# Speech Emotion Recognition Using Local and Global Features

Yuanbo Gao[1], Baobin Li[1(✉)], Ning Wang[2], and Tingshao Zhu[3]

[1] School of Computer and Control, University of Chinese Academy of Sciences, Beijing 100190, China
libb@ucas.ac.cn
[2] Beijing Institue of Electronics Technology and Application, Beijing 100091, China
[3] Institute of Psychology Chinese Academy of Sciences, Beijing 100101, China
tszhu@psych.ac.cn

**Abstract.** Speech is an easy and useful way to detect speakers' mental and psychological health, and automatic emotion recognition in speech has been investigated widely in the fields of human-machine interaction, psychology, psychiatry, etc. In this paper, we extract prosodic and spectral features including pitch, MFCC, intensity, ZCR and LSP to establish the emotion recognition model with SVM classifier. In particular, we find different frame duration and overlap have different influences on final results. So, Depth-First-Search method is applied to find the best parameters. Experimental results on two known databases, EMODB and RAVDESS, show that this model works well, and our speech features are enough effectively in characterizing and recognizing emotions.

**Keywords:** Speech · Emotion · SVM · EMODB · RAVDESS

## 1 Introduction

As a natural and effective way of human communication, speech is an important biometric feature classifying speakers into categories ranging from age, identity, idiolect and sociolect, truthfulness, cognitive health. In particular, it is also one of the most expressive modalities for human emotions, and how to recognize emotion automatically from human speech has been widely discussed in the fields of the human-machine interaction, psychology, psychiatry, behavioral science, etc. [1,2].

Finding and extracting good and suitable features is one of challenging and important tasks in the study of speech emotion recognition. In 1989, Cummings et al. reported that the shape of the glottal pulse varies with different stressed conditions in a speech signal [3]. Seppänen et al. extracted pitch, formant and energy features on MediaTeam emotional speech corpus, and got the 60% accuracy [4]. In 2005, Luengo et al. used prosodic features to recognize speech emotion based on Basque Corpus with the three different classifiers, and the best result


Y. Zeng et al. (Eds.): BI 2017, LNAI 10654, pp. 3–13, 2017.
https://doi.org/10.1007/978-3-319-70772-3_1

was up to 98.4% by using GMM classifier. Origlia et al. extracted 31 dimensional pitch and energy features, and got almost 60% accuracy for a multilingual emotional database including four European languages in 2010 [5].

Moreover, prosodic and spectral features including the linear predictor cepstral coefficients (LPCC) [6] and mel-frequency cepstral coefficients (MFCC) [7] have been also widely used in speech emotion recognition [8,9]. In 2000, Bou-Ghazale and Hansen [10] found that features based on cepstral analysis, such as LPCC and MFCC, outperformed ones extracted by linear predictor coefficients (LPC) [11] in detecting speech emotions. In 2003, New et al. proved that the performance of log-frequency power coefficient was better than LPCC and MFCC when using hidden Markov model as a classifier [12]. Wu et al. proposed modulation spectral features for the automatic recognition of human affective information from speech, and an overall recognizing rate of 91.6% was obtained for classifying seven emotions [13].

In addition, voice quality features are also related to speech emotion including format frequency and bandwidth, jitter and shimmer, glottal parameter, etc. [14,15]. Li et al. extracted jitter, shimmer features mixed with MFCC features as voice quality parameters to identify emotions on SUSAS database, and compared to MFCC, the accuracy increased by 4% [14]. Lugger et al. combined prosodic features with voice quality parameters for speech emotion recognition and the best accuracy was up to 70.1% with two ways of combining classifiers [15].

Recently, the combination of many kinds of features has been widely used for automatic speech emotion recognition. In 2012, Pan et al. found the combination of MFCC, mel-energy spectrum dynamic coefficients and Energy, obtained high accuracies on both Chinese emotional database (91.3%) and Berlin emotional database (95.1%) [16]. Chen et al. extracted the energy, zero crossing rate (ZCR), pitch, the first to third formants, spectrum centroid, spectrum cut-off frequency, correlation density, fractal dimension, and five Mel-frequency bands energy for every frame. Average accuracies were 86.5%, 68.5% and 50.2%, respectively [17]. In 2013, Deng et al. merged ZCR, MFCC and harmonics-to-noise features, and used the autoencoder method to classify emotions [18]. In 2014, Han et al. proposed to utilize deep neutral networks to extract MFCC and pitch-based features from raw data on Emotional Dyadic Motion Capture to distinguish excitement, frustration, happiness, neutral and surprise emotions, and the best accuracy obtained by DNN-KELM was 55% [19].

