Signal, Image and Video Processing (2022) 16:1253–1261
https://doi.org/10.1007/s11760-021-02076-0

ORIGINAL PAPER

Multi-classification speech emotion recognition based on two-stage bottleneck features selection and MCJD algorithm

Linhui Sun[1] · Yiqing Huang[1] · Qiu Li[1] · Pingan Li[1]

Received: 30 December 2020 / Revised: 24 October 2021 / Accepted: 27 October 2021 / Published online: 12 January 2022

Abstract
Feature extraction and classification decision play an important role in speech emotion recognition. To improve the performance of the multi-classification speech emotion recognition (SER) system, a two-stage bottleneck features selection model and a novel multi-classifier joint decision (MCJD) algorithm are proposed. In two-stage bottleneck features selection model, firstly, bottleneck features at different hidden layers are extracted from deep neural network (DNN) and fused using genetic algorithm (GA). Secondly, principal component analysis (PCA) is used to eliminate the dimension disaster caused by high-dimensional feature vectors. In addition, to make up for the shortcomings of single SVM classifier in SER, we use different feature sets to train multiple SVM classifiers based on classification targets. The final recognition result is obtained by joint decision of SVMs according to MCJD algorithm. Five-fold cross-validation is used, and an average accuracy of 84.89% is achieved using the two-stage bottleneck features selection model and traditional support vector machines (SVM) classifier. Then, using the MCJD algorithm, the average SER rate of the multi-classification SER system for seven kinds of emotions is 87.08% on Berlin Database, which further improves the performance of SER system and shows the effectiveness of our method.

Keywords Deep neural network (DNN) · Speech emotion recognition · Bottleneck features selection · Multi-classifier joint decision algorithm

1 Introduction

With the wide application of human–computer interaction technology, people are longer satisfied with the current intelligence capability of computers. Among the many ways for computers to recognize human emotional states, speech-based emotion recognition is a very effective approach and has become a new research hotspot. At present, speech emotion recognition (SER) technology has been widely used in education, information, medicine, criminal investigation, entertainment, etc. [1]. However, SER is facing a great many difficulties: First, the emotion itself is difficult to be defined.

Second, it is still unclear which features are the most useful features in distinguishing emotions. Third, one speech signal may contain more than one emotion, so it is difficult to determine which emotion is the dominant one. Fourth, the emotional expression can be influenced by gender, age, environment and even culture. A lot of research on these challenging problems has done [2].

The extraction of speech emotion features is a crucial process in SER. At present, the effective feature parameters include voice quality features, prosodic features, spectral features and bottleneck features [3, 4]. Bottleneck features from deep neural network (DNN) contain deeper information of speech signal, which can get excellent performance in the field of SER [5–12]. In previous studies [5], we proposed an SER method based on DNN-decision tree support vector machines (SVM). For diverse emotion groups, different DNN networks were trained to extract the bottleneck features. Compared to the traditional SVM and DNN-SVM classi-

✉ Linhui Sun
sunlh@njupt.edu.cn

Yiqing Huang
1218012224@njupt.edu.cn

Qiu Li
andyliqiu@163.com

Pingan Li
lpa@njupt.edu.cn

[1] College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, Jiangsu, China

fication method, the method made the SER rate increase. Long et al. proposed a model that jointed a bottleneck feature extraction network and attention model [13]. In the proposed model, bottleneck features were extracted by deep belief network and the encoder-decoder model was based on attention. Experiments were conducted on the TIMIT corpus, showing that the phoneme error rate was 17.80% on the test set and the average training iterations decreased. Wöllmer et al. introduced a novel context-sensitive feature extraction method for spontaneous speech recognition by combining bidirectional long short-term memory modelling and bottleneck feature generation [14]. Evaluations showed that this method prevailed over recently published architectures for feature-level context modelling. These works prove that bottleneck features are effective for SER. Inspired by these, we extract bottleneck features through DNN in the experiment.

