Speech Emotion Recognition Based on DNN-Decision Tree SVM Model

Linhui Sun , Bo Zou , Sheng Fu , Jia Chen , Fu Wang

PII: S0167-6393(19)30127-X
DOI: https://doi.org/10.1016/j.specom.2019.10.004
Reference: SPECOM 2671

To appear in: *Speech Communication*

Received date: 3 April 2019
Revised date: 29 August 2019
Accepted date: 16 October 2019

Please cite this article as: Linhui Sun , Bo Zou , Sheng Fu , Jia Chen , Fu Wang , Speech Emotion Recognition Based on DNN-Decision Tree SVM Model, *Speech Communication* (2019), doi: https://doi.org/10.1016/j.specom.2019.10.004

HIGHLIGHTS

- We propose to combine the multi-level decision idea and deep learning to finish speech emotion recognition. The idea of multi-level classification is mainly realized by tree structure.
- The decision tree SVM structure is firstly constructed by computing the confusion degree of emotion, and then different DNN networks are trained for diverse emotion groups to extract the bottleneck features that are used to train each SVM in the decision tree.
- The new method can effectively solve the problem that recognition rate decreases due to the increase of emotional categories compared to the traditional SVM and DNN-SVM classification method.

Speech Emotion Recognition Based on DNN-Decision Tree SVM Model

Linhui Sun, Bo Zou, Sheng Fu, Jia Chen and Fu Wang

College of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China

(e-mail:sunlh@njupt.edu.cn;zb12387@163.com;fusheng9445@163.com; 1302416675@qq.com;15062201099@163.com).

Corresponding author at: College of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China

E-mail address: sunlh@njupt.edu.cn (L. Sun)

***Abstract*--Motivated by the development of DNN technology, a speech emotion recognition method based on DNN-decision tree SVM model is proposed. The proposed method can not only excavate the deep emotion information of the speech signal, but also extract more distinctive emotion features from the easily confused emotions. In this method, the decision tree SVM structure is firstly constructed by computing the confusion degree of emotion, and then different DNN networks are trained for diverse emotion groups to extract the bottleneck features that are used to train each SVM in the decision tree. Finally, speech emotion classification is realized based on this model. This model is assessed by using the Chinese Academy of Sciences Emotional Corpus. The experiment results show that the average emotion recognition rate based on the proposed method is 6.25% and 2.91% higher than traditional SVM and DNN-SVM classification method, respectively. It is proved that this method can effectively reduce the confusion between emotions, thus improving the speech emotion recognition rate.**

***Keywords*--Speech emotion recognition; emotional confusion; decision tree SVM; deep neural network; bottleneck feature.**

1 Introduction

In recent years, with the rapid development of artificial intelligence, voiceprint, iris, fingerprint, face and other biometrics has attracted wide attention (Khokher et al., 2015). As an important information carrier for people's emotional and cognitive communication, speech signal is the most basic and natural way of communication in people's life and work. It not only contains its own semantic information, but also carries the speaker's identity information and emotional state. With the further improvement of computer processing ability and the growth of people's demand for intelligent life, emotion recognition has been more and more widely used in human-computer interaction (Zhang et al., 2013; Sui et al., 2017; Le et al.,2014; Mustafa et al.,2018; Rázuri et al., 2015), and attracted increasingly attention from researchers at home and abroad.

Intelligent speech emotion recognition mainly simulates and understands human emotion through computer, and then matches the target emotion using prosodic features, spectrum-based correlation features, sound quality features contained in speech information, thus completing the task of emotion recognition. The speech emotion recognition system mainly includes three parts: speech data preprocessing, emotion feature extraction and emotion classifier (Lu et al., 2018). Because humans are affected by their own physical factors and external environmental conditions in the process of speaking, the emotions will have diversity and variability. Therefore, robust classification model and speech emotion features with discriminative information are the two important factors in emotion recognition.

In general, the main feature parameters used in speech emotion recognition system can be divided into traditional features and depth features. Huang et al. (2012) fused energy, zero-crossing rate and fundamental frequency to finish speech emotion recognition, and achieved an average recognition rate of 56.46% in Berlin emotion corpus with seven kinds of emotion. Wang et al. (2015) proposed a new Fourier parameter model using the first-order and second-order differences for speaker-independent speech emotion recognition. It obtained good performances on the Germany database (EMODB) and Chinese language database (CSAIA). In the multiple speech emotion recognition, since various types of features have different contributions to the final recognition results, it is necessary to select the most appropriate acoustic characteristics. Huang et al. (2010) used Fisher discriminant coefficient to select appropriate 10-dimensional features from 84-dimensional features for each emotion pair for five kinds of emotions recognition. Based on the idea of tree structure, Sun et al. (2019) proposed to use Fisher feature selection method to remove the redundant information. A three-layer model based on the acoustic features and feature selection was proposed in Li and Akagi (2019) to improve multilingual speech emotion recognition performance. In recent years, with the improvement of computer hardware, it is possible to design and implement deep neural networks. A major application of DNN is a tool to extract speech feature parameters (Li and Xu, 2017; Mao et al., 2014; Panagiotis et al., 2018), which contain deep information of speech signal. Then the acquired features can be used to train other models. For example, in Li and Xu (2017), the bottleneck features were extracted using DNN network, and then applied to train SVM model for speech emotion recognition. Mao et al. (2014) introduced the convolutional neural networks (CNN) to learn the distinctive features of speech emotions. Firstly, the speech was converted into a fixed-scale spectrogram, and then input into the CNN to learn the deep characteristics automatically. Finally, the recognition task was carried out based on this method. The experiment result showed that it had better

robustness and stability when the noise and speech signals were mutated. Panagiotis et al. (2018) proposed a new continuous speech emotion recognition model. CNN was used to extract emotion features from the original signal and stack a two-layer long and short-term memory (LSTM). The system fully considered the context of the speech signal and it was superior to other systems in terms of consistency coefficient and recognition rate. Good classifiers are also critical to system identification rates. The main classifiers used in speech emotion recognition are Gauss Mixture Model (GMM) (Lanjewar et al., 2015), Artificial Neutral Network (ANN) (Xu et al., 2012; Zhou et al., 2017; Trentin et al., 2015), Support Vector Machine (SVM) and the fusion of different classifiers. The GMM classifier mainly estimates the sample probability density distribution, and uses the maximum likelihood method to classify test samples. It has a large amount of computation and achieve good results under small batch and ideal conditions. ANN has powerful associative memory function and are used for non-linear binary classification problems. Based on statistical learning theory, SVM has shown many unique advantages in solving small sample, nonlinear and high-dimensional pattern recognition (Gupta and Mehra, 2015; Torres et al., 2017). Therefore, more and more researches are gradually using SVM for classification. The work in Zhu et al. (2017) attempted to combine Deep Belief Network(DBN) extracting the deep features and SVM instead of using only one of them. Gender-dependent experiments were conducted using an emotional speech database created by the Chinese Academy of Sciences. The result showed that DBN could work very well for small training databases if it is properly designed.

