Journal of Ambient Intelligence and Humanized Computing (2019) 10:1787–1798
https://doi.org/10.1007/s12652-017-0644-8

ORIGINAL RESEARCH

Feature fusion methods research based on deep belief networks for speech emotion recognition under noise condition

Yongming Huang[1,2] · Kexin Tian[1,2] · Ao Wu[1,2] · Guobao Zhang[1,2]

Received: 19 September 2017 / Accepted: 27 November 2017 / Published online: 2 December 2017

Abstract
The speech emotion recognition accuracy of prosody feature and voice quality feature declines with the decrease of signal to noise ratio (SNR) of speech signals. In this paper, we propose novel sub-band spectral centroid weighted wavelet packet Cepstral coefficients (W-WPCC) for robust speech emotion recognition. The W-WPCC feature is computed by combining the sub-band energies with sub-band spectral centroids via a weighting scheme to generate noise-robust acoustic features. And deep belief networks (DBNs) are artificial neural networks having more than one hidden layer, which are first pre-trained layer by layer and then fine-tuned using back propagation algorithm. The well-trained deep neural networks are capable of modeling complex and non-linear features of input training data and can better predict the probability distribution over classification labels. We extracted prosody feature, voice quality features and wavelet packet Cepstral coefficients (WPCC) from the speech signals to combine with W-WPCC and fused them by DBNs. Experimental results on Berlin emotional speech database show that the proposed fused feature with W-WPCC is more suitable in speech emotion recognition under noisy conditions than other acoustics features and proposed DBNs feature learning structure combined with W-WPCC improve emotion recognition performance over the conventional emotion recognition method.

Keywords Speech emotion recognition · Weighted wavelet packets Cepstral coefficients (W-WPCC) · Feature fusion · Deep belief networks (DBNs)

1 Introduction

Speech emotion recognition is useful in various applications where natural human–computer interaction is needed. For example, in the design of interactive movies and online games (Caponetti et al. 2011), in call-centers to help with call processing according to perceived urgency (Morrison et al. 2007; Petrushin 2000), in intelligent automobile systems to assess driver's mental state and ensure safety (Malta et al. 2009) and in healthcare service to help diagnosing depression and suicide risk (France et al. 2000).

The task of speech emotion recognition is to recognize underlying emotional state of a speaker from speech signal. To accomplish this task, the extraction of relevant features that efficiently characterize emotions is an important but challenging step. Most of the widely used features in speech emotion recognition are acoustic features and can be grouped into two categories: prosodic features and spectral features (Zeng et al. 2009). Prosodic features (e.g., pitch, intensity, and speaking rate) have been reported to deliver important emotional cues (Brisson et al. 2014; Bahreini et al. 2016; Crumpton and Bethel 2015; Idris and Salam 2015), and have been extensively studied in previous work (Lee and Narayanan 2005; Schuller et al. 2004; Vlasenko et al. 2007). Spectral features [e.g., Mel-frequency Cepstral coefficients, and linear predictor Cepstral coefficients (Atal 1974)], on the other hand, have been studied to a lesser extent due to their lack of intuitive correlation with emotional states. However, since spectral features characterize speech signal in the frequency domain, they can provide emotion information from another perspective such as spectral energy distribution (Guzman et al. 2013), and show promising prospects in speech emotion recognition.

✉ Yongming Huang
huang_ym@163.com

[1] Laboratory of Measurement and Control of Complex Systems of Engineering, Southeast University, Nanjing, China

[2] Ministry of Education, School of Automation, Southeast University, Nanjing 210096, China

In most of the existing studies, acoustic analysis of speech signals is usually based on fourier transform (FT) or short time fourier transform (STFT). However, fourier-based analysis methods may not be suitable for speech signal analysis since speech signal is inherently non-stationary. As an alternative, the discrete wavelet transform (DWT) and wavelet packet transform (WPT), which are appropriate tools for non-stationary signal analysis (Mallat 2009), has received increasing research attention in the field of speech analysis during the past decade. As an extension version of DWT, WPT provides a family of frequency axis partition methods, and makes it easy to mimic the critical band (CB) structure of human auditory system (Fastl and Zwicer 1999) and enrich the way conventional Mel-scale filterbank divides the frequency band. Based on the CB structure of human auditory system, Karmakar et al. (2007) proposed a criterion for the design of optimal wavelet packet (WP) filterbank structure in speech and audio signal analysis. Another Mel filter-like WP structure developed by Farooq and Datta (2001) shows superiority over MFCC in unvoiced phoneme classification. Some studies have focused on the use of DWT and WPT in speech emotion recognition in the literature. Sarikaya and Gowdy (1997) proposed a new set of speech features based on Dyadic Wavelet Transform for stress classification and confirmed that the proposed features are suitable for this task. In (Kandali et al. 2009), Kandali et al. extracted different sets of features based on wavelet packet transform, and the proposed features show promising performance in text and speaker independent emotion recognition tasks. In this paper, we also base our study on WPT as it provides considerable advantages for speech signal analysis.

As we all know, wavelet packet (WP) is efficient in providing flexible and adaptive frequency band division methods (Stephane 2009), and is a prominent technique for quasi-periodic and non-stationary signal processing, such as speech processing. In this paper, the weighted WP-based acoustic features are proposed to combine with deep learning for speech emotion classification.

Although a lot of studies have been devoted to emotion recognition with clean speech, the real-world applications are always required to work in environments with various levels of noise. As a result, it is important to take noise robustness into consideration in the design of emotion recognition systems. However, few studies so far have been conducted to deal with this issue. In (Iliev and Scordilis 2011) Iliev and Scordilis studied the robustness of a set of glottal airflow features. With additive white Gaussian noise at the signal-to-noise ratio (SNR) of 10 dB, performances of 53 and 47% for four- and six-emotion tasks were achieved, respectively. You et al. (2006) proposed an enhanced Lipschitz embedding method for dimensionality reduction of acoustic features and evaluated 6-class emotion recognition performance of the proposed system under different SNR levels. Their results show promising prospects for robust speech emotion recognition; however, there is still a lack of investigation of acoustic features that are inherently noise-robust and are effective in emotion classification at the same time. As a result, this issue is worth further exploring.

