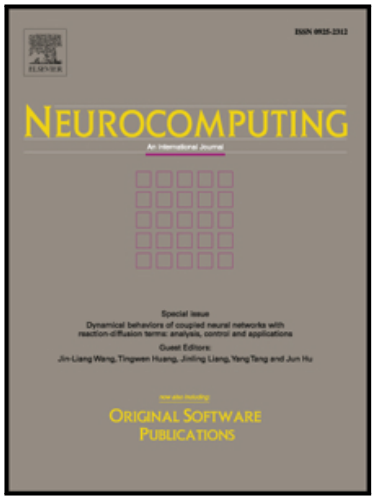
Communicated by Dr. Nianyin Zeng

Accepted Manuscript



Speech emotion recognition based on feature selection and extreme learning machine decision tree

Zhen-Tao Liu, Min Wu, Wei-Hua Cao, Jun-Wei Mao, Jian-Ping Xu, Guan-Zheng Tan

PII: S0925-2312(17)31356-5
DOI: 10.1016/j.neucom.2017.07.050
Reference: NEUCOM 18746

To appear in: *Neurocomputing*

Received date: 15 April 2017
Revised date: 3 July 2017
Accepted date: 31 July 2017

Please cite this article as: Zhen-Tao Liu, Min Wu, Wei-Hua Cao, Jun-Wei Mao, Jian-Ping Xu, Guan-Zheng Tan, Speech emotion recognition based on feature selection and extreme learning machine decision tree, *Neurocomputing* (2017), doi: 10.1016/j.neucom.2017.07.050

Speech emotion recognition based on feature selection and extreme learning machine decision tree☆

Zhen-Tao Liu[a,b], Min Wu[a,b], Wei-Hua Cao[a,b,*], Jun-Wei Mao[a,b], Jian-Ping Xu[a,b], Guan-Zheng Tan[c]

[a]*School of Automation, China University of Geosciences, Wuhan 430074, China*
[b]*Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, Wuhan 430074, China*
[c]*School of Information Science and Engineering, Central South University, Changsha 410083, China*

Abstract

Feature selection is a crucial step in the development of a system for identifying emotions in speech. Recently, the interaction between features generated from the same audio source was rarely considered, which may produce redundant features and increase the computational costs. To solve this problem, feature selection method based on correlation analysis and Fisher is proposed, which can remove the redundant features that have close correlations with each other. To improve the recognition performance of the feature subset after proposal feature selection further, an emotion recognition method based on extreme learning machine (ELM) decision tree is proposed according to the confusion degree among different basic emotions. A framework of speech emotion recognition is proposed and the classification experiments based on proposed classification method by using Chinese speech database from institute of automation of Chinese academy of sciences (CASIA) are performed. And the experimental results show that the proposal achieved 89.6% recognition rate on average. By proposal, it would be fast and efficient to discriminate emotional states of different speakers from speech, and it would make it possible to realize the interaction between speaker-independent and

☆This work was supported by the National Natural Science Foundation of China under Grants 61403422 and 61273102, the Hubei Provincial Natural Science Foundation of China under Grant 2015CFA010, the 111 project under Grant B17040, and the Fundamental Research Funds for National University, China University of Geosciences (Wuhan).

*Corresponding author: Wei-Hua Cao, email: weihuacao@cug.edu.cn

computer/robot in the future.

Keywords: Speech emotion recognition, Feature selection, Correlation analysis, Decision tree, Extreme learning machine

1. Introduction

To understand and convey each other's intentions in a natural way, human-computer interaction (HCI) has been paid more attentions in recent years [1]. The primary problem that the HCI faces is how to master the ability of identifying emotional information accurately, which is similar to the emotional intelligence capability in human-robot interaction [2]. As a fast and easy-understand way for communication, speech signal was considered to identify emotions [3]. It is believed that speech conveys not only syntactic and semantic contents of the linguistic sentences but also the emotional states of humans [4]. Thus, human emotion recognition using speech signal is feasible, which mainly studies on how to identify the emotional or physical states of humans from his/her voice automatically [5].

In speech emotion recognition, one of the central research issues is how to select an optimal feature set from speech signals [6]. Most of the previous works on speech emotion recognition have been devoted on the analysis of speech prosodic features and spectral information [7]. And some novel feature parameters are used for speech emotion recognition, such as the Fourier parameters [8]. Although there are many acoustic parameters have been proven to contain emotional information, little success has been achieved in realizing such a set of features that performs consistently over different conditions [9]. Thus, most researchers prefer to use mixing feature set that is composed of many kinds of features containing more emotional information [10]. However, using mixing feature set may give rise to high dimension and redundancy of speech features, thereby it makes the learning process complicated for most machine learning algorithms and increases the likelihood of overfitting. Therefore, feature selection is indispensible to reduce the dimensions redundancy of features.

Speech Emotion Recognition (SER) can be regarded as a static or dynamic classification problem, which makes SER a superb test bed for investigating and adopting various deep learning architectures [11, 12]. However, for the application of speech emotion recognition technology, most of them are based on the application environment of small-scale and small sample. Deep

learning that consumes a large number of tagged data isn't fully applicable in speech emotion recognition. Thus, the study of speech emotion recognition based on machine learning algorithm still occupies a very important position [3, 4, 13]. And Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), and several other one-level standard techniques have been explored for the classifier [14]. Satisfactory results are obtained by standard classifiers, however their performance improvements are usually limited. Fusion, combination, and ensemble of classifiers could represent a new step towards better emotion recognition systems [15].

In last ten years, some works using a combination of standard methods have been presented. Two classification methods (i.e., stacked generalization and unweighted vote) are applied to emotion recognition of six emotional classes in [16],by which the classification performance is improved modestly compared to the traditional method. A two-stage classifier for five emotions is proposed in [17], in which a support vector machine (SVM) is used to classify five emotions into two groups, after that, HMMs are used to classify emotions within each group. In [18], Bayesian logistic regression and SVM classifiers in a binary decision tree are used. A binary multi-stage classifier guided by the dimensional emotion model is proposed in [19]. A true comparison among the results of the previously mentioned methods is very difficult because they have used different corpus, training/test partitions, etc. [20]-[22].

Although the mentioned methods above are useful for recognizing specific emotions, there is no sufficiently effective method to describe complicated emotional states and the current speech emotion recognition technology is still immature in people's real life, because of two main reasons. The one reason is that the redundance of speech emotion information leads to the fall of final recognition rate and the long training time of sample data, and the data processing of many speech information slows down the real-time feedback of the designed system. Another reason is that the general efficiency of the algorithms based on speaker-independent features which can be applied to speech emotion recognition is not high and it also affects the practicability of speech emotion recognition technology.

In view of the above analysis, a framework of speech emotion recognition is proposed to reduce the influence of individual differences and enhance the practicability of speech emotion recognition. Firstly, a relative appropriate speech emotional feature set with speaker-independent characteristics and rich emotional information is selected based on previous studies, and then the feature selection method based on correlation analysis and Fisher criterion

is proposed to reduce the feature redundancy. Finally, ELM decision tree based on the confusion degree of basic emotion is presented for the selected representative speech emotional feature set. Based on the above presented methods, a framework of speech emotion recognition from feature extracting to emotion classification is proposed.