In multiple emotion recognition, the overall recognition rate decreases due to the increase of confusion between emotions, and the same kind of emotion features have different contributions to distinguishing different emotion. However, the decision tree can classify multiple emotions at different levels, which can effectively reduce the confusion between emotions. Therefore, in order to take into account the advantages of DNN and decision tree SVM, we propose a speech emotion recognition model based on DNN-decision tree SVM. In this method, the decision tree SVM structure is firstly constructed by computing the confusion degree of emotion, and then different DNN networks are trained for different emotion classifications to extract the bottleneck features, which are used to train each SVM in the decision tree. Finally, the model is applied to speech emotion recognition to improve the system's performance. In Trabelsi et al. (2016), they study the combination of GMM method with different generative models and SVM for the robust emotion recognition. Moreover, the binary decision tree model and SVM were also applied to the emotion recognition (Garg et al., 2013; Lee et al., 2011), which have shown superior performance.

The rest of the paper is organized as follows: Section 2 describes the proposed emotion recognition system, the extraction of bottleneck features and the decision tree SVM model. The attained results in the current work and their discussion are demonstrated in Section 3. Finally, Section 4 gives the concluding remarks of this work.

2 Proposed emotional recognition method

In emotion recognition system, we need to extract some characteristic parameters that can reflect emotion information, and then use these characteristic parameters to train the classification model. Finally, the trained model is used for recognition. Therefore, the selection of feature parameters directly affects the system performance. The traditional speech features commonly used in emotion recognition tasks can be divided into three types: prosodic features, spectral-based features and sound quality features (Liang et al., 2016; Kachele et al., 2014). These traditional features can represent the shallow features of the speech signal, but cannot describe the emotional features in a deeper level. The DNN can extract deep bottleneck features to mine deeper information of speech signals. In this paper, deep bottleneck features are used for emotion recognition. With the increase of the emotion category, the degree of confusion between emotions also increases. It is difficult for traditional SVM classifier to make accurate judgments, thus reducing the recognition rate of the system. The tree structure SVM integrates the idea of multi-level classification, which can effectively reduce the confusion degree. Therefore, in order to improve the recognition rate of multi-classification speech emotion system, we propose an emotion recognition method based on DNN-decision tree SVM. The block diagram is shown in Fig.1.

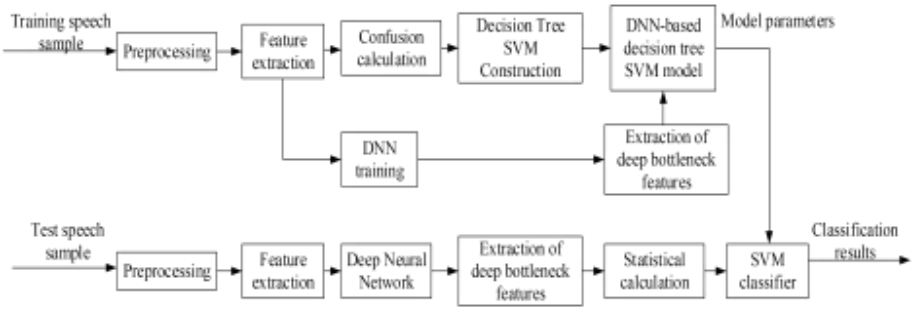


Fig.1. Emotion recognition based on DNN-decision decision tree SVM model

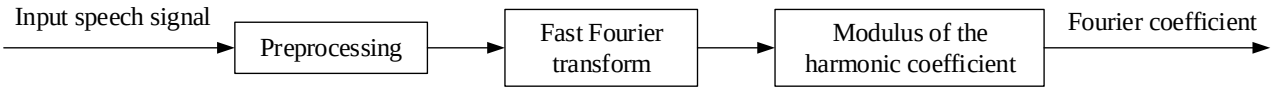


Fig. 2. Flow chart of Fourier coefficient extraction

In the training stage, the speech signal is preprocessed by pre-emphasis and framing to extract the Fourier coefficient of the speech signal. Then MFCC coefficient and the traditional SVM are combined with the proposed decision tree construction algorithm to create a suitable tree classification structure. Lastly, according to the emotion groups, different deep networks are trained to extract the speech bottleneck features that are used to train each SVM in the decision tree. In the test stage, bottleneck features are extracted from the test samples using DNN and corresponding statistics are also calculated. These statistics are input into the trained model for classification decision.

2.1 Extraction of speech emotion bottleneck feature

In the previous work, the Fourier coefficients (Busso et al., 2009; Yang and Lugger, 2010) were often applied to speech emotion recognition and achieved good results. In this paper, we use the Fourier coefficient as the characteristic parameter of training DNN. The flow chart of the extraction process is shown in Fig.2. The preprocessing mainly includes endpoint detection, framing, windowing and so on.

In deep network structure, when the number of neurons in one hidden layer is far less than that in other hidden layers, the layer is called bottleneck layer. The corresponding characteristic parameters from bottleneck layer are called bottleneck features. It can not only extract the deep emotion information of speech signal, but also reduce the dimension of the input feature parameters and the computational complexity. The model structure is shown in Fig.3. Firstly, the training speech samples are preprocessed, and then the Fourier coefficients of the signals are extracted as the input of DNN. When the training of DNN is completed, all the layers of the network after the bottleneck layer are removed. Thus, the network for extracting the bottleneck features is completed.