In this paper, we extract different features including MFCC, intensity, pitch, Line spectral pairs (LSP) and ZCR features, which have been proved to be effective, and then, a smooth and normalization operation is needed to reduce the noise. The output with delta regression coefficients are regarded as local features. In the end, we use 15 statistics methods to extract global features based on local features. In particular, we find the features of samples not only can effect the results of classification, but the frame duration and overlap are also two key factors, so the Depth First Search (DFS) is used to select frame duration and overlap and find the most appropriate combination between features and frame

duration and overlap, which works well on two public databases: EMODB and RAVDESS.

This paper is organized as follows. Section 2 introduces and describes the information of two databases: EMODB and RAVDESS. Section 3 presents the process of extracting features including local and global features. Section 4 shows the classified results, and the average accuracy of cross-validation is 87.3% for EMODB and 79.4% for RAVDESS respectively. Finally, the conclusion and future work are outlined in Sect. 5.

## 2 Materials and Methods

### 2.1 Database

In this paper, two databases are considered: Berlin emotional corpus (EMODB) [20] and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [21]. EMODB has been built by Institute of Communication Science at the Technical University of Berlin, which contains 535 utterances of 10 actors (5 males, 5 females) with 7 different emotions—angry, anxiety/fear, happiness, sadness, disgust, boredom and neutral. The distributions of each class are shown in Table 1.

**Table 1.** The distributions of EMODB Database

| Emotion | Angry | Boredom | Disgust | Fear | Happy | Neutral | Sad |
|---|---|---|---|---|---|---|---|
| Number | 127 | 81 | 46 | 69 | 71 | 79 | 62 |

RAVDESS includes the audio-visual of 24 performers (12 male, 12 female) speaking and singing the same two sentences with various emotions, where the speech recordings include angry, calm, disgust, fearful, happy, neutral, sad, surprise eight emotions, and the song recording have six emotions except for digust and surprise emotions. RAVDESS contains over 7000 files (audio-only, video-only, full audio-video). We only use audio utterances with normal intensity, and the number of each class is 92.

### 2.2 Features for Speech Emotion Recognition

We choose frame duration and overlap by DFS. The range of frame durations is from 20 ms to 100 ms, and every frame durations acts as a father node. Every overlaps is a child node and start at 10 ms, increasing by 1 ms each time until equal to the half of its father node.

Every speech sample is divided into a sequence of frames. A hamming window is applied to remove the signal discontinuities at the ends of each frame. Within each window, we extract pitch, LSP, MFCC, intensity and ZCR features, and combine these features with a smooth operation. The output with their delta regression coefficients are our speech features. The following subsections will describe these features in details.

**Local Features.** The local features are comprised of pitch, MFCC, LSP, Intensity and ZCR features.

*Pitch Features.* Each frame $x^i(t)$ pass a hamming window, and a short-time Fourier transform function is applied on it,

$$H(w) = \sum_{n}^{N} x^i(t) * Ham(len(x^i(t))) * e^{-jwn}, \tag{1}$$

where $x^i$ means the $i$th frame's sample points, N means the length of $i$th frame. Then, we use above results to compute the autocorrelation function (ACF) and Cepstrum coefficients,

$$C^i = \sum_{w}^{N} log^{|H(w)|} * e^{jwn}, R^i = \sum_{w}^{N} |H(w)|^2 * e^{jwn}, \tag{2}$$

where $C^i$ denotes the $i$th frame's Cepstrum coefficient and $R^i$ is $i$th frame's ACF. Pitches are computed as follows,

$$\begin{aligned} C_p^i &= \frac{f}{idx(max(C^i[f*0.02 : f*0.2])) + f*0.02 - 1}, \\ R_p^i &= \frac{f}{idx(max(R^i[f*0.02 : f*0.2])) + f*0.02 - 1}, \end{aligned} \tag{3}$$

where $C_p^i$ and $R_p^i$ denote the $i$th frame pitch computed by Cepstrum coefficient and ACF, respectively, $f$ stands for the sampling rate, and the $idx$ means the index of values.

*MFCC Features.* MFCC features are commonly used for human speech analysis. MFCCs use frequency bands based on the Mel-spaced triangular frequency bins, then a Discrete Cosine Transform (DCT) is applied to calculate the desired number of cepstral coefficients. The first twelve cepstral coefficients are extracted as our MFCC features.