In order to make full use of the information expressed by bottleneck features at different hidden layers, we can extract the deep bottleneck features and the shallow bottleneck features from DNN, and then combine them. However, the method of direct combination of speech features will produce high-dimensional speech emotion feature data, which may have the problem of feature redundancy because of the strong correlation between different features. Also, too high-dimensional feature data will increase training time and consume computer resources. By selecting the optimal features that have a strong contribution to various kinds of emotions of speech, redundant features can be eliminated, which reduce the feature dimension [15–18]. Researchers have also done many works on feature dimensionality reduction. Ke et al. proposed PCA-continuous hidden Markov model (PCA-CHMM), which improved the performance of SER [15]. Zhang et al. adopted kernel isometric mapping (KIsomap) for feature extraction on spoken emotion recognition tasks and KIsomap achieved better performance than locally linear embedding (LLE) and isometric mapping (Isomap) [17]. Ingo Siegert et al. presented a corpus similarity measure based on PCA-ranked features, which improved cross-corpus emotion recognition [18]. In this paper, we use genetic algorithm (GA) for parameter optimization to combine the speech features, and then use principal component analysis (PCA) for feature dimension reduction, which is a two-stage feature selection model. In addition, the main classifiers used in SER are K-nearest neighbour (KNN), Gauss mixture model (GMM), artificial neutral network (ANN) and SVM. Ratna Kanth et al. presented the construction of binary SVMs and its significance for efficient SER [19]. On the test set, the average accuracy for the binary SVMs and the multiclass SVM was 92.25% and 77.07%, respectively. On the same test set using the combinator algorithm, the fused model produced an overall accuracy of 87.86%. Rahul et al. used GMM and KNN models for the recognition of six emotional categories on Berlin emotion database and showed the comparison of the two algorithms for performance analysis which was supported by the confusion matrix [20]. In this paper, a multi-classification SER system based on a novel MCJD algorithm is proposed to make up for the deficiency of a single classifier in recognizing speech emotion.

This paper is organized as follows. In Sect. 2, we introduce the two-stage bottleneck features selection model. In Sect. 3, the proposed SER classification system based on novel MCJD algorithm is given. In Sect. 4, we describe the experiments and results. Finally, we draw a conclusion in Sect. 5.

2 Two-stage bottleneck features selection model

In the research of traditional SER, the traditional acoustic emotional features are usually used for experiments, and good results can be achieved in some scenes. However, in the field of SER, bottleneck features are also excellent, which is the deep expression of speech emotion. In this paper, we select the bottleneck features extracted from DNN as the speech features. What's more, feature fusion can overcome the shortcoming that single feature cannot describe the emotional information comprehensively, improving the classification accuracy. The contribution of each feature to emotion classification is different, so we cannot simply connect the various features, but give the bigger weight to the features with the larger contribution.

Here, a two-stage bottleneck features selection method is proposed, which is shown in Fig. 1. In the first stage feature selection, appropriate bottleneck features from DNN are extracted and fused using GA for SER. In the second stage feature selection, PCA, an optimal feature subset selection technique, is used to eliminate the curse of dimensionality.

Firstly, speech samples are divided into training set and test set. On the training set, the DNNs with different bottleneck layer are trained, and deep and shallow bottleneck features are extracted, respectively. Then, the fused bottleneck features with initial weights are taken as the input of GA. GA is used to search for the optimal weights to realize the bottleneck features fusion. Then, PCA is used to filter the fused features. Finally, the emotional features for classification are obtained and used to train the classifier.

On the test set, the deep and shallow bottleneck features are extracted by the trained DNN model, and then combined with the weights obtained by GA in the training process. Thus, the fused features on the test set are obtained. Then, the features after dimension-reduction with PCA are used as the input of the trained SVM. Finally, the result of speech emotion recognition is achieved.

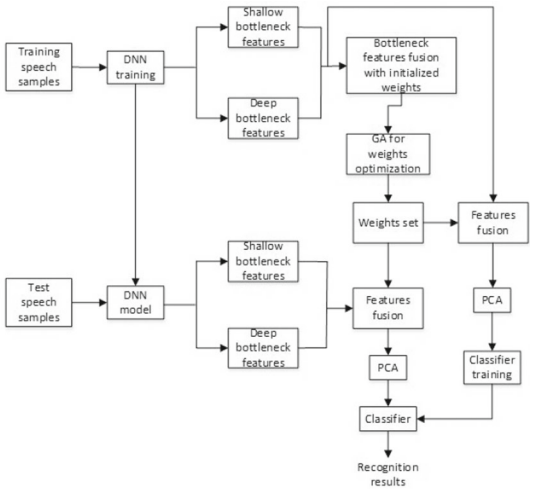

Fig. 1 Two-stage bottleneck features selection method

2.1 DNN model for bottleneck features

DNN is applied to computer vision, speech processing, natural language processing and other fields, which has achieved great success [21, 22]. DNN can be used to extract the emotional features from speech, which is an indirect application of DNN. Using the DNN model to directly complete the speech emotion classification task is a direct application. However, the direct speech emotion classification effect of DNN is not very ideal. In this paper, DNN is used to extract the bottleneck features. The designed DNN network has five layers, the Fourier coefficients are extracted as the input of DNN, through a series of steps.

First, the speech signal needs some pre-processing operations, such as pre-emphasis, framing windowing and endpoint detection. Then, the harmonic coefficients in the speech are obtained through Fourier transform processing. When the harmonic coefficients are calculated, the modulus of the harmonic coefficients can be calculated to obtain the Fourier coefficients. There are three hidden layers, among which the number of neuron nodes in the middle layer is much smaller than that in other hidden layers, so the middle hidden layer is the so-called bottleneck layer. The output layer is a softmax layer. After two processes of pre-training and fine-tuning, DNN fully learns the emotional information contained in the speech signals and mines the structural information hidden in the speech.