While the sub-band energy-based acoustic features such as MFCC and the WP sub-band energy to be adopted in this paper provide good representations of speech spectral information, they are quite sensitive to noise (Zeng et al. 2008) and therefore are less efficient in distinguishing different emotional states from noisy speech. It has been proved by Paliwal (1998) that spectral sub-band centroids are relatively robust to noise and exhibit properties similar to formant frequencies. The noise-robust property of sub-band spectral centroids provides us with new ideas for improving robustness of acoustic features in emotion recognition. In this paper, we explore the combination of sub-band spectral centroids and WP sub-band energies to develop noise-robust features for the emotion recognition task. For this purpose, an auto-regressive (AR) method is adopted for sub-band spectral estimation and the derived sub-band spectral centroids are combined with WP sub-band energies via a weighting scheme. On this basis, a novel acoustic feature named sub-band spectral centroid weighted wavelet packet Cepstral coefficients (W-WPCC) is extracted for robust speech emotion recognition.

Usually, in a classification task like emotion recognition, fold cross-validation is applied to obtain the generalization performance. In reality, any SER system will have to deal with many unknown speakers. To get more realistic performance estimates, one should therefore apply a scheme like leave-one-speaker-out cross-validation. Such “speaker independent” cross-validation (SI-CV) will guarantee that no data from any test speaker is used for training the classifier (Ali Hassan et al. 2013).

By applying standard cross-validation (SI-CV) to speech taken from a single, uniform database, as is commonly done in almost all research in this area, many variables like microphone, room acoustics and language remain constant. However, this will not be the case if training and test data are recorded in different environments or two separate databases are used as training and test datasets. In such a scenario, along with different speakers, the two datasets will also have different acoustic environments or recording channels. It may even be that the languages spoken are different. These differences will have adverse effects on the real-world performance of the SER classifier, since its training will not have prepared it for data subsequently encountered in use. Several recent studies (Shamiand and Verhelst 2007; Tahon et al. 2015; Tahon and Devillers 2016; Shah et al. 2015; Deng et al. 2014) have reported results on inter-database emotion recognition by training on one or more available databases and testing on a left-out database. To improve the generalization capabilities of the SER classifier in such conditions, one should explicitly compensate for the speaker and acoustic differences between the training and test datasets.

In our previous work (Yongming et al. 2014a, b), we attempted to adopt different wavelet packet basis to experiment based on the support vector machine and found the wavelet packet basis which achieved the best performance in speech emotion recognition. In this paper, based on our previous work (Yongming et al. 2014a, b), we tried to find out the best DBNs feature learning model parameters. And then we tested the combinations of four different kinds of features, prosody feature, voice quality features, WPCC and W-WPCC to find out whether the performance could benefit from employing deep learning and improve the noise-robust property.

2 Related work

In the very recent years, a lot of studies have been devoted to emotion recognition with clean speech, the real-world applications are always required to work in environments with various levels of noise. As a result, it is important to take noise robustness into consideration in the design of emotion recognition systems. However, few studies so far have been conducted to deal with this issue. For the noise in the recording environment, the current researches mostly start with the noise reduction in the preprocessing. For example, Schuller et al. (2004) studied the influence of additive noise with different signal–noise levels on the speech recognition accuracy Zhou et al. (2001). In the recognition of stress and emotion, researchers added speech enhancement preprocessing to the background noise existing in emotional speech data. Yan Yonghong of China Institute of Acousticsproposed a method of emotion recognition in noisy environment. In (2011) Iliev and Scordilis studied the robustness of a set of glottal airflow features. With additive white Gaussian noise at the signal-to-noise ratio (SNR) of 10 dB, performances of 53 and 47% for four- and six-emotion tasks were achieved, respectively. You et al. (Mingyu et al. 2006) proposed an enhanced Lipschitz embedding method for dimensionality reduction of acoustic features and evaluated 6-class emotion recognition performance of the proposed system under different SNR levels. Their results show promising prospects for robust speech emotion recognition; however, there is still a lack of investigation of acoustic features that are inherently noise-robust and are effective in emotion classification at the same time. As a result, this issue is worth further exploring.

Recently, the applications of DBN or deep learning (DL) make breakthroughs in many difference areas (Bengio 2012). DBN represents a series of multi-layer architecture NNs that training with the greedy layer-wise unsupervised pre-training algorithms (Hinton and Salakhutdinov 2006; Bengio 2009). By applying the greedy layer-wise unsupervised pre-training mechanism, DBN can reconstruct the raw data set, in other words, DBN can "Learn" features from the original data (Wang and He 2004). Some people (Lee et al. 2014; Feng and Zheng 2015) study time delay systems, in order to reduce system turmoil, make the system stable and obtain less conservative results. Zhang et al. (2017) carried out experiments which showed that the DBN-based approach has good potential for practical usage and suitable feature fusions will further improve the performance of speech emotion recognition. Zhu et al. (2017) proposed a novel classification method that combines DBN and SVM instead of using only one of them. And the intelligent models, like classifiers usually can obtain higher accuracy and better generalization with the learned features.

3 Feature set

We proposed to construct a high-dimensional feature set, which is fused with prosody feature, voice quality features and spectral feature. In this paper, fundamental frequency (F_0), and power are extracted as prosody features and the first, second and third formants with their bandwidths are extracted as voice quality features. At the same time, wavelet packet Cepstral coefficients (WPCC) and the proposed sub-band spectral centroid weighted wavelet packet Cepstral coefficients (W-WPCC) compose the spectral feature. In this section, the WPCC and W-WPCC is described in detailed.

3.1 WPCC

In this paper, WPT is adopted for speech signal analysis instead of the widely used short-time Fourier transform for the reasons given in Sect. 1. For discrete input signal x(n) obtained with a sampling rate of f_s, before its WP coefficients can be calculated, it should first be associated to an approximation of a signal $\bar{x}(t)$ at the resolution $2^{J_0} = f_s^{-1}$, with decomposition coefficients $a_0(n)$ that satisfy

$$x(n) = f_s^{1/2} a_0(n) \approx \bar{x}\left(n \cdot f_s^{-1}\right). \tag{1}$$

Let $\mathcal{T}$ denote the binary tree structure, and each node in $\mathcal{T}$ is denoted as (j, p) in this paper, where j is the depth of the node in the tree and p ($0 \leqslant p \leqslant 2^j - 1$) is the number of nodes on its left at the same depth j. The root node (0, 0) of $\mathcal{T}$ is associated with coefficients $a_0(n)$. By applying (h, g) on the WP coefficients d_j^p at node $(j, p) \in \mathcal{T}$, the node is split into two child nodes $(j+1, 2p)$ and $(j+1, 2p+1)$, which are associated with WP coefficients d_{j+1}^{2p} and d_{j+1}^{2p+1}, respectively.

$$d_{j+1}^{2p}(n) = \sum_{r=-\infty}^{+\infty} h(r-2n)\, d_j^p(r), \tag{2}$$

$$d_{j+1}^{2p+1}(n) = \sum_{r=-\infty}^{+\infty} g(r-2n)\, d_j^p(r). \tag{3}$$

where $d_0^0 = a_0$.

