Improving Emotion Recognition Performance by Random-Forest-Based Feature Selection

Olga Egorow(✉), Ingo Siegert, and Andreas Wendemuth

Cognitive Systems Group, Otto von Guericke University, 39016 Magdeburg, Germany
olga.egorow@ovgu.de

Abstract. As technical systems around us aim at a more natural interaction, the task of automatic emotion recognition from speech receives an ever growing attention. One important question still remains unresolved: The definition of the most suitable features across different data types. In the present paper, we employed a random-forest based feature selection known from other research fields in order to select the most important features for three benchmark datasets. Investigating feature selection on the same corpus as well as across corpora, we achieved an increase in performance using only 40 to 60% of the features of the well-known emobase feature set.

Keywords: Speech emotion recognition · Feature selection
Random forest

1 Introduction

Speech is a carrier of different kinds of information – besides the pure semantic content of an utterance, there are several layers underneath [14]. In human-human interaction (HHI), the interlocutors try to extract this additional information, often using multiple channels – simply speaking, by listening not only to what is said but also how it is said. One such layer of information is the emotional layer – the same sentence can have different meanings depending on its emotional toning. This can be transferred to the domain of human-computer interaction (HCI) to enable computer systems to understand the emotional level in order to make HCI more natural and pleasant for the user.

Unfortunately, the recent performance boost in speech recognition provided by deep learning did not improve the performance of emotion recognition alike: Although there are first attempts to implement end-to-end approaches [24], they are still in their infancy and rely on multimodal data. As long as the required massive data amounts are not yet available for audio-based emotion recognition, it is necessary to explore the existing possibilities and to look for other ways to improve the performance of current systems. One such way is the extraction and selection of the most suitable features.

A. Karpov et al. (Eds.): SPECOM 2018, LNAI 11096, pp. 134–144, 2018.
https://doi.org/10.1007/978-3-319-99579-3_15

Since the Interspeech 2009 Emotion challenge [21], the *emobase* feature set (as described in detail in [10]) is often used as a go-to feature set for various acoustic recognition systems: e.g. dialogue performance [19], user state detection [8], physical pain detection [17], etc. It contains 988 features based on 19 functionals of 26 Low-Level-Descriptor (LLDs) and their deltas: Mel-Frequency Cepstral Coefficient (MFCC), Line Spectral Pairs (LSPs), intensity, fundamental frequency, and other – there are also larger versions of this set such as the *2010 emobase* version and the *emo large* version containing 1582 and 6552 features, respectively. Besides these large feature sets, there are also relatively small ones, such as the *GeMaps* set [9], containing 18 LLDs (based on frequency and spectrum) and their derivatives, resulting in a total of only 62 features for the minimalistic and 88 features for the extended set.

Although widely used, these sets are not perfect. So, the 988 features of emobase are often used to classify relatively small amounts of samples. The GeMaps set on the other hand, while having not as many features, does not achieve the same performance as emobase [9].

In the present study, we want to examine two questions. Our first research question is whether the emotion recognition performance achieved using the emobase feature set is the best possible, or whether the same or even better performance can be achieved with less features using a data-driven feature selection process. Our second question is whether the same features are important for different data types. To investigate these questions, we employ a Random Forest (RF)-based feature ranking procedure on three different corpora and conduct classification experiments using same-corpus as well as cross-corpus features.

1.1 Literature Review

As early as 2003, Kwon et al. have deducted that the extraction of good features is more important to the emotion recognition task than the choice of the optimal classifier [13]. The most frequently used features comprise prosodic and spectral information. One problem concerning such features is that their values depend on the individual speaker's voice characteristics. Possible solutions are the calculation of speaker-independent features, such as the changes instead of the absolute values [15], or different normalisation methods [3]. Some research questions have already been answered: For example, it was shown that suprasegmental features perform better than segmental ones [22] or that features are not language-independent [26]. The choice of the best suitable features was also addressed in different investigations. So, Bitouk et al. used spectral features to classify emotions on two corpora and investigate the influence of different feature selection techniques, but none of the employed methods lead to clear gains [2]. Gharavian et al. presented a sophisticated feature selection approach based on fast correlation-based filters and genetic-algorithm-based optimisation to achieve 5% absolute improvement in terms of accuracy [11]. Unfortunately, the authors opted for a training and test set evaluation procedure instead of a true Leave-One-Speaker-Out (LOSO) setting and therefore did not report on differences between the speakers. Besides the usually employed prosodic and spectral features, there are also approaches investigating novel feature sets – for instance based on the Fourier parameters [25] and wavelets [18].

