Sun et al. EURASIP Journal on Audio, Speech, and Music Processing (2019) 2019:2
https://doi.org/10.1186/s13636-018-0145-5

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Decision tree SVM model with Fisher feature selection for speech emotion recognition

Linhui Sun*, Sheng Fu and Fu Wang

Abstract

The overall recognition rate will reduce due to the increase of emotional confusion in multiple speech emotion recognition. To solve the problem, we propose a speech emotion recognition method based on the decision tree support vector machine (SVM) model with Fisher feature selection. At the stage of feature selection, Fisher criterion is used to filter out the feature parameters of higher distinguish ability. At the emotion classification stage, an algorithm is proposed to determine the structure of decision tree. The decision tree SVM can realize the two-step classification of the first rough classification and the fine classification. Thus the redundant parameters are eliminated and the performance of emotion recognition is improved. In this method, the decision tree SVM framework is firstly established by calculating the confusion degree of emotion, and then the features with higher distinguish ability are selected for each SVM of the decision tree according to Fisher criterion. Finally, speech emotion recognition is realized based on this model. The decision tree SVM with Fisher feature selection on CASIA Chinese emotion speech corpus and Berlin speech corpus are constructed to validate the effectiveness of our framework. The experimental results show that the average emotion recognition rate based on the proposed method is 9% higher than traditional SVM classification method on CASIA, and 8.26% higher on Berlin speech corpus. It is verified that the proposed method can effectively reduce the emotional confusion and improve the emotion recognition rate.

Keywords: Speech emotion recognition, Emotional confusion, Decision tree SVM algorithm, Feature selection, Fisher criterion

1 Introduction

In recent years, speech emotion recognition has been widely applied in the field of human-computer interaction [1–3]. Emotion recognition helps machine understand and learn human emotions. However, the performance of the emotion recognition is still far from the expectation of researchers. In speech emotion recognition, there are mainly two difficulties [4] that are how to find effective speech emotion features, and how to construct a suitable speech emotion recognition model. In previous studies, some effective feature parameters were extracted for emotional recognition tasks. Zhao et al. adopted the pitch frequency, short-term energy, formant frequency, and chaotic characteristics to construct 144 dimensional emotion feature vector for recognition [5]. Cao et al. combined the feature parameters such as energy, zero crossing rate, and first-order derivative for speech emotion recognition, and encouraging results were obtained in comparison with other methods [6]. In [7], the first 120 Fourier coefficients of the speech signal were extracted, and the recognition rate of 79.51% was obtained using Germany Berlin speech emotion database with 6 emotions. In [8], some new harmonic and Zipf-based features for better speech emotion characterization in the valence dimension were proposed for better emotional class discrimination. In [9], Prosody features and voice quality information were combined in emotion recognition. The methods mentioned above improve the performance of emotion recognition by feature fusion. However, feature fusion may lead to high dimension

* Correspondence: sunlh@njupt.edu.cn
College of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China

and redundancy of features, so it is vital to filter out the characteristic parameters of higher distinguish ability. Fisher criterion is a classical linear decision method, which can achieve satisfying results in selecting features. Huang et al. used the Fisher discriminant coefficient to screen out 10 dimensional features from 84 dimensional features for the identification of 5 emotions [10], which increased the emotion recognition by 8%.

To further improve the performance of speech emotion recognition, an effective emotion recognition model needs to be constructed. Currently, some classifiers are extensively used in speech emotion recognition, including Gaussian mixture model (GMM) [11], artificial neural network (ANN) [12], support vector machine (SVM), etc. Among them, the SVM has a unique advantage in solving nonlinear, small sample, and high dimensional pattern recognition problems, so it is widely used in speech emotion recognition [13, 14]. In [15], Zhang et al. proposed an improved leaping algorithm to optimize the SVM classifier, and this algorithm was applied to speech emotion recognition. In [16], an integrated system of hidden Markov model (HMM) and SVM, combining advantages on capability of dynamic time warping of HMM and pattern recognition of SVM, had been proposed to implement emotion classification, which achieved an 18.3% improvement compared to the method using HMM in the experiment of speaker independent emotion classification. Work in [17] applied the GMM-MAP/SVM generative models and discriminative models to speech emotion recognition, which increased the average emotion recognition by 6.1% compared to the method using SVM. In addition, the binary decision tree SVM recognition model had also been applied to multiple emotion recognition, which obtained good performance [18, 19].

In our study, we found that the statistical variables of multi frames emotional features were better than emotional features extracted by frame in emotion recognition. Due to the diversity of information represented by different features, combining various features can achieve better performance, which has become the current research hotspot [20–22]. Besides, in multiple emotion recognition, based on the ability of various features to discriminate emotional types, we can filter out an optimal feature set to eliminate the redundant parameters.

The overall recognition rate will reduce due to the increase of emotional confusion in multiple speech emotion recognition. Inspired by the above methods, for multiple speech emotion recognition, we proposed a speech emotion recognition method based on the decision tree SVM model with Fisher feature selection. In this method, a high-performance decision tree SVM classifier is established by calculating the degree of emotional confusion, to realize the two step classification of the first rough classification and the fine classification. For each of decision tree SVM, we filter out the feature parameters of higher distinguish ability by Fisher criterion to gain an optimal feature set. Finally, this model is used for speech emotion recognition. Thus, a better emotional classification performance can be obtained.

The contributions made in this paper include (1) adopt Fisher criterion to remove the redundant features to improve emotion recognition performance; (2) propose an algorithm to determine the structure of decision tree dynamically, and construct the system frameworks on the CASIA Chinese speech emotion corpus and the EMO-DB Berlin speech corpus; and (3) combine Fisher criterion with decision tree SVM, and adopt genetic algorithm to optimize the parameters of SVM to further improve the emotion recognition rate.

The rest of this paper is organized as follows. In Section 2, we present the idea of decision tree SVM model with Fisher feature selection. Section 3 introduces the experiment and the analysis of results. In Section 4, we summarize our paper and discuss the future work for speech emotion recognition.

2 Decision tree SVM model with Fisher feature selection

2.1 Emotion features

To effectively recognize emotions from speech signals, we need to extract some feature parameters

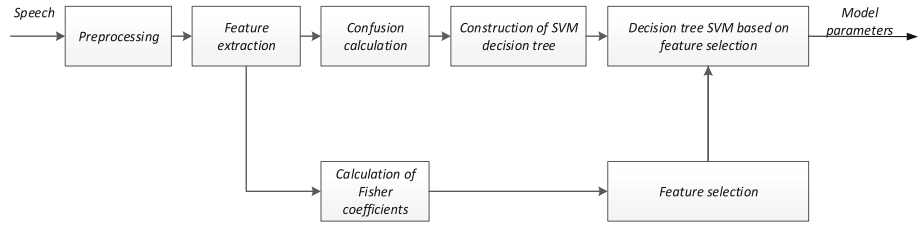

Fig. 1 Construction flow chart of decision tree SVM with feature selection

Table 1 Recognition confusion matrix of six emotions (%)

Emotions	Angry	Happy	Fear	Neutral	Surprise	Sad
Angry	77	8	1	1.5	11	1.5
Happy	7	70.5	3	2.5	14.5	2.5
Fear	2	3	56	2.5	2	34.5
Neutral	0.5	2.5	2.5	92	1.5	1
Surprise	11.5	8	3	1	76.5	0
Sad	1	4	29	2	0	64

that can reflect the emotional information in the speech signal, and then use these parameters to train the model which is used for the emotion recognition. The quality of selected feature parameters affects the recognition rate of the system directly. Traditional speech features used for emotion recognition tasks are mainly divided into prosodic features, spectrum-based features, and voice quality features [23]. The prosodic features include speech rate, pitch period, amplitude energy, etc. The spectrum-based features include linear predictor coefficient (LPC), Mel-frequency cepstral coefficient (MFCC), etc. The voice quality features include formant frequency and glottis parameters. In addition, some basic parameters, such as Fourier coefficients [24, 25], are often used in speech emotion recognition.