A binary tree where each node has either zero or two children is called an admissible binary tree. By splitting the WP tree nodes in a particular way, a corresponding admissible binary tree structure is achieved. However, the resulting WP tree is not frequency ordered and a frequency ordering step should be taken to change the position of node (*j, p*) to (*j, q*) with $q = \mathrm{G}[p]$. After frequency ordering, the WP tree is frequency ordered and each node (*j, q*) has a frequency support of

$$I_j^q = \left[-(q+1)\pi 2^{-j},\ -q\pi 2^{-j}\right] \cup \left[q\pi 2^{-j},\ (q+1)\pi 2^{-j}\right]. \tag{4}$$

Let $L(\mathcal{T}) = \{(j_m, p_m)\}_{1\leqslant m\leqslant M}$ denote the set of leaf nodes of an admissible binary tree $\mathcal{T}$, the corresponding tree-structured WPT decomposes the original signal into a set of sub-band components with frequency supports $\left\{I_{j_m}^{q_m}\right\}_{1\leqslant m\leqslant M}$, where $q_m = \mathrm{G}[p_m]$; and this process can be viewed as an M-channel filter-bank followed by a set of aggregated down-samplers (Deng et al. 2014).

3.2 W-WPCC

3.2.1 W-WPCC feature extraction

The nature of the wavelet packet Cepstral coefficients (WPCC) is a kind of spectral features. Considering the complementarity between WPCC and sub-band spectral centroid which possesses noise-robustness, we combine them by some strategies to construct new speech emotion features possessing great noise-robustness.

We adopt a weighted method to realize the combination of sub-band spectral centroid and wavelet packet Cepstral coefficients, which forms the W-WPCC. On one hand, different sub-band energies of the speech signal combined with white noise, can be confused easily because of the affection of white noise. Relative to the sub-band energy, the sub-band spectral centroid is less affected by the noise. As a result, weighted wavelet packet Cepstral coefficients possess great robustness to the white noise of the speech signal. On the other hand, as a new description of the frequency domain distribution of the speech signal, sub-band energy weighted by sub-band spectral centroid contains important emotion information, which can distinguish emotion classes and construct speech emotion features.

3.2.2 Computation of W-WPCC

The algorithm for extracting the sub-band spectral centroid weighted wavelet packet Cepstral coefficient features (W-WPCC) is described as followed in Fig. 1.

4 Feature fusion for robust speech emotion recognition

In this section, we explored the problem of fusing different speech emotion features. WPCC and W-WPCC belong to the kind of spectral features. Since the prosody features, voice quality features, spectral features characterize the speech signals from different angles and they have complementarity between each other. As a result, combining different speech emotion features to compensate for the lack of a single feature on their emotion recognition ability by the complementarity is a method to improve the accuracy of the speech emotion recognition.

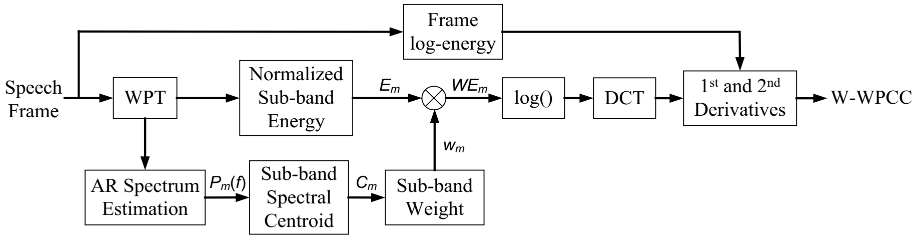
The features fused by DBNs are used for the emotion recognition. Figure 2 showed the proposed feature fusion block diagram.

5 Deep belief network

5.1 Restricted Boltzmann machine

The restricted Boltzmann machine (RBM) is a two-layer networking with one visible layer and one hidden layer. Figure 3 gives an illustration of RBM architecture. As shown in Fig. 3, the standard type of RBM has binary-valued *m* hidden and *n* visible neurons, and consists of a matrix of weights $W = \left(w_{i,j}\right)$ (size $m \times n$) associated with the connection between hidden neurons h_j and visible neuron v_i. The word "restricted" means that there is no connection between any two neurons in the same layer (Fig. 4).

Fig. 1 Block diagram of the W-WPCC feature extraction



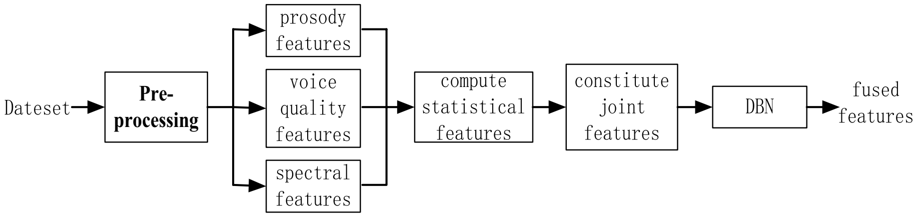


Fig. 2 Feature fusion block diagram

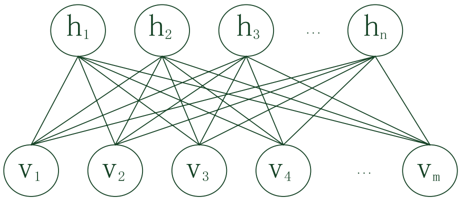


Fig. 3 The illustration of RBM architecture

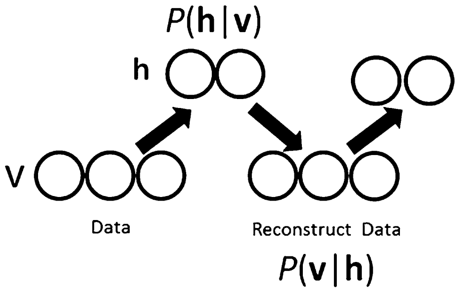


Fig. 4 The model of CD learning algorithm

5.1.1 The learning target of RBM

The learning target of RBM is Maximizing likelihood. It's a kind of model based on energy and the energy of joint configuration between visible variable $\mathbf{v}$ and hidden variable $\mathbf{h}$ is:

$$E(\mathbf{v},\mathbf{h};\theta) = -\sum_{ij} W_{ij}v_ih_j - \sum_{i} b_iv_i - \sum_{j} a_jh_j, \tag{5}$$

where W represents the weight of the side between the visible unit and the hidden unit, b and a represent the offset of the visible unit and the hidden unit respectively.