According to the different positions of the bottleneck layer in the hidden layer, 3 DNNs used in this paper. They are named DNN1, DNN2 and DNN3, respectively (DNN1 refers to the first hidden layer as the bottleneck layer), used to extract different bottleneck features. Bottleneck features of all training speech are acquired, and five global statistics, including mean, variance, median, maximum and minimum, are calculated for the extracted features. In the test phase, the trained DNN was used to extract the bottleneck features in the same way.

2.2 Feature fusion with GA

The expression of emotion in speech varies with different language, gender, age and culture. The difference of these factors will lead to the diversification of acoustic emotion distribution. At present, there is no direct identification method for which features are more distinguishable in different speech emotion databases. In this paper, a feature fusion strategy based on GA is adopted. Bottleneck features based on different hidden layers are fused in a weighted way. The weights are optimized by GA to obtain the contribution weights of different features. Then, features are fused in a weighted way to obtain the emotional feature set. Finally, they are input into the classifier for training and testing to get the final recognition result.

In SER systems, different features contribute differently to SER. The greater the contribution, the greater the weight. In this paper, a weight set is calculated to measure the contribution of different features. Suppose the feature set of the ith feature is represented as x_i, and the contribution weight of the ith feature is w_i. In this paper, deep bottleneck feathers and shallow bottleneck feathers are fused. The fused feature set can be expressed as:

$$X = [w_1 x_1 \ , \ w_2 x_2] \tag{1}$$

GA is an advanced algorithm to solve optimization problems. It has many advantages, such as good convergence, strong adaptability, high optimization efficiency and good optimization results. The procedures of GA are as follows [23, 24]. When GA is used to search for the optimization of $\{w_1, w_2\}$. X needs to be used as the input of GA, and then GA is randomly initialized to generate multiple groups of $\{w_1, w_2\}$ individuals. These individuals constitute the initialized population $P(0)$, and then according to the rules of GA, these $\{w_1, w_2\}$ need to be encoded. This paper adopts the binary encoding method, assuming that the binary string sequence is the binary encoding form of $\{w_1, w_2\}$, these binary string sequences are called chromosomes in GA, and $\{w_1, w_2\}$ represents an individual. Then decode the encoded chromosome to obtain the weight parameters that need to be optimized in the individual, and put the optimized parameters into the fitness function to calculate the individual fitness value. The fitness function adopted in this section is the average recognition rate of all kinds of emotions, that is, after calculating the fitness function, the average recognition rate corresponding to each weight set can be obtained, among which the classifier used in the fitness function is SVM. Then judge whether it reaches the number of iterations. If it reaches, terminate the search and output the value of the

optimization weight. Otherwise, continue the selection operation, select two individuals from the population to copy according to the fitness value, and then determine whether the two selected individuals need to perform the crossover operation through the crossover probability in GA.

2.3 Feature dimensionality reduction with PCA

The feature dimension will become larger after fusion, which results in long training time of the SVM model, and previous research shows that recognition rate will be reduced due to feature redundancy. Therefore, before classification, we choose PCA to filter the fused features.

The goal of PCA is to map high-dimensional data to low-dimensional space through a certain linear projection, and maximize the variance of the data in the projection dimension, thereby using fewer data dimensions and retaining more features of the original data points. In plain English, if all the points are mapped to the same point, almost all the information is lost, while if the variance is as large as possible, the data points are scattered to retain more information. It can be proved that PCA is an excellent linear dimensionality reduction method with the least loss of original data information. Actually, it is closest to the original data, which explains why PCA does so well to some extent.

3 Proposed multi-classification SER system based on novel MCJD algorithm

The feature fusion method based on GA effectively fuses bottleneck features by weighting and obtains good SER performance on the EMO-DB database. However, in this feature fusion strategy, the optimization method of GA is based on the average recognition rate of all kinds of emotions as the optimization goal. In fact, this method does not take into account the uniqueness of each type of emotion, only fusing features with the same weight set. That is, the same feature set is used when classifying different emotions. In order to select the most appropriate feature set for each emotion and get better SER performance, we propose a multi-classification SER system based on the MCJD algorithm.