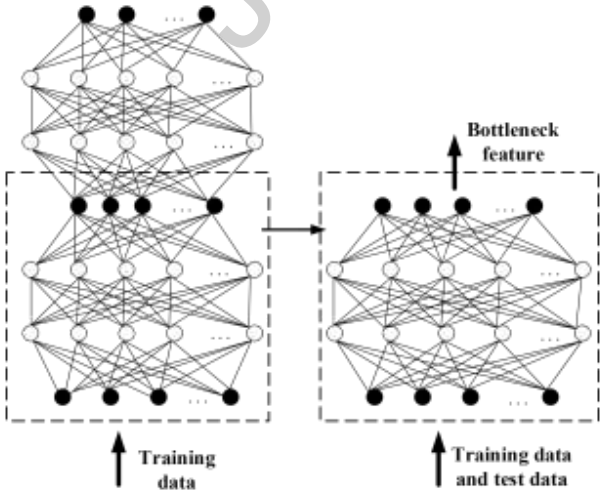


Fig. 3. DNN model of extracting bottleneck features

The emotional information of the speech signal is abundant, so it is difficult to comprehensively reflect the emotional information of speech by dividing the signal into 20-30 ms speech segments.

In this paper, we use the Fourier coefficients extracted from multi-frame as the input of DNN, and then extract the deep bottleneck feature as the feature parameter and calculate the five statistics (minimum, maximum, standard deviation, mean and median) of the bottleneck feature in units of one sentence. Finally, the statistics are used to train SVM classifiers.

2.2 Training of Deep Neural Network

DNN is a traditional multi-layer perceptron (MLP) with multiple (more than two) hidden layers and each layer of DNN is realized by Restricted Boltzmann Machine (RBM) (Hinton et al., 2012). The training process can be divided into two stages: pre-training and fine-tuning. In the pre-training stage, each RBM is trained unsupervised by training data. In this process, training data is used as the input of the first RBM and the input of other RBM is the output of the previous RBM. In the fine-tuning stage, when the RBM pre-training is completed, the parameters of back propagation (BP) neural network are initialized by the parameters obtained from the RBM pre-training. Then the BP neural network is applied to realize fine-tuning the parameters of each RBM in a top-down manner, so as to complete the training of model.

2.2.1 Pre-training based on RBM

RBM is a generative neural network model with random characteristics (Zhang et al., 2015). It is essentially an undirected graph model consisting of a visible layer and a hidden layer. RBM itself is a two-layer connectionist system with no connection between units in the same layer. The value types of neurons in the visible layer are generally real values or binary. In the hidden layer, the values of neurons are generally binary numbers obeying Bernoulli distribution. According to the energy theory, the relationship between the visible layer v and the hidden layer h of each RBM is assigned an energy value, which is defined as:

$$E(\mathrm{v},h\,|\,\theta)=-\sum_{i=1}^{n}\sum_{j=1}^{m}W_{ij}v_ih_j-\sum_{i=1}^{n}b_iv_i-\sum_{j=1}^{m}a_jh_j \qquad \textbf{(1)}$$

where $\theta=\{\mathbf{w},\mathbf{a},\mathbf{b}\}$ is the parameter set of RBM model. $\mathbf{w}$, $\mathbf{a}$ and $\mathbf{b}$ denote the connection weight vector, visible layer bias vector and hidden layer bias vector, respectively. Additionally, the w_{ij}, a_i and b_i are their corresponding components. n, m are the number of visible and hidden units, respectively. When the energy value between the visible layer and the hidden layer is calculated, the probability distribution assigned to the visible-hidden unit pair can be defined as

follows:

$$p(v,h\,|\,\theta)=\frac{e^{-E(v,h|\theta)}}{Z(\theta)} \quad (2)$$

where $Z(\theta)$ is the normalization term (partition function) defined by $Z(\theta)=\sum_{v,h}e^{-E(v,h|\theta)}$. The optimization objective of RBM is to maximize the probability distribution of nodes in the visible layer. In the process of training, Contrastive Divergence (CD) (Fischer and Igel, 2011) is used to estimate the parameters of RBM.

In the pre-training process of DNN network, the training data is used as the input of the first RBM. After the training of the first layer finished, the output value of the hidden layer is used as the input of the visible layer of the next RBM, thus completing the training of the second layer of RBM. In this way, the output of the hidden layer of the former layer is used as the input of the visible layer of the next layer, thus realizing the pre-training process of all RBM. After all the RBM training is completed, each RBM is superimposed in a hierarchical relationship to obtain a trained multi-layer network structure.

2.2.2 *Fine adjustment based on BP algorithm*

When the DNN pre-training is finished, the parameters of the pre-training are taken as initialization parameters of the network, and then a softmax function output layer is added to form a complete DNN network. Each output node of DNN corresponds to a category, so as to supervise training.

In the parameter fine adjustment process, the model parameters $\{w, b\}$ are learned by classical BP algorithm. The model parameter updating formula is as follows:

$$\mathbf{W}_{t+1}^{l}\leftarrow\mathbf{W}_{t}^{l}-\varepsilon\Delta\mathbf{W}_{t}^{l} \quad (3)$$

$$\mathbf{b}_{t+1}^{l}\leftarrow\mathbf{b}_{t}^{l}-\varepsilon\Delta\mathbf{b}_{t}^{l} \quad (4)$$

where ε is the learning rate, $\mathbf{W}_t^l, \mathbf{b}_t^l$ are the weight matrix and bias vectors of the l layer after the t iterations of the network update. In the process of using BP algorithm to optimize parameters, we adopt Cross Entropy (CE) function as the tuning function and estimate the model parameters by calculating the minimum cost function. The calculation formula is as follows:

$$L=-\sum_{s}d_s\log p_s \quad (5)$$

where d_s is the correct category value. When s belongs to the category of training data, the d_s is equal to one, otherwise d_s is equal to zero.