With the energy of joint configuration between $\mathbf{v}$ and $\mathbf{h}$, we can get the joint probability:

$$P_\theta(\mathbf{v},\mathbf{h}) = \frac{1}{Z(\theta)}\exp\left(-E(\mathbf{v},\mathbf{h};\theta)\right), \tag{6}$$

where $Z(\theta)$ represents the normalization factor.

According to the formula (5), we can turn the previous one into:

$$P_\theta(\mathbf{v},\mathbf{h}) = \frac{1}{Z(\theta)}\exp\left(\sum_{i=1}^{D}\sum_{j=1}^{F} W_{ij}v_ih_j + \sum_{i=1}^{D} v_ib_i + \sum_{j=1}^{F} h_ja_j\right). \tag{7}$$

We hope to maximize the likelihood function $P(v)$ of observation data. As a result, $P(v)$ can be obtained by calculate the edge distribution with the formula (7):

$$P_\theta(v) = \frac{1}{Z(\theta)}\sum_{\mathbf{h}}\exp[\mathbf{v}'W\mathbf{h} + \mathbf{a}'\mathbf{h} + \mathbf{b}'\mathbf{v}], \tag{8}$$

where '$'$' represents the vector transpose.

We can get the parameter of RBM by maximizing the value of $P(v)$, which equals maximizing $\log\left(P(v)\right) = L(\theta)$:

$$L(\theta) = \frac{1}{N}\sum_{n=1}^{N}\log P_\theta\left(v^{(n)}\right), \tag{9}$$

where $N$is the dimension of vector $\mathbf{v}$.

5.1.2 Learning method of limited Boltzmann machine

We can maximize $L(\theta)$ by the method of stochastic gradient descent. First, we need to figure out $L(\theta)$ derivative of W:

$$\frac{\partial L(\theta)}{\partial W_{ij}} = \frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial W_{ij}}\log\left(\sum_{\mathbf{h}}\exp\left[\mathbf{v}^{(n)\prime}W\mathbf{h} + \mathbf{a}'\mathbf{h} + \mathbf{b}'\mathbf{v}^{(n)}\right]\right). \tag{10}$$

After simplification:

$$\frac{\partial L(\theta)}{\partial W_{ij}} = E_{P_{data}}[v_ih_j] - E_{P_\theta}[v_ih_j], \tag{11}$$

where $\mathrm{E}_{P_\theta}[v_ih_j] = \sum_{\mathbf{v},\mathbf{h}} v_ih_jP_\theta(\mathbf{v},\mathbf{h})$.

The $E_{P_{data}}[v_ih_j]$ in the formula (11) can be easy to figure out that it only needs to calculate the average of v_ih_j among all data sets. But $\mathrm{E}_{P_\theta}[v_ih_j]$ need large amount of calculation to solve, because it involves all combination of $2^{|\mathbf{v}|+|\mathbf{h}|}$ between $\mathbf{v}$ and $\mathbf{h}$.In order to solve the calculation problem of $\mathrm{E}_{P_\theta}[v_ih_j]$, Hinton and some other people came up with contrastive divergence (CD), which is an efficient learning algorithm. The basic idea is shown in the picture below:

First, get the state of $\mathbf{h}$ according to the data of $\mathbf{v}$. Then, reconstruct the visible vector $\mathbf{v}_1$ by $\mathbf{h}$. After that, a new hidden vector $\mathbf{h}_1$ is generated by $\mathbf{v}_1$. When $\mathbf{v}$ is given, the activation status of each hidden unit $\mathbf{h}$ is independent. Otherwise, when $\mathbf{h}$ is given, the activation status of each visible unit $\mathbf{v}_i$ is independent as well.

That is:

$$P(\mathbf{h}|\mathbf{v}) = \prod_j P(h_j|\mathbf{v}). \tag{12}$$

The formula (12) can be factorization into:

$$P(h_j = 1|\mathbf{v}) = \frac{1}{1 + \exp\left(-\sum_i W_{ij}v_i - a_j\right)}. \tag{13}$$

At the same time: $P(\mathbf{v}|\mathbf{h}) = \prod_i P(v_i|\mathbf{h})$.

The formula (13) can be factorization into:

$$P(h_j = 1|\mathbf{v}) = \frac{1}{1 + \exp(-\sum_i W_{ij}v_i - a_j)} \tag{14}$$

The reconstructed visible vector $\mathbf{v}_1$ and hidden vector $\mathbf{h}_1$ is a sampling of $P(\mathbf{v}, \mathbf{h})$. The sample collection of multiple samplings can be seen as an approximate of $P(\mathbf{v}, \mathbf{h})$. Thus, the formula (11) can be solved.

5.2 BP algorithm

In 1986, Rumelhart came up with reverse propagation learning algorithm, which is also called backpropagation (BP) algorithm. BP algorithm is proposed to solve weighted coefficient optimization of multilayer forward neural networks. Therefore, BP algorithm usually hints that the neural network topology is a non-feedback multi-layer forward network. So, sometimes non-feedback multi-layer forward network is also called BP algorithm.

This algorithm can revise the weight coefficient of each layer in the network, so it applies to the learning of multilayer network. BP algorithm is one of the most extensive neural network learning algorithm at present.

5.2.1 The principle of BP algorithm

BP algorithm is used for feed-forward multilayer networks. Figure 5 showed the structure of it.

It contains the input layer, the output layer and the middle layer between the input and output layers. The middle layer has monolayer or multilayer, which is also called hidden layer Neurons in the hidden layer are also called hidden units. Although the hidden layer does not connect with the outside, the state can affect the relationship between input and output. That is to say, the change of the weight of the hidden layer will change the performance of the entire multilayer neural network.

5.2.2 The steps of the BP algorithm

When the back propagation algorithm is applied to the feed-forward multilayer network, the weight coefficient of the networkW$_{ij}$ can be recursively obtained by the following steps.

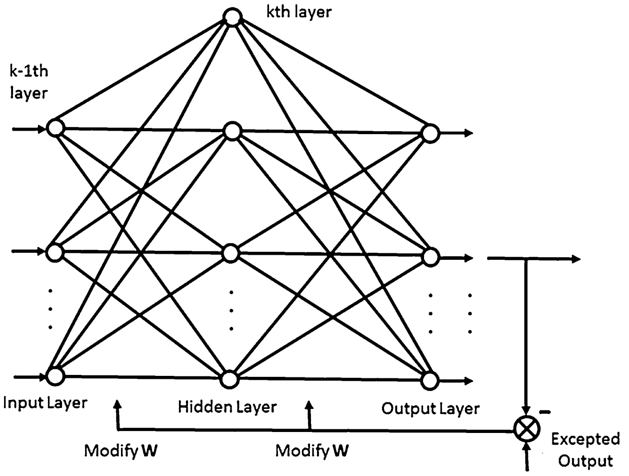


Fig. 5 The structure of learning algorithm for feed-forward multilayer network

Note the situation that there is a neuron on each part, then $i = 1, 2, ..., n$, $j = 1, 2, ..., n$. For the n neuron on the k layer, so there are n weight coefficients $W_{i1}, W_{i2}, \ldots, W_{in}$, in addition that take one more W_{in+1} in use of expressing threshold θ_i; and when enter sample X, take $X = (X_1, X_2, ..., X_n, 1)$.