The proposed classification system for SER is shown in Fig. 2. After the speech signals are preprocessed and input to the two-stage feature selection model, different features for SVMs are obtained according to the different optimization goals. SVM0 focuses on the overall performance of seven emotions, so the fitness function of GA is the average recognition rate of seven emotions. The obtained emotion feature set is more suitable for the classification task of seven emotions. For SVM1-SVM7, we take the average recognition rate of a single emotion as the fitness function of GA to obtain the contribution weights of the emotion feature. Then, we input optimized fusion features to SVMs to train SVM1–SVM7 classification models corresponding to anger, happiness, boredom, neutral, sadness, fear and disgust, respectively. Finally, the novel MCJD algorithm is used to obtain the final classification results.

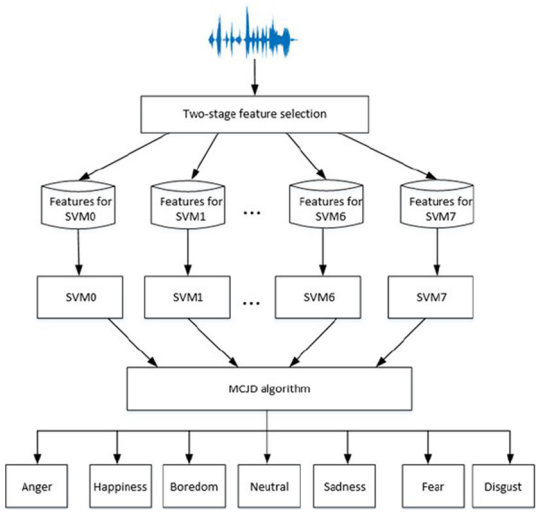

Fig. 2 Proposed multi-classification SER system based on MCJD algorithm

In the SER system, emotion features are extracted, fused and dimensionally reduced by the two-stage bottleneck features selection model. Then, the suitable emotion feature set is fed on SVM0–SVM7, respectively. Finally, based on the novel MCJD algorithm, speech emotion classification is realized. The MCJD algorithm pseudo code is shown in Table 1. First, check the output of SVM0, if the output of SVM0 is angry, then find SVM1 (SVM1 is the classifier with the highest emotion recognition rate of angry as the training target, so in SVM0–SVM7, SVM1 has the best recognition effect on angry speech). If the output of SVM1 is also angry, then the classification result is angry. If not, then the classification result is the emotion identified by most classifiers in SVM1–SVM7. If the output of SVM0 is the other six emotions, the decision is also made in this way.

Activated by voting mechanism, MCJD algorithm is an improvement of voting based on several basic classifiers, which is an ensemble learning model that follows the principle of minority obeying the majority. The variance is reduced through the integration of multiple classifier models so as to improve the robustness of the whole model. In the MCJD algorithm, we first make use of the particularity of the trained basic classifiers (SVM0–SVM7) for pre-judgement, and then use the voting method for joint decision to get the final result, which can further improve the robustness of the classification model.

Table 1 Pseudo code of MCJD

Pseudo code of MCJD
Begin
Case result of SVM0 is
Case1: Anger
If output of SVM1 is Anger, then classification result is Anger
Else output the emotion identified by most classifiers in SVM1–SVM7
EndIf
Case2: Happiness
If output of SVM2 is Happiness, then classification result is Happiness
Else output the emotion identified by most classifiers in SVM1–SVM7
EndIf
Case7: Disgust
If output of SVM1 is Disgust, then classification result is Disgust
Else output the emotion identified by most classifiers in SVM1–SVM7
EndIf
EndCase
End

4 Experiments and results

In this section, we evaluate the system performance of SER by conducting experiments on the EMO-DB emotional corpus. Firstly, we introduce the dataset. Secondly, experimental data and preparation are discussed. Thirdly, results with the proposed method are displayed.

4.1 Emotional corpus and data preparation

4.1.1 EMO-DB corpus

In the field of SER, no matter how advanced the technology is, it is necessary to have an appropriate emotion corpus to conduct relevant experiments. At present, there are a variety of emotional corpus in this research field, which are usually divided into two categories: discrete and dimensional. Due to the difficulty in recording the dimensionality emotion corpus and the rarity of the corresponding corpus, this paper uses discrete corpus for experimental simulation.

Berlin Emotional Corpus (EMO-DB) was recorded by Berlin University of Technology in Germany. The corpus includes seven different emotions: anger, happiness, boredom, neutral, sadness, fear, and disgust. It is obtained by ten professional actors (five men and five women) who perform different emotion simulations on ten sentences (five long sentences and five short sentences). The sentences in the corpus are 800 in total, and more than 500 of them are obtained after screening, sampling at 48 kHz (then compressed to 16 kHz), and finally quantizing at 16 bit. The selection of corpus text follows the principle of semantic neutrality and no emotional inclination, and it is a daily colloquial style without too much written language modification. Voice recording is completed in a professional studio. Actors are required to recall their real experience or experience before performing a specific emotional segment to brew their emotions, so as to enhance their feelings.