A series of contrast experiments are performed to verify the effectiveness of the proposed method on CASIA Chinese speech database. The experiments of ELM decision tree and support vector machine (SVM) decision tree are carried out, respectively using the speech emotional feature set without feature selection. In contrast, the same experiments are carried out respectively using the speech emotional feature set after feature selection based on correlation analysis and Fisher criterion. In addition, comparison experiments between the proposal and two kinds of traditional classification methods (i.e., k-nearest neighbours (KNN) and BP neural network) are performed using the feature set after feature selection.

The rest of paper is organized as follows. Framework of speech emotion recognition is proposed in Section 2. Analysis and extraction of speech emotional features are presented in Section 3. Feature selection of speech emotional features is given in Section 4. The binary classification method based on ELM decision tree is introduced in Section 5. Experimental results and analysis are given in Section 6.

2. Framework of speech emotion recognition

Speech emotion recognition mainly consists of three processes, i.e., feature extraction, feature selection, and emotion classification. To improve the accuracy, a framework of speech emotion recognition is proposed in this paper as shown in Fig. 1. Firstly, initially feature set including speaker-independent features and speaker-dependent features is obtained after feature exaction. Secondly, the extracted features are compared by correlation analysis and optimal feature subset is obtained after feature selection. Finally, speech emotion is classified into several categories using the selected features.

To make it more clear, we introduce each process in Fig. 1 respectively. In the feature extraction, the mixing feature set including speaker-independent features (i.e., emotional features such as average change rate of fundamental frequency that could eliminate the impact of individual differences) and speaker-dependent features (i.e., emotional features such as rhythm characteristics that are easily influenced by the speaker's personal characteristics) is

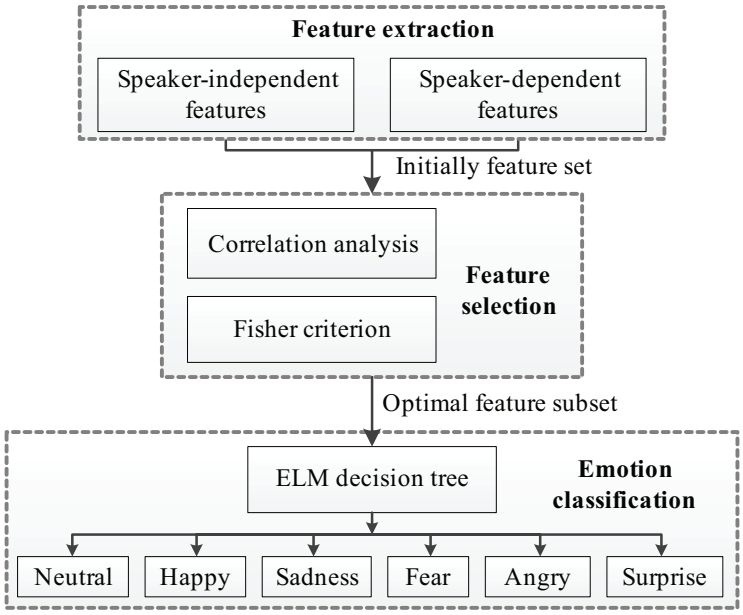


Figure 1: Flowchart of speech emotion recognition.

obtained from preprocessed speech signal samples. In the feature selection, correlation analysis and Fisher criterion are used to gain the lower redundancy features and reduce the dimension of selected feature set, in which correlation analysis method consists of Distance analysis, Partial correlation analysis, and Bivariate correlation analysis. In the emotion classification, a decision tree is constructed for speech emotion recognition by comparing the value of confusion degree among six kinds of basic emotions (i.e., neutral, angry, surprise, happy, fear, and sadness), in which ELM is adopted as the binary classifier.

3. Analysis and extraction of speech emotional feature

In present studies, most of the speech emotion recognition can only be trained for specific people of a certain emotion database, thus there are problems that the speech emotion recognition system does not have strong generality and the emotion recognition accuracy of the speaker who is not included in the training emotion database is greatly reduced. Therefore, how to eliminate the individual difference in speech is the key step to improve the

recognition rate of speech emotion recognition and the universality of speech emotion recognition system.

3.1. Speaker-dependent features

It is generally believed that the emotional features such as rhythm characteristics are easily influenced by the speaker's personal characteristics and data distribution of feature set will change greatly for different speakers with similar emotional states. And these features are summarized as speaker-dependent features of the speech, which usually contain rich personal emotional information [3], as shown in Table 1, including fundamental frequency, short-time energy, MFCC coefficient, spectral energy dynamic coefficients, etc. These emotional features contain vast quantities of emotional information, and they reflect personal speech characteristics of the speaker. Even in the same emotional state, the speed and the volume of different speaker may have great difference.

Table 1: Speaker-dependent features.

Speaker-dependent features	Specific characteristics	
Prosodic features	Fundamental frequency characteristic	Fundamental frequency; Fundamental frequency maximum; Four bit value of fundamental frequency
	Energy related features	Short-time maximum amplitude; Short-time average energy; Short-time average amplitude
	Time length correlation characteristic	Short-time average cross zero ratio; Speed of speech
Quality characteristics	Breath sound; Brightness; Throat sound; The maximum value and the average value of the first, two and three formant frequencies	
Spectrum characteristics	12 order MFCC coefficients; Spectral energy dynamic coefficients of 12 equally spaced frequency bands	

3.2. Speaker-independent features

In order to suppress the individual characteristics of different people and eliminate the differences in the numerical values of individualized speech emotion characteristics, the change rate is introduced to reflect the changes in

the process of speaking. Derivative is adopted to obtain speaker-independent speech emotional feature [3]. In [21], the researchers determine speaker-independent speech emotional feature that includes the average change rate of fundamental frequency, the change rate of short-time energy frequency, and so on. According to the characteristics of these features, the main speaker-independent features are summarized in Table 2, which contain a certain amount of emotional information.

Table 2: Speaker-independent features.

Speaker-independent features	Specific characteristics	
Prosodic features	Fundamental frequency characteristic	Average change rate of fundamental frequency; Fundamental frequency standard deviation; Four bit point frequency change rate
	Energy related features	The average rate of short-time energy and Short-time energy amplitude
	Time length correlation characteristic	Partial time ratio
Quality characteristics	The average variation and standard deviation of the first, second and third formant frequency ratio; Each sub-point value of the first, second and third formant frequency change rate	
Spectrum characteristics	First order difference MFCC coefficient; Second order difference MFCC coefficient	

The reason of speaker-independent features is in its infancy is its emotional information isn't as rich as speaker-dependent features. The final contribution of the recognition rate is less than the speaker-dependent features, but the recognition effect is better than the speaker-dependent features under the condition that the speaker is change. Taking the fact that the contribution on the recognition rate of speaker-independent features is relatively small into account, this paper extracted speaker-independent features and part of speaker-dependent features for the follow-up research.