*LSP Features.* LSP is a way of uniquely representing the LPC-coefficients. It decomposes the $p$ order linear predictor $A(z)$ into a symmetrical and anti-symmetrical part denoted by the polynomial $P(z)$ and $Q(z)$, separately.

The $p$ order linear prediction system can be written as the Eq. (4) in z-domain, and a prediction error $A(z)$ is produced by (5),

$$x(n) = \sum_{k=1}^{p} a_k x(n-k) \longrightarrow x(z) = \sum_{k=1}^{p} a_k x(z), \tag{4}$$

$$A(z) = 1 - \sum_{k=1}^{p} a_k * z^{-k}. \tag{5}$$

Two polynomials $P(z)$ and $Q(z)$ are given by

$$P(z) = A(z) + z^{-(p+1)}A(z^{-1}), \quad Q(z) = A(z) - z^{-(p+1)}A(z^{-1}). \tag{6}$$

LSP parameters are expressed as the zeros of P(z) and Q(z), and we choose the first eight parameters as our LSP features.

*Intensity and ZCR features.* Intensity, one of the major vocal attribute, can be computed by

$$I_i = \frac{1}{N}\sum_{n}^{N} Ham^2(x^i(n)). \tag{7}$$

ZCR is the rate at which the signal changes from positive to negative or back. It has been used in both speech recognition, being a key feature to classify percussive sounds.

**Smoothing and Normalization.** Smoothing is an useful tool to reduce the noise and slick data contours. In smoothing, the data points of a signal are modified so that individual points that are higher than the immediately adjacent points, are reduced, and points that are lower than the adjacent points are increased. In this paper, we adopt a moving average filter of 3-length to smooth data,

$$s_i = \frac{x_{i-1} + x_i + x_{i+1}}{3}. \tag{8}$$

It is essential to normalize data for a strong emotion recognition system. The goal of normalization is to eliminate speaker and recording variability while keeping the emotional discrimination. In this paper, the min-max method that limits the value of each feature ranging between 0 and 1 is adopted for feature scaling. For every sample, the normalized feature is estimated by the following

$$s_i = \frac{s_i - min(s_i)}{max(s_i) - min(s_i)}. \tag{9}$$

**Global Features.** Global features are calculated as statistics of all speech features extracted from an utterance. The majority of researchers have agreed that global features are superior to local ones in terms of classification accuracy and classification time. Therefore, the maximum, minimum, mean, maximum position, minimum position, range (maximum-minimum), standard deviation, skewness, kurtosis, linear regression coefficient (slope and offset), linear regression error (linear error and quadratic error), quartile and inter-quartile range are computed based on local features, and combined into the global features.

Every frame consists of pitch, MFCC, LSP, intensity and ZCR features. After smoothing these features, the delta regression coefficients are merged, then, the global features are computed based on local features. Figure 1 shows the block diagram for the processing of feature extraction. We extract these features with the help of OpenSMILE [22].

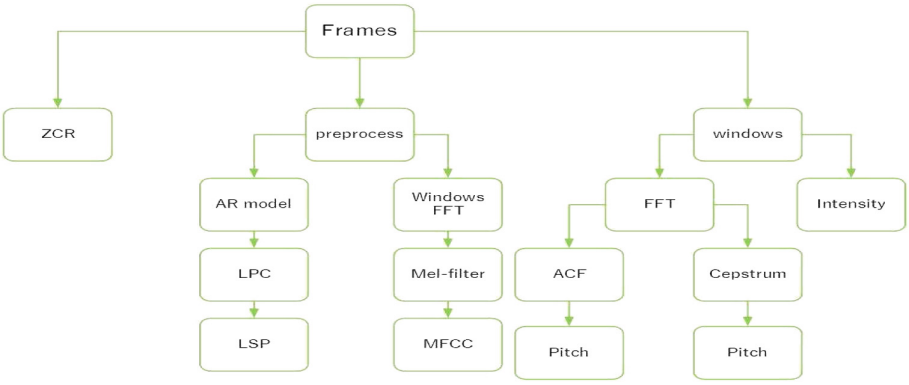
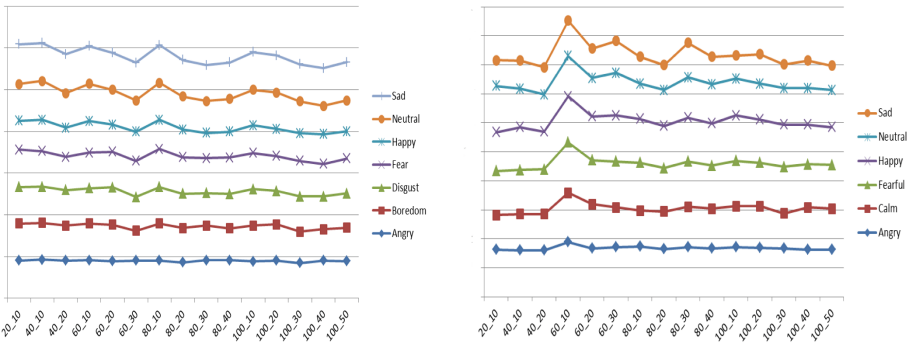