4.1.2 Experimental data and preparation

The emotion recognition model is SVM using the tool of LIBSVM in the environment of MATLAB R2018a. In the experiment, the CPU model of the computer is Intel Core i5-8250U, and the graphics card model is NVIDIA GeForce MX150.

During the experiment, each speech sample in the emotional corpus is all preprocessed, and the frame length and frame shift are set to 256 points and 128 points, respectively. The emotional feature parameter selected in the experiment is the Fourier coefficient, that is, the entire voice is studied from the frequency domain. After frame splitting, each frame has 256 points. The emotional feature parameters of five consecutive frames are spliced to form a 1280-dimensional vector, which is the input of the whole DNN. Due to the uneven characteristic parameters, the characteristic parameters are normalized in the experiment. The DNN used in this paper consists of five layers. The input layer is a 1280-dimensional vector, and there are three hidden layers in the middle. One of the hidden layers will be set as the bottleneck layer. The number of neurons in this layer is set to 100, while the number of neurons in the other hidden layers is set to 1280, and the output layer is a softmax layer with the same size as the classification, that is, the last layer has seven neurons. In fact, softmax can also be used as a classifier to recognize seven kinds of emotions. Therefore, only by taking this as a constraint, we can get the bottleneck characteristics to represent the emotions. Then calculate the corresponding five global statistics (mean, maximum, minimum, variance, median) to train DNN. After the bottleneck features are extracted by DNN, the features of different bottleneck layers are fused and selected for training and testing SVM.

During training DNNs, batch size, learning rate, and network structure are important. According to the previous research and experiments on related parameters, the best effect is obtained when the batch size is 2. The purpose of setting the batch size is to allow the model to select batches of data for processing during the training process. When there are a lot of samples in the training set, directly inputting these data into the neural network will lead to a large amount of calculation. Moreover, when all samples are input into the network at the same time, it is often difficult to determine a global optimal learning rate to make the training effect the best. At the same time, the learning rate is set to 0.005. If the learning rate is too high, the training may not converge or even diverge. If the learning rate is too low, the training will

Table 2 Average recognition rate for fused bottleneck features with GA

Fused bottleneck features with GA	Average recognition rate (%)
DNN1 + DNN2	81.11
DNN1 + DNN3	80.71
DNN2 + DNN3	79.07
DNN1 + DNN2 + DNN3	77.27

Table 3 Average recognition rate with two-stage bottleneck features selection

Two-stage bottleneck features selection	Average recognition rate (%)
DNN1 + DNN2	84.89
DNN1 + DNN3	84.34
DNN2 + DNN3	83.87
DNN1 + DNN2 + DNN3	82.79

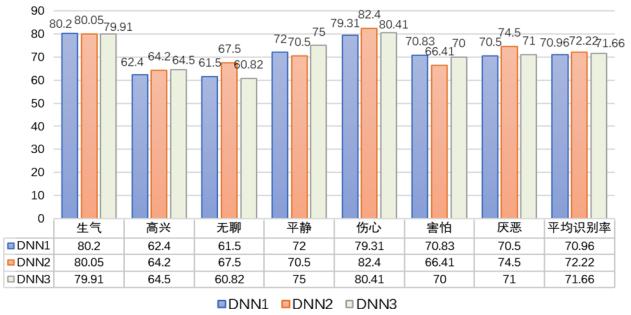

Fig. 3 SER rate of bottleneck features at different hidden layers (%)

definitely converge, but it will take a long time. The number of iterations is set to 50 times.

4.2 Results with the proposed method

4.2.1 Comparison results using fused bottleneck features with GA

Firstly, we test the emotion recognition rate of the bottleneck features at different hidden layers on the EMO-DB database, which is shown in Fig. 3, and the average recognition rates of DNN1, DNN2 and DNN3 are 70.96%, 72.22% and 71.66%, respectively. We can see that the system using the features of DNN2 performs best. For sadness and anger, single-layer bottleneck features can achieve good recognition results, but the recognition rate of happiness is low and the average recognition rate needs to be further improved. Therefore, we use GA to fuse deep bottleneck features and shallow bottleneck features to improve the performance of speech emotion recognition.

For each segment of speech, we use GA to fuse deep bottleneck features and shallow bottleneck features, and then calculate five statistical values of fused bottleneck features to obtain 1000-dimensional emotion features. Then, emotion features are fed to traditional SVM classifier. The experimental results are shown in Table 2. It can be seen from Table 2, when the bottleneck features of DNN1 and DNN2 are fused, the emotion recognition rate reached 81.11%, which is 10.15% higher than that of DNN1 and 8.89% higher than that of DNN2. However, after we fuse the features of the bottleneck layers of DNN1, DNN2, and DNN3, the emotion recognition rate is only 77.27%, which is not as high as we expected. It is not always correct that the more features, the higher recognition rate. On the contrary, more features will cause feature redundancy and affect the emotion recognition rate.