Feature parameters are usually extracted by frame. Since a single frame contains less information, most researchers use feature parameters to calculate statistical variables in multiple frames for emotion recognition tasks. In this paper, five kinds of features are adopted, including MFCC, energy, Fourier coefficients, pitch frequency, and zero-crossing rate, and five statistical variables (i.e., maximum, minimum, mean, standard deviation, and median) of multi-frame features are calculated and applied to recognition tasks.

2.2 Support vector machine model

Support vector machine (SVM) proposed in the 1990s is a kind of machine learning method which is applied in many areas. For the nonlinear separable problem, its basic idea is that the input space is mapped into a high dimensional feature space by nonlinear transformation, and the optimal hyperplane is found in the new space. The optimal hyperplane not only needs to ensure that different categories can be discriminated correctly, but also the maximum categorization interval between them should be promised. Thus, the generalization capability of the support vector machine is stronger. In another word, looking for a hyperplane with a maximum interval is the goal of training SVM.

Table 2 Confusion degree of six emotions (%)

Emotions	Angry	Happy	Fear	Neutral	Surprise
Happy	7.5	–	–	–	–
Fear	1.5	3	–	–	–
Neutral	1	2.5	2.5	–	–
Surprise	11.25	11.25	2.5	1.25	–
Sad	1.25	3.25	31.75	1.5	0

Table 3 Confusion degree between angry, happy, and surprise (%)

Emotions	Angry	Happy
Happy	9	–
Surprise	12	11.5

The target function corresponding to the nonlinear separable support vector machine is given by:

$$\min\left(\frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\omega}+C\sum_{i=1}^{N}\xi_i\right) \tag{1}$$

$$\text{s.t.}\quad y_i\left(\boldsymbol{\omega}^T\boldsymbol{x}_i+b\right)\geq 1-\xi_i,\xi_i\geq 0,i=1,2,\ldots,N$$

where $\boldsymbol{\omega}$ represents the weight coefficient vector, and b is a constant. C denotes the penalty coefficient to control the penalty degree for misclassified samples and balance the complexity of the model and loss error. ξ_i represents the relaxation factor to adjust the number of misclassified samples that allowed exit in the process of classification.

When the SVM is used to solve the classification problems, two strategies can be adopted. One is ONE-TO-ALL, and another is ONE-TO-ONE. According to previous studies, ONE-TO-ONE classification strategies have an advantage in speed [26]. Therefore, the paper uses the ONE-TO-ONE strategies. Kernel functions are the key for SVM. The kernel functions commonly used include linear kernel function, polynomial kernel function, radial basis function (RBF), and multilayer perceptron kernel function. Based on the previous experiments, the best RBF kernel function is used in this paper.

2.3 Construction strategy of decision tree SVM

In multi-classified emotion recognition, the overall recognition rate is reduced due to the increase of confusion between emotions. To solve this problem, this paper establishes a decision tree SVM by calculating the degree of emotional confusion, and uses the decision tree SVM as a classifier for speech emotion recognition.

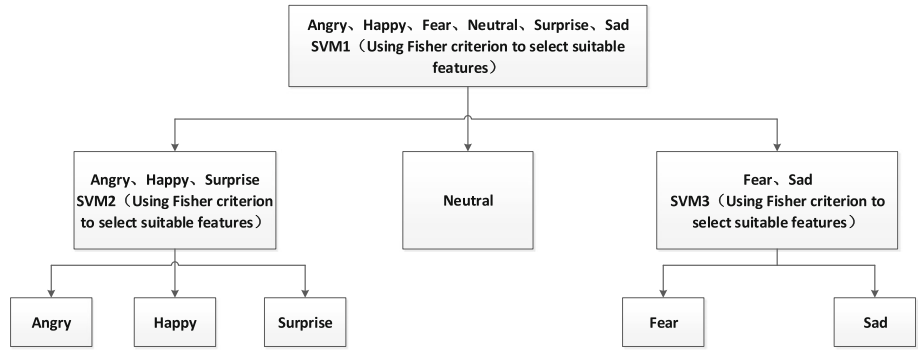

Fig. 2 System framework for CASIA

At first, we defined $E = \{e_1, e_2, e_3, \ldots, e_n\}$ as the emotional state set, where the number of the state is n. Defined the confusion degree between the ith emotion e_i and the jth emotion e_j is $I_{i,j}$, which represents the average of the probability that the ith emotion is misjudged as the jth emotion and the jth emotion is misjudged as the ith emotion [27]. The formula is given by:

$$I_{i,j} = \frac{P(r = j|\boldsymbol{x} \in e_i) + P(r = i|\boldsymbol{x} \in e_j)}{2} \quad (2)$$

where $\boldsymbol{x}$ represents the test sample, and r represents the result of classification corresponding to $\boldsymbol{x}$.

The proposed decision tree SVM algorithm is as follows:

(a) Calculate the emotional confusion matrix using traditional SVM method, only the MFCC parameters are used to train SVM.

(b) Set an appropriate initial threshold P at primary classification. The emotions in which the confusion degree exceeds the threshold P are classified into a same group. If $I_{a,b} > P$, a and b will be divided into one group. If $I_{a,b} > P$, $I_{b,c} > P$, a, b, and c will be divided into one group. If the confusion degree between a certain emotion and other emotions is less than the threshold, this certain emotion will be grouped separately.

(c) Calculate the confusion degree between ungrouped emotions and other emotions according to Eq. (2), and then move to step (b) to divide the ungrouped emotional states into existing groups or a new group.

(d) Calculate the number of emotional states in each group. If the number is greater than 2, the threshold needs to be increased by P, and move to step (a); otherwise, move to step (e).

(e) All emotions are categorized and ended.