**Fig. 1.** Block diagram of processing of features extraction




(a) (b)

**Fig. 2.** Frame Duration

## 3 Results/Discussion

In the speech emotion recognition, many classifiers including GMM, artificial neutral networks (ANN), and support vector machine (SVM), has been used. Specially, SVM classifier is widely used in many pattern recognition applications, and outperforms other well-known classifiers. This study use a linear kernel SVM with sequential minimal optimization (SMO) to build model.

In particular, by DFS, we find the best durations between two database are different. Every broken line in Fig. 2 means the change of accuracy in different duration and overlap for corresponding emotion. For German Database, Fig. 2(a) shows that the best frame duration is 40 ms and overlap is 10 ms while Fig. 2(b) tells us that the best frame duration is 60 ms and overlap is 10 ms in terms of RAVDESS Database.

### 3.1 Classification Results for EMODB

For EMODB database, the number of samples is 535. We adopt the ten-fold cross validation to avoid overfitting, and the results are shown in Table 2. It can be seen the average accuracy of cross-validation is 87.3%. The previous results on EMODB database such as in [23], the average accuracy was 83.3%. In [24], they aggregated features from MPEG-7 descriptors, MFCCs and Tonality. The best results was 83.39% by using SVM classifier. In [25], they extracted 1800 dimensional features, and average accuracy was 73.3% for classifying six categories ignoring the disgust emotion.

**Table 2.** Confusion Matrix on the EMODB (%)

| | Angry | Boredom | Disgust | Fear | Happy | Neutral | Sad |
|---|---|---|---|---|---|---|---|
| Angry | 92.9 | 0 | 0 | 2.3 | 4.6 | 0 | 0 |
| Boredom | 0 | 87.5 | 0 | 0 | 0 | 4.9 | 7.4 |
| Disgust | 2.1 | 2.1 | 86.9 | 4.2 | 4.2 | 0 | 0 |
| Fear | 5.7 | 0 | 1.4 | 85.5 | 4.3 | 0 | 2.8 |
| Happy | 14.1 | 0 | 1.4 | 9.8 | 74.6 | 0 | 0 |
| Neutral | 0 | 6.4 | 0 | 0 | 0 | 93.6 | 0 |
| Sad | 0 | 9.6 | 0 | 0 | 0 | 1.7 | 90.3 |

Table 3 shows the results of each feature (e.g. pitch, MFCC, etc.) and MFCC+LSP to compare with our original features. The reason why we choose MFCC+LSP feature is that MFCC and LSP are the best two types of features in term of classifying accuracy.

**Table 3.** Comparisons the accuracy of each feature and MFCC+LSP with fusion features (%)

| | Angry | Boredom | Disgust | Fear | Happy | Neutral | Sad |
|---|---|---|---|---|---|---|---|
| ZCR | 79.5 | 52.5 | 36.9 | 51.4 | 30.9 | 51.2 | 56.4 |
| Intensity | 81.1 | 61.2 | 26.9 | 42.6 | 11.2 | 48.7 | 53.2 |
| Pitch | 83.4 | 72.5 | 47.8 | 55.8 | 36.6 | 57.5 | 80.6 |
| LSP | 81.8 | 83.7 | 69.5 | 70.5 | 59.1 | 81.2 | 66.1 |
| MFCC | 86.6 | 85.0 | 80.4 | 79.4 | 64.7 | 83.7 | 83.8 |
| MFCC+LSP | 92.1 | 88.7 | 84.7 | 82.3 | 74.6 | 87.5 | 79.0 |
| Original | 92.9 | 87.5 | 86.9 | 85.5 | 74.6 | 93.6 | 90.3 |

### 3.2 Classification Results for RAVDESS

RAVDESS database composed by six emotions which are angry, calm, fearful, happy, neutral and sad. Each emotions has 92 samples, and we adopt the same processes as EMODB. The results are shown in Table 4. The average accuracy of angry is 94.5%, calm is 84.7%, fearful is 86.9%, happy is 79.3%, neutral is 69.5% and sad is 60.8%. In [26], it has the accuracy 82% for angry, 96% for happy, 70% for neutral, 58% for sad, 84% for calm and 88% for fearful.