4.2.2 Effect of feature dimension reduction

In the experiment, we investigate the effect of feature dimension reduction using PCA on recognition. PCA function in the Libsvm toolbox is used to reduce the feature dimension. The most important parameter is the threshold, which is the degree of interpretation of the original variable (a number between 0 and 100). The principal component can be selected through this threshold. Its default value is 90, that is, the selected principal component can reach the degree of interpretation of the original variable by 90% by default.

The specific dimensional reduction effects of PCA on the EMO-DB corpus are shown in of Fig. 4. It can be seen from the experimental results that the fused bottleneck features with different dimensions get the different performance of speech emotion recognition. In general, the system recognition rate is increase first, and then decrease with the increase in feature dimensions. It can be seen from Fig. 4 that the fused bottleneck features from DNN1 and DNN2 achieve the best performance when the feature dimension is reduced from 1000 to 165. The highest average recognition rate for the seven categories of emotions is 84.89%, when the threshold is set to 90. That is to say, the reduced 165-dimensional features can explain 90% of the information covered by 1000-dimensional features, which greatly reduces the redundancy and improves the accuracy of recognition.

4.2.3 Comparison results using two-stage bottleneck features selection model

In order to reduce the computational complexity, PCA feature screening method is used to reduce the dimension of the feature set before the feature set is input into SVM classifier, which not only reduces the time of classifier training, but also improves the performance of speech emotion recognition.

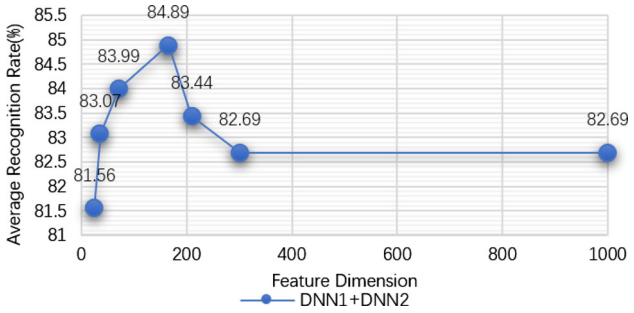

Fig. 4 Specific dimensional reduction effects of PCA(DNN1+ DNN2)

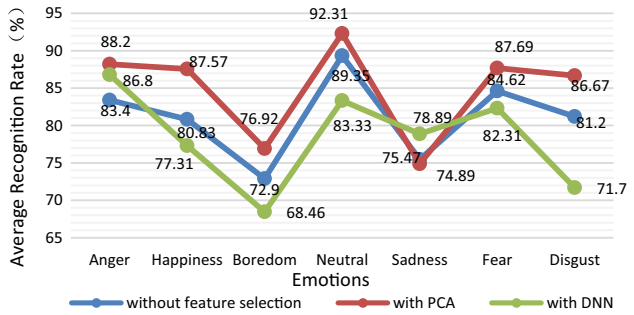

Fig.5 Comparison results of features selection

The average classification results of fused features from different DNNs are shown in Table 3, after two-stage bottleneck features selection model is used. Compared with Table 2, after dimensionality reduction by PCA, the emotion recognition rate of the combined bottleneck features at different hidden layers is increased. Among them, the fused bottleneck features from DNN1 and DNN2 have the best performance in emotion recognition rate. It is 84.89%, which is 3.78% higher than that without dimensionality reduction.

Using fused features from DNN1 and DNN2, with and without PCA, the classification results of 7 emotions are shown in Fig. 5. As shown in Fig. 5, except for the sadness emotion, the recognition rate of the other six emotions is greatly improved. In particular, the recognition rate of the happiness emotion is improved by 6.74%. This experimental result verifies the effectiveness of PCA in bottleneck features selection.

Obviously, DNN also has the ability of feature selection. The node number of the bottleneck layer is far less than that of other hidden layers. We use DNN to fuse and select features to compare with the two-stage bottleneck feature selection model. The experimental results are shown in Fig. 5, and the average recognition rate is 78.4%, which is 6.49% lower than the proposed two-stage bottleneck feature selection model.