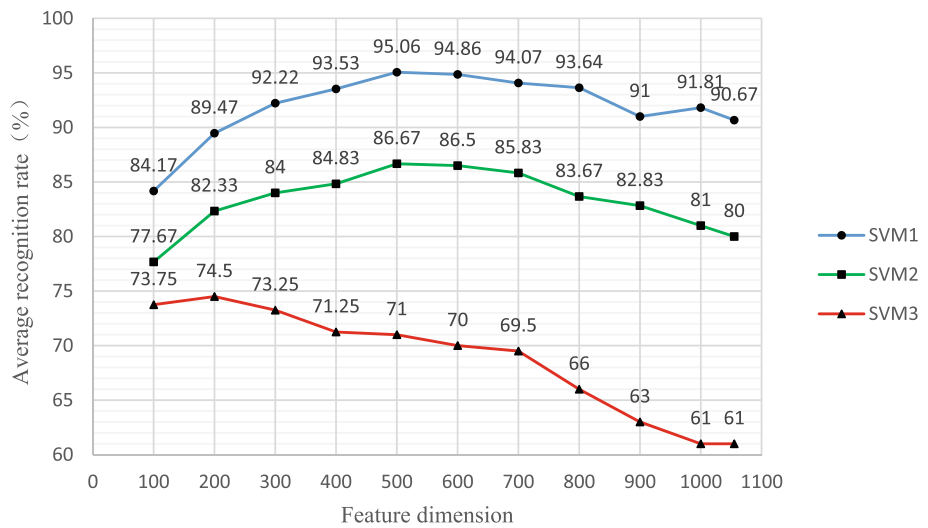

Fig. 3 Average recognition rate of SVMs with different feature dimensions

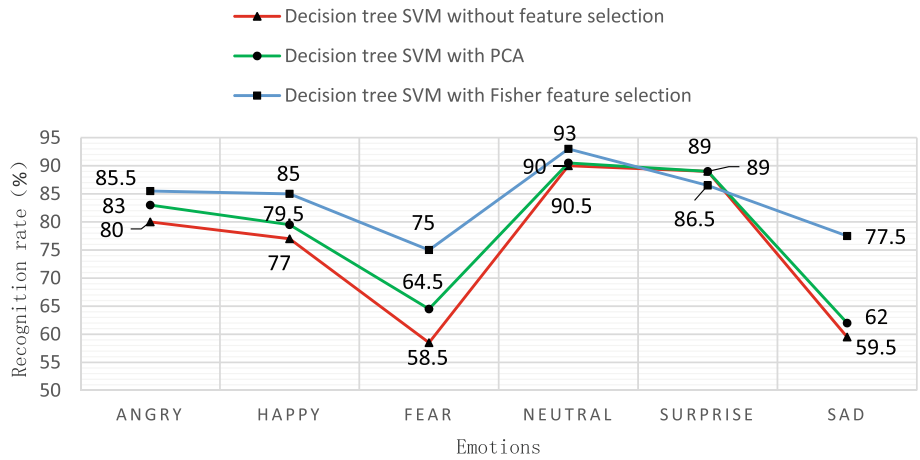

Fig. 4 Comparison of speech emotion recognition rates

In this section, we introduce an algorithm to obtain the decision tree SVM model. This algorithm can determine the depth of decision tree dynamically, which means the decision tree structure of various corpus is different. At first, we need to decide the initial threshold P based on a large number of comparison experiments. After determining the initial threshold, the threshold of the remaining layer is also determined. For the second layer, the threshold is $2P$. For the third layer of decision tree, the threshold is $3P$, and so on. When the threshold of each layer is determined, the emotional states can be classified. When the confusion degree between emotions is greater than the threshold in this layer, these emotions are divided into a same group. If their confusion degree is below the threshold, these emotions need not be grouped, but are directly classified by the SVM in this layer. In this way, various corpus can obtain the optimal structure of decision tree model.

2.4 Feature selection strategy for decision tree SVM

In order to improve the recognition rate of multiple classification speech emotion recognition, we propose a speech emotion recognition method based on the decision tree SVM model and Fisher feature selection. In this method, the speech signal is preprocessed by pre-emphasize and framing, and the MFCC coefficients of the speech signal are extracted. Then, the confusion matrix between emotions is obtained by using MFCC coefficient and traditional SVM, and the confusion degree between emotions is calculated based on the confusion matrix. Finally, the decision tree SVM is constructed based on confusion degree and the strategy of decision tree SVM. When the SVM decision tree is constructed, Fisher discriminant coefficient can be obtained by calculating the mean and variance of each dimension feature parameters. Feature parameters of higher distinguish ability are selected for each SVM in the decision tree according to the Fisher discriminant coefficient, which is used for training. The specific flow chart is shown in Fig. 1.

In speech emotion recognition, due to the difference in the ability to discriminate emotional states for various features, it is vital to select appropriate features to discriminate different sets of emotions. In an ideal feature space, the distance between different categories should be as large as possible, and the distance between the same categories should be as small as possible, so that we can classify effectively. For the feature selection of the extracted features, the mean and variance of the feature are used as a criterion to measure the characteristics of the feature. Assume that the feature matrix F_P of Pth emotion is given by:

Table 4 Average recognition rate of three methods on CASIA

Methods	Traditional SVM without feature selection	Decision tree SVM without feature selection	Decision tree SVM with PCA	Decision tree SVM with Fisher feature selection
Average recognition rate (%)	74.75	75.67	78.08	83.75

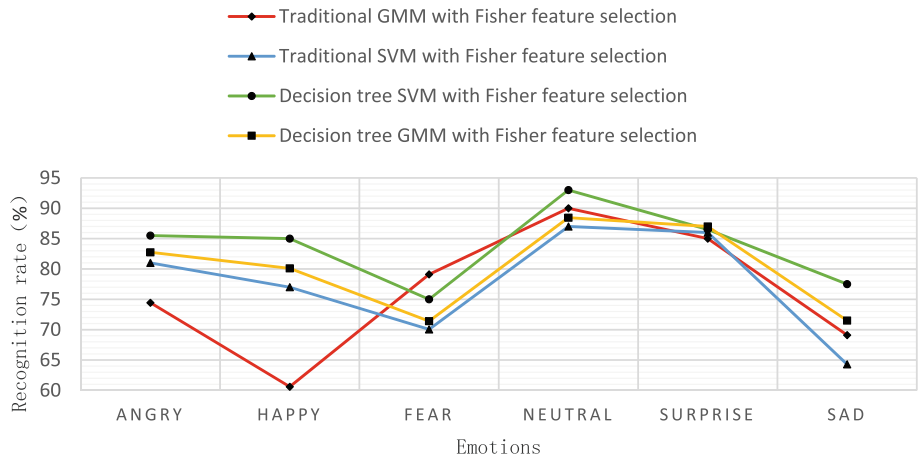

Fig. 5 Comparison of speech emotion recognition rates

$$F_P = \begin{bmatrix} X_{11}^P & X_{12}^P & X_{13}^P & X_{14}^P & X_{15}^P & X_{16}^P & \dots & X_{1N}^P \\ X_{21}^P & X_{22}^P & X_{23}^P & X_{24}^P & X_{25}^P & X_{26}^P & \dots & X_{2N}^P \\ X_{31}^P & X_{32}^P & X_{33}^P & X_{34}^P & X_{35}^P & X_{36}^P & \dots & X_{3N}^P \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ X_{M1}^P & X_{M2}^P & X_{M3}^P & X_{M4}^P & X_{M5}^P & X_{M6}^P & \dots & X_{MN}^P \end{bmatrix} \tag{3}$$

where M and N are the number and the dimension of the feature parameters respectively. $\boldsymbol{T}_{id} = \{X_{1d}^i, X_{2d}^i, X_{3d}^i, \dots, X_{Md}^i\}$ represents the set of the d-dim feature of the ith type emotion. For the d-dim feature, the Fisher discriminant coefficient is defined as:

$$f(d) = \frac{(\mu_{1d} - \mu_{2d})^2}{\sigma_{1d}^2 + \sigma_{2d}^2} \tag{4}$$

where $\mu_{id}$and σ_{id}^2 denote the mean and variance of the vector $\boldsymbol{T}_{id}$ respectively. The size of Fisher discriminant coefficient can reflect the degree of dissimilarity between different classes and the similarity between the same classes. For distinguishing emotional states, the larger the Fisher discriminant coefficient of the feature is, the greater the emotion contribution the feature makes. For multiple classifications, Fisher's discriminant coefficient is calculated as:

$$f(d) = \frac{1}{\mathrm{C}_Q^2} \sum_{0<i<j\le Q} \frac{\left(\mu_{id} - \mu_{jd}\right)^2}{\sigma_{id}^2 + \sigma_{jd}^2} \tag{5}$$

where Q is the total number of emotional states.

