Journal of System Simulation

Volume 32 | Issue 12 Article 11

12-16-2020

Human-computer Interaction Speech Emotion Recognition Based on Random Forest and Convolution Feature Learning

Wang Jing
1. College Of Optical And Electronical Information Changchun University Of Science And Technology,Changchun 130000,China; ;

Hongyan Liu
2. Network management center of China Mobile Communication Group Jilin Co.,Ltd,Changchun 130012,China; ;

Fangfang Liu
1. College Of Optical And Electronical Information Changchun University Of Science And Technology,Changchun 130000,China; ;

Qingqing Wang
3. Jilin Animation Institute,Changchun 130012,China;

Human-computer Interaction Speech Emotion Recognition Based on Random Forest and Convolution Feature Learning

Abstract

Abstract: Focus on the different speech features of different types of people in the automatic speech emotion recognition of emotional robots,*a random forest for speech emotion recognition is proposed,and a preliminary simulation experiment of emotional social robot system based on convolution feature learning is carried out.*The results show that the emotional robot can track in real time,the seven basic emotions of excitement,anger,sadness,happiness,surprise,fear and neutrality.By using non personalized speech emotion features,the original personalized speech emotion features are supplemented,and the general emotion and special emotion are extracted.For emotional robot,using these indicators has a certain application prospect in the simulation experiment and application experiment.

Keywords

emotional robot, automatic speech emotion recognition, random forest, convolution feature learning

Recommended Citation

Wang Jing, Liu Hongyan, Liu Fangfang, Wang Qingqing. Human-computer Interaction Speech Emotion Recognition Based on Random Forest and Convolution Feature Learning[J]. Journal of System Simulation, 2020, 32(12): 2388-2400.

第 32 卷第 12 期
2020 年 12 月
系统仿真学报©
Journal of System Simulation
Vol. 32 No. 12
Dec., 2020

Human-computer Interaction Speech Emotion Recognition Based on Random Forest and Convolution Feature Learning

Wang Jing[1], *Liu Hongyan*[2], *Liu Fangfang*[1], *Wang Qingqing*[3*]

(1. College Of Optical And Electronical Information Changchun University Of Science And Technology, Changchun 130000, China;
2. Network management center of China Mobile Communication Group Jilin Co., Ltd, Changchun 130012, China;
3. Jilin Animation Institute, Changchun 130012, China)

Abstract: Focus on the different speech features of different types of people in the automatic speech emotion recognition of emotional robots*, a random forest for speech emotion recognition is proposed, and a preliminary simulation experiment of emotional social robot system based on convolution feature learning is carried out.* The results show that the emotional robot can track in real time, the seven basic emotions of excitement, anger, sadness, happiness, surprise, fear and neutrality. By using non personalized speech emotion features, the original personalized speech emotion features are supplemented, and the general emotion and special emotion are extracted. For emotional robot, using these indicators has a certain application prospect in the simulation experiment and application experiment.

Keywords: emotional robot; automatic speech emotion recognition; random forest; convolution feature learning

基于随机森林和卷积特征学习的人机交互语音情感识别

王静[1]，刘洪岩[2]，刘芳芳[1]，王青青[3*]

(1. 长春理工大学光电信息学院，吉林 长春 130000；2. 中国移动通信集团吉林有限公司网络管理中心，吉林 长春 130012；
3. 吉林动画学院，吉林 长春 130012)

摘要：针对情绪机器人的自动语音情感识别在不同类型人群之间的语音特征差异，提出了*一种用于语音情感识别的随机森林，结合卷积特征学习对情绪化社交机器人系统进行了初步的仿真实验*，结果表明情绪机器人能够实时跟踪兴奋、愤怒、哀伤、高兴、惊讶、恐惧、中性 7 种基本情绪。通过采用非个性化的语音情感特征，补充了原始的个性化语音情感特征，实现了对通用性情感和特殊性情感的提取，对于情感机器人来说，利用这些指标在模拟实验和应用实验中都具有一定的应用前景。

关键词：情绪机器人；自动语音情感识别；随机森林；卷积特征学习
中图分类号：TP391.9 文献标识码：A 文章编号：1004-731X (2020) 12-2388-13
DOI: 10.16182/j.issn1004731x.joss.20-FZ0494E

Introduction

Human language has a very strong expressive power. Usually, people are willing to use emotion, intonation and language style to express the potential intention in the information. For the intelligent voice in human-computer interaction system, recognizing the information behind these languages, which is emotional information, can make the human-computer interaction experience better, and