**Table 4.** Confusion Matrix on the RAVDESS (%)

| | Angry | Calm | Fearful | Happy | Neutral | Sad |
|---|---|---|---|---|---|---|
| Angry | 94.5 | 5.5 | 0 | 0 | 0 | 0 |
| Calm | 2.5 | 84.7 | 1.0 | 0 | 3.2 | 8.6 |
| Fearful | 0 | 1.0 | 86.9 | 3.2 | 1.0 | 7.6 |
| Happy | 0 | 0 | 7.6 | 79.3 | 9.7 | 3.2 |
| Neutral | 0 | 4.2 | 1.0 | 6.5 | 69.5 | 18.4 |
| Sad | 0 | 3.2 | 11.9 | 3.2 | 20.6 | 60.8 |

Table 5 shows the results of each feature (e.g. pitch feature, MFCC features, etc.) and MFCC+LSP features to compare with our original features.

**Table 5.** Comparison the accuracies between different screening methods on RAVDESS (%)

| | | | | | | |
|---|---|---|---|---|---|---|
| ZCR | 39.1 | 18.4 | 25 | 5.4 | 39.1 | 20.6 |
| Intensity | 60.8 | 41.3 | 25.0 | 40.2 | 31.5 | 17.3 |
| Pitch | 52.1 | 44.5 | 44.5 | 31.5 | 16.3 | 10.8 |
| LSP | 70.6 | 51.1 | 59.7 | 59.7 | 53.2 | 28.2 |
| MFCC | 71 | 59.7 | 64.1 | 64.1 | 63.2 | 30 |
| MFCC+LSP | 94.5 | 82.6 | 88.1 | 80.6 | 69.5 | 59.6 |
| Original | 94.5 | 84.7 | 86.9 | 79.3 | 69.5 | 60.8 |

### 3.3 SFFS

Moreover, we use the sequential floating forward search (SFFS) which is simple, fast, effective and widely accepted technique to select features. The central idea of SFFS is that we choose the best feature firstly by forward tracking, then exclude a number of features by backtracking, this process will be repeated until the number of features which have been selected is unchange. For EMODB corpus, the number of features which are selected by using SFFS is 35 and the accuracy

of Happy improved by 1.4%. For RAVDESS corpus, the dimension reduce to 27. The accuracy of Happy improved by 1.3% and Fearful improved by 1.2%. The results for EMODB corpus and RAVDESS corpus are described in Tables 6 and 7 respectively.

**Table 6.** Comparison the accuracies between different screening methods on EMODB (%)

| | Angry | Boredom | Disgust | Fear | Happy | Neutral | Sad |
|---|---|---|---|---|---|---|---|
| SFFS | 92.9 | 87.5 | 86.9 | 85.5 | 76 | 93.6 | 90.3 |
| Original | 92.9 | 87.5 | 86.9 | 85.5 | 74.6 | 93.6 | 90.3 |

**Table 7.** Comparison the accuracies between different screening methods on EMODB (%)

| | Angry | Calm | Fearful | Happy | Neutral | Sad |
|---|---|---|---|---|---|---|
| SFFS | 94.5 | 84.7 | 88.1 | 80.6 | 69.5 | 60.8 |
| Original | 94.5 | 84.7 | 86.9 | 79.3 | 69.5 | 60.8 |

## 4 Conclusions

In this paper, we extract two types of features for emotion recognition based on two well-known databases: EMODB and RAVDASS. Then we carry out the fusion, smooth and normalization operation on these features, and the accuracies obtained on the two databases are very promising using SVM classifier, but we still not find a useful reduce dimensional method to get better results. In addition, the limitation of this paper is that we only used two speech emotional corpora. In the future, we will use more databases to identify the effect of our method and build more persuasive model using some methods of deep learning like Bidirectional Long ShortTerm Memory (BLSTM) networks and Gaussian RBM, and we also improve our reducing methods to get better results so that we can further throw invalid features, and use less features to get better results.

**Acknowledgments.** The research was supported in part by NSFC under Grants 11301504 and U1536104, in part by National Basic Research Program of China (973 Program2014CB744600).