3 Experiments

In this paper, to evaluate the effectiveness of proposed method, two different corpus are employed: the CASIA Chinese speech emotion corpus and the EMO-DB Berlin speech corpus. The CASIA Chinese speech emotion corpus is recorded and provided by the institution of Automation, Chinese Academy of Sciences. The CASIA corpus contains six kinds of basic emotions: Angry, Happy, Fear, Neutral,

Table 5 Emotional confusion matrix of EMO-DB (%)

Emotions	Angry	Happy	Boring	Neutral	Sad	Fear	Disgust
Angry	73.6	17.1	0.8	1	0.6	1.5	5.4
Happy	6.4	67.9	0.9	0.8	0	15	9
Boring	2	1.1	81.5	4.6	6	1.2	3.6
Neutral	0.9	1.2	1.5	91.7	2	1.5	1.2
Sad	0.8	0	10	1.5	85.6	2.1	0
Fear	9.5	1.2	3	0.7	5.4	79.2	1
Disgust	10.8	1	2	4	0	5	77.2

Table 6 Confusion degree of seven emotions (%)

Emotions	Angry	Happy	Boring	Neutral	Sad	Fear
Happy	11.75	–	–	–	–	–
Boring	1.4	1	–	–	–	–
Neutral	0.95	1	3.05	–	–	–
Sad	0.7	0	8	1.75	–	–
Fear	5.5	8.1	2.1	1.1	3.75	–
Disgust	8.1	5	2.8	2.6	0	3

Table 7 Confusion degree between angry, happy, fear, and disgust (%)

Emotions	Angry	Happy	Fear
Happy	13.85	–	–
Fear	6.7	12.4	–
Disgust	2	8.4	1.8

Surprise, and Sad. The EMD-DB consists of seven basic emotions: Angry, Happy, Fear, Neutral, Boring, Disgust, and Sad. The EMO-DB corpus consists of 535 speech utterances, and all of these utterances are used in the experiments. The experiments carried out in this paper are all based on a tenfold cross-validation method. In other words, the samples are randomly divided into ten parts equally, among which 9/10 of samples are used for training and 1/10 of samples are used for testing. The experiment is repeated ten times, and the final recognition result is the average of these results. This paper selects SVM as an emotion recognition model and uses the LIBSVM toolbox developed by Professor Lin Zhiren of Taiwan University to realize the training and testing of SVM. The development tool of Matlab2013a is adopted to extract emotion features, and LIBSVM is installed in the environment of Visual Studio 2010.

Before extracting the parameters of the speech signal, this paper first performs endpoint detection on the speech signal, and the speech signal is framed with the frame length of 256 points and the frame shift of 128 points. The feature parameters of the experiment include the first 160 Fourier coefficients, the amplitude energy, the pitch frequency, the zero crossing rate, the 24 order MFCC, and the first-order difference. The statistical variables include the maximum, the minimum, the mean, the median, and the variance. In addition, all 1055 dimensional feature parameters are normalized.

3.1 CASIA Chinese speech emotion corpus

3.1.1 Decision tree SVM model for CASIA

According to the proposed decision tree SVM algorithm, the emotional confusions among emotions need be calculated, and the confusion matrix of six emotions is shown in Table 1, using MFCC parameters and traditional SVM. According to Eq. (2), the confusion degree among emotions is calculated, which is shown in Table 2. The initial threshold P is set to 7% by a large number of experiments using CASIA corpus (the solution of optimal initial threshold will be introduced in Section 3.3). From Table 2, we can find that the confusion degree between Angry and Surprise is 11.25%, while that of Happy and Surprise is 11.25%. Both of them are more than 7% of the initial classification threshold. According to the decision tree SVM construction algorithm, Angry, Happy, and Surprise are divided into the first group. The confusion degree between Fear and Sad is 31.75%, so Fear and Sad are divided into the second group. Since the confusion degree between Neutral and other emotions is less than 7%, Neutral is classified into the third group. At this point, the SVM, which implement the three major classifications, is recorded as SVM1.

Table 8 Confusion degree between angry, happy, and fear (%)

Emotions	Angry	Happy
Happy	16.45	–
Fear	10.27	11.85

According to step d in the decision tree SVM construction algorithm, Angry, Happy, and Surprise need reclassified. The confusion degree between these emotions is obtained by the construction algorithm as shown in Table 3. As shown in Table 3, the value of confusion degree between Angry and Surprise is 12%, that of Happy and Surprise is 11.5%, and that of Happy and Angry is 9%. All of them are lower than the threshold which is 14% in second time classification. So we use SVM to classify these three basic emotional states directly, and the SVM is recorded as SVM2. The SVM that is used to classify the second group (Fear and Sad) is recorded as SVM3. Meanwhile, each SVM adopts one-to-one strategy and radial basis kernel functions. Finally, the SVM decision tree can be obtained, as shown in Fig. 2.

3.1.2 Feature selection by Fisher criterion

In feature selection, in order to screen out the feature parameters of higher distinguish ability from 1055 dimensional feature parameters to train the SVM in decision tree, the Fisher coefficients of each dimensional features are calculated. And then, the Fisher coefficients are ordered from largest to smallest, and features with larger Fisher coefficients are selected. For SVM1, the first 100 to 1000 dimensional (step size is 100) features with the larger Fisher coefficient is tested for the 3 basic emotional states, and the average recognition rates are shown in Fig. 3. It can be seen that the average recognition rate gradually increases when the dimensionality of the features selected ranges from 100 to 500 for SVM1. And when the dimension is higher than 500, the average recognition rate is reduced. That is, when using 500 dimensional features with the larger Fisher coefficient, the average recognition rate is the highest, which can

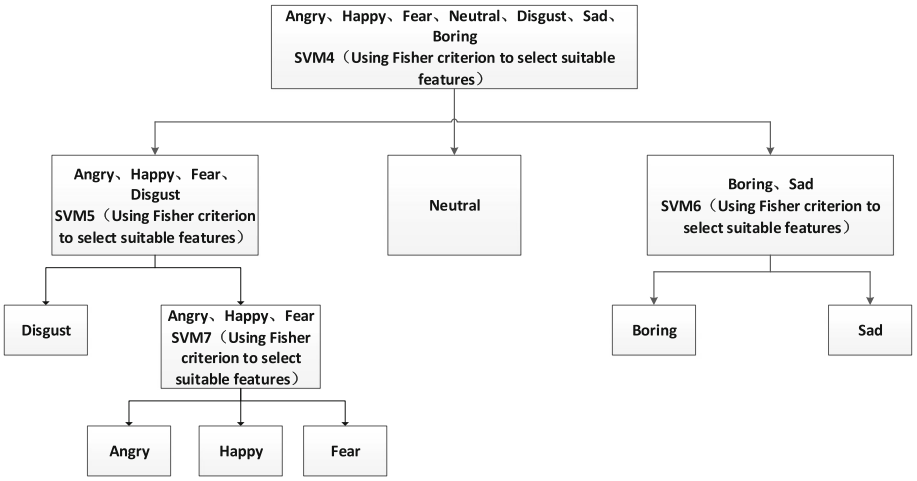

Fig. 6 System framework for EMO-DB

reach 95.06%, even higher than the average recognition rate (90.67%) using all 1055 dimensional features. Thus, we select the first 500 dimensional feature parameters to train SVM1. Similarly, in the same way, the first 500 dimensional feature parameters with the highest correct recognition rate can be selected to train SVM2. For SVM3, which distinguishes Fear and Sad, we select the top 200 dimensional features with the highest correct recognition rate as the feature parameters to train SVM3.

3.1.3 Comparison results on CASIA

To verify the validity of the proposed method for speech emotion recognition, we conduct two comparison experiments. The first experiment is used to verify the effectiveness of our feature selection method, and the second experiment is used to verify the superiority of decision tree SVM model.