Received date: 2020-03-28 Revised date: 2020-07-25;
Biographies: Wang Jing (1981-), female, Beijing, Master's degree, associate professor, Research direction is artificial intelligence; Liu Hongyan (1985-), male, Shenyang, Bachelor, Assistant Engineer, Research direction is artificial intelligence; Liu Fangfang(1982-), female, Changchun, Master's degree, associate professor, research direction is artificial intelligence.

can better understand the user’s intention. The purpose of SER is to detect emotion from speech. Because it can extract the speaker’s emotional state from the speaker's voice and analyze the specific semantic information from the emotional speech, it plays a key role in the research of human-computer interaction and machine intelligence, and is also the key premise of human-computer interaction. Through continuous research on SER, its importance to computer development and social life has become increasingly prominent, and has been widely used in interactive films, emotional speech translation, automatic psychological testing[1], video interactive games[2], psychological assistant therapy[3], intelligent nursing for the elderly[4]. It has gradually become a research hotspot in the field of human-computer interaction[5].

In speech human-computer interaction, the input perceptual speech may have different speaker characteristics and similar acoustic properties. Therefore, the most important thing in automatic speech emotion recognition is how to extract special speech signals from speech. The traditional method is to extract the low-level frame level descriptor[6] from speech, and then use a series of statistical aggregation functions to solve it, such as mean value, maximum, variance, etc. The statistical feature vector can roughly describe the time variation and contour of the low-level frame level descriptor, which is considered to be closely related to voice emotion. Then, through the reasoning model of deep learning[7], as well as recurrent neural network, it can help in speech emotion recognition. For example, deep neural network is applied to infer speech emotion from the generated statistical vector[8].

However, these advanced methods still have limitations, such as insufficient use of context information. In the perception of human language emotion, it is necessary to integrate the conventional and effective context from the whole language, so as to automatically capture the special emotional voice. Human emotional perception is mainly based on the vocalization at the lexical level, and the voice with emotion can be activated more than neutral voice in the human cortex. Therefore, this paper mainly studies the following three aspects:

(1) How to generate special information features with resolution for emotion perception according to people’s gender, age, occupation and other information, and put forward a speech emotion recognition model which combines various features;

(2) How to identify different emotional states and classify some emotions which are difficult to recognize to some extent;

(3) According to the research scheme, the speech emotion robot system is tested.

The experimental results show that the emotion robot can track and distinguish seven common emotions: excitement, anger, sad, happy, surprise, fear and neutral in real time.

In the second section of this paper, the key research methods will be analyzed. Section 3 introduces simulation experiments and analysis, including further applications. In Section 4, the paper is summarized.

1 Analysis of key methods of research

In this paper, random forest and convolution feature learning is applied to speech emotion recognition, including excitement, anger, sadness, happiness, surprise, fear and neutral emotions. The personalized features and non-personalized features of emotion are integrated together. Research shows

https://dc-china-simulation.researchcommons.org/journal/vol32/iss12/11
DOI: 10.16182/j.issn1004731x.joss.20-FZ0494E

that people's identity information, including gender, age, occupation, has a certain impact on emotional expression. Therefore, according to the gender, age, occupation and other recognition information, this paper proposes a convolutional neural network for one-dimensional time convolution layer information feature learning[9]. This is a kind of stack that keeps the original structure of input signal through residual structure in the process of progressive feature extraction, so as to achieve the purpose of retaining important speech information[10]. On this basis, the random forest algorithm is used to analyze and recognize speech features by using decision tree[11] and Bootstrap method. In addition, in each of the multiple classifications, a separate classification is adopted to deal with the difficult emotion.

1.1 Feature extraction

Low level frame level descriptors are extracted from speech signals as input acoustic features to solve the extraction of speech emotional features. In addition, the derivative-based non-personalized speech emotion features are adopted, which can make up for the deficiency of personalized emotional features, and achieve universal emotional features. At the same time, the convolutional neural network layer with multiple head self-attention is used to extract emotional salient features and merge the semantic similar hidden output to generate super segment features with lower temporal resolution[12]. The speech emotion feature set is calculated by Open Smile, and the common features such as F0, ZCR, RMS energy, harmonic, noise ratio and MFCC 1~12 are extracted and analyzed. Features which have little influence on different individuals are regarded as general popular features, and the statistical values of these characteristics are calculated[13]. Through the above calculation and analysis, personality and non-personality characteristics can be obtained.