2.3 Construction of DNN-decision tree SVM model

Firstly, we define the set of emotional states A= {anger, happy, fear, neutral, surprise, sad}, where the number of emotional states in A is six. The degree of similarity between emotions is defined as the degree of emotional confusion, which is expressed by $I_{m,n}$. The $I_{m,n}$ is used to represent the arithmetic mean of the probability that the m-th emotion A_m is misjudged as the n-th emotion A_n and the n-th emotion A_n is misjudged as the m-th emotion A_m(Zhan et al., 2013). The calculation formula of the above process is as follows:

$$I_{m,n}=\frac{P(s=n\,|\,x\in A_m)+P(s=m\,|\,x\in A_n)}{2} \quad (6)$$

where x is the test sample and s is the classification result of the test sample x .

The specific steps of the DNN-decision tree SVM construction algorithm (Wang et al., 2018) are as follows (taking four emotions a, b, c, d as example).

Step1: Feature parameters are extracted from the input signals, and input into the SVM classifier. The confusion matrix is obtained from the judgment results and the confusion degree between two different emotions is calculated according to formula (6).

Step2: Set the first classification threshold to P. If $I_{d,i}<P$ ($i\in\{a,b,c\}$), divide d into the first category separately. If $I_{a,b}>P, I_{b,c}>P$, then divide the emotions a, b into one group, and b, c are also sorted into another group. That is to say, the three emotions a, b and c are divided into the second category.

Step3: If the number of emotions contained in a category is greater than two, a second subdivision is required. When a new round of judgment is added, the threshold will increase P automatically. According to the Step2, the number of emotions in the second category is three, so the threshold is increased to 2P. The confusion matrix and degree of confusion of a, b and c are recalculated and judge again according to Step2. If $I_{a,b}>2P, I_{b,c}>2P$, then a, b, c are classified into a big category, which is the final grouping result. Otherwise, according to the calculated confusion, a, b and c are further divided into smaller subclasses.

Step4：After all the emotions to be grouped are correctly divided, according to the emotion groups, different deep networks are trained to extract the speech bottleneck features.

Step5: Each SVM in the decision tree is trained using the statistics of the bottleneck features. The algorithm ends.

In the test process, bottleneck features of the test samples are extracted using DNN and corresponding statistics are also calculated. These statistics are input into the trained DNN-decision tree SVM model for the final classification.

3 Experiments and results analysis

3.1 Experiment preparation

We select the Chinese emotion database of the Chinese Academy of Sciences as the experiment corpus. Because of the complexity of the Chinese language, high accuracy emotion recognition of Chinese speech is challenging. The database was recorded by two male and two female professional speakers with six emotions: angry, happy, fear, neutral, surprise and sad. Each participant reads 50 times with six kinds of emotion. Therefore, there are 1200 (4×50×6) sentences in total. The sampling rate is 16-kHz, 16-b quantization. We choose six different emotion sentences, of which 160 files are used as training samples and 40 files are

used as test samples. In the experiment, we choose the SVM as emotional classifier and adopt the LIBSVM toolbox developed by Lin Zhiren of Taiwan University. The kernel function is Radial Basis Function (RBF). Experiment environment is Matlab2013a and the installation environment of LIBSVM is Visual Studio 2012.

During the process of extracting the speech signal parameters, the endpoint detection is carried out in advance and the speech is divided into frames in the form of 256 frame lengths and 128 frame shifts. The feature parameters used in this work are the Fourier coefficients mentioned in section 2.1. Each frame contains 256 points. Five consecutive frames of feature parameters are joined together to form a 1280-dimension long vector. Then the feature parameters are normalized and used as the input of DNN. The DNN model we used consists of seven layers and each layer adopt "s" activation function. First, the input layer is a long vector of 1280 dimensions. After the input layer, there are five hidden layers, one of which is set as the bottleneck layer with 100 neurons. The number of neurons in other hidden layers is 1280. Lastly, the output layer is a softmax layer and the size is the same as the category, which is six in our experiments. We use the CD algorithm to update the RBM weights and adopt the BP algorithm to adjust the parameters. When DNN training is completed, the bottleneck features of different emotions are extracted separately and their five global statistical variables (minimum, maximum, standard deviation, mean, and median) are calculated, which are used to train each classifier SVM. In the following experiments, we perform the gender-independent emotion recognition.

3.2 *Experimental simulation and analysis*

3.2.1 *Calculation of emotions confusion*

Firstly, we use the traditional MFCC and SVM to conduct experiments on six emotions in the corpus. The recognition rates of the six kinds of emotions are shown in Table 1.

According to Eq. (6) in Section 2.3, the confusion degree is calculated for the above confusion matrix. The confusion degree between the six emotions is obtained as shown in Table 2, and the threshold is set to 7% at the initial classification (the selection of initial threshold will be discussed in Section 3.2.3).

According to the results calculated in Table 2, since the confusion degree between happy and surprise, surprise and anger are both 11.25%, which is greater than the threshold of 7% set by the initial classification. Therefore, the happy, surprise, anger are divided into the first class based on the decision tree construction algorithm proposed in Section 2.3.

Similarly, the confusion between sad and fear is 31.75%, which is also greater than 7%, so they are merged into the second class. The confusion degree between neutral and the other five categories is less than the threshold, so the neutral should be categorized as the third category alone. The first round of classification is completed by SVM1. According to Step3 in the construction algorithm, when the number of emotions in a large class is greater than two, further subdivision is needed. We calculate the confusion degree of happy, surprise, anger and the result is shown in Table 3. At the second round, the threshold automatically increases P on the basis of the previous round. From Table 3, it can be seen that the confusion degree between two groups is less than 2P (14%). Thus, the three kinds of emotion are directly classified by SVM2. Because there are only two emotions in the second class, so we need not to divide them again and SVM3 is constructed to classify the two emotions. Finally, the DNN-decision tree SVM structure is obtained as shown in Fig.4.

3.2.2 *Influence of deep network parameters on system performance*

In order to verify the influence of the parameters on the performance of the model and find the optimal network parameters, we use the bottleneck features and traditional SVM to test the six emotions in the corpus. By changing different network parameters, we verify its effect and find the optimal system parameters.