4. Feature selection of speech emotional features

If the performance of a feature is overall evaluated in advance based on a metric, it will be very valuable for the selection of subsequent feature. The

most significant part of the feature evaluation is the correlation analysis between the feature or feature subsets, and the common evaluation method is to use the relevancy as an evaluation criterion to select a feature or feature subset [23]. In this paper, the feature selection contains correlation analysis for speech emotion recognition and feature dimension reduction based on Fisher criterion. The speech emotional features with stronger ability of signifying emotion are preliminarily selected through distance-analysis, partial-correlation analysis, and binary correlation analysis, after that, the Fisher criterion is used for dimension reduction of emotional features.

4.1. Correlation analysis process

Firstly, Euclidean distance analysis is carried out in the correlation analysis, by which the features are divided into several clusters. Then, the partial correlation analysis of each feature in the group are carried out. After the results of the reflection of each feature in the group are obtained, the Spearman rank correlation analysis of the similar emotion is respectively carried out for the similar features to obtain the specific correlation analysis results, thus the representative speech emotion characteristics are selected. The whole analysis model is shown in Fig. 2.

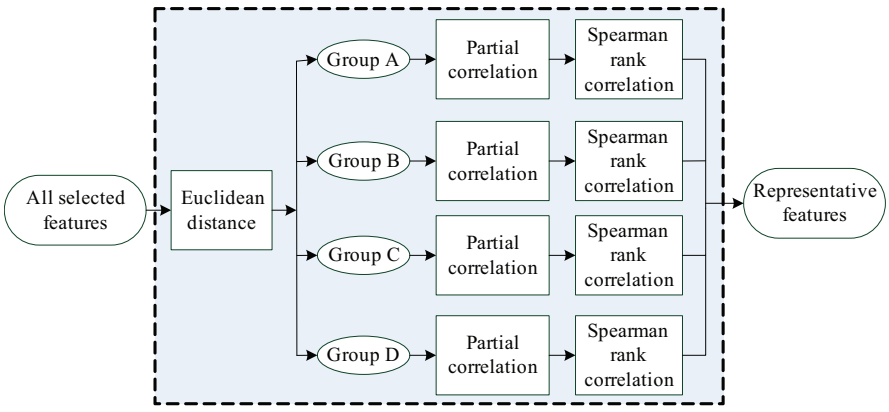


Figure 2: Correlation analysis for speech emotion recognition.

According to the results of the correlation analysis, the redundant speech emotional features are discarded, and the representative speech emotional features that reflects emotion obviously are input into the following feature selection algorithm to reduce the computational complexity.

4.1.1. Distance analysis

Emotional state is characterized by several speech emotional features, but it is difficult to use so many features at the same time to recognize the individual emotions. Thus, we need to know what kinds of features have a greater impact on the emotions, and control these features. First of all, all feature variables are classified to determine which variables are relatively close, and then assign these similar characteristics into a group according to the distance analysis. In this paper, owing to the equal interval and no similar features of speech emotion, the Euclidean distance [23] is used for distance analysis. Euclidean distance represents the true distance Q between two points in the n dimensional space.

$$\begin{aligned} Q &= \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \\ &= \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}, \end{aligned} \tag{1}$$

where x_n, y_n are points in the n dimensional space. And the features with the closest proximity are obtained as the similar features.

4.1.2. Partial correlation analysis

In fact, many emotional features reflect the same characteristics of a class of emotional states. And there may be mutual restraint or mutually reinforcing relationship between features, which make it difficult to determine the influence of speech features on the emotional state. Therefore the features that may affect other features should be removed or controlled before correlation analysis between the features and emotions.

Analysis of above process refers to the Partial correlation analysis [24] that is known as a net correlation analysis, which is the process of analyzing linear correlation between two variables under the control of other variables affected by the condition of linear. When the number of the control variables is zero, the partial correlation coefficient is the correlation coefficient. Assuming that there is a group of independent variable $\{X_1, X_2, ..., X_n\}, i, j = 1, 2, ..., n$, Partial correlation coefficient calculated as follows. Firstly, the correlation

matrix of simple Partial correlation coefficient p_{ij} is calculated as

$$R = (\rho_{i*j})_{n*n} = \begin{pmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1\mathrm{n}} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2\mathrm{n}} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{\mathrm{n}1} & \rho_{\mathrm{n}2} & \cdots & \rho_{\mathrm{nn}} \end{pmatrix}. \tag{2}$$

Then inverse matrix of R is calculated as

$$R^{-1} = (\lambda_{i*j})_{n*n} = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1\mathrm{n}} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2\mathrm{n}} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{\mathrm{n}1} & \lambda_{\mathrm{n}2} & \cdots & \lambda_{\mathrm{nn}} \end{pmatrix}. \tag{3}$$

At last, Partial correlation coefficient of X_i and Y_j is calculated as

$$\gamma_{ij} = \frac{-\lambda_{ij}}{\sqrt{\lambda_{ii}\lambda_{jj}}}. \tag{4}$$

Partial correlation coefficient represents two variables interdependence when other elements exist in the model.

4.1.3. Bivariate correlation analysis

To calculate the correlation coefficient between the two variables, Pearson product-moment correlation coefficient [25], Spearman rank correlation [26], and Kendall coefficient of concordance [27, 28] could be used. However, speech emotion features that are extracted from the sub-frame speech signal in each interval are completely rank discrete variables, for which the Spearman rank correlation method is employed in this paper.

Spearman rank correlation coefficient is an index that measures the statistical correlation between two variables, and it is used to evaluate the relationship between two variables under monotonic function. In the absence of duplicate data, if a variable is strictly monotonic function of another variable, the Spearman rank correlation coefficient between the two variables is +1 or -1, called variable completely Spearman correlation. Assuming that initial data has been arranged in order from large to small, recording x_i, y_j are data's location after arrangement, x_i, y_j are rank of variables X and Y, thus $d_{ij} = x_i - y_j$ are rank difference of x_i , y_j.

If there is no same rank, it can be calculated by

$$\rho_S = 1 - \frac{6\sum {d_{ij}}^2}{n(n^2-1)}. \tag{5}$$

If there is same rank, then it is needed to calculate the linearly dependent coefficient as

$$\rho_S = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - Y)}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}. \tag{6}$$

Spearman rank correlation coefficient is often referred to non-parametric correlation coefficient. At first, as long as the X and Y have a monotonous or non-linear function, then X and Y are completely Spearman correlation, which is different from Pearson correlations that are entirely related only when two variables have a linear relationship. Second, Spearman rank correlation coefficient is to obtain an accurate distribution of samples between X and Y without knowing the joint probability density function of them.

4.2. Speech emotional feature dimension reduction based on Fisher criterion

When using the statistical methods for pattern recognition, there are always many issues involved the data dimension, the method that can work in low-dimensional space is difficult to be applied into the high-dimensional space. And the method of solving low-dimensional issue is generally low computational complexity, efficient, and convenient [29]. After obtaining the representative features through correlation analysis, these extracted data sets represented by a set of high-dimensional speech features are used to train and test pattern classifiers. The well-known curse of dimensionality often emerges, thus removing irrelevant features, as an important preprocessing step to a classifier, is needed [30].