1.2 Convolution feature learning

At the same time, a convolutional neural network is proposed for learning information features of one-dimensional time convolution layer to obtain emotions[14]. As shown in Fig. 1, for convolution neural network, each group of convolution layers adopts a two-step structure, increasing the receiving field and reducing the time resolution. For layer I convolution layer, the output calculation is as follows.

$$g_i^n \times f_i^n(\mathrm{p}) = \sum\nolimits_{\mathrm{x+y=p}} g_i^n(\mathrm{x}) f_i^n(\mathrm{y}) \qquad (1)$$

Where f_i^n is the filter of g_i^n, and the field of p is the feature mapping in the convolution layer[15]. In particular, in order to optimize the training efficiency, the special function change method of batch normalization is used in the convolution layer. In addition, the perception of human voice is based on consonant and vowel syllables and lasts from 100 ms to 250 ms[16]. In this study, for the input acoustic features with a displacement length of 5 ms, n=4 convolution layers are used for feature learning, as shown in Fig. 1, resulting in a learned super segment feature of 200 ms close to the human perception granularity[17].

1.3 Speech emotion recognition based on random forest algorithm

In the process of analysis, if there are n decision trees and X speech feature variables, then the generation of random forest algorithm is mainly divided into three steps[18-19]:

(1) Firstly, n subsets samples are randomly selected and N decision trees are selected. Then, the

bootstrap method is used to analyze the training number. The training number is composed of several subsets and affects the growth of the decision tree. The training number sample contains OOB data, which is used for training and testing model.

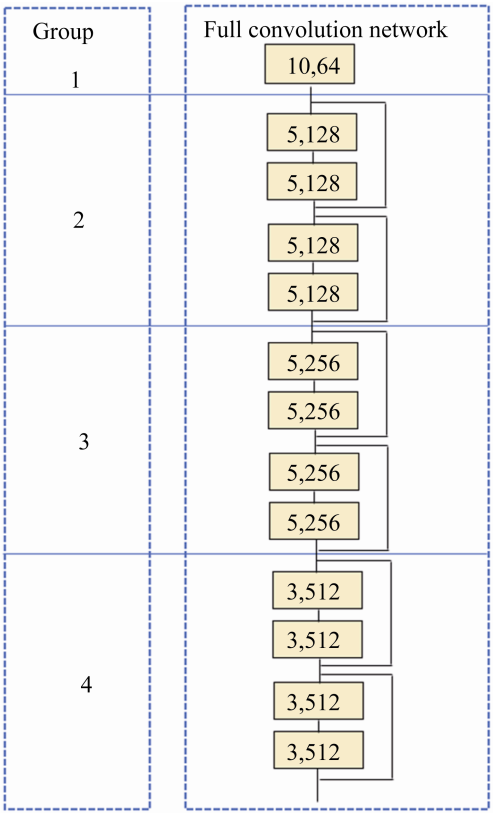


Fig. 1　Convolution feature learning

(2) Suppose K features are selected randomly, and then these feature variables with the best classification function are split optimally to get the split nodes. In the process of random forest growth, the number of variables do not change. Repeat this process until the entire decision tree is complete.

(3) Each decision tree grows to the maximum without any pruning. The random forest is composed of the generated decision tree. The output of the whole random forest is analyzed by the results obtained after the decision tree grows to the maximum. Because the random forest can classify the tested voice data, the voice data category with the most votes is the output result. The calculation decision is as follows:

$$Output(x) = argmax\sum\nolimits_{i=1}^{\mathrm{M}} I(S_i(x) = T) \qquad (2)$$

Among them, Output(x) is the output of each category of voice data, I is the voice set function, $S_i(x)$ is the single decision tree model, T is the target marker, and represents the emotion type. Bootstrap method uses Gini coefficient to extract training subset randomly. The results show that the smaller the Gini coefficient is, the better the selectivity is. If K samples are selected, Gini coefficient is defined as follows in the test speech data set a, where the probability of class I is $\mathrm{P_i}$:

$$Gini(A) = \sum\nolimits_{i=1}^{k} p_i(1-p_i) = 1 - \sum\nolimits_{i=1}^{k} P_i^2 \qquad (3)$$

In (3), if the probability of the first speech data sample is set to P, the Gini coefficient formula can be simplified, as shown in formula (4):

$$Gini(A) = 2P(1-P) \qquad (4)$$

Next, the index Ginib corresponding to each set Branch is designed to facilitate the search, as shown in formula (5):

$$Gini_b\left(A\right) = \frac{A_1}{A} Gini\left(A_1\right) + \ldots + \frac{A_n}{A} Gini(A_n) \qquad (5)$$

Here, Gini represents the coupling degree between sets. The smaller Gini value is, the lower the coupling degree between sets is. The nth subset of set A is represented by An. Through calculation, the Gini value of each speech feature can be sorted, and the difficulty degree of recognition can be obtained.