Steps of the algorithm is show as follow:

Step 1 Set up initial values to W_{ij}. Set a small random number except 0 to each weight coefficient W_{ij}, $W_{i,n+1} = -\theta$.

Step 2 Input a sample:$X = (X_1, X_2, ..., X_n, 1)$, and corresponding excepted output $Y = (Y_1, Y_2, ..., Y_n)$.

Step 3 Calculate output of each layer:

Towards X_i^k, which represents the output of the ith neurons in the kth layer, there are some conditions that $U_i^k = \sum_{j=1}^{n+1} W_{ij}X_j^{k-1}$, $X_{n+1}^{k-1} = 1$, $W_{i,n+1} = -\theta$ and X_i^k=$f(U_i^k)$.

Step 4 Calculate the error of each layerd_i^k.

For the output layer $k = m$, the error can be calculated by the following equation:

$$d_i^m = X_i^m(1 - X_i^m)(X_i^m - Y_i). \tag{15}$$

For the other output layers, the errors can be calculated by the following equation:

$$d_i^k = X_i^k\left(1 - X_i^k\right) \cdot \sum_l W_{li} \cdot d_l^{k+1}. \tag{16}$$

Step 5 Modify W_{ij} and threshold θ_i with the following equation:

$$\Delta W_{ij}(t + 1) = -\eta d_i^k \cdot X_j^{k-1} + \alpha \Delta W_{ij}(t). \tag{17}$$

With the following equation as the condition:

$$W_{ij}(t + 1) = W_{ij}(t) - \eta \cdot d_i^k \cdot X_j^{k-1} + \alpha \Delta W_{ij}(t). \tag{18}$$

Among that:

$$\Delta W_{ij}(t) = -\eta \cdot d_i^k \cdot X_j^{k-1} + \alpha \Delta W_{ij}(t-1) \\ = W_{ij}(t) - W_{ij}(t-1). \quad (19)$$

Step 6 After finding out the weight coefficient of each layer, we can judge whether the calculation meet the given indicators.

If the results meet the demand, the algorithm ends up; else the algorithm processes return to Step 3 to carry out.

During the learning process, the given samples $X_p = (X_{p1}, X_{p2}, ..., X_{pn}, 1)$and expected output $Y_p = (Y_{p1}, Y_{p2}, ..., Y_{pn})$ need to be carried out until the calculations meet the requests of input and output.

5.3 Training process of DBN

As shown in Fig. 6, deep belief networks (DBNs) are probabilistic generative models stacked up by many layers of restricted Boltzmann machines, in which latent units are typically assigned with stochastic binary values.

The DBNs combine the acoustic features to a high-dimensional feature, which describe the relationships between speech emotion features. Besides, DBNs has powerful ability to learn relationships between features in high-dimensional space.

DBN is divided in two steps in the process of training the model:

The first step is pre-training Unsupervised train each layer of RBM network respectively to ensure that when eigenvectors map to different feature space, they can all retain the information of characteristics as much as possible.

The second step is fine tune Set up the BP network at the last level of the DBN.

The output feature vector of the RBM is received as its input feature vector, and supervised train the classifier. And each layer of RBM network can only ensure that the weight of the self layer to the layer of the eigenvector mapping achieves the optimal, but cannot ensure the entire DBN eigenvector mapping. As a result, the back-propagation network also spread the error message from top to bottom to each layer of RBM to adjust the entire DBN network slightly. The process of RBM network training model can be regarded as the initialization of a deep BP network weight parameter, which makes the DBN overcome the shortcomings of the BP network due to the random initialization of the weight parameter and easy to fall into the local optimal and long training time.

Figure 7 showed the process of DBN model training.

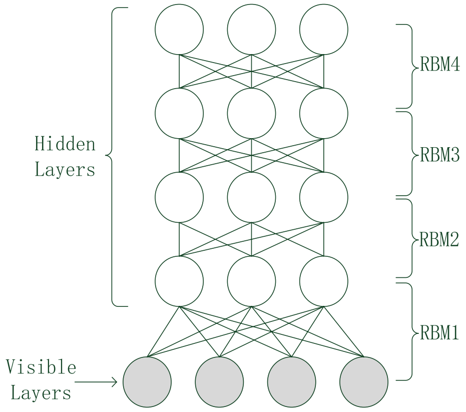


Fig. 6 Schematic diagram of DBNs

6 Proposed system

In this section, we described details of the proposed speech emotion recognition system with emphasis on the feature fusion. The following subsections give detailed description of each part of the proposed system. Framework of the proposed speech emotion recognition system is shown in Fig. 8.

6.1 Dataset division

The whole emotional speech database is divided into four parts. The DBNs training dataset tree-pruning dataset, training dataset and test dataset are used for DBNs training, WP

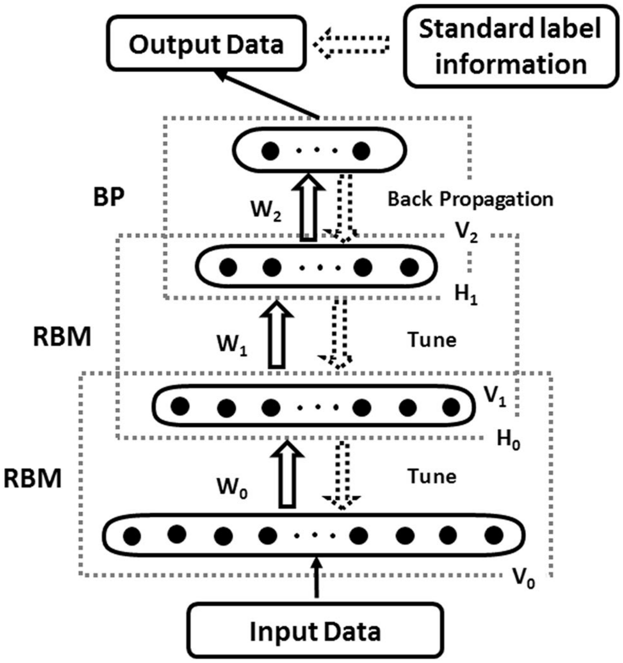


Fig. 7 Process of DBN model training

tree pruning, classifier training and emotion recognition respectively, and are not overlapped with each other.