Table 4 Optimal weights for SVMs

SVMs	Fitness function	Optimal weights	
		$K1$	$K2$
SVM0	Average recognition rate of 7 emotions	0.1280	0.8404
SVM1	Recognition rate of Anger	0.7797	0.5699
SVM2	Recognition rate of Happiness	0.1007	0.5850
SVM3	Recognition rate of Boredom	0.8846	0.0657
SVM4	Recognition rate of Neutral	0.7338	0.0330
SVM5	Recognition rate of Sadness	0.3219	0.7386
SVM6	Recognition rate of Fear	0.0185	0.6293
SVM7	Recognition rate of Disgust	0.2674	0.9439

Table 5 Emotion recognition rate of proposed method

Emotion	Accuracy (%)
Anger	90.05
Happiness	88.17
Boredom	80.77
Neutral	93.33
Sadness	80.3
Fear	90.2
Disgust	86.71
Average	87.08

4.2.4 Performance of proposed method

One classifier does not necessarily perform well for each emotion, so we train different SVM classifiers according to the characteristics of different emotions, and then combine these SVM classifiers by MCJD algorithm. In other words, considering the difference emotions, the recognition rate of each emotion category is taken as the optimal search target of GA to obtain the emotion feature set that contributes the most to the certain emotion category. Optimal weights searched by GA for SVMs are given in Table 4.

As can be seen from Table 4, the optimal weight set of each SVM classifier is different. Taking SVM0 as an example, the weight set obtained by GA is {0.1280, 0.8404}, so the feature set of SVM0 is $[0.1280 \times 1, 0.8404 \times 2]$. Through these different feature sets, different SVMs are trained, and correspondingly, different test results are obtained naturally.

Table 5 shows the emotion recognition rates of the proposed system based on two-stage bottleneck features selection and MCJD algorithm. Though there is slightly insufficiency for the recognition of boredom and sadness, the recognition rate of the MCJD algorithm is 2.19% higher than that of the traditional SVM, which proves that to some extent the proposed classification system based on novel MCJD algorithm can obtain better recognition results than traditional SVM classification.

Table 6 SER rate at each stage

Bottleneck features			Feature selection	Classifier		SER rate (%)
DNN1	DNN2	DNN3	PCA	SVM	MCJD	
✓				✓		70.96
	✓			✓		72.22
		✓		✓		71.66
✓	✓			✓		81.11
✓		✓		✓		80.71
	✓	✓		✓		79.07
✓	✓	✓		✓		77.27
✓	✓		✓	✓		84.89
✓		✓	✓	✓		84.34
	✓	✓	✓	✓		83.87
✓	✓	✓	✓	✓		82.79
✓	✓		✓		✓	87.08

Table 7 Comparison of proposed SER method with state-of-the-art methods

Methods	SER (%)
Low level feature with GMMs model [25]	82.82
Log-mel spectrogram with SNN [26]	84.3
MFCC, chromagram, etc. with CNN [27]	86.1
Speech spectrograms + CNN + BiLSTM [28]	85.57
Our method	87.08

Table 6 summarizes the SER rate at each stage. As we can see, the average recognition rate of the proposed method is 87.08%, which proves the effectiveness of the two-stage bottleneck features selection model and the MCJD algorithm.

In addition, the comparative study of the proposed method with the state-of-the-art works on EMO-DB is illustrated in Table 7, which better proves the superiority of our method.

5 Conclusion

In this paper, in order to get the best bottleneck features, a two-stage bottleneck features selection model is proposed, which not only eliminates the redundant features, but also gets the features that are most suitable for the classification target. To further improve the emotion recognition rate, a multi-classification SER system based on a novel MCJD algorithm is presented. In the system, eight SVMs are acquired by different feature sets and the classification result is obtained by the MCJD algorithm. Finally, an average recognition rate of 87.08% is achieved on the test set, which proves that bottleneck features are very effective for SER, especially for angry, happiness, neutral, fear and disgust. At the same time, it also shows that the bottleneck features still contain redundant information, and PCA is an excellent feature screening method. It also proves that the MCJD algorithm effectively compensates for the deficiency of single SVM classifier, which has a certain correction effect on the classification of emotions. However, the recognition rate of boredom and sadness is not ideal, which needs further study in the future.

Acknowledgements This work is supported by the National Natural Science Foundation of China (No. 61901227), and the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No. 19KJB510049).

References

1. Zhang, Z., Coutinho, E., Deng, J., et al.: Cooperative learning and its application to emotion recognition from speech. IEEE/ACM Trans. Audio Speech Lang. Proc. **23**(1), 115–126 (2015)
2. Tahon, M., Devillers, L.: Towards a small set of robust acoustic features for emotion recognition: challenges. IEEE/ACM Trans. Audio Speech Lang. Proc. **24**(1), 16–28 (2016)
3. Sun, L., Fu, S., Wang, F.: Decision tree SVM model with Fisher feature selection for speech emotion recognition. J Audio Speech Music Proc. **2019**, 2 (2019)
4. Chuang, Z.J., Wu, C.H.: Emotion recognition using acoustic features and textual content. In: *2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763), Taipei*, vol. 1, pp. 53–56 (2004).
5. Sun, L., Zou, B., Fu, S., et al.: Speech emotion recognition based on DNN-decision tree SVM model. Speech Commun. **115**, 29–37 (2019)
6. Liu, G., He, W., Jin, B.: Feature fusion of speech emotion recognition based on deep learning. In: *2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC), Guiyang*, pp. 193–197 (2018)
7. Hifny, Y., Ali, A.: Efficient Arabic emotion recognition using deep neural networks. In: *ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton*, pp. 6710–6714 (2019)