(1) In order to verify the impact of batch size on network performance, the batch values are set to 5, 10, 15, 20, 25 and 30 to carry out six groups of experiments. The DNN structure is 1280-1280-100-1280-1280-6, and the learning rate in the pre-training phase is 0.005. The fine-tuning stage is 0.01, and the number of iterations is 200. The average recognition of

Table 1 Speech emotion recognition results based on traditional MFCC and SVM (%)

Emotion category	Anger	Happy	Fear	Neutral	Surprise	Sad
Anger	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

Table 2 Confusion degree of six emotions

Emotion category	Anger	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 among three emotions of anger, happy and surprise (%)

Emotion category	Anger	Happy
Happy	9	—
Surprise	12	11.5

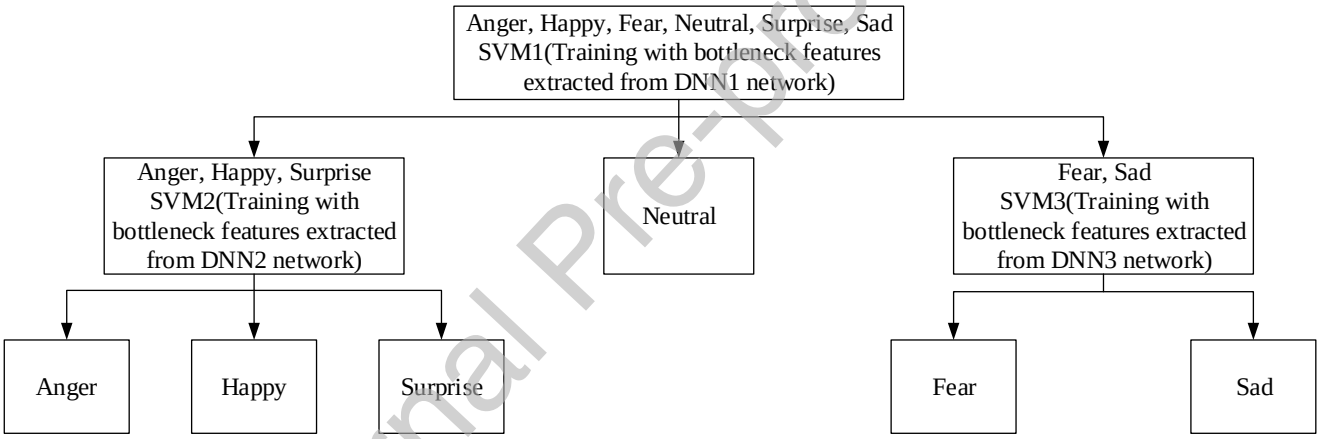


Fig. 4. Structure diagram of the DNN-decision tree SVM model

six emotions is calculated respectively and the results are shown in Table 4. It can be seen from Table 4 that when the batch value is greater than 10, the average recognition rate of the system will decrease with the increase of the batch value. The result verify that large batch may cause network overfitting, thus reducing the average recognition rate of the whole system. Therefore, when we train the DNN network, appropriate batches should be selected according to the actual situation. The average recognition rate of the system is the highest when the batch value is 10, so we set the batch size is 10 in the subsequent experiment.

(2) In order to find the influence of the network structure on the system effect, we conduct a set of comparative experiments on the change of hidden layer position. In this paper, the first hidden layer, second hidden layer, third hidden layer, fourth hidden layer and the fifth hidden layer are respectively set as bottleneck layers (i.e., the network structure is:1280-100-1280-1280-1280-1280-6, 1280-1280-100-1280-1280-1280-6, 1280-1280-1280-100-1280-1280-6, 1280-1280-1280-1280-100-1280-6, 1280-1280-1280-1280-1280-100-6) to perform the comparative experiments. The recognition rates are shown in Table 5. According to Table 5, when the bottleneck layer is set to the third hidden layer, the average recognition rate of the system is 71.25%, which is higher than the other case. Therefore, we set the third hidden layer as the bottleneck layer in subsequent experiments.

(3) Learning rate is also an important parameter of the model and its value directly affects the convergence speed. In order to study the impact of different learning rates on the system, experiments are conducted with the learning rates set to 0.001, 0.003, 0.005, 0.008 and 0.01. The recognition results are shown in Table 6. The results indicate that the average recognition rate of the system increases as the learning rate decreases. When the learning rate is 0.001, the performance of the system is the best. The main reason is that when the learning rate is smaller, the network has more sufficient training and good convergence. Thus, these bottleneck features

Table 4 Average recognition rate with different batch size (%)

Batch size	5	10	15	20	25	30
Average recognition rate	70.42	71.25	69.78	67.92	67.14	66.34

Table 5 Average recognition rate of bottleneck layer at different positions (%)

Location of the bottleneck layer	first hidden layer	second hidden layer	third hidden layer	fourth hidden layer	fifth hidden layer
Average recognition rate	70.23	70.76	71.25	66.67	60

Table 6 Average recognition rate at different learning rates (%)

Learning rate	0.001	0.003	0.005	0.008	0.01
Average recognition rate	72.92	71.89	71.25	69.58	68.75

Table 7 Average recognition rate of different initial thresholds (%)

Initial threshold	1%~2%	3%,4%,6%	5%,7%~9%	10%~15%	16%~31%	≥ 32%
Average recognition rate	67.35	73.62%	75.83	71.18%	70.82%	67.51%

contain more abundant emotion information and can better distinguish all kinds of emotions.

3.2.3 The influence of different initial threshold P

In the proposed recognition system, we need to construct a suitable tree structure. As mentioned in Section 2.3, the initial threshold should be given firstly. The different values P may construct distinct tree structures, thus affecting the final performance. Once the optimal initial threshold P is determined, the structure of the tree is uniquely determined. Therefore, it's essential to select an optimal initial threshold at first. In the Section 3.2.1, we calculate the confusion degree of six emotions. The result are shown in Table 2. It can be seen that the value of confusion degree varies from 1% to 32%. Therefore, we set the threshold step size to 1% for a group of comparative experiments. The emotion recognition rate of each threshold is displayed in the Table 7.

According to the obtained results, when the initial threshold P is set to 5% or in the interval 7%~9%, the overall average recognition rate is the highest. Therefore, in the process of constructing tree structure, the threshold is set to 7%.

3.2.4 Comparative experiment of bottleneck layer features and traditional features

In order to verify the superiority of bottleneck feature, we conduct a comparative experiment of speech emotion recognition based on traditional feature and bottleneck feature.