1. We compared our method with two other feature selection methods, the decision tree SVM with 1055 dimensional features and the decision tree SVM model with PCA dimension reduction method. In our experiment, 1055 dimensional alternative features are used as feature parameters, and the decision tree SVM mentioned in this paper are used as speech emotion recognition classifier. The recognition rates of various emotions are shown in Fig. 4. Except for a slight drop in surprise using our method compared

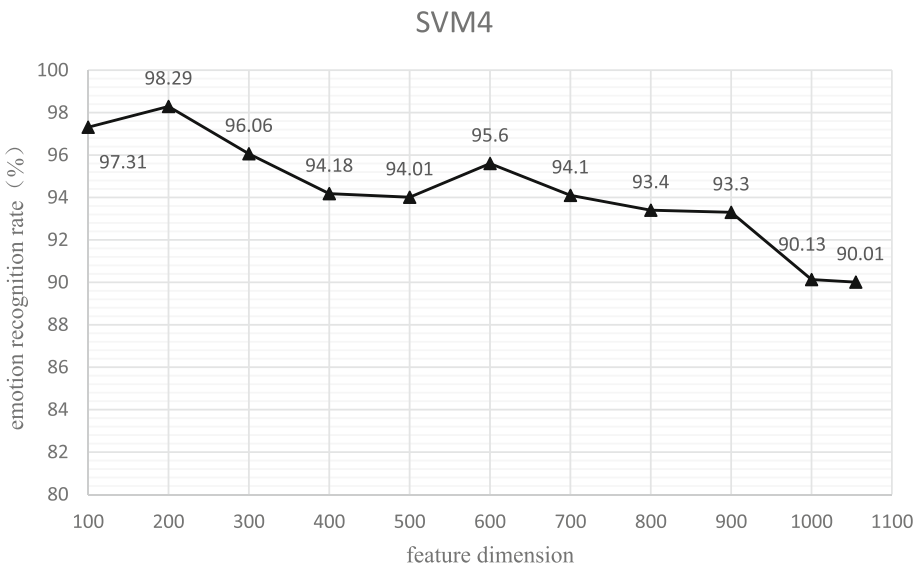

Fig. 7 Average recognition rate of SVM4 with different feature dimensions

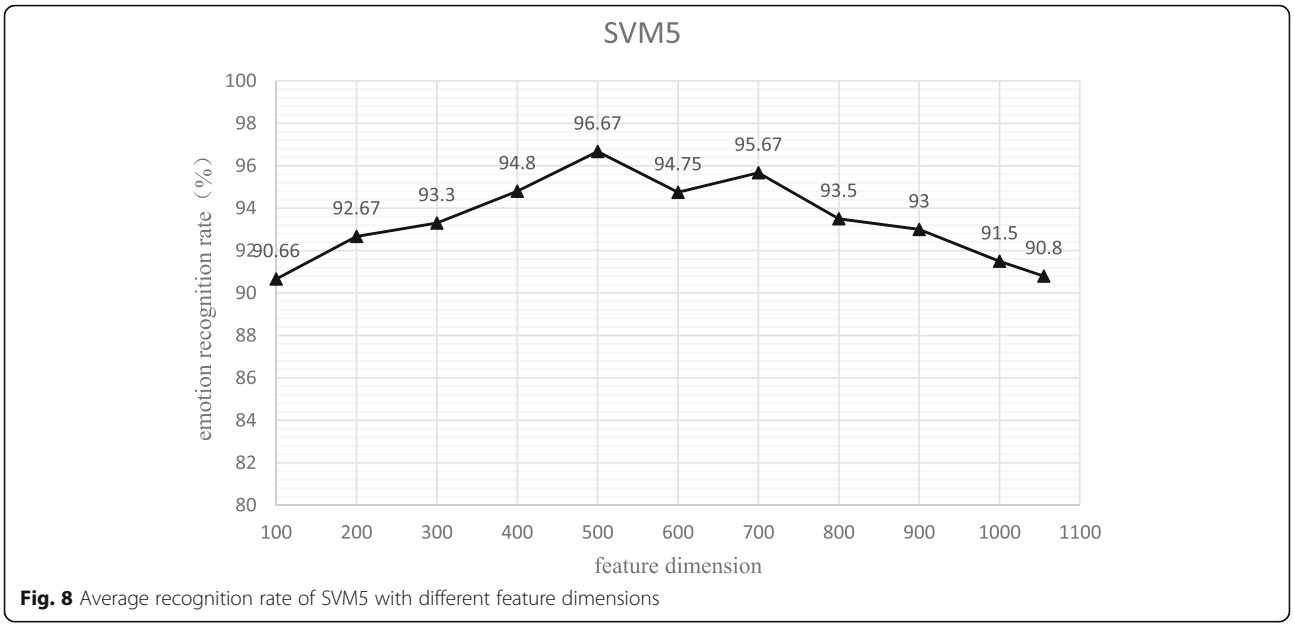

Fig. 8 Average recognition rate of SVM5 with different feature dimensions

with the decision tree SVM with 1055 dimensional features, all other emotional recognition rates have been improved and the overall recognition rate is improved. The results in Table 4 show that the average recognition rate of decision tree SVM without feature selection is 75.67%. When Fisher criterion is used to our framework, the average recognition rate is increased to 83.75%, which is 9% higher than traditional SVM without feature selection (average recognition rate 74.75%) and 8.08% higher than decision tree SVM without feature selection. Moreover, after the feature selections, the dimensionality of feature is reduced, and the computational complexity of the recognition system dropped as well.

In this experiment, we also compared the PCA dimension reduction method with Fisher criterion feature selection method, where the decision tree SVM is used as a classifier. When the principal component analysis (PCA) method is used to reduce