In the present study, we investigate the performance of RF-based feature selection on three benchmark emotional datasets in a LOSO setting and compare the features selected for different data types.

2 Datasets

In order to be able to answer our research questions in the most possible generalised way, we employed three famous benchmark corpora with different languages, emotion types and recording conditions.

The *Audiovisual Interest Corpus* (AVIC) [20] is a dataset built around a product presenter in an English commercial presentation. The recordings were made in an office environment and contain three levels of interest (loi1 - loi3) as classes.

The *Berlin Emotional Speech Database (emoDB)* [5] is a studio-recorded German dataset containing recordings of ten emotionally neutral sentences with seven emotions: anger, boredom, disgust, fear, joy, neutral, and sadness.

The *Speech Under Simulated and Actual Stress* (SUSAS) dataset [12] contains acted and spontaneous emotional utterances of English speakers in four different conditions: neutral, medium stress, high stress and screaming. Some of the utterances also contain field noise.

An overview over the details of the corpora is given in Table 1.

Table 1. Characteristics of the selected corpora.

Property	AVIC	emoDB	SUSAS
Quality	Office	Studio	Noisy
Language	English	German	English
Emotion type	Spont	Acted	Mixed
# Speakers	21 (10f)	10 (5f)	7 (3f)
# Emotions	3	7	4
# Samples	3002	535	3593

3 Feature Selection with Random Forests

In order to find the optimal amount of features, we first ranked the features according to their importance for the classification task using RF. We then analysed the obtained feature rankings and compared them for different speakers of the same corpus as well as between the different corpora. In the last step, we compared the classification performance using an increasing number of features to find an optimum.

3.1 Feature Extraction

For feature extraction, we used the *emobase* feature set of the openSMILE toolkit mentioned above, providing 988 spectral and prosodic features extracted on utterance level (cf. [10] for details). In order to establish comparability of the features among different speakers, we standardised the data to zero mean and unit variance.

3.2 Feature Ranking

In order to select the most important features, it is necessary to rank the features according to their importance. One possibility for this is a feature ranking routine based on RF – an ensemble learning method combining a typically high number of binary decision trees [4]. In each decision tree, each node samples a random subset of features and chooses the feature that is suited best to split the data into classes based on the impurity measure (e.g. the Gini index or information gain). By iterating this process, the features can be ranked according to their ability to decrease the impurity. A detailed explanation can be found in [7, 23]. The method was tested for several applications, for example in the field of spectroscopy analysis [16].

To realise this feature ranking procedure, we used the random forest implementation provided by KNIME [1]. The procedure consists of three steps as illustrated in Fig. 1. In the first step, a random forest containing a high number of trees with k levels each (k can be a low number since the most relevant features are close to the root) is built on the training portion of the data in order to obtain two statistical values for each feature f: the number of models M_i which use f as split on a tree level i, and the number of times T_i f was in the feature sample for the level i. Their quotient summed up over all levels is the score S_f for each f:

$$S_f = \sum_{i=0}^{k} \frac{M_i}{T_i}$$

In a second step, a random score S_{rand_f} is generated by calculating the score in the same way, but now with randomly shuffled labels – this is done in order to eliminate a bias that might be contained in the data.

In order to balance the influence of randomness, both S_f and S_{rand_f} are calculated ten times and then averaged. The new score S_{new_f} is then obtained in a final third step by subtracting S_{rand_f} from S_f: $S_{new_f} = S_f - S_{rand_f}$. The features are then sorted according to their final scores, the ranking indicating their importance.

In order to avoid overfitting to the data, this procedure is executed in a LOSO manner: For each speaker, the feature ranking is performed only on the data of all the other speakers, excluding the data of the current speaker, which is reserved for later testing.