We adopt the 256-dimensional Fourier coefficient and SVM model to complete the speech emotion recognition, and the results of the six emotions are shown in the third column of Table 8. The bottleneck features of speech signals are extracted by using DNN with the structure of 1280-1280-100-1280-1280-1280-6(according to the experiment results of Section 3.2.2, the batch size of the network is set to 10, the learning rate is 0.001 and the number of iterations is 200), the recognition results are shown in the fourth column of Table 8. As can be seen from Table 8, compared with the traditional features and SVM-based system, the method based on bottleneck features and SVM can partly improve the recognition rate of four emotions of happy, anger, fear and sad. What's more, the average recognition rate of the six emotions increases from 69.58% to 72.92%, when the dimension of feature parameters decreases. This fully demonstrates that DNN can mine the deeper emotion features of the speech signal and compress the information in the bottleneck layer. Therefore, the bottleneck feature is more distinguishable than the traditional feature, which improves the performance of the whole system.

3.2.5 Comparative experiment of DNN-SVM and DNN-decision tree SVM model

In Section 3.2.1, we have constructed the DNN-decision tree SVM recognition system shown in Fig.4. In the training stage, the batch size of all three networks is 10, and learning

rate is 0.001. The bottleneck layer is in the third layer and the number of iteration is 200. The number of neurons in the bottleneck layer is also an important factor. As the number of neurons varies, the extracted features will change to some

Table 8 Comparison of recognition results between traditional features and bottleneck features (%)

Feature parameters		Traditional features	Bottleneck features
Emotion category	Happy	57.5	69
	Anger	72.5	73
	Fear	52.5	60.5
	Neutral	92.5	92.5
	Surprise	82.5	77.5
	Sad	60	65
Average recognition rate		69.58	72.92

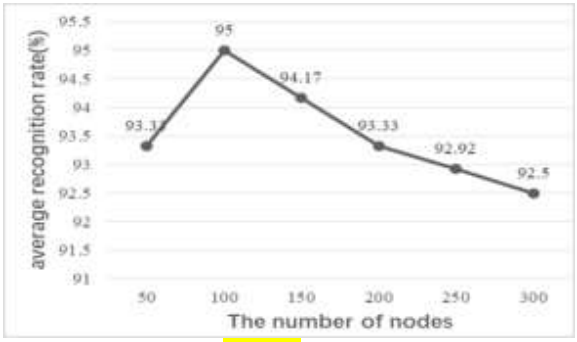
Table 9 Speech emotion recognition results based on DNN-decision tree SVM (%)

Emotion category	Happy	Anger	Fear	Neutral	Surprise	Sad
Recognition rate	80	70	60	92.5	85	67.5
Average recognition rate				75.83		

extent, which affects the whole system performance. When the number of neurons is too large, bottleneck features with noise may interfere the system, thus reducing the recognition rate. When the number of neurons in the bottleneck layer is too small to be much smaller than that of input layer, the bottleneck features can not represent the information of input features adequately, so that some effective classification information is lost, resulting in bad performance of the whole system. In our proposed DNN-decision tree model, there are three DNNs to extract the speech bottleneck features. In order to ensure that the final recognition rate can achieve a better result, we optimize the number of bottleneck nodes in each DNN network. For the three different DNNs, the number of nodes in the bottleneck layer is set to 50, 100, 150, 200, 250, 300, to find out the optimal number of nodes. The relationship between the average recognition rate of DNN1, DNN2 and DNN3 and the number of nodes is shown in Fig. 5, Fig. 6 and Fig. 7, respectively.

According to Fig. 5, Fig. 6 and Fig. 7, in the DNN1 and DNN3, when the number of bottleneck nodes is 100, the average recognition rate reaches a peak of 95% and 66.25% respectively. As the number of nodes increase or decrease, the recognition rate shows a trend of gradual decline. In the DNN2, when the number of nodes is less than 150, the average recognition rate of happy, angry and surprise gradually increases with the increase of the number of bottleneck nodes. The recognition rate reaches a peak of 79.16% when the node is set to 150.

Fig. 5. Relationship between the average recognition rate of three major classes and the number of nodes in the bottleneck layer (DNN1)



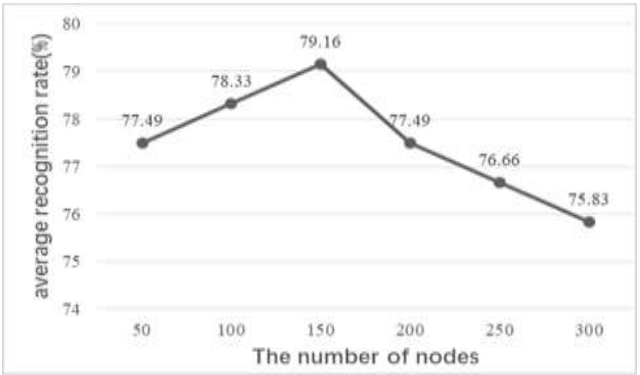


Fig. 6. (happy, anger, surprise) Relationship between the average recognition rate of three emotions and the number of bottleneck nodes (DNN2)

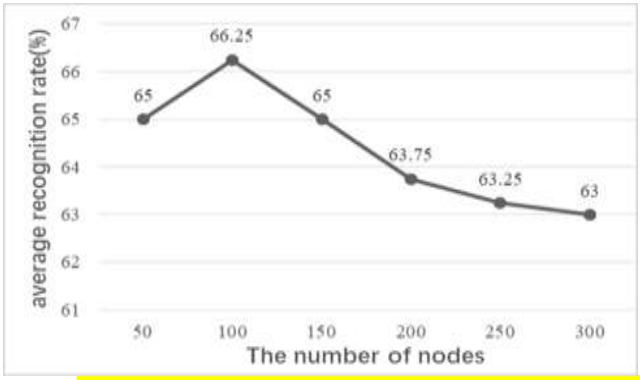


Fig. 7. (fear, sad) Relationship between the average recognition rate of two emotions and the number of bottleneck nodes (DNN3)

We propose to train three different DNNs to extract bottleneck features for different emotion grouping. In the process of DNN1 training, the input data are the speech feature parameters of six kinds of emotions and they are divided into three big classes. Combining with the optimal number of nodes, the structure of DNN1 is 1280-1280-100-1280-1280-1280-3. The training data of DNN2 are the speech feature parameters of happy, anger and surprise. Similarly, the DNN2 structure is 1280-1280-100-1280-1280-1280-2. After the training of each DNN is completed, the bottleneck features extracted from DNN are used to train corresponding SVM classifier. The recognition rates of the six emotions obtained in the experiment are shown in Table 9 and Fig. 8, which demonstrate the experiment results based on traditional feature-SVM, DNN-SVM and DNN-decision tree SVM.