The traditional linear dimensionality reduction methods including principal component analysis (PCA) and Fisher criterion have been successfully used for reducing the dimensionality of emotional speech features [31]. As one of the linear dimensionality reduction mentions, Fisher criterion, overcomes the problems that PCA is not able to extract the discriminant embedded information from high-dimensional emotional features. Four groups of comparative experiments including Fisher + SVM, PCA + SVM, Fisher + ANN, PCA + ANN were realized and the experimental results proved that Fisher is

better than PCA for dimension reduction [32]. According to above comparison, Fisher criterion is employed for dimension reduction of speech emotional features in this paper.

4.2.1. Fisher criterion

To obtain the each characteristic components ability of distinguishing emotions, the Fisher rate of each characteristic component is calculated. Then we compare the values of ratios to select the optimal features. Fisher criterion function [33]-[35] is represented as

$$\lambda_{Fisher} = \frac{\sigma_{between}}{\sigma_{within}}, \tag{7}$$

where λ_{Fisher} is the Fisher rate of feature components, $\sigma_{between}$ represents variance between class of feature components, in other words, different phonetic features of the mean variance. σ_{within} represents variance inside features component, that is, the mean of the same speech characteristics within the same component of the variance. $\sigma_{between}$, σ_{within} are calculated by Eq. (8) and Eq. (9).

$$\sigma_{between} = \sum_{i=1}^{M} \left({m_k}^{(i)} - m_k \right)^2, \tag{8}$$

where i is the symbol of specific class. M is the number of feature class. k represents the dimensions of feature, and also represents the k class for emotion. ${m_k}^{(i)}$ represents the mean of all samples of the i class feature and k dimension. m_k represents all the mean of all samples of the k dimension.

$$\sigma_{within} = \sum_{i=1}^{M} \frac{1}{n_i} \sum_{j=1, c \in w_i}^{n_i} \left({c_k}^{(ij)} - {m_k}^{(i)} \right)^2, \tag{9}$$

where n_i represents the sample number of the i class, j represents the symbol of samples, w_i represents the sample gather of i class, ${c_k}^{(ij)}$represents the k dimensional value of the i class feature of j sample.

4.2.2. Procedure of Fisher criterion

The procedure of Fisher criterion is designed according to Fisher criterion and the specification of feature class.

Step 1: Input the feature class S to criterion arithmetic;

Step 2: Calculate the Fisher rate of each dimension according to Fisher criterion arithmetic;

Step 3: Calculate the sum of Fisher rate class λ_{Fisher};

Step 4: Order the feature of S according to Fisher from large to small, and select the sequence of S to T according to the sequenced order until $\lambda_{Fisher} < \frac{1}{n_i} SUM$, n is the number of S element;

Step 5: Acquire the feature class T after selection.

5. Speech emotion recognition based on ELM decision tree

5.1. Extreme learning machine

Extreme learning machine(ELM) is a single-hidden layer feedforward Neural Network which is improved from the gradient algorithm [36, 37]. Compared with other gradient learning algorithms, there is no need for ELM to refresh parameters in Neural Network by repeated iteration, and it is only required to initialize connection weights between the input layer and output layer randomly before training the data, as well as the bias parameters in the hidden layer [38]. After setting the number of neurons in the hidden layer, unique optimal solution can be obtained. This kind of learning algorithm can reduce the data processing time and improve the learning speed. Fig. 3 shows the single-hidden layer feed forward Neural Network which is composed of the sequential connection of input layer, hidden layer, and output layer.

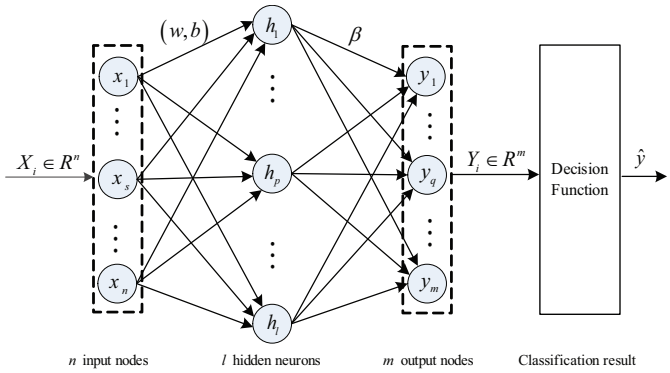


Figure 3: Single-hidden layer feed forward neural network.

In the structure of ELM, there are N random samples (X_i, t_i), where $X_i = [x_{1i}, x_{2i}, \cdots, x_{ni}]^T \in R^n$, $t_i = [t_{1i}, t_{2i}, \cdots, t_{qi}, \cdots, t_{mi}]^T \in R^m$. A single

hidden layer Neural Network with L hidden nodes can be expressed as

$$\sum_{i=1}^{L} \beta_i g(W_i \cdot X_j + b_i) = o_j, j = 1, \cdots, N, \tag{10}$$

wherein, $g(x)$ is the activation function of hidden neuron, $W_i = [w_{i,1}, w_{i,2}, \cdots, w_{i,n}]^T$ is the input weight, and β_i is the output weight, b_i is the i^{th} neuron's bias in hidden layer. The target of single-hidden feedforward Neural Network is to minimize the error between actual output and desired output, which can be expressed as

$$\sum_{j=1}^{N} \|o_j - t_j\| = 0. \tag{11}$$

Namely, there are proper β_i, W_i, b_i which can satisfy the Eq. (11). This formula can be represented as

$$\mathbf{H}\beta = \mathbf{T}, \tag{12}$$

where $\mathbf{H}$ is the output matrix of hidden layer, β is the output weight, and $\mathbf{T}$ is the expected output. To train the single hidden Neural Network, the proper $\widehat{w}_j$, $\widehat{b}_j$, $\widehat{\beta}_j$ need to be found, which is equal to solve the minimal cost function as

$$E = \sum_{i=1}^{Q} \left[\sum_{j=1}^{l} \beta_i G(w_j \cdot X_i + b_j) - T_i \right]^2. \tag{13}$$

In the ELM algorithm, once the input weight W_i and the bias of the hidden layer b_i are randomly determined, the output matrix of the hidden layer $\mathbf{H}$ is uniquely determined. The training of ELM can be transformed into solving a linear system in Eq. (12).

The above single ELM is a universal classifier which can realize both classification and regression. In field of classification, ELM is sufficient for binary classification and multi-classification. It does not need iterative learning which can enhance the learning speed greatly. ELM algorithm can be defined as following steps.

Step 1: Give the training set $\psi = (x_i, t_i)|x_i \ inR^n, t_i \in R^m, i = 1, \cdots, N$, activation function $G(x)$, the number of hidden neuron L;

Step 2: Randomly assign value of input weight w_i and the bias b_i;

Step 3: Calculate the hidden layer output matrix $\mathbf{H}$;

Step 4: Calculate the output weight β: β=$\mathbf{H}$†$\mathbf{T}$, where the $\mathbf{H}$†is Moore-Penrose generalized inverse of hidden output matrices $\mathbf{H}$.