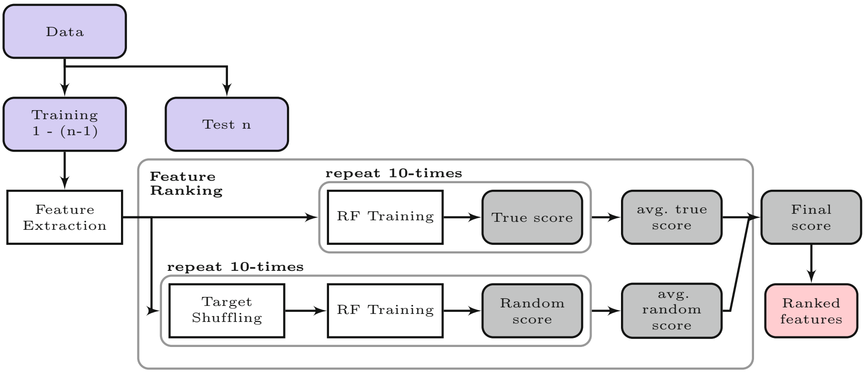


Fig. 1. An overview over the RF-based feature ranking procedure.

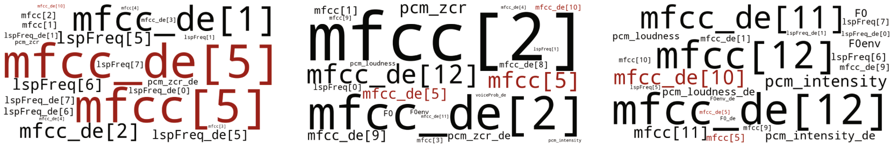
3.3 Comparison of Feature Rankings

One of our research questions was to investigate whether there are generally important features carrying emotional information or whether the most important features differ depending on the data. In order to answer this question we compared the feature rankings obtained on the three employed datasets and conduct several Pearson's correlation tests – between the feature selection rankings of different speakers of the same corpus for intra-corpus comparison as well as between the feature selection rankings of different corpora for inter-corpus comparison.

Intra-corpus Comparison. In order to test whether the feature rankings are consistent for all speakers within a corpus, we compared the LOSO rankings by conducting Pearson's correlation tests.

For *AVIC*, the Pearson's correlation coefficient r between the feature rankings of the individual speakers lies between 0.95 and 0.98 ($\overline{r} = 0.97 \pm 0.008$), leading to the conclusion that the feature rankings of the speakers are very similar. Our idea was now to construct an average feature ranking for the whole corpus by averaging the feature rankings over all speakers, F_{AVIC}. Naturally, the Pearson's correlation between F_{AVIC} and the feature rankings of the individual speakers is just as high as between the speakers, with values between 0.96 and 0.99 ($\overline{r} = 0.98 \pm 0.008$). The LLDs occurring most frequently in the top 100 are illustrated in Fig. 2a.

For *EmoDB*, we implemented the same procedure. Here the correlations between the speakers are about as high as for *AVIC*, with r values between 0.95 and 0.98 ($\overline{r} = 0.98 \pm 0.01$) indicating that the feature rankings are consistent. Also, in the same way as for AVIC, we constructed a new average feature ranking F_{EmoDB}. Again, r between F_{EmoDB} and the feature rankings of the individual speakers is between 0.97 and 0.99 ($\overline{r} = 0.99 \pm 0.006$). The LLDs occurring most frequently in the top 100 are illustrated in Fig. 2b.



(a) AVIC **(b)** EmoDB **(c)** SUSAS

Fig. 2. Word clouds of the LLDs most frequently occurring in the top 100 of the feature rankings for AVIC, EmoDB and SUSAS. The LLDs occurring for all three corpora are written in red. (Color figure online)

Finally, we repeated this procedure for *SUSAS*. The correlations between the feature rankings of the individual speakers are slightly lower than for EmoDB, with r values between 0.87 and 0.96 ($\overline{r} = 0.92 \pm 0.03$) but still sufficiently high to conclude that the feature rankings are consistent. The correlations between the average feature ranking F_{SUSAS} and the individual rankings are between 0.92 and 0.98 ($\overline{r} = 0.96 \pm 0.02$). The LLDs occurring most frequently in the top 100 are illustrated in Fig. 2c.