As can be seen from Fig. 8, the recognition rate of emotion based on DNN-decision tree SVM is higher than the other two methods to a certain extent. Especially for the happy emotion, the system recognition rate of the proposed method is 22.5% and 11% higher than that of the traditional SVM and DNN-SVM, respectively. Moreover, surprise and sad also increase by a few percentage points as well. The experimental results in Table 8 and Table 9 indicate that the average emotion recognition rate based on the proposed method is 6.25% and 2.91% higher than traditional SVM and DNN-SVM classification method, respectively. It verifies the validity of the proposed method, which can improve the performance of emotion recognition system.

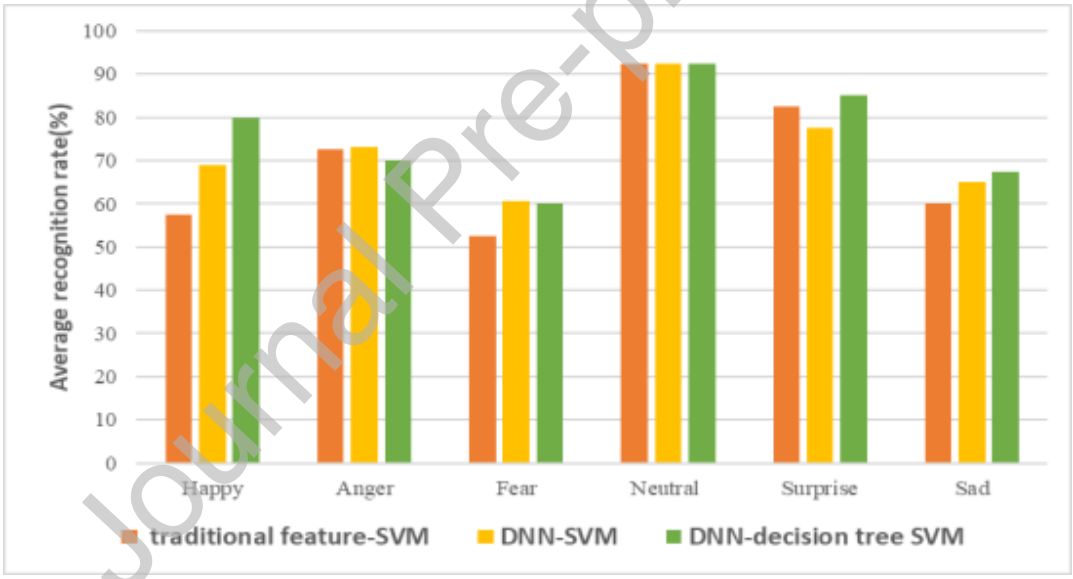


Fig. 8. Emotion recognition rate of the three methods

4 Conclusion

In order to improve the performance of speech emotion recognition for multiple emotions effectively, we present a new speech emotion recognition method based on DNN-decision tree SVM from two aspects: how to find more distinctive speech emotion features and establish an effective recognition model. In the proposed approach, we first establish a decision tree SVM framework by calculating the degree of emotional confusion. Then we train different DNNs for diverse emotion groups to extract bottleneck features used to train the each SVM classifier in the decision tree. Finally, we use this model for speech emotion recognition. The experiments show that the average recognition rate of the proposed method based on DNN-decision tree SVM can reach 75.83%, which is 6.25% and 2.91% higher than traditional SVM and DNN-SVM classification method, respectively. Since the confusion between fear and sad is still very large during the experiment, the next research direction is to find more distinctive features or models, so as to further improve the performance of recognition system.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 61901227, 61671252), and the Natural Science Foundation of the Jiangsu Higher

Education Institutions of China (No. 19KJB510049).

Conflict of interest statement

We would like to submit the enclosed manuscript entitled “Speech Emotion Recognition Based on DNN-Decision Tree SVM Model”, which we wish to be considered for publication in “Speech Communication”. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted., and manuscript is approved by all authors for publication. I would like to declare on behalf of my co-authors that the work described was original research that has not been published previously, and not under consideration for publication elsewhere, in whole or in part. All the authors listed have approved the manuscript that is enclosed.

References

Busso, C., Lee, S., Narayanan, S., 2009. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. IEEE Transactions on Audio Speech & Language Processing. 17(4), 582-596.

Fischer, A., Igel, C., 2011. Bounding the Bias of Contrastive Divergence Learning. Neural Computation. 23(3), 664-673.

Garg, V., Kumar, H., Sinha, R., 2013. Speech based Emotion Recognition based on hierarchical decision tree with SVM, BLG and SVR classifiers. Communications. IEEE, pp, 1-5.

Gupta, S., Mehra, A., 2015. Speech emotion recognition using SVM with thresholding fusion. 2nd International Conference on Signal Processing & Integrated Networks. pp, 570-574.

Hinton, G., Deng, L., Yu, D., et al., 2012. Deep neural networks for acoustic modeling in speech recognition: The Shared Views of Four Research Groups. IEEE Signal Processing Magazine. 29(6), 82-97.

Huang, C., Jin, Y., Wang, Q., et al., 2010. Speech emotion recognition based on decomposition of feature space and information fusion. Signal Processing. 26(6), 835-842.

Huang, Y., Zhang, G., Li, X., et al., 2012. Small sample size speech emotion recognition based on global features and weak metric learning. Acta Acustica. 37(3), 330-338.