5.2. Classification method of speech emotion recognition

SVM and ELM algorithm are originally designed for the classification of the two values [39], it's vital to construct a suitable multi-class classifier for multi-class recognition problem of speech emotion classification. In this paper, "one to many" method is adopted, which is based on the binary decision tree hierarchical speech emotion recognition method. This method is based on the confusion degree between a class of emotions and other categories of emotions. The confusion degree between the two groups is

$$D_{L1,L2} = \frac{\sum_{i=1}^{i=m} \sum_{j=1}^{j=n} D_{ij}}{mn}, \tag{14}$$

where D_{ij} represents the existing emotional marked group, $D_{L1,L2}$ represents the average value of the confusion degree between two groups. The smaller confusion degree between two groups is, the greater difference between the emotion groups is, which means that it's easier to distinguish. Conversely the larger confusion degree represents opposite case.

In the decision tree, when the emotion category of the upper node is selected, the confusion degree among the rest of the emotional state will change. Thus, we need to calculate the confusion degree between the emotion and the remaining emotional state at each level of the decision tree. And according to the method in [40, 41], we can get the confusion degree between different emotions in the decision tree, as shown in Table 3.

Table 3: Confusion degree among basic emotions in each level of decision tree.

Layer \ Emotion	Happy	Sadness	Surprise	Angry	Fear	Neutral
First layer	1.462	1.734	2.23	2.578	2.96	3.723
Second layer		1.541	2.178	2.781	3.522	4.231
Third layer			2.934	3.843	4.765	5.589
Fourth layer				3.693	8.421	8.423
Fifth layer					10.8	10.8

According to the confusion degree between emotional categories, a decision tree based on binary classifiers can be obtained. The decision tree is constructed according to the confusion degree between an emotional state and the rest of the emotional states. In order to reduce the cumulative loss

of the decision tree, the emotion with small confusion degree is placed on higher node of the binary tree to identify and the emotion with large confusion degree is placed on the lower nodes of the decision tree, as shown in Fig. 4, in which 'happy', 'sadness', 'surprise', 'angry', 'fear' and 'neutral' are identified sequentially according to the confusion degree between each basic emotion in each level of decision tree.

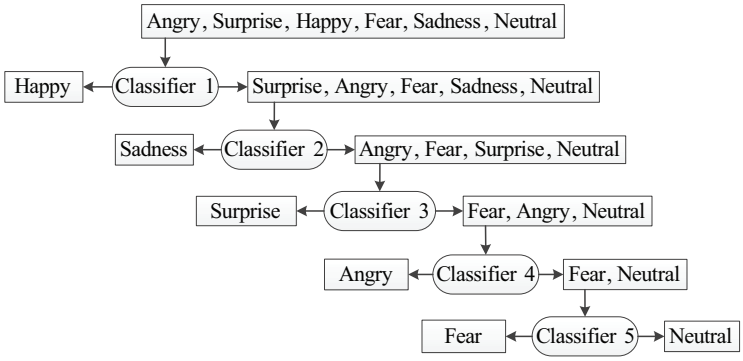


Figure 4: Binary decision tree.

In this paper, a binary classification method based on ELM decision tree is proposed to realize speech emotion recognition, and the experiment of SVM decision tree classification is performed as well to compare with ELM decision tree. How to choose the activation functions and the optimal parameters for ELM and SVM are critical. Since the ELM in the decision tree is used as a binary classifier, the sigmoid function is selected as activation function. In addition, the neurons number of hidden layer in ELM is determined by comparing the experimental results of multi-groups ELM classification with different number of neurons.

Radial basis function (RBF) is a kernel function that only needs to determine a kernel parameter. Therefore, the Gaussian kernel function that is the most commonly used RBF is adopted for the SVM classifier in this paper. The parameters in SVM including penalty parameters c and kernel functions parameters g, are optimized by the cross validation method in the LIBSVM toolbox[42].

6. Experimental results and discussion

In this paper, the general process of the speech emotion recognition including the following steps. First of all, the development tools of Praat

phonetic software and MATLAB R2012b are used to realize the extraction of speech emotional features, in which the speaker-independent features and the speaker-dependent features are extracted separately. Then, these extracted emotional data are analyzed by correlation analysis to get rid of the redundant information, and the Fisher criterion is adopted to realize the finally feature selection of the representative speech emotional feature using SPSS 22.0 statistical software. Finally, the recognition algorithm based on ELM decision tree is used for realize speech emotion recognition, which is realized by Microsoft Visual Studio 10.0 program and Matlab R2012b program. The experiments are carried out on the computer of the 32-bit windows7 system, and CPU is the dual-core Intel CORE i5, clocked is 2.4 GHz, and running memory is 3.16 G.

In the early stages of the experiments, feature extraction and feature selection are carried out based on the proposed method, respectively. And the optimized parameters in both SVM and ELM are obtained by trial and error using the speech subset consists of 240 emotional speech randomly selected from CASIA. In the classification experiment, six groups of classification experiments using the speech emotional feature set are performed. The experiments of ELM decision tree and SVM decision tree are carried out using the speech emotional feature set without feature selection, respectively. In contrast, the same experiments are performed using the speech emotional feature set after feature selection based on correlation analysis and Fisher criterion, respectively. In addition, comparison experiments between the proposal and two kinds of traditional classification methods (i.e., KNN and BP neural network) are carried out using the feature set after feature selection.

6.1. Speech database

CASIA Chinese emotion corpus recorded by the Institute of Automation [43], Chinese Academy of Sciences, is recorded by four people (i.e., two men and two women) in a clean recording environment (SNR is about 35 dB), deducting 300 basic emotional short utterances, adopting 16 kHz sampling, 16 bit quantified. It contains Mandarin utterances of five basic emotions (i.e., surprise, happy, sad, angry, and fear) and neutral.

In the experiment, the training samples are randomly selected from the CASIA Chinese emotion database, in which there are 10 samples of each speaker with each emotion, i.e., 360 samples of emotional speech in total. For the test samples, each speakers speech samples are used as a test set (i.e., four test sets in total), and there are 20 samples of each speaker with

each emotion, i.e., each test set includes 120 samples of emotional speech in total.

6.2. Experiment using feature set without feature selection

After the pre-processing of the speech signals, the speech emotional feature set in the test training sample are extracted, in which the feature set consists of selected speaker-independent speech emotional features and a small amount of speaker-dependent speech emotional features under consideration of emotion information of speaker-independent features is sufficient. And the selection result is shown in Table 4, in which the emotional feature set without feature selection is consisted of 34 kinds of speech emotion features in total.

Table 4: Emotional feature set without feature selection.

Feature type	Statistical characteristics	Characteristic number	Total
Fundamental frequency class	Fundamental frequency; Fundamental frequency maximum; Fundamental frequency variation range; Fundamental frequency change rate; Fundamental frequency standard deviation	5	34
Energy class	Short-time average energy; Short-time average amplitude; The average change rate of short-time energy and short-time energy	4	
Time length class	Short-time average zero crossing rate; Silent part time and sound partial time ratio	2	
Resonance peak class	The maximum and mean value of the first, two and third formant frequencies; The average frequency and the standard deviation of the first, second and formant frequency	15	
Spectrum class	1-6 order MFCC coefficients; First, Second order differential MFCC coefficients	8	

For these features, the speech emotional feature data are extracted from the test samples to compose the set of feature vectors, then these set are respectively input into the ELM decision tree and SVM decision tree to realize the classification experiment. The time of accomplish speech emotion recognition and the accuracy of the two groups of experiment are shown in Table 5, in which the recognition time is the time from the input of the

selected feature vector set to the output of the result. As shown in Table 5, when using the feature set without feature selection, the average recognition accuracy of ELM decision tree is up to 88.251%, which is about 1.2% higher than SVM decision tree. And recognition time of ELM decision tree is about 1.2 s less than SVM decision.