Inter-corpus Comparison. In the second step of our analysis, we compared the inter-corpus results in order to find whether the feature rankings are similar between the different types of data used. For this, we calculated the Pearson's correlation coefficients between the previously constructed average feature rankings F_{emoDB}, F_{SUSAS} and F_{AVIC}. In contrast to the intra-corpus comparison presented above, the results lead to the conclusion that there are no correlations between the feature rankings of the different corpora.

For the correlation between F_{EmoDB} and F_{AVIC}, the r value is 0.18. For the correlation between F_{EmoDB} and F_{SUSAS}, r is even lower, 0.14. For F_{SUSAS} and F_{AVIC}, r is negative, -0.07. These results are shown in Fig. 2: There are only two LLDs shared by all three datasets (MFCC[5]and its derivative as well as the derivative of MFCC[10]). This means that, unfortunately, the feature rankings are not universally transferable for different types of data. However, there are similarities – different MFCCs seem to be the most important features, since they occur relatively often in the top 100 features for all three datasets.

3.4 Selecting the Optimal Number of Features

In the next part, we searched for an optimal number of features for each of the corpora. For this, we classified the data using an increasing number of features, starting with 50 features with the highest RF-scores and then consecutively adding 50 more features with decreasing scores in each step, until we reached the full 988 emobase feature set. In order to avoid overfitting, we again used a LOSO validation setting. For each feature subset, we calculated the Unweighted Average Recall (UAR) over all classes and speakers. The UARs achieved during this optimisation procedure are shown in Fig. 3. Here, AVIC and EmoDB show

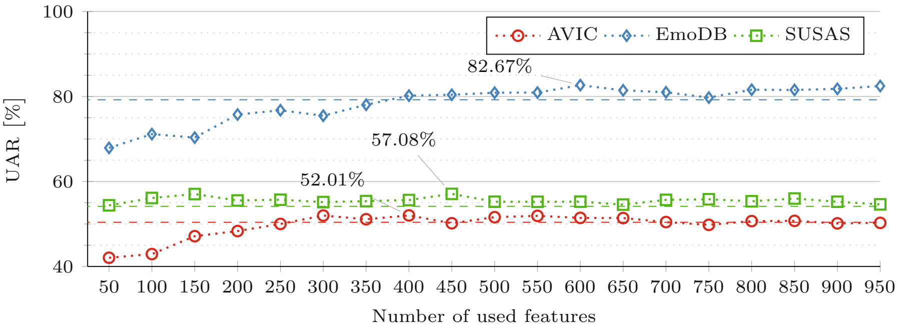


Fig. 3. The UAR of the classification performance on the three datasets depending on the number of selected features. The results achieved using the full number of features are indicated by the dashed line

similar results: after starting with a rather low UAR value for low numbers of features, the UAR rises rapidly and stays at a stable value. However, for SUSAS the number of features seems to have less influence, since the UAR does not change as much as for the other two corpora.

4 Classification Using Previously Selected Features

After selecting the optimal number of features, we conducted classification experiments in order to evaluate and compare the performance of the selected features to the full emobase feature set.

4.1 Classification Setup

For the classification, we again implemented the LOSO procedure as described above. Since we obtained between 7 and 21 models for each corpus, we decided against parameter fine-tuning and employed default employed default Support Vector Machine (SVM) parameters as provided by the LibSVM library [6]. For evaluation, we computed the unweighted average f-measure (UAF) as the harmonic mean of the unweighted average recall and precision over all classes of one speaker, and then the unweighted average over all speakers. In order to include variations over speakers, we report the average values as well as the standard deviation as performance measures.

