vidrascu, L., & Lamel, L. (2005). Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, *18*(4), 407–422.

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

Esposito, A., Esposito, A. M., & Vogel, C. (2015). Needs and challenges in human computer interaction for processing social emotional information. *Pattern Recognition Letters*, *66*, 41–51.

Fayek, H., Lech, M., Cavedon, L. (2015). Towards real-time speech emotion recognition using deep neural networks. In 2015 9th international conference on signal processing and communication systems (ICSPCS) (pp. 1–5). IEEE.

Gorham-Rowan, M. M., & Laures-Gore, J. (2006). Acoustic-perceptual correlates of voice quality in elderly men and women. *Journal of communication disorders*, *39*(3), 171–184.

Haq, S., Jackson, P. J., Edge, J. (2008). Audio-visual feature selection and reduction for emotion classification. In Proceedings of international conference on auditory-visual speech processing (AVSP), Tangalooma, Australia (2008)

Hunter, G., Kebede, H. (2012). Formant frequencies of British English vowels produced by native speakers of farsi. In Acoustics 2012

Ingale, A. B., & Chaudhari, D. (2012). Speech emotion recognition. *International Journal of Soft Computing and Engineering (IJSCE)*, *2*(1), 235–238.

Jackson, P., Haq, S. (2014). Surrey audio-visual expressed emotion(savee) database.

Kamińska, D., & 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.

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: A review. *International Journal of Speech Technology*, *15*(2), 99–117.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. *R News*, *2*(3), 18–22.

Liu, H., & Motoda, H. (2007). *Computational methods of feature selection*. Boca Raton: CRC Press.

Lüsi, I., Escarela, S., Anbarjafari, G. (2016). Sase: Rgb-depth database for human head pose estimation. In Computer vision–ECCV 2016 workshops (pp. 325–336). Springer

Millhouse, T., Clermont, F., Davis, P. (2002). Exploring the importance of formant bandwidths in the production of the singer's formant. In Proceedings of the 9th Australian SST (pp. 373–378).

Neiberg, D., Elenius, K., Laskowski, K. (2006). Emotion recognition in spontaneous speech using gmms. In Interspeech (pp. 809–812)

Nordhausen, K. (2013). Ensemble methods: Foundations and algorithms by Zhi-Hua Zhou. *International Statistical Review*, *81*(3), 470–470.

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

Palm, G., Glodek, M. (2013). Towards emotion recognition in human computer interaction. In Neural nets and surroundings (vol. 19, pp. 323–336). Springer.

Petrushin, V. A. (2000). Emotion recognition in speech signal: experimental study, development, and application. *Studies*, *3*, 222–225.

Pribil, J., & Pribilova, A. (2013). Determination of formant features in czech and slovak for gmm emotional speech classifier. *Radioengineering*, *22*(1), 52–59.

Puts, D. A., Hodges, C. R., Cárdenas, R. A., & Gaulin, S. J. (2007). Men's voices as dominance signals: Vocal fundamental and formant frequencies influence dominance attributions among men. *Evolution and Human Behavior*, *28*(5), 340–344.

Rabiei, M., Gasparetto, A. (2014). A system for feature classification of emotions based on speech analysis; applications to human-robot interaction. In 2014 second RSI/ISM international conference on robotics and mechatronics (ICRoM) (pp. 795–800). IEEE

Refaeilzadeh, P., Tang, L., Liu, H. (2009). Cross-validation. In Encyclopedia of database systems (pp. 532–538). Springer (2009)

Rodriguez, J. J., Kuncheva, L. I., & Alonso, C. J. (2006). Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(10), 1619–1630.

Scherer, K. R. (2013). Vocal markers of emotion: Comparing induction and acting elicitation. *Computer Speech & Language*, *27*(1), 40–58.

Scherer, K. R., Sundberg, J., Tamarit, L., & Salomão, G. L. (2015). Comparing the acoustic expression of emotion in the speaking and the singing voice. *Computer Speech & Language*, *29*(1), 218–235.

Schuller, B., Seppi, D., Batliner, A., Maier, A., Steidl, S. (2007). Towards more reality in the recognition of emotional speech. In IEEE international conference on Acoustics, speech and signal processing, 2007. ICASSP 2007 (vol. 4, pp. IV–941). IEEE.

Sebe, N., Lew, M. S., Sun, Y., Cohen, I., Gevers, T., & Huang, T. S. (2007). Authentic facial expression analysis. *Image and Vision Computing*, *25*(12), 1856–1863.

Stiefelhagen, R., Fügen, C., Gieselmann, P., Holzapfel, H., Nickel, K., Waibel, A. (2004). Natural human-robot interaction using speech, head pose and gestures. In 2004 IEEE/RSJ international conference on intelligent robots and systems, 2004 (IROS 2004). Proceedings (vol. 3, pp. 2422–2427). IEEE.

Sun, N., Zheng, W., Sun, C., Zou, C., Zhao, L. (2006). Facial expression recognition based on boostingtree. In Advances in neural networks-ISNN 2006 (pp 77–84). Springer.

Townsend, J. T. (1971). Theoretical analysis of an alphabetic confusion matrix. *Perception & Psychophysics*, *9*(1), 40–50.

Vlasenko, B., Schuller, B., Wendemuth, A., Rigoll, G. (2007). Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing. In Affective computing and intelligent interaction (pp. 139–147). Springer.

Vogt, T., André, E., Wagner, J. (2008). Automatic recognition of emotions from speech: A review of the literature and recommendations for practical realisation. In Affect and emotion in human-computer interaction (vol. 4868, pp. 75–91). Springer.

Wu, C. H., & Liang, W. B. (2011). Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing*, *2*(1), 10–21.

Yoon, W. J., Park, K. S. (2007). A study of emotion recognition and its applications. In: Modeling decisions for artificial intelligence (pp. 455–462). Springer.

Zeng, Z., Hu, Y., Roisman, G. I., Wen, Z., Fu, Y., Huang, T. S. (2007). Audio-visual spontaneous emotion recognition. In Artifical intelligence for human computing (pp. 72–90). Springer.

Zhang, S., Zhao, X., Lei, B. (2013). Speech emotion recognition using an enhanced kernel isomap for human-robot interaction. International Journal of Advanced Robotic Systems. doi:10.5772/55403.

Zhou, Z. H. (2012). *Ensemble methods: Foundations and algorithms*. Boca Raton: CRC Press.