This paper classifies some emotions in random forest algorithm. There are n different emotional states, and two kinds of relatively difficult to identify emotions are extracted. It can be concluded that Y random forests are needed.

$$Y_R = 2\left|\frac{n}{2}\right| - 1, k = 1, 2, \ldots, k \qquad (6)$$

https://dc-china-simulation.researchcommons.org/journal/vol32/iss12/11
DOI: 10.16182/j.issn1004731x.joss.20-FZ0494E

Through the establishment of emotion corpus containing seven kinds of basic emotions, we can improve the recognition efficiency of different kinds of emotions and increase the accuracy of recognition. The random forest algorithm for emotion classification includes the following four steps[20].

(1) OpenSmile is used to preprocess speech signal and extract feature data.

(2) Using fuzzy clustering method, considering the influence of recognition information on emotion, the training data set is classified by clustering analysis. `

(3) Train the RF classifier. According to the similarity between experience and pressure, five decision tree classifiers are used to determine the structure of random forest. According to the experiments on the benchmark database, in most cases, the accuracy of excitement and happiness is relatively low, and the accuracy of sadness and fear is relatively low. Therefore, the two kinds of emotions are trained separately from the other three kinds of emotions. The structure of classification is as follows: Random Forest 1 distinguishes excitement, happiness and other emotions, random forest 2 distinguishes excitement and happiness, and random forest 3 areas divide the remaining five emotions. Random forest 4 distinguishes sadness from fear, and random forest 5 distinguishes anger, surprise and neutrality. The random forest structure of speech emotion recognition is shown in Fig. 2. Due to the similarity of some emotions, it is easy to confuse them. Therefore, in the training stage, excitement and happiness, sadness and fear are trained separately.

(4) The trained random forest is used to classify 7 basic emotions, and L subclass is integrated into the classification results[21].