6.2 Wavelet packet filter-bank construction

An optimal filter-bank structure is obtained using the tree-pruning algorithm. The optimal WP filter-bank structure is then applied on the training and test samples to calculate WP-based acoustic features. With the fast tree-pruning algorithm (Yongming et al. 2014a, b), a sequence of WP admissible trees with different number of leaf nodes is obtained, and correspondingly the set of filter-bank structures with different number of sub-bands. The obtained WP filter-bank structures are then used to calculate emotion-discriminative acoustic features from original speech signal.

6.3 Pre-processing

Before feature extraction, conventional speech signal processing operations including pre-emphasis, frame blocking and windowing are performed first. The speech signal is first pre-emphasized by a high-pass FIR filter 1–0.9375z^{-1} to spectrally flatten the signal and make it less susceptible to finite precision effects later in the signal processing (Wang and He 2004). The pre-emphasized speech signal is then blocked into frames of K samples with an overlap of K' between adjacent frames. Here we use $K = 256$ and $K' = K/2$. And each individual frame is multiplied by a Hamming window to reduce ripples in the spectrum.

6.4 DBNs training

In the pre-training process of neural networks, the learning rate was set to 0.02, mini-batch size was set to 256 and the weight cost was set to 0.0002. In the training process of one RBM, the momentum was started at 0.5 and raised to 0.9.

The input layer is the 594-dimensional acoustic features, which is formed by combining 429-dimensional W-WPCCs, 66-dimensional prosody features and 99-dimensional voice quality features (with the window context of 11 frames). The number of dimensions of output layer is 141.

A softmax regression is added to the output layer to ensure that the output probabilities sum up to 1.0, after which the network was trained discriminatively using back propagation algorithm. About 15% of the training data was selected as CV set during fine-tuning. We set mini-batch size to 128 and learning rate to 0.8 at the beginning. After each iteration, the performance of the system was evaluated on the CV set. If the performance showed not enough improvements, the learning rate was halved for the next iteration.

6.5 Classifier

Support vector machine (SVM) is adopted for speech emotion classification in this paper. The implementation of the SVM classifier is provided by a publicly available Matlab toolbox named LIBSVM Matlab Toolbox (Chang and Lin 2011).

7 Experiment

7.1 Emotional speech database and experimental setup

The proposed speech emotion recognition system is evaluated on the Berlin emotional speech database (Burkhardt et al. 2005), which contains seven simulated emotions (anger, boredom, disgust, fear, joy, neutral and sadness). In this paper, six emotions (no disgust) with a sum of 489 utterances are used for the classification task. 20% of the database is randomly selected to form the tree-pruning dataset, 20% of the database is randomly selected to train the DBNs structure and we apply fivefold cross validation on the remaining 60% utterances to assess the classification performance.

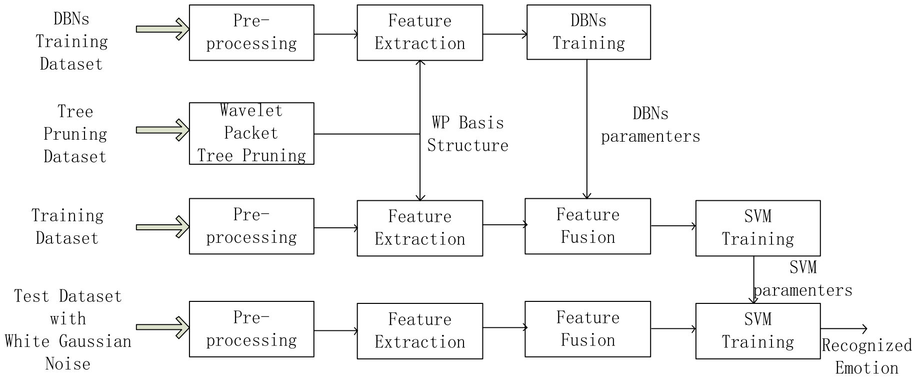


Fig. 8 Block diagram of the proposed system

7.2 Experiments with noisy speech

The set of experiments considers noise-robustness of the fused feature. In these experiments, noisy speech is obtained by adding white Gaussian noise at 10 db SNR levels to the test data and the training data is kept clean. To study the importance of the size of the hidden layers in the speech emotion recognition, we changed the size of the hidden layers in the experiments. At the same time, different kinds of combinations of prosody features, voice quality features, WPCC and W-WPCC are also computed and used as features in this set of experiments to study the role of the proposed features in emotion recognition of noisy speech.

7.3 Experimental results

To evaluate performance of the extracted feature, a set of experiments are conducted and the results are presented in our previous work (Yongming et al. 2014a, b). WP filterbank structures generated by coif 3 achieved the highest accuracy rate in speech emotion recognition compared with other wavelet packets.

We believed the recognition rate of the system varies when the size of the hidden layers is changed. So we changed the size of the hidden layers to 512 and 2048 in experiments. Figure 9 summarizes the effects of different layer size on the performance of system under the noisy condition (Figs. 10, 11).

According to the above figure, the system with a layer size of 1024 and 2048 gained better results. In the recognition system architecture with hidden units of 2048, pretraining the first two layers of RBMs was enough for this emotion recognition task and the results stayed stable when increasing the number of hidden layers. We found the best performance occurred when the size of the networks was 1024 and the number of hidden layers was five, so we fixed them and then investigated the performance of varying the acoustic features that input to the neural networks.

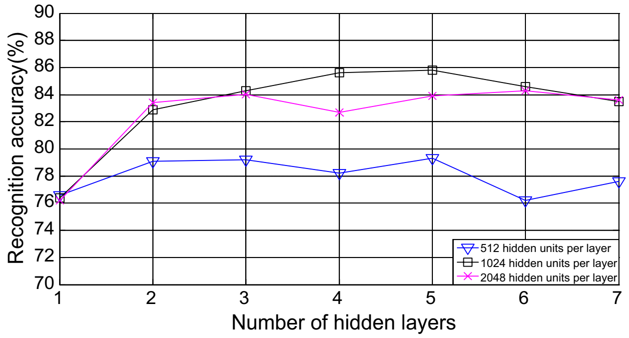


Fig. 9 Emotion recognition rate using fusion features, with different sizes of hidden layers and hidden units per layer