Kachele, M., Zharkov, D., Meudt, S., et al., 2014. Prosodic, Spectral and Voice Quality Feature Selection Using a Long-Term Stopping Criterion for Audio-Based Emotion Recognition. Proceedings of 22nd International Conference on Pattern Recognition. pp, 803-808.

Khokher, R., Singh, R. C., Kumar, R., 2015. Footprint Recognition with Principal Component Analysis and Independent Component Analysis. Macromolecular Symposia. 347(1), 16-26.

Lanjewar, R. B., Mathurkar, S., Patel, N., 2015. Implementation and Comparison of Speech Emotion Recognition System Using Gaussian Mixture Model (GMM) and K-Nearest Neighbor (K-NN) Techniques. Procedia Computer Science. 49(1), 50-57.

Le, B. V., Lee, S., 2014. Adaptive Hierarchical Emotion Recognition from Speech Signal for Human-Robot Communication. Tenth International Conference on Intelligent Information Hiding & Multimedia Signal Processing. IEEE, pp, 807-810.

Lee, C. C., Mower, E., Busso, C., et al., 2011. Emotion recognition using a hierarchical binary decision tree approach. Speech Communication. 53(2011), 1162-1171.

Liang, R., Zhao, L., Tao, H., et al., 2016. Speech emotion recognition algorithm based on pseudo-selective attention mechanism. Acta Acustica. 41(04), 537-544.

Li, S., Xu, L., 2017. Research on emotion recognition algorithm based on spectrogram feature extraction of bottleneck feature. Computer Technology and Development. 27(5), 82-86.

Li, X., Akagi, M., 2019. Improving multilingual speech emotion recognition by combining acoustic features in a three-layer model. Speech Communication. 110(2019), 1-12.

Lu, G., Yuan, L., Yang, W., et al., 2018. Speech Emotion Recognition Based on Long-term and Short-term Memory and Convolutional Neural Network. Journal of Nanjing University of Posts and Telecommunications (NATURAL SCIENCE EDITION). 38(05), 63-69.

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

Mustafa, M. B., Yusoof, M, A. M., Don, Z. M., et al., 2018. Speech emotion recognition research: an analysis of research focus. International Journal of Speech Technology. 21(1), 137-156.

Panagiotis, T., Zhang, J., Bjorn, W., 2018. End-to-End Speech Emotion Recognition Using Deep Neural Networks. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. pp, 5089-5093.

Rázuri, J. G., Sundgren, D., Rahmani, R., et al., 2015. Speech emotion recognition in emotional feedback for Human-Robot Interaction. International Journal of Advanced Research in Artificial Intelligence. 4(2), 20-27.

Sui, X., Zhu, T., Wang, J., 2017. Speech emotion recognition based on local feature optimization. Journal of University of Chinese Academy of Sciences. 34(4), 431-438.

Sun, L., Fu, S., Wang, F., 2019. Decision tree SVM model with Fisher feature selection for speech emotion recognition. EURASIP Journal on Audio, Speech, and Music Processing. 2019(2).

Torres, V. C., Álvarez, L. M., Álvaro O. G., 2017. SVM-based feature selection methods for emotion recognition from multimodal data. Journal on Multimodal User Interfaces. 11(1), 1-15.

Trabelsi, I., Amami, R., Ellouze, N., 2016. Automatic Emotion Recognition using Generative and Discriminative Classifiers in the GMM Mean Space. 2nd International Conference on Advanced Technologies for Signal and Image Processing. IEEE, pp, 767-770.

Trentin, E., Scherer, S., Schwenker, F., 2015. Emotion recognition from speech signals via a probabilistic echo-state network. Pattern Recognition Letters. 66(2), 4-12.

Wang, F., Sun, L., Su, M., et al., 2018. Speech emotion recognition of decision tree SVM based on parameter optimization. Computer Technology and Development. 28(07), 63-67.

Wang, K., An, N., Li, B., et al., 2015. Speech Emotion Recognition Using Fourier Parameters. IEEE Transactions on Affective Computing. 6(1):69-75.

Xu, C., Cao, T., Feng, Z., et al., 2012. Multi-Modal Fusion Emotion Recognition Based on HMM and ANN. Communications in Computer & Information Science. 332(5), 541-550.

Yang, B., Lugger, M., 2010. Emotion recognition from speech signals using new harmony features. Signal Processing. 90(5), 1415-1423.

Zhan, Y., Mao, Q., Lin, Q., et al., 2013. Visual speech emotion recognition. Beijing: Science Press.

Zhang, C., Ji, N., Wang, G., 2015. Restricted Boltzmann Machines. Chinese Journal of Engineering Mathematics. 28(2), 159-173.

Zhang, S., Li, Li., Zhao, Z., 2013. Research progress on speech emotion recognition in human-computer interaction. Journal of Circuits and Systems. 18(2), 440-451.

Zhou, X., Guo, J., Bie. R., 2017. Deep learning based affective model for speech emotion recognition. Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress. IEEE, pp, 841-846.

Zhu, L., Chen, L., Zhao, D., et al., 2017. Emotion Recognition from Chinese Speech for Smart Affective Services Using a Combination of SVM and DBN. Sensors. 17(7), 1694-1708.

AUTHOR BIOGRAPHY

Linhui Sun is an associate professor in Nanjing University of Posts and Telecommunications. She received her B.S. from Jilin University in 2002, and received her M.S. and Ph.D. both from Nanjing University of Posts and Telecommunications in 2005 and 2013 respectively. Her research interests speech signal processing and modern speech communication.

Bo Zou received his B.S. from Hubei University of Automotive Technology in 2017. He is currently a M.S. student of Signal and Information Processing in Nanjing University of Posts and Telecommunications. His research interests mainly include speaker recognition and speech emotion recognition

Sheng Fu received his B.S. from Nanjing University of Posts and Telecommunications in 2017. He is currently a M.S. student of Signal and Information Processing in Nanjing University of Posts and Telecommunications.

Jia Chen received his B.S. from North University of China in 2016 and received his M.S. from Nanjing University of Posts and Telecommunications in 2019. His research interests mainly include speech emotion recognition.

Fu Wang received his B.S. from Qingdao Agricultural University in 2015 and received his M.S. from Nanjing University of Posts and Telecommunications in 2018. His research interests mainly include speech emotion recognition.