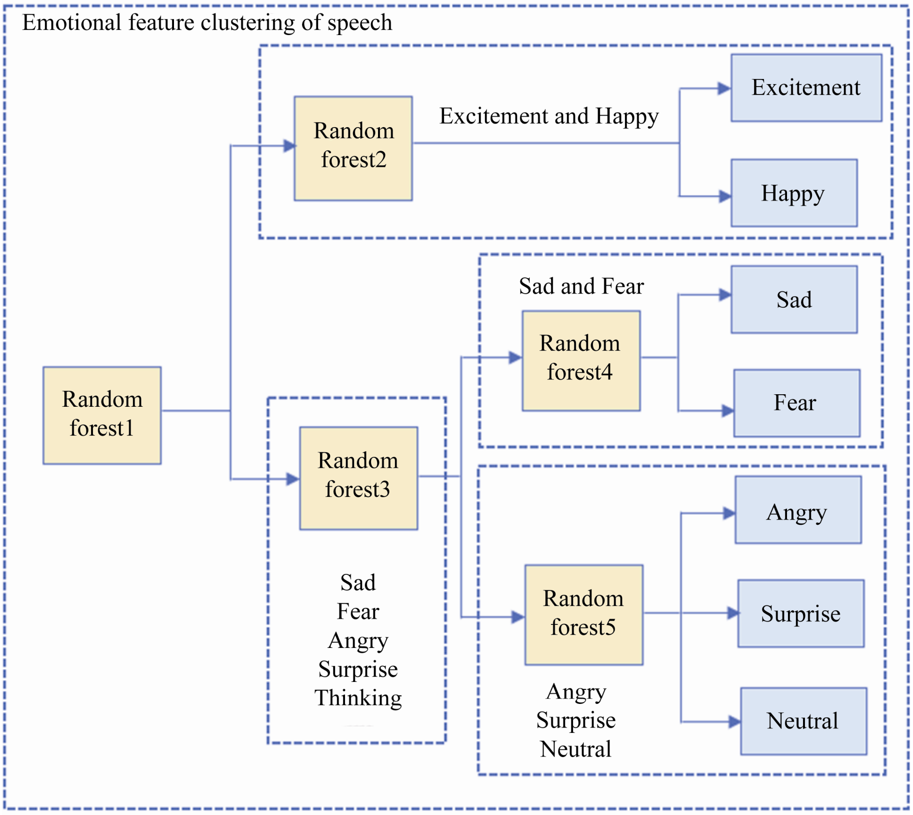


Fig. 2 Random forest structure of speech emotion recognition

1.4 Feature dimension reduction algorithm

In the process of speech feature extraction, a large number of speech extraction data are obtained, but the dimension of these data is relatively high, which will have a great impact on the training time of data, the real-time performance of classification and the consumption of computer resources. Therefore, in order to improve the real-time performance of speech recognition, it is necessary to reduce the dimension during the experiment. There are many dimensionality reduction methods currently. Among them, MDS, which can better preserve the internal structure of data and has relatively small amount of calculation, is selected. This dimension reduction method is very convenient to retain some similar emotional data in the process of dimensionality reduction. The principle of the algorithm is to reduce the samples and variables in the high-dimensional space to the low-dimensional space, and then analyze and classify these samples and variables in the low-dimensional space. The biggest feature of the algorithm is that it can retain the similarity between emotional data.

If n-dimensional sample matrix A is obtained in feature extraction, the distance between relative position coordinates (x, y) can be found in low-dimensional space, and can be calculated by formula (7):

$$\sigma_{xy}=\sum\nolimits_{s=1}^{n}\sqrt{(r_{xs}-r_{ys})^2} \tag{7}$$

Then make the distance of eigenvector as close as possible to σ_{xy}, Next, the reduced dimension matrix Y can be solved by formula (8):

$$\hat{X}=-\frac{1}{2}(E-\frac{1}{m})\sigma_{xy}{}^{2}YY^{\mathrm{T}} \tag{8}$$

The diagonal matrix composed of the eigenvalues of $\hat{X}$ constitutes X. The eigenvalues of $\hat{X}$ are sorted according to the order of magnitude. The larger eigenvalue and the corresponding eigenvector K are selected, and the formula (9) is used:

$$Y=KX^{\frac{1}{2}} \tag{9}$$

In this way, the reduced dimension sample matrix Y can be obtained, and the similarity between the data is retained to ensure the real-time performance of emotion recognition.

2 Experimental results and discussion

2.1 Database settings

In the experiment, IEMO-CAP and RID are used to evaluate the performance. IEMO-CAP is used to compare the performance of the proposed stochastic forest structure and the latest method of speech emotion recognition. RID is used to evaluate the robustness and effectiveness of the proposed structure in the actual Interaction scenario[22].

The IEMO-CAP database contains a 24-hour English conversation with 7 381 sentences, which are divided into seven emotional categories: excitement, anger, sad, happy, surprise, fear and neutral. In the experiment, we combine the class marked “excitement” with the class marked “happy”, and combine the class marked "sad" with the class marked “fear” to form a database of three classes marked {happy, sad, angry}. Each class contains {2 215, 2 384, 2 782} discourse. Rid is collected from real human-computer interaction, written and marked by users. The database contains 435 675 utterances, which are divided into seven emotional categories: excitement, anger, sad, happy, surprise, fear and neutral. Each class contains {61 128, 59 294, 69 387, 63 825, 58 267, 62 139, 61 635} utterances. IEMO-CAP and rid are randomly shifted and divided

DOI: 10.16182/j.issn1004731x.joss.20-FZ0494E

into three parts according to the proportion of 7:1:1 for training, verification and testing. The expression forms of voice emotion are divided into personality and non-personality. The speech emotion sets of the two features are extracted by OpenSmile, and four kinds of basic feature analysis in Tab.1 are obtained. These feature analysis is calculated from 12 statistical data, representing 12 equal frequency spectrum energy dynamic coefficients. The difference between non personalized features and personalized features is that non personalized features can grasp the influence of different people on speech recognition by analyzing the change rate of dynamic coefficients.

Tab. 1 Use OpenSmile to extract speech emotion features

Index	Maximum	Minimum	Mean	Slope
RMS energy	7.23E-03	1.56E-02	2.14E+01	1.93E–01
Zero current ratio	4.76E+00	–2.23E+01	2.45+01	2.19E+00
Harmonic noise ratio	4.23E+01	–2.16E+01	2.13E+01	2.34E+00
MFCC	7.22E+00	–3.78E+01	4.29E+01	1.89E–01

2.2 Environment settings

After preprocessing the speech emotion feature set, 2 923 514 feature vectors are obtained from two different genders. Each feature vector corresponds to an emotional tag, 1=exception, 2=angry, 3=sad, 4=happy, 5=surprise, 6=fear, 7