4.2 Classification Performance

The classification results are shown in Fig. 4 – we report the classification performance for each dataset, the baseline performance using all 988 emobase features and the performance using the previously selected features. Furthermore, we also report the results using cross-corpus feature selection. For this, we performed the

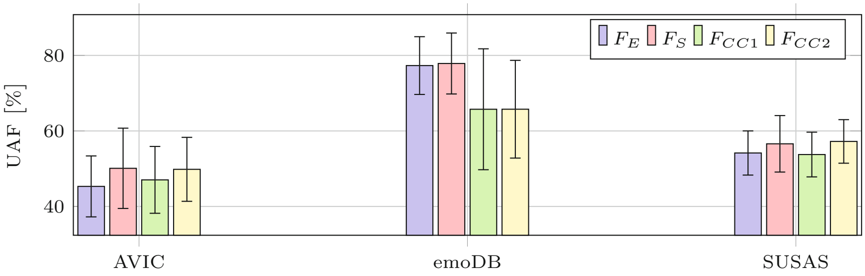


Fig. 4. The UAF of the classification performance for the emobase feature set F_E, the best feature selection set F_S, the cross-corpus feature set with the lower correlation F_{CC1} and with the higher correlation F_{CC2}

classification on one dataset using the feature set obtained on another one. Since we used three corpora, this procedure results in two additional values per corpus: F_{CC1} denotes the results using the feature set with the lower correlation coefficient (as obtained in Sect. 3), F_{CC2} the results with the higher correlation coefficient.

The classification with feature selection outperforms the classification using the full emobase feature set for all three corpora by several percent absolute – but the improvements lie within the standard deviation of the average values of the speakers. However, the results show that for all three corpora, a performance improvement can be achieved using between 40 and 60% less features than the original feature set. This is an interesting finding since feature extraction as well as classification are resource-intensive tasks, where a reduction of the processing overhead can be a real benefit – for example in the domain of mobile applications.

Regarding the performance of the different feature sets across corpora, we can observe that the results are almost as expected: except for SUSAS, the "alien" feature sets obtained by feature selection on another corpus do not perform as good as the one obtained on the same corpus. Furthermore, F_{CC2} outperforms F_{CC1} in all cases (albeit marginally as for emoDB), which corresponds to the higher correlation between F_{CC2} and F_S compared to F_{CC1} and F_S. The only exception is SUSAS, where the F_{CC2} works about 0.7% better than F_S.

Based on these results, we can conclude that RF-based feature selection is a viable method to improve emotion recognition performance for different types of data.

5 Conclusion

The first question we aimed to investigate in this study was whether the number of features used for emotion recognition can be reduced achieving the same or even better performance. We have shown that by applying RF-based feature selection, we can reduce the number of features roughly by half and obtain an even better performance than using the full emobase set – furthermore, by using three different corpora we have shown that this result is independent of the type of emotions, language and recording conditions.

The second research question was whether there are inter-corpus similarities in the selected features. Here our finding is that the most important features are not consistent over different corpora, and therefore the feature selection needs to be done for each emotion recognition task separately. However, different MFCCs are among the most important features of all three corpora indicating that there is a common ground of acoustic information.

There are two main directions for further research. The first interesting question here is to investigate further feature sets – besides larger versions of the emobase feature set including up to 6552 features also novel and less frequently used features such as the Fourier parameters and wavelet-based features are of interest. The second open question is to consolidate feature classes according to the type of material used – in this investigation, we have seen that features important for EmoDB differ from those for AVIC. The question is whether these differences are based on the type of emotions, on the emotional classes, on the recording conditions, or on some still unknown factors. This needs to be further investigated in order to understand the relations between the features and the information on the emotional status of the speaker contained in them.

Acknowledgements. This work has been sponsored by the German Federal Ministry of Education and Research in the program Zwanzig20 – Partnership for Innovation as part of the research alliance 3Dsensation (grant number 03ZZ0414). It was also supported by the project Intention-based Anticipatory Interactive Systems (IAIS) funded by the European Funds for Regional Development (EFRE) and by the Federal State of Sachsen-Anhalt, Germany (grant number ZS/2017/10/88785).