Table 5: Experimental results of ELM decision tree and SVM decision tree using feature set without feature selection.

Group	ELM decision tree		SVM decision tree	
	Average recognition rate (%)	Average recognition time (s)	Average recognition rate (%)	Average recognition time (s)
Wang(M)	88.251	1.232	87.634	2.478
Liu(W)	87.112	1.248	84.879	2.62
Zhao(M)	86.892	1.237	85.113	2.593
Zhao(W)	86.793	1.251	83.745	2.947

6.3. Experiment using feature set after correlation analysis and Fisher criterion

In this experiment, the speech emotional feature data is extracted from the test samples that are firstly divided into two feature groups that contain the speaker-independent features and the speaker-dependent features. And after that, correlation analysis are carried out respectively. Then, the corresponding feature set in each feature group through the correlation analysis are consist of a new representative speech emotional feature set with lower output redundancy.

For the two feature group through correlation analysis that contain the speaker-independent samples and the speaker-dependent samples, the Fisher criterion is used to select the speech emotional features respectively, after that, the emotional feature set after feature selection is consisted. The feature selection results are shown in Table 6, in which the feature set after correlation analysis and Fisher criterion consists of 20 kinds of speech emotional features that contain fundamental frequency class, energy class, time length class, resonance peak class, and spectrum class. The category in this feature set is reduced by 14 compare with the initially speech emotional feature set. Similarly, the classification experiments of ELM decision tree and

Table 6: Emotional feature set after feature selection.

Feature type	Statistical characteristics	Characteristic number	Total
Fundamental frequency class	Fundamental frequency; Fundamental frequency range; Average change rate of fundamental frequency	3	20
Energy class	Short-time average energy; Short-time average amplitude; Short-time energy, Short-time energy average amplitude change rate	4	
Time length class	Short-time average zero crossing rate; Silent part time and sound partial time ratio	2	
Resonance peak class	The mean value and dynamic range of the first and third formant frequencies; The average change rate in the first and third formant frequencies	6	
Spectrum class	1-4 order MFCC coefficients; First order differential MFCC coefficients	5	

SVM decision tree using emotional feature set after feature selection are respectively performed, the time of accomplishing speech emotion recognition and the accuracy of the results are shown in Table 7.

Table 7: Experimental results of ELM decision tree and SVM decision tree using feature set after feature selection.

Group	ELM decision tree		SVM decision tree	
	Average recognition rate (%)	Average recognition time (s)	Average recognition rate (%)	Average recognition time (s)
Wang(M)	90.431	1.647	87.733	3.021
Liu(W)	89.217	1.652	88.972	3.104
Zhao(M)	89.86	1.639	86.539	2.987
Zhao(W)	88.893	1.657	85.548	3.11

As shown in Table 7, when using the feature set after feature selection, the average recognition accuracy of ELM decision tree is up to 90.431%, which is about 2.7% higher than SVM decision tree. And recognition time of ELM decision tree is about 1.3 s less than SVM decision. In the same time, the results have some difference compare with the result in Table 5 owing to adopt the feature set after feature selection. And the experiments of two kinds of traditional classification methods (i.e., KNN and BP neural

network) are performed using the feature set after feature selection, and the experimental results are shown in Table 8, in which classification accuracy of ELM decision tree is about 2.4% higher than other classification algorithms.

Table 8: Experimental results of different classification method using feature set after feature selection.

Classification method	Average Recognition Rate(%)
ELM decision tree	89.6
SVM decision tree	87.2
BPNN	82.3
KNN	80.7

6.4. Comparative analysis of experimental results

In the flowing contrastive figures, "ELM" represents the classification experiment of ELM decision tree using the feature set without feature selection, "SVM" represents the classification experiment of SVM decision tree using feature set without feature selection, "C-ELM" represents the classification experiment of ELM decision tree using feature set after the feature selection, "C-SVM" represents the classification experiment of SVM decision tree using feature set after the feature selection. The data of average recognition rate of the above experiment results in the Table 5 and the Table 7 are plotted as shown in Fig. 5, from which the superiority of ELM decision tree from the comparison of the experiments respectively adopted different recognition algorithm (i.e., ELM decision tree and SVM decision tree) and same feature set are clearly illustrated.

According to the comparative results based on the different feature set whether the correlation analysis and the Fisher criterion are adopted, the recognition rate of the experiments using the feature set after feature selection is relatively lower. This is mainly because the extracted speech feature set in this paper is not enough, the feature vector dimension that finally input recognition algorithm is insufficient after the feature selection based on correlation analysis and Fisher criterion, which affects the final recognition accu-

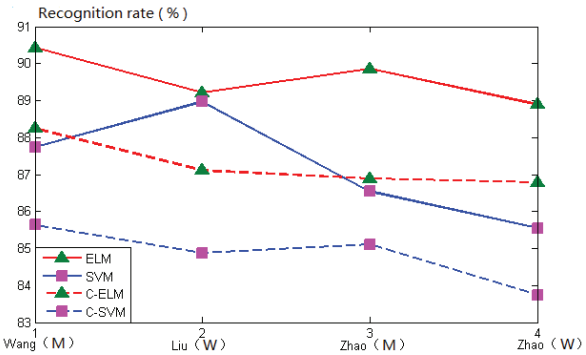


Figure 5: Comparison of average recognition rate of four groups of experiments.

racy. Another reason is that the emotion information of speaker-independent speech emotion features is insufficient.

The data of average recognition time of the above experiment results in Table 5 and Table 7 are plotted as shown in Fig. 6, from which the superiority of ELM decision tree in recognition speed from the comparison of the experiments adopted different recognition algorithm (i.e., ELM decision tree and SVM decision tree) and same feature set are clearly illustrated. According to the comparative results based on decision tree adopted different binary classifiers, the ELM is more suitable for the decision tree algorithm based on the confusion degree between emotional categories.

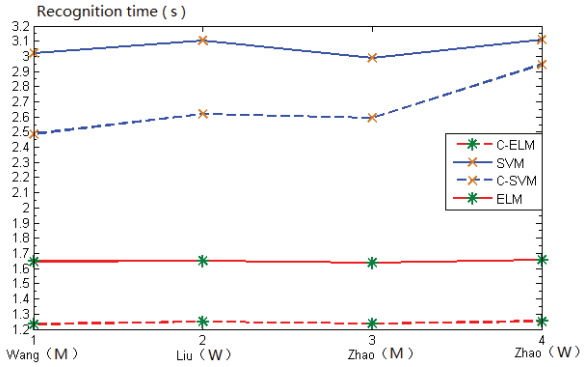


Figure 6: Comparison of average recognition time of four groups of experiments.

Similarly, according to the comparative results based on the different feature set, the recognition time of the experiments using feature set after

feature selection consist of correlation analysis and Fisher criterion is relatively lower, by which the utility of the feature selection based on correlation analysis and the Fisher criterion is fully verified. This is because mainly because the feature selection of initially selected speech emotion features based on correlation analysis and Fisher criterion reduces information redundancy of the relative emotional features, and the number of the selected features is reduced, so that the operation time of the recognition algorithm can be shortened effectively, by which the validity of the feature selection is demonstrated.

7. Conclusion

In this paper, a framework of speech emotion recognition from feature extraction to emotion classification was proposed according to the fact that speech emotion recognition is difficult to be applied in human-computer interaction, in which the feature selection method based on correlation analysis and Fisher criterion, and the ELM decision tree recognition method based on confusion degree among basic emotions were proposed. And the validity of the proposal was verified through a series of contrast experiments respectively, which contained four groups of experiment under different experimental condition. The ELM is more suitable for the decision tree algorithm through the experimental comparison of recognition rate or recognition time under different experimental condition. And the utility of the feature selection based on correlation analysis and the Fisher criterion was fully verified through the experimental comparison of recognition rate under different emotional feature set.

In future research, some aspects of the experiment would be improved. Firstly, some novel speech emotional features that contain sufficient emotional information such as Teager energy operator feature [44] would be extracted. In addition, other feature extraction methods, e.g., deep learning [45], would be adopted in our future research. Secondly, we will be studying on feature selection based on evolutionary computation, by which the information of emotional labels in feature set can be fully utilized. Thirdly, different speech databases could be taken into account to verify the practicability of the proposed method.

On account of the practicability excellent results that the proposed approach achieved, the proposal could be applied into human-robot interaction, internet teaching system, virtual reality environment etc., and the proposed

method in this paper would be used in the multi-modal emotion recognition system that is developed in [46].

References

References

[1] M. Song, M. You, N. Li, A robust multimodal approach for emotion recognition, *Neurocomputing.* 71 (10) (2008) 1913-1920.

[2] P. Salovey, J.D. Mayer, Emotional intelligence, *Magination, Cognition, and Personality.* 9 (3) (1990) 185-211.

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

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

[5] WS. Wang, JB. Wu, Speech emotion recognition in emotional feedback for Human-Robot Interaction, *International Journal of Advanced Research in Artificial Intelligence.* 4 (2) (2011) 20-27.

[6] B. Schuller, S. Steidl, A. Batliner, et al, The interspeech 2010 paralinguistic challenge, *In Interspeech.* (2010) 2794-2797.

[7] A.B. Ingale, D.S. Chaudhari, Speech emotion recognition, *International Journal of Soft Computing and Engineering.* 2 (1) (2012) 235-238.

[8] K. Wang, N. An, B.N. Li, Speech emotion recognition using Fourier parameters, *IEEE Transactions on Affective Computing.* 6 (1) (2015) 69-75.

[9] F. Eyben, K. Scherer, B. Schuller, et al, The Geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing, *IEEE Transactions on Affective Computing.* 7 (2) (2016) 190-202.

[10] M. Tahon, L. Devillers, Towards a small set of robust acoustic features for emotion recognition: Challenges, *IEEE/ACM Transactions on Audio, Speech, and Language Processing.* 24 (2016) 16-28.

[11] W. Liu, Z. Wang, X. Liu, et al. A survey of deep neural network architectures and their applications, *Neurocomputing.* 234 (2016) 11-26.

[12] N. Zeng, Z. Wang, H. Zhang, et al. Deep belief networks for quantitative analysis of a gold immunochromatographic strip, *Cognitive Computation.* 8 (4) (2016) 684-692.

[13] N. Zeng, H. Zhang, W. Liu, et al. A switching delayed PSO optimized extreme learning machine for short-term load forecasting, *Neurocomputing.* 240 (2017) 175-182.

[14] A.I. Iliev, M.S. Scordilis, J.P. Papa, Spoken emotion recognition through optimum-path forest classification using glottal features, *Computer Speech Language.* 24 (3) (2010) 445-460.

[15] El.M. Ayadi, M.S. Kamel, F. Karray, Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure, *Knowledge-Based Systems.* 63 (3) (2014) 68-81.

[16] D. Morrison, R. Wang, L.C.D. Silva, Ensemble methods for spoken emotion recognition in call-centres, *Speech Communication.* 49 (2) (2007) 98-112.

[17] L. Fu, X. Mao, L. Chen, Speaker independent emotion recognition based on SVM/HMMs fusion system, *International Conf. on Audio, Language and Image Processing.* (2008) 61-65.

[18] C. Lee, E. Mower, C. Busso, Emotion recognition using a hierarchical binary decision tree approach, *Interspeech 2009, Brighton, UK.* 43 (2) (2009) 320-323.

[19] Z. Xiao, E. Dellandrẽa, W. Dou, Recognition of emotions in speech by a hierarchical approach, *International Conference on Affective Computing and Intelligent Interaction.* (2009) 312-319.

[20] Z.T. Liu, K. Li, D.Y. Li, Emotional feature selection of speaker-independent speech based on correlation analysis and Fisher. *The 34th Chinese Control Conference.* 35 (15) (2015) 3780-3784.

[21] Q.R. Mao, X.L. Zhao, Y.Z. Zhan, Extraction and analysis for non-personalized emotion features of speech, *Advances in information Sciences and Service Sciences(AISS).* 3 (10) (2011) 255-263.

[22] J. Rybka, A. Janicki, I. Giannoukos, Comparison of Speaker Dependent and Speaker Independent Emotion Recognition, *International Journal of Applied Mathematics and Computer Science.* 23 (4) (2013) 797-808.

[23] J. Aparicio, J.T. Pastor, On how to properly calculate the euclidean distance-based measure in DEA, *Optimization.* 63 (3) (2014) 421-432.

[24] L. Chen, S.K. Zheng, Studying Alternative Splicing Regulatory Networks through Partial Correlation Analysis, *Genome Biology.* 10 (1) (2009) 1-20.

[25] T.R. Derrick, B.T. Bates, J.S. Dufek, Evaluation of time-series data sets using the Pearson product-moment correlation coefficient, *Medicine and science in sports and exercise.* 26 (7) (1994) 919-928.

[26] J.H. Zar, Significance testing of the Spearman rank correlation coefficient, *Journal of the American Statistical Association.* 67 (339) (2015) 578-580.

[27] R. Baumgartner, R. Somorjai, R. Summers, Assessment of cluster homogeneity in fMRI data using Kendalls coefficient of concordance, *Magnetic Resonance Imaging.* 17 (10) (1999) 1525-1532.

[28] P. Sedgwick, Statistical Question Spearmans Rank Correlation Coefficient *BMJ-British Medical Journal.* 349 (5) (2014) 27-37.

[29] L. Chen, S.K. Zheng, Spoken Emotion Recognition Using Local Fisher Discriminant Analysis, *The 10th IEEE International Conference on Signal Processing Proceedings (ICSP2010).* 1 (3) (2010) 538-540.