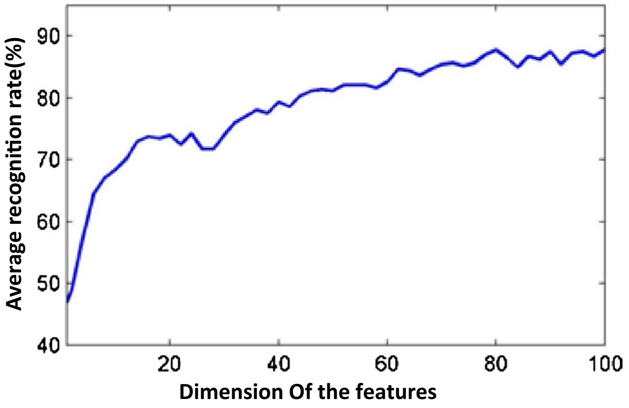


Fig. 10 Different mixed feature dimensions for speech emotion recognition

We can see from the above figures that the emotion recognition accuracy with the fused features, which combined with prosody features, voice quality features, WPCC and W-WPCC, achieved the best performance under noisy conditions. A highest classification accuracy of 86.60% is achieved for the proposed feature fusion method. So we fixed the size of the networks, the number of hidden layers and the fused features to compare with the conventional emotion recognition method that extracting WPCC features and the recognition as we did in our previous work (Yongming et al. 2014a, b).

From Tables 1, 2 and 3, we can know that the recognition accuracy of our proposed method has been improved compared with conventional emotion recognition method. In a conclusion, the acoustic feature set with W-WPCC fused by the DBNs, whose size of the networks was 1024 and number of hidden layers was five, improved 5.48% recognition accuracy under noisy conditions.

8 Conclusion and future work

In this paper we explored the sub-band spectral centroid weighted wavelet packet Cepstral coefficients (W-WPCC) based acoustic feature fusion approach combined with DBNs for speech emotion recognition under noisy conditions. We tried different sizes of hidden layers (512, 1024, 2048), different number of DBNs layers and different types of acoustic features combinations to model the emotion recognition system. The emotion recognition system using deep learning performed better than the conventional systems just using SVM [q] as the classifiers under noisy conditions. Future work also includes investigating more robust deep learning models. Apart from this, seeking for robust feature representation is also considered as part of

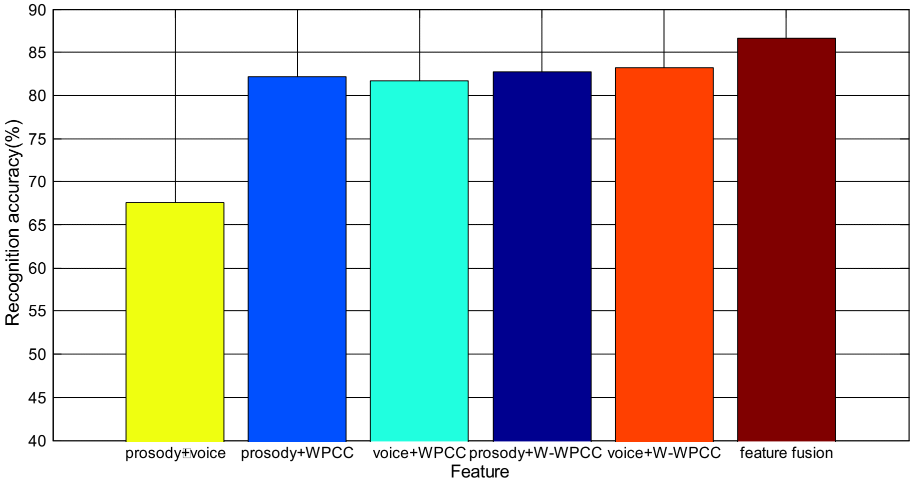


Fig. 11 Emotion recognition rate using combinations of prosody features, voice quality features, WPCC and W-WPCC

Table 1 Confusion matrix with conventional emotion recognition method

Emotion	Anger	Boredom	Fear	Joy	Neutral	Sadness
Anger	**89.22%**	0.00%	2.94%	7.84%	0.00%	0.00%
Boredom	0.00%	**84.62%**	1.54%	0.00%	6.15%	7.69%
Fear	1.82%	1.82%	**76.36%**	7.27%	5.45%	7.27%
Joy	28.07%	0.00%	14.04%	**57.89%**	0.00%	0.00%
Neutral	0.00%	4.76%	3.17%	0.00%	**88.89%**	3.17%
Sadness	0.00%	16.00%	0.00%	0.00%	2.00%	**82.00%**

Using confusion matrix with conventional emotion recognition method, the recognition rates of Anger, Boredom, Fear, Joy, Neutral and Sadness were 89.22%, 84.62%, 76.36%, 57.89%, 88.89%, 82.00%

Average recognition rate: 81.12%

Table 2 Confusion matrix with proposed emotion recognition method

Emotion	Anger	Boredom	Fear	Joy	Neutral	Sadness
Anger	**94.46%**	0.00%	1.96%	3.58%	0.00%	0.00%
Boredom	0.00%	**89.53%**	1.54%	0.00%	2.17%	6.76%
Fear	1.82%	1.82%	**81.54%**	4.76%	3.76%	6.29%
Joy	23.75%	0.00%	11.73%	**65.52%**	0.00%	0.00%
Neutral	0.00%	3.72%	1.16%	0.00%	**93.47%**	1.64%
Sadness	0.00%	11.73%	0.00%	0.00%	1.07%	**87.34%**

Using confusion matrix with proposed emotion recognition method, the recognition rates of Anger, Boredom, Fear, Joy, Neutral and Sadness were up to 94.46%, 89.53%, 81.54%, 65.52%, 93.47%, 87.34%

Table 3 Confusion matrix with different emotion recognition method

Eight fetures	Recognition rate (%)						
	Anger	Boredom	Fear	Joy	Neutral	Sadness	Average
Cadence	68.76	58.71	41.53	33.06	52.84	55.33	53.8
Acoustic	76.17	55.67	39.09	46.23	45.87	58.84	56.1
WPCC	75.55	68.56	60.59	52.59	68.12	70.41	67.1
W-WPCC	74.21	70.23	63.45	57.98	69.05	71.54	68.7
Mixed	74.84	71.91	66.32	63.38	69.98	72.68	70.4

the ongoing research, as well as efficient classification techniques for automatic speech emotion recognition.

Acknowledgements An earlier version of this paper was presented at the International Conference on Network, Communication and Computing (ICNCC 2016).This work was supported by open Fund of Jiangsu Province Natural Science Foundation (No. BK20140649) and National Natural Science Foundation (No. 61503081, No. 61473079).