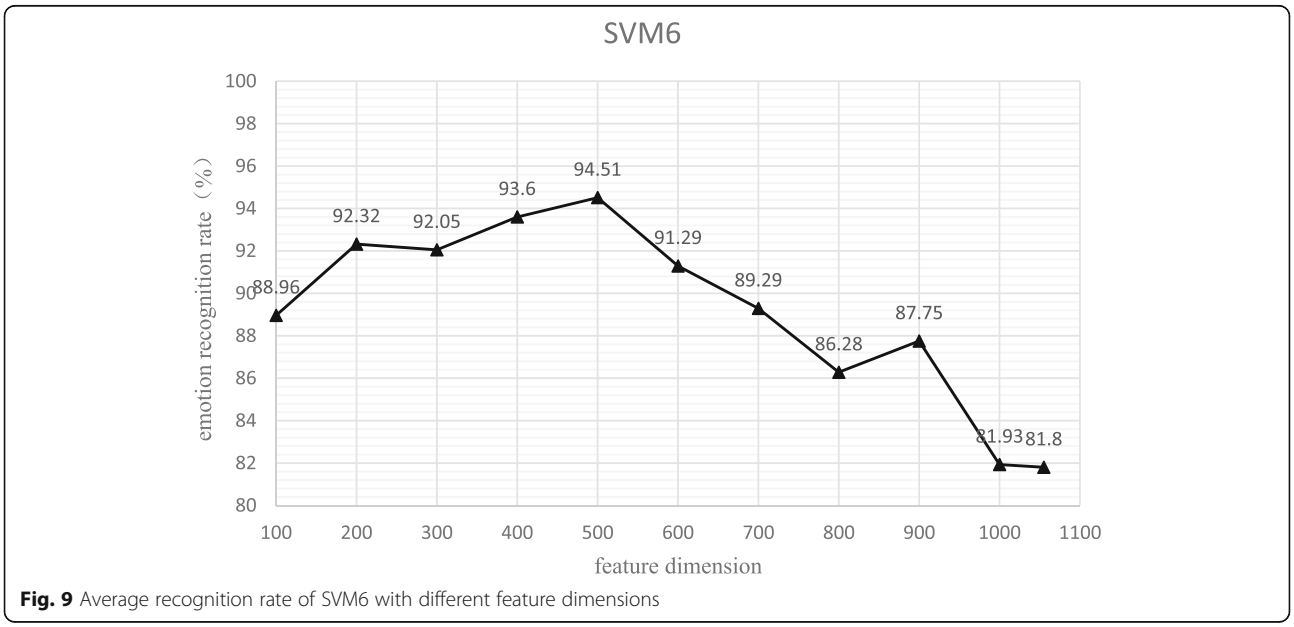

Fig. 9 Average recognition rate of SVM6 with different feature dimensions

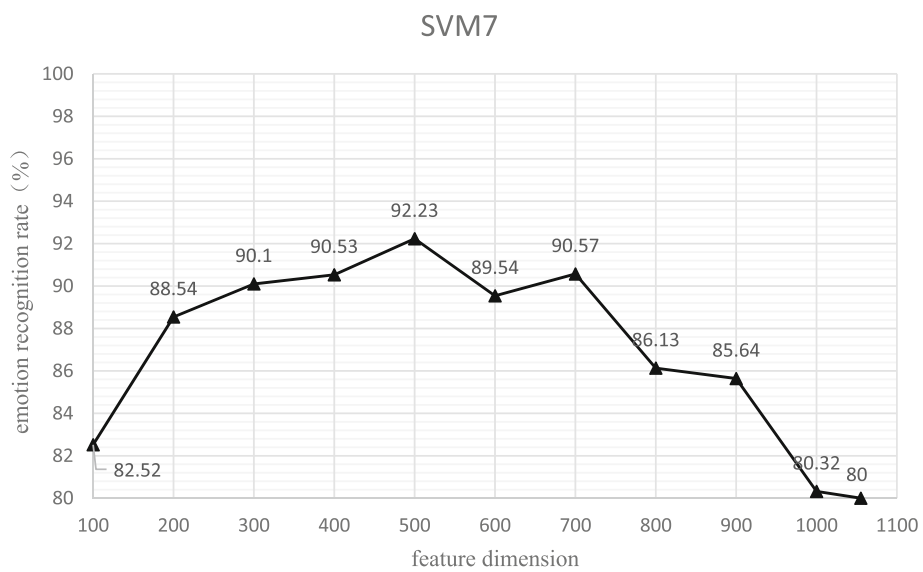

Fig. 10 Average recognition rate of SVM7 with different feature dimensions

the dimensionality of feature parameters, the feature parameters with a dimensionality of 104 are used for experiments. The experiment results show that the method based on Fisher criterion is more effective than the method based on PCA dimension reduction. The reason for higher recognition rate is the features that can better discriminate emotional states in some cases are remained by Fisher criterion, and features with lower distinguish ability are excluded. Therefore, the overall recognition improved greatly.

2. To verify the validity of the decision tree SVM model for speech emotion recognition, we have made the following comparison, as shown in Fig. 5. All of the following methods are adopt Fisher feature selection strategy to screen out the features with higher distinguish ability. Four kinds of classifier are used in experiment: traditional GMM, decision tree GMM, traditional SVM, and decision tree SVM. GMM refers to the linear combination of several Gauss distribution functions, which can establish a probability model for each kind of emotions and achieve classification [28]. Traditional GMM is trained with the same feature set as traditional SVM, and the number of Gaussian components is

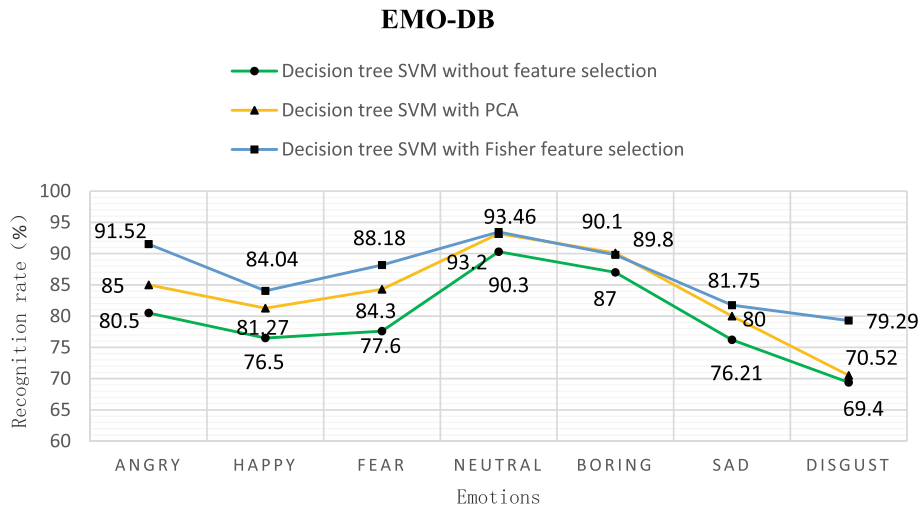

Fig. 11 Average recognition rate of different feature selection methods

Table 9 Average recognition rate of feature selection methods on EMO-DB (%)

Methods	Traditional SVM without feature selection	Decision tree SVM without feature selection	Decision tree SVM with PCA	Decision tree SVM with Fisher feature selection
Average recognition rate (%)	78.6	79.64	83.48	86.86

set to 8. From Fig. 5, we can see that the average recognition rate of traditional GMM is 76.77%, which is close to that of traditional SVM. In decision tree GMM method, GMM, like SVM, is used to classify emotions at decision tree nodes, and the features after Fisher selection are used to train decision tree GMM. The average recognition rate of decision tree GMM is 80.2%, which is 3.43% higher than that of traditional GMM. The decision tree SVM can obtain better performance, with an average recognition rate 83.75%, which is 3.55% higher than that of the decision tree GMM. The results indicate that the decision tree framework built by our algorithm can actually improve the performance of speech emotion recognition. By comparing with the result of three other classifiers, the proposed method based on the decision tree SVM can obtain better performance. The reason is that using the decision tree SVM established by the confusion degree, the confusable emotional states are divided into one group, and then the fine classification is conducted in this group. Hence, the confusion between emotions is reduced and the average recognition rate of all emotions is improved. Compared with the result of decision tree GMM, the proposed method based on the decision tree SVM also achieves higher recognition rate. SVM has excellent generalization ability and high robustness, so decision tree SVM can achieve better performance.

3.2 EMO-DB emotional corpus

3.2.1 System framework for EMO-DB

To verify the effectiveness of our proposed framework, we also conduct experiments on Berlin database. According to construction strategy of decision tree SVM, the confusion degree among emotions need be calculated. Tables 5 and 6 show the emotional confusion, and the initial threshold P can be set to 6% (the solution of the optimal initial threshold will be introduced in Section 3.3). From Tables 5 and 6, we can know that the confusion degree between Angry and Happy is 11.75%, while that of Angry and Disgust is 8.1%, and Happy and fear is 8.1%. All of them are more than the initial classification threshold, so Angry, Happy, Fear, and Disgust are divided into the first group. For the rest of emotions, with the exception of Boring and Sad, all the emotional confusion degree is below 6%, so Boring and Sad are divided into the second group, and the Neutral is divided into the third group.

Based on the algorithm proposed in Section 2.3, we need to reclassify the emotions in first group, and Table 7 shows the confusion degree among the first

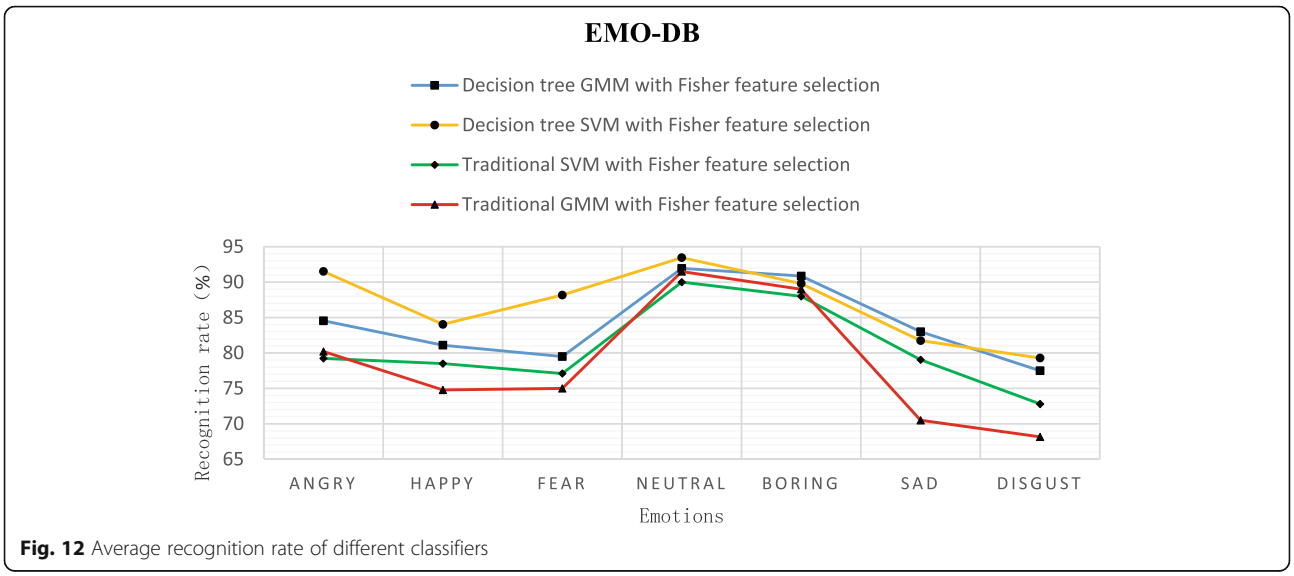

Fig. 12 Average recognition rate of different classifiers

Table 10 Average recognition rate of different classifiers on EMO-DB (%)