References

1. Berthold, M.R., et al.: KNIME: The konstanz information miner. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R. (eds.) Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78246-9_38
2. Bitouk, D., Verma, R., Nenkova, A.: Class-level spectral features for emotion recognition. Speech Commun. **52**(7–8), 613–625 (2010)
3. Böck, R., Egorow, O., Siegert, I., Wendemuth, A.: Comparative study on normalisation in emotion recognition from speech. In: Horain, P., Achard, C., Mallem, M. (eds.) IHCI 2017. LNCS, vol. 10688, pp. 189–201. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-72038-8_15
4. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
5. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A database of German emotional speech. In: Proceedings of the INTERSPEECH-2005, pp. 1517–1520 (2005)
6. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. Trans. Intell. Syst. Technol. **2**, 1–27 (2011)

7. Chen, Y.W., Lin, C.J.: Combining SVMs with various feature selection strategies. In: Guyon, I., Nikravesh, M., Gunn, S., Zadeh, L.A. (eds.) Feature Extraction: Foundations and Applications, pp. 315–324. Springer, Berlin Heidelberg (2006). https://doi.org/10.1007/978-3-540-35488-8_13
8. Egorow, O., Wendemuth, A.: Detection of challenging dialogue stages using acoustic signals and biosignals. In: Proceedings of the 24th International Conference on Computer Graphics, Visualization and Computer Vision, pp. 137–143 (2016)
9. Eyben, F., et al.: The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. Trans. Affect. Comput. **7**(2), 190–202 (2016)
10. Eyben, F., Wöllmer, M., Schuller, B.: OpenEAR - introducing the Munich open-source emotion and affect recognition toolkit. In: Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 1–6. IEEE (2009)
11. Gharavian, D., Sheikhan, M., Nazerieh, A., Garoucy, S.: Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network. Neural Comput. Appl. **21**(8), 2115–2126 (2012)
12. Hansen, J., Bou-Ghazale, S.: Getting started with SUSAS: A speech under simulated and actual stress database. In: Proceedings of the EUROSPEECH-1997, pp. 1743–1746 (1997)
13. Kwon, O.W., Chan, K., Hao, J., Lee, T.W.: Emotion recognition by speech signals. In: Proceedings of the 8th European Conference on Speech Communication and Technology (2003)
14. Levinson, S.C., Holler, J.: The origin of human multi-modal communication. Phil. Trans. R. Soc. B **369**(1651), 20130302 (2014)
15. Mao, Q., Zhao, X., Zhan, Y.: Extraction and analysis for non-personalized emotion features of speech. Adv. Inf. Sci. Serv. Sci. **3**(10), 255–263 (2011)
16. Menze, B.H., et al.: A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics **10**(1), 213 (2009)
17. Oshrat, Y., Bloch, A., Lerner, A., Cohen, A., Avigal, M., Zeilig, G.: Speech prosody as a biosignal for physical pain detection. In: Proceedings of Speech Prosody, pp. 420–424 (2016)
18. Palo, H.K., Mohanty, M.N.: Wavelet based feature combination for recognition of emotions. Ain Shams Eng. J. (2017, in Press)
19. Ramanarayanan, V., et al.: Using vision and speech features for automated prediction of performance metrics in multimodal dialogs. ETS Research Report Series 1 (2017)
20. Schuller, B., Müller, R., Hörnler, B., Höthker, A., Konosu, H., Rigoll, G.: Audiovisual recognition of spontaneous interest within conversations. In: Proceedings of the 9th International Conference on Multimodal interfaces, pp. 30–37. ACM (2007)
21. Schuller, B., Batliner, A., Steidl, S., Seppi, D.: Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. Speech Commun. **53**(9–10), 1062–1087 (2011)
22. Schuller, B., Wöllmer, M., Eyben, F., Rigoll, G.: The role of prosody in affective speech, linguistic insights, studies in language and communication. Lang. Commun. **97**, 285–307 (2009)
23. Silipo, R., Adae, I., Hart, A., Berthold, M.: Seven techniques for dimensionality reduction. Technical report, KNIME (2014)

24. Tzirakis, P., Trigeorgis, G., Nicolaou, M.A., Schuller, B.W., Zafeiriou, S.: End-to-end multimodal emotion recognition using deep neural networks. J. Sel. Top. Signal Process. **11**(8), 1301–1309 (2017)
25. Wang, K., An, N., Li, B.N., Zhang, Y., Li, L.: Speech emotion recognition using fourier parameters. Trans. Affect. Comput. **6**(1), 69–75 (2015)
26. Yang, C., Ji, L., Liu, G.: Study to speech emotion recognition based on TWINsSVM. In: Proceedings of the 5th International Conference on Natural Computation, vol. 2, pp. 312–316. IEEE (2009)