References

Ali Hassan R, Damper, Niranjan M (2013) On acoustic emotion recognition: compensating for covariate shift. IEEE Trans Audio Speech Lang Process 21(7):1458–1468

Atal BS (1974) Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J Acoust Soc Am 55(6):1304–1312

Bahreini K, Nadolski R, Westera W (2016) Towards multimodal emotion recognition in e-learning environments. Inter Learning Environ 24(3):590–605

Bengio Y (2009) Learning deep architectures for AI. Now Publ Inc 2(1):67–76

Bengio Y (2012) Deep learning of representations for unsupervised and transfer learning. J Mach Learning Res Proc Track 27(2), 17–36

Brisson J, Martel K, Serres J, Sirois S, Adrien JL (2014) Acoustic analysis of oral productions of infants later diagnosed with autism and their mother. Inf Ment Health J 35(3):285–295

Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W, Weiss B (2005) A database of german emotional speech. In: proceeding interspeech 2005, ISCA, pp 1517–1520

Caponetti L, Buscicchio CA, Castellano G (2011) Biologically inspired emotion recognition from speech. Eurasip J Adv Signal Process 2011(1):1–10

Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol (TIST) 2(3), pp 1–27

Crumpton J, Bethel CL (2015) A survey of using vocal prosody to convey emotion in robot speech. Int J Social Robot 8(2):271–285

Deng J, Xia R, Zhang Z, Liu Y (2014) Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition. Icassp IEEE international conference on acoustics, pp 4818–4822

Farooq O, Datta S (2001) Mel filter-like admissible wavelet packet structure for speech recognition. Signal Process Lett IEEE 8(7):196–198

Fastl H, Zwicer E (1999) Psychoacoustics Facts and Models[M], 2nd edn. Springer, New York

Feng Z, Zheng WX (2015) On extended dissipativity of discrete-time neural networks with time delay. IEEE Trans Neural Netw Learning Syst 26(12):3293–3300

France DJ, Shiavi RG, Silverman S et al (2000) Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans Biomed Eng 47(7):829–837

Guzman M, Correa S, Munoz D et al (2013) Influence on spectral energy distribution of emotional expression. J Voice 27(1):129.e1–129.e10

Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313(5786):504–507

Idris I, Salam MS (2015) Voice quality features for speech emotion recognition. J Info Assur Secur 10(4):183–191

Iliev AI, Scordilis MS (2011) Spoken emotion recognition using glottal symmetry. Eurasip J Adv Sig Process 2011(1):1–11

Kandali AB, Routray A, Basu TK (2009) Vocal emotion recognition in five native languages of Assam using new wavelet features. Int J Speech Technol 12(1):1–13

Karmakar A, Kumar A, Patney RK (2007) Design of optimal wavelet packet trees based on auditory perception criterion. Ieee Signal Process Lett 14(4):240–243

Lee CM, Narayanan SS (2005) Toward detecting emotions in spoken dialogs. IEEE Trans Speech Audio Process 13(2):293–303

Lee TH, Park MJ, Park JH, Kwon OM, Lee SM (2014) Extended dissipative analysis for neural networks with time-varying delays. IEEE Trans Neural Netw Learning Syst 25(10):1936–1941

Mallat SA (2009) Wavelet tour of signal processing, 3rd edn. Academic Press, Burlington

Malta L, Miyajima C, Kitaoka N et al. (2009) Multimodal estimation of a driver's spontaneous irritation. Intelligent vehicles symposium, 2009 IEEE, pp 573–577

Mingyu Y, Chun C, Jiajun B et al. (2006) Emotion recognition from noisy speech. In: multimedia and expo, IEEE international conference on 2006, pp 1653–1656

Morrison D, Wang RL, De Silva LC (2007) Ensemble methods for spoken emotion recognition in call-centres. Speech Commun 49(2):98–112

Paliwal KK (1998) Spectral subband centroid features for speech recognition. Acoustics, speech and processings. Proceedings of the IEEE international conference on 1998, pp 617–620

Petrushin V (2000) Emotion recognition in speech signal experimental study, development, and application. ICSLP 2000, Beijing, pp 222–225

Sarikaya R, Gowdy JN (1997) Wavelet based analysis of speech under stress[C]. Southeastcon '97. engineering new century., proceedings IEEE, pp 92–96

Schuller B, Rigoll G, Lang M (2004) Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture[C]. Acoustics, speech, and signal processing, proceedings (ICASSP '04). IEEE international conference on 2004, pp I-577–580

Shah M, Chakrabarti C, Spanias A (2015) Within and cross-corpus speech emotion recognition using latent topic model-based features. Eurasip J Audio Speech Music Process 2015(1):1–17

Shamiand M, Verhelst W (2007) Anevaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. Speech Commun 49(3):201–212

Stephane M (2009) A wavelet tour of signal processing, 3rd edn. Academic Press, Burlington

Tahon M, Devillers L (2016) Towards a small set of robust acoustic features for emotion recognition: challenges. IEEE ACM Trans Audio Speech Lang Process 24(1):16–28

Tahon M, Sehili MA, Devillers L (2015) Cross-corpus experiments on laughter and emotion detection in HRI with elderly people. In: International Conference on Social Robotics, vol 31. Springer, pp 633–642

Vlasenko B, Schuller B, Wendemuth A et al (2007) Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing[C]. Affect Comp Intell Interact Proc 781:139–147

Wang X, He Q (2004) Enhancing generalization capability of svm classifiers with feature weight adjustment. International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, vol 3213. Springer, Heidelberg, pp 1037–1043

Yongming H, Ao W, Guobao Z, Yue L (2014a) Speech emotion recognition based on coiflet wavelet packet Cepstral coefficients. Chinese conference on pattern recognition, pp 436–443

Yongming H, Guobao Z, Yue L, Ao W (2014b) Improved emotion recognition with novel task-oriented wavelet packet features, vol 8588. In: 10th international conference, ICIC 2014, Taiyuan, China, August 3–6, pp 706–714

Zeng ZH, Tu JL, Pianfetti BM et al (2008) Audio-visual affective expression recognition through multistream fused HMM[J]. IEEE Trans Multimed 10(4):570–577

Zeng ZH, Pantic M, Roisman GI et al (2009) A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans Pattern Anal Mach Intell 31(1):39–58

Zhang WS, Zhao DH, Chai Z, Yang LT, Liu X, Gong FM, Yang S (2017) Deep learning and SVM-based emotion recognition from Chinese speech for smart affective services. Softw Pract Exp 47(8):1127–1138

Zhou GJ, Hansen JHL, Kaiser JF (2001) Nonlinear feature based classification of speech under stress. IEEE Trans Speech Audio Process 9(3):201–216

Zhu LZ, Chen LM, Zhao DH, Zhou JH, Zhang WS (2017) Emotion recognition from chinese speech for smart affective services using a combination of SVM and DBN. Sensors 17(7):1694