8. Tzirakis, P., Zhang, J., Schuller, B. W.: End-to-end speech emotion recognition using deep neural networks. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary*, , pp. 5089–5093 (2018).
9. Kim, E., Shin, J.W.: DNN-based emotion recognition based on bottleneck acoustic features and lexical features. In: *ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton*, pp. 6720–6724 (2019)
10. Lee, K.H., Kyun Choi, H., Jang, B.T.: A study on speech emotion recognition using a deep neural network. In: *2019 International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island*, pp. 1162–1165 (2019)
11. Harár, P., Burget, R., Dutta, M.K.: Speech emotion recognition with deep learning. In: *2017 4th International Conference on Signal Processing and Integrated Networks (SPIN), Noida*, pp. 137–140 (2017)
12. Wu, A., Huang, Y., Zhang, G.: Feature fusion methods for robust speech emotion recognition based on deep belief networks. In: *Proceedings of the Fifth International Conference on Network, Communication and Computing (ICNCC '16)*. Association for Computing Machinery, New York, pp. 6–10 (2018)
13. Long, X., Qu, D. Joint bottleneck feature and attention model for speech recognition. In: *Proceedings of 2018 International Conference on Mathematics and Artificial Intelligence (ICMAI '18)*. Association for Computing Machinery, New York, pp 46–50 (2018)
14. Wöllmer, M., Schuller, B.: Probabilistic speech feature extraction with context-sensitive Bottleneck neural networks. Neurocomputing **132**, 113–120 (2014)
15. Ke, X., Cao, B., Bai, J. *et al*: Speech emotion recognition based on PCA and CHMM. In: *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Chongqing*, pp. 667–671 (2019).
16. Jagini, N.P., Rao R.R.: Exploring emotion specific features for emotion recognition system using PCA approach. In: *2017 International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai*, pp. 58–62 (2017)
17. Zhang, S., Lei, B., Chen, A. et al.: KIsomap-based feature extraction for spoken emotion recognition. In: *IEEE 10th International Conference on Signal Processing Proceedings, Beijing*, pp. 1374–1377 (2010)
18. Siegert, I., Böck, R., Wendemuth, A.: Using a PCA-based dataset similarity measure to improve cross-corpus emotion recognition. Comput. Speech Lang. **51**, 1–23 (2018)
19. Kanth, N. R., Saraswathi, S.: Efficient speech emotion recognition using binary support vector machines & multiclass SVM. In: *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Madurai*, pp. 1–6 (2015)
20. Lanjewar, R.B., Mathurkar, S., Patel, N.: Implementation and comparison of speech emotion recognition system using Gaussian mixture model (GMM) and K-nearest neighbor (K-NN) techniques. Proc. Comput. Sci. **49**, 50–57 (2015)
21. Sarikaya, R., Hinton, G.E., Deoras, A.: Application of deep belief networks for natural language understanding[J]. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(4), 778–784 (2014)
22. Orłowski, T.: Application of deep belief networks in image semantic analysis and lossy compression for transmission. In: *2013 Signal Processing Symposium (SPS), Serock*, pp. 1–5 (2013)
23. Sim, K.B., Jang, I.H., Park, C.H.: The development of interactive feature selection and GA feature selection method for emotion recognition. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) *Knowledge-Based Intelligent Information and Engineering Systems. KES 2007. Lecture Notes in Computer Science*, vol 4694. Springer, Berlin (2007)
24. Le, B.V., Bang, J., Lee, S.: Hierarchical emotion classification using genetic algorithms. In: *Proceedings of the Fourth Symposium on Information and Communication Technology (SoICT '13)*. Association for Computing Machinery, New York, pp. 158–163 (2013)
25. Daneshfar, F., Kabudian, S.J.: Speech emotion recognition using discriminative dimension reduction by employing a modified quantum-behaved particle swarm optimization algorithm. Multimed. Tools Appl. **79**(1), 1261–1289 (2020)
26. Ntalampiras, S.: Speech emotion recognition via learning analogies. Pattern Recogn. Lett. **144**, 21–26 (2021)
27. Issa, D., Demirci, M.F., Yazici, A.: Speech emotion recognition with deep convolutional neural networks. Biomed. Signal Process. Control **59**, 101894 (2020)
28. Mustaqeem, M., Sajjad, M., Kwon, S.: Clustering based speech emotion recognition by incorporating learned features and deep BiLSTM. *IEEE Access* (2020)