## References

1. Minker, W., Pittermann, J., Pittermann, A., Strauß, P.M., Bühler, D.: Challenges in speech-based human-computer interfaces. Int. J. Speech Technol. **10**(2–3), 109–119 (2007)

2. Ntalampiras, S., Potamitis, I., Fakotakis, N.: An adaptive framework for acoustic monitoring of potential hazards. EURASIP J. Audio Speech Music Process. **2009**, 13 (2009)
3. Cummings, K.E., Clements, M.A., Hansen, J.H.: Estimation and comparison of the glottal source waveform across stress styles using glottal inverse filtering. In: Proceedings of the IEEE Energy and Information Technologies in the Southeast. Southeastcon 1989, pp. 776–781. IEEE (1989)
4. Seppänen, T., Väyrynen, E., Toivanen, J.: Prosody-based classification of emotions in spoken finnish. In: INTERSPEECH (2003)
5. Origlia, A., Galatà, V., Ludusan, B.: Automatic classification of emotions via global and local prosodic features on a multilingual emotional database. In: Proceeding of the 2010 Speech Prosody. Chicago (2010)
6. Atal, B.S.: Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J. Acoust. Soc. Am. **55**(6), 1304–1312 (1974)
7. Davis, S.B., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. **28**(4), 357–366 (1980)
8. Ververidis, D., Kotropoulos, C., Pitas, I.: Automatic emotional speech classification. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP 2004), vol. 1, IEEE I-593 (2004)
9. Fernandez, R., Picard, R.W.: Classical and novel discriminant features for affect recognition from speech. In: Interspeech, pp. 473–476 (2005)
10. Bou-Ghazale, S.E., Hansen, J.H.: A comparative study of traditional and newly proposed features for recognition of speech under stress. IEEE Trans. Speech Audio Process. **8**(4), 429–442 (2000)
11. Rabiner, L.R., Schafer, R.W.: Digital processing of speech signals (prentice-hall series in signal processing) (1978)
12. Nwe, T.L., Foo, S.W., De Silva, L.C.: Speech emotion recognition using hidden markov models. Speech Commun. **41**(4), 603–623 (2003)
13. Wu, S., Falk, T.H., Chan, W.Y.: Automatic speech emotion recognition using modulation spectral features. Speech Commun. **53**(5), 768–785 (2011)
14. Li, X., Tao, J., Johnson, M.T., Soltis, J., Savage, A., Leong, K.M., Newman, J.D.: Stress and emotion classification using jitter and shimmer features. In: IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP 2007, vol. 4, IEEE IV-1081 (2007)
15. Lugger, M., Janoir, M.E., Yang, B.: Combining classifiers with diverse feature sets for robust speaker independent emotion recognition. In: 2009 17th European Signal Processing Conference, pp. 1225–1229. IEEE (2009)
16. Pan, Y., Shen, P., Shen, L.: Speech emotion recognition using support vector machine. Int. J. Smart Home **6**(2), 101–108 (2012)
17. Chen, L., Mao, X., Xue, Y., Cheng, L.L.: Speech emotion recognition: features and classification models. Digit. Signal Process. **22**(6), 1154–1160 (2012)
18. Deng, J., Zhang, Z., Marchi, E., Schuller, B.: Sparse autoencoder-based feature transfer learning for speech emotion recognition **7971**, 511–516 (2013)
19. Han, K., Yu, D., Tashev, I.: Speech emotion recognition using deep neural network and extreme learning machine. In: Interspeech, pp. 223–227 (2014)
20. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B.: A database of german emotional speech. Interspeech **5**, 1517–1520 (2005)

21. Livingstone, S., Peck, K., Russo, F.: Ravdess: the ryerson audio-visual database of emotional speech and song. In: 22nd Annual Meeting of the Canadian Society for Brain, Behaviour and Cognitive Science (CSBBCS) (2012)
22. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM international conference on Multimedia, pp. 1459–1462. ACM (2010)
23. Kotti, M., Paternò, F.: Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema. Int. J. Speech Technol. **15**(2), 131–150 (2012)
24. Lampropoulos, A.S., Tsihrintzis, G.A.: Evaluation of MPEG-7 Descriptors for Speech Emotional Recognition (2012)
25. Wang, K., An, N., Li, B.N., Zhang, Y., Li, L.: Speech emotion recognition using fourier parameters. IEEE Trans. Affect. Comput. **6**(1), 69–75 (2015)
26. Zhang, B., Essl, G., Provost, E.M.: Recognizing emotion from singing and speaking using shared models. In: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 139–145. IEEE (2015)