Methods	Traditional GMM with Fisher feature selection	Traditional SVM with Fisher feature selection	Decision tree GMM with Fisher feature selection	Decision tree SVM with Fisher feature selection
Average recognition rate (%)	78.45	80.66	84.06	86.86

group. According to the algorithm, the second level threshold is 12%. So at this level, Happy, Fear, and Angry are divided in a group, and the Disgust can be separated out. In the third level, the threshold is set to 18%, as show in Table 8. All the confusion is below the threshold, so the last three emotions are divided directly. Finally, the decision tree SVM model can be obtained, which is showed in Fig. 6.

3.2.2 Feature selection by Fisher criterion

In Section 2.4, Fisher feature selection method has stated clearly that it is vital to select appropriate features to discriminate different sets of emotions. In an ideal feature space, the distance between different categories should be as large as possible, so that different emotions can be classified easily. Figures 7, 8, 9, and 10 show the feature dimension that can achieve the best performance for each SVM. For SVM4, Fig. 7 shows the change of average recognition rates. When the feature dimension is 200, emotion recognition rate reaches the highest performance, which is 98.29%. So, 200 dimensional features are used to train SVM4. In the same way, we can obtain the best dimensionality for the rest of SVM. For SVM5, SVM6, and SVM7, after the feature selection, the features with lower Fisher coefficient are removed, and finally 500 dimensional features are remained. So the best feature dimension is 500 for SVM5, SVM6, and SVM7.

3.2.3 Comparison results on EMO-DB

In this section, we also conduct two comparison experiments on Berlin emotional speech corpus to prove the effectiveness of our proposed framework. In the first experiment, different feature selection methods are used. The results in Fig. 11 and Table 9 show that our proposed framework achieves an average recognition rate of 86.86%, that is 8.26% higher than traditional SVM without feature selection (average recognition rate 78.6%), a 7.22% improvement compared to decision tree SVM without feature selection, and a 3.38% improvement compared to decision tree SVM with PCA. Fisher feature selection method can remove the features that are unrelated with emotional information, so the recognition rate can be improved, which has been proved by the experiment results.

Figure 12 and Table 10 show the experiment results of different classifier model. Decision tree GMM with Fisher feature selection can obtain an average recognition rate of 84.06%, which is 5.61% higher than that of traditional GMM. In our proposed framework, the average recognition rate reach 86.86%, which is 6.2% higher than traditional SVM with Fisher feature selection. Although the recognition rate of Boring and Sad in our framework is slightly lower than the results of decision tree GMM with Fisher feature selection, our proposed framework can achieve better performance in overall recognition. The experimental results in Table 10 also imply that our decision tree algorithm is effective in improving the recognition rate.

3.3 The optimal initial threshold

In our algorithm, the structure of decision tree is decided by the initial threshold *P*, as described in Section 2.3. When the optimal initial threshold *P* is determined, the structure of decision tree can be uniquely determined. So it is necessary to determine the optimal initial threshold first. Table 2 and Table 6 show the emotional confusion of CASIA Chinese speech emotion corpus and EMO-DB Berlin speech corpus respectively. From these tables, we know that the initial threshold can range from 1 to 32% for CASIA Chinese speech emotion corpus, and for the EMO-DB Berlin speech corpus, the initial threshold range is 1% to 12%. In our experiments, the threshold step interval is set to 1%, and the threshold is increased layer by layer in our decision tree construction algorithm. When some initial thresholds do not change the classification results, same decision tree

Table 11 Average recognition rate of different initial thresholds on CASIA

Initial threshold	1%, 3%	2%, 4%, 6%	5%, 7~11%	12~31%	≥ 32%
Average recognition rate (%)	76.15	83.65	83.75	81.53	76.12

Table 12 Average recognition rate of different initial thresholds on EMO-DB

Initial threshold	1~5%	6%	7~8%	9~11%	≥ 12%
Average recognition rate (%)	86.39	86.86	84.36	83.73	80.47

Table 13 Parameter values of each SVM after genetic algorithm optimization

The name of SVM	*C*	*g*
SVM1	42.44	0.0591
SVM2	10.34	0.0162
SVM3	48.31	0.1068
SVM4	41.22	0.045
SVM5	8.71	0.02
SVM6	30.3	0.1
SVM7	28.54	0.12

structure is obtained, which results in the same emotion recognition rate. When other initial thresholds change the classification results, different decision tree structures are obtained, which may lead to different emotion recognition rates.

We get the emotion recognition rate corresponding to each threshold, as shown Table 11 and Table 12. According to the results, the optimal initial threshold can be obtained. For CASIA Chinese speech corpus, when the initial threshold is set to 5% or the value in 7~11%, we can achieve the best emotion recognition performance. For the EMO-DB Berlin speech corpus, the best performance can be obtained when the initial threshold is set to 6%. In our experiments, the optimal initial threshold is set to 7% when CASIA Chinese speech corpus is used. In the same way, the optimal initial threshold is set to 6% when EMO-DB Berlin speech corpus is used.

3.4 SVM parameters optimization by genetic algorithm

In the training process of SVM, the penalty factor *C* and kernel function parameter *g* have some effects on the recognition results. The previous experiments of the proposed method do not optimize these two parameters, so this part adopts the genetic algorithm [29] to optimize the parameters of each SVM, and the parameter values of each SVM are shown in Table 13.

Table 14 Emotion recognition rates of decision tree SVM with parameter optimization and feature selection (CASIA) (%)

Emotions	Angry	Happy	Fear	Neutral	Surprise	Sad
Angry	90	5.5	0.5	1	3	0
Happy	2.5	88.5	1	1.5	6	0.5
Fear	1.5	2	74.5	0.5	1.5	20
Neutral	2	3.5	1.5	92.5	0	0.5
Surprise	5	4.5	1	0.5	87	2
Sad	0.5	1.5	18.5	0	1.5	78
Average recognition rate			85.08			

Then, the optimized parameters are applied to the feature selection decision tree SVM for experimentation. The results on two different corpus are shown in Tables 14 and 15. The results show that using the optimized parameter speech emotion recognition system can further increase the average recognition rate. There is a 1.33% improvement on CASIA and a 0.69% improvement on EMO-DB, which verified that the speech emotion recognition performance can be further improved by optimizing penalty factor *C* and kernel function parameter *g*.