[30] S.Q. Zhang, X.M. Zhao, Dimensionality reduction-based spoken emotion recognition, *Multimedia Tools and Applications.* 63 (3) (2013) 615-646.

[31] C.M. Lee, S.S. Narayanan, Toward Detecting Emotions in Spoken Dialogs, *IEEE transactions on speech and audio processing.* 13 (2) (2005) 293-303.

[32] L. Chen, S.K. Zheng, Speech Emotion Recognition: Features and Classification Models, *Digital Signal Processing.* 22 (6) (2012) 1154-1160.

[33] W. Malina, On an extended Fisher criterion for feature selection, *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 5 (1981) 611-614.

[34] J. Yang, J. Yang, Why can LDA be performed in PCA transformed space?, *Pattern recognition.* 36 (2) (2003) 563-566.

[35] S.Q. Zhang, B.C. Lei, A.H. Chen, Spoken Emotion Recognition Using Local Fisher Discriminant Analysis, *The 10th IEEE International Conference on Signal Processing Proceedings.* (2010) 538-540.

[36] G.B. Huang, Q.Y. Zhu, C.K. Siew, Extreme learning machine: a new learning scheme of feedforward neural networks, *Neural Networks, 2004.Proceedings. 2004 IEEE International Joint Conference on. IEEE.* (2004) 985-990.

[37] G.B. Huang, Q.Y. Zhu, C.K. Siew, Extreme learning machine: theory and applications, *Neural computing.* (70) (1) (2006) 489-501.

[38] F. Schwenker, E. Trentin, Pattern classification and clustering: a review of partially supervised learning approaches, *Pattern Recognition Letters.* 37 (2014) 4-14.

[39] S. Thapanee, W. Sartra, Speech Emotion Recognition using Support Vector Machines, *The 5th International Conference on Knowledge and Smart Technology (KST).* 99 (20) (2013) 86-91.

[40] Q.R. Mao, Y.Z. Wang, Y.Z. Zhan, Speech emotion recognition method based on selective features and decision binary tree, *Journal of Computational Information Systems.* 4 (4) (2008) 1795-1801.

[41] Q.R. Mao, Y.Z. Wang, Y.Z. Zhan, Speech emtion recognition method based on improved decision tree and layered feature selection, *International Journal of Humanoid Robotics.* 7 (2) (2010) 245-261.

[42] C.C. Chang, C.J. Lin, Chang C C, Lin C J. LIBSVM: A library for support vector machines, *ACM.* 2011.

[43] J. Tao, F. Liu, M. Zhang, H.B. Jia, Design of speech corpus for mandarin text to speech, *The Blizzard Challenge 2008 workshop.* (2008).

[44] R. Sun, I. Elliot Moore, Investigating Glottal Parameters and Teager Energy Operators in Emotion Recognition, A multimodal emotional communication based humans-robots interaction system, *Affective Computing and Intelligent Interaction.* (2011) 425-434.

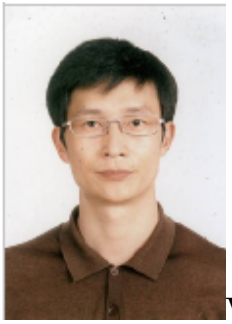
[45] Q. Mao, M. Dong, Z. Huang, et al. Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks, *IEEE Transactions on Multimedia.* 16 (8) (2014) 2203-2213.

[46] Z.T. Liu, F.F. Pan, M. Wu, A multimodal emotional communication based humans-robots interaction system, *The 35th Chinese Control Conference.* (2016) 6363-6368.

Zhen-Tao Liu received the B.E. and M.E. degrees from Central South University, Changsha, China, in 2004 and 2008, respectively, and Dr. E. degree from Tokyo Institute of Technology, Tokyo, Japan, in 2013. From 2013 to 2014, he was with Central South University, Changsha, China. Since Sept. 2014, he has been with School of Automation, China University of Geosciences, Wuhan, China. His research interests include fuzzy systems, affective computing, and intelligent robot. He is a member of CAAI (Chinese Association for Artificial Intelligence) and SOFT (Japan Society for Fuzzy Theory and Systems). He is an editor of Int. J. of Advanced Computational Intelligence and Intelligent Informatics. He received Young Researcher Award of Int. J. of Advanced Computational Intelligence and Intelligent Informatics in 2014, Best Paper Award in ASPIRE League Symposium 2012, and Excellent Presentation Award in IWACIII 2009.

Min Wu received the B.S. and M.S. degrees in Engineering from Central South University, Chang-sha, China, in 1983 and 1986, respectively, and the Ph.D. degree in Engineering from the Tokyo Institute of Technology, Tokyo, Japan, in 1999. From 1986 to 2014, he was a Faculty Member of the School of Information Science and Engineering at Central South University as a Full Professor. From 1989 to 1990, he was a Visiting Scholar with the Department of Electrical Engineering, Tohoku University, Sendai, Japan, and from 1996 to 1999, a Visiting Research Scholar with the Department of Control and Systems Engineering, Tokyo Institute of Tech-

nology, Tokyo. From 2001 to 2002, he was a Visiting Professor at the School of Mechanical, Materials, Manufacturing Engineering and Management, University of Nottingham, Nottingham, U.K. In 2014, he moved to the China University of Geosciences, Wuhan, China, where he is currently a Professor in the School of Automation. His current research interests include robust control and its applications, process control, and intelligent control. Dr. Wu is a Member of the Chinese Association of Automation. He received the IFAC Control Engineering Practice Prize Paper Award in 1999 (together with M. Nakano and J. She).



Wei-Hua Cao received his B.S., M.S., and Ph.D. degrees in Engineering from Central South University, Changsha, China, in 1983, 1986, and 2007, respectively. He is a Professor in School of Automation, China University of Geosciences. He was a Visiting Scholar with the Department of Electrical Engineering, Alberta University, Edmonton, Canada, from 2007 to 2008. His research interest covers intelligent control and process control. He is also a member of the Institute of Electrical and Electronics Engineers (IEEE).

Jun-Wei-Mao received the B.E. degree from China Three Gorges University, Yichang, China, in 2016. He is currently a master student in the School of Automation, China University of Geosciences. His main research interests include speech emotion recognition and human-robot interaction system.

Jian-Ping Xu received the B.E. degree from China University of Geosciences, Wuhan, China, in 2015. He is currently a master student in the School of Automation, China University of Geosciences. His research interests include speech signal processing and speech emotion recognition.

Guan-Zheng TAN, received his B. Sc. degree from the Department of Aeronautical Engine, Nanjing Aeronautical Institute, Nanjing, China, in 1983; received his M.Sc. degree from the Department of Automatic Control, National University of Defense Technology, Changsha, China, in 1988; received his Ph.D. degree from the Department of Mechanical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 1992. From October 1, 2004 to September 30, 2005, He was a visiting professor with the School of Computer Science, University of Birmingham, United Kingdom. He is currently a professor with the School of Information Science and Engineering, Central South University, Changsha, China. His main research areas are artificial intelligence and robotics.