4 Conclusion

In order to find the best speech emotional features and establish an effective recognition model in speech emotion recognition, we propose a speech emotion recognition method based on decision tree SVM with Fisher feature selection. Based on the strategy, we have constructed the system frameworks on the CASIA Chinese speech emotion corpus and the EMO-DB Berlin speech corpus. Firstly, the decision tree SVM framework is built by calculating the degree of emotional confusion. Then according to Fisher feature selection method, feature parameters with higher distinguish ability are selected to train each SVM in the decision tree. Thus, the feature dimension is reduced and the computational complexity of the recognition system is decreased. Experiments show that for speech emotion recognition, the decision tree SVM with feature selection strategy proposed in this paper can achieve a recognition rate of 83.75% on CASIA, which is 9% higher than traditional SVM and 8.08% higher than decision tree SVM without feature selection. The results verify that the feature selection method is very effective for emotional recognition based on the proposed decision tree SVM. The conclusion can also be verified on Berlin speech corpus. Subsequently, when the genetic algorithm is used to optimize the penalty factor *C* and the kernel function parameter *g* for

Table 15 Emotion recognition rates of decision tree SVM with parameter optimization and feature selection (EMO-DB) (%)

Emotions	Angry	Happy	Fear	Neutral	Boring	Sad	Disgust
Angry	91.8	3.4	1.2	0.25	0.9	0.8	1.65
Happy	6.4	85	5.5	0.24	0.82	0.5	1.54
Fear	3.5	2.74	88.3	0.42	1.8	2.01	1.23
Neutral	1.2	2.75	0.81	94.5	0.14	0.3	0.2
Boring	2.1	1.85	1.74	0.9	89.85	1.85	1.71
Sad	0.57	1.5	2.4	1.5	8.4	83	2.45
Disgust	5.5	4.75	4.2	1.2	1.5	2.4	80.45
Average recognition rate			87.55				

each SVM in the decision tree, the system can further increase the average recognition rate of the six emotions by 1.33% on CASIA.

From the experiment results, we can find that the confusion between Fear and Sad on CASIA is still relatively high. Therefore, it is necessary to carry out targeted research on the two emotions types to search for more effective feature parameters in future study. At the same time, we need to find more effective feature selection strategies for feature selection.

Acknowledgements
Not applicable

Funding
This work is supported by the National Natural Science Foundation of China (61671252, 61501251, 61571233), the Natural Science Foundation of Jiangsu Province (BK20140891).

Availability of data and materials
Not applicable

Authors' contributions
LS designed the core methodology of the study, carried out the implement, and drafted the manuscript. SF and FW carried out the experiments and drafted the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Received: 7 June 2018 Accepted: 13 December 2018
Published online: 07 January 2019

References

1. Clavel, C., Vasilescu, I., Devillers, L., et al. (2008). Fear-type emotion recognition for future audio-based surveillance systems. *Speech Communication, 50*(6), 487–503.
2. Ramakrishnan, S., & Emary, I. M. M. E. I. (2013). Speech emotion recognition approaches in human computer interaction. *Telecommunication Systems, 52*(3), 1467–1478.
3. Cynthia, B., & Aryananda, L. (2002). Recognition of affective communicative intent in robot-directed speech. *Kluwer Academic Publishers, 12*(1), 83–104.
4. Li, Z., & Huang, C. W. (2014). Key technologies in practical speech emotion recognition. *Journal of Data Acquisition and Processing, 29*(2), 157-170.
5. Zhang, X. D., Huang, C. W., Zhao, L., & Zou, C. R. (2014). Recognition of practical speech emotion using improved shuffled frog leaping algorithm. *Chinese Journal of Acoustics, 34*(4), 441–456.
6. Cao, W. H., Xu, J. P., & Liu, Z. T. (2017). Speaker-independent speech emotion recognition based on random Forest feature selection algorithm. *Chinese control conference (CCC). 2017 36th Chinese. IEEE*, 10995–10998.
7. Wang, K., An, N., Li, B. N., et al. (2015). Speech emotion recognition using Fourier parameters. *IEEE Transactions on Affective Computing, 6*(1), 69–75.
8. Xiao, Z., Dellandrea, E., Dou, W., et al. (2010). Multi-stage classification of emotional speech motivated by a dimensional emotion model. *Multimedia Tools and Applications, 46*(1), 119–145.
9. Lugger, M., & Yang, B. (2007). The relevance of voice quality features in speaker independent emotion recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP. IEEE, 4*, IV-17–IV-20.
10. Huang, C. W., Y, J. I. N., WANG, Q. Y., et al. (2010). Speech emotion recognition based on decomposition of feature space and information fusion. *Signal Processing, 26*(6), 835–842.
11. Nwe, T. L., Foo, S. W., & Silva, L. C. D. (2003). Speech emotion recognition using hidden Markov models. *Speech Communication, 41*(4), 603–623.
12. Mao, X., Chen, L., & Fu, L.(2009). Multi-level speech emotion recognition based on HMM and ANN. *Computer science and information engineering, 2009 WRI World Congress on IEEE*, 2009: 225–229.
13. Sinith, M. S., Aswathi, E., Deepa, T. M., Shameema, C. P., & Rajan, S. (2015). Emotion recognition from audio signals using support vector machine. *IEEE Recent advances in intelligent computational systems (RAICS), Trivandrum, 2015*, 139–144.
14. Chandrakala, S., & Sekhar, C. C. (2009). Combination of generative models and SVM based classifier for speech emotion recognition. *International Joint Conference on Neural Networks. IEEE Press*, 1374–1379.
15. Zhang, X. D., Huang, C. W., Li, Z., et al. (2014). Recognition of practical speech emotion using improved shuffled frog leaping algorithm. *Chinese Journal of Acoustics, 33*(4), 441–441.
16. Fu, L., Mao, L., & Chen, L. (2008). Speaker independent emotion recognition based on SVM/HMMS fusion system. *International conference on audio, language and image processing. IEEE*, 2008:61–65.
17. Huang, Y., Zhang, G., Dong, F., et al. (2013). Speech emotion recognition using stacked generative and discriminative hybrid models. *ACTA ACUSTICA, 38*(2), 231–240.
18. Lee, C. C., Mower, E., Busso, C., et al. (2009). Emotion recognition using a hierarchical binary Decisi-on tree approach. *INTERSPEECH,* 53(9), 1162–1171.
19. Garg, V., Kumar, H., & Sinha, R. (2013). Speech based emotion recognition based on hierarchical decision tree with SVM, BLG and SVR classifiers. *Communications. IEEE*, 2013: 1–5.
20. Song, P., Zhang, X., S, O., et al. (2016). Cross-corpus speech emotion recognition using transfer semi-supervised discriminant analysis. *Chinese spoken language processing (ISCSLP), 2016 10th international symposium on. IEEE*, 1–5.
21. Schuller, B., & Burkhardt, F. (2010). Learning with synthesized speech for automatic emotion recognition. In *Proc. of the 2010 IEEE Int'l Conf. on Acoustics Speech and Signal Processing (ICASSP)* (pp. 5150–5153). Dallas: IEEE Computer Society.
22. Liang, R., Zhao, L., Tao, H., et al. (2016) Speech emotion recognition algorithm based on the selective attention mechanism. *ACTA ACUSTICA*, 41(4), 537–544.
23. Jiang, X. Q., Tian, L., & Cui, G. H. (2006). Statistical analysis of prosodic parameters and emotion recognition of multilingual speech. *ACTA ACUSTICA, 31*(3), 217.
24. Busso, C., Lee, S., & Narayanan, S. (2009). Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE Transactions on Audio, Speech and Language Processing, 17*(4), 582–596.
25. Yang, B., & Lugger, M. (2010). Emotion recognition from speech signals using new harmony features. *Signal Processing, 90*(5), 1415–1423.
26. Fan, R. E., Chen, P. H., & Lin, C. J. (2005). Working set selection using second order information for training support vector machines. *Journal of Machine Learning Research, 6*(4), 1889–1918.
27. Ren, H., Ye, L., & Li, Y. (2017). X J Sha. *Speech emotion recognition algorithm based on multi-layer SVM classification. Application Research of Computers, 34*(6), 1–4.
28. Hosseini, Z., Ahadi, S. M., & Faraji, N. (2014). Speech emotion classification via a modified Gaussian mixture model approach. *International Symposium on Telecommunications. IEEE*, 487–491.
29. Saini, L. M., Aggarwal, S. K., & Kumar, A. (2009). Parameter optimisation using genetic algorithm for sup-port vector machine-based price-forecasting model in national electricity market. *IET Generation Transmission and Distribution, 4*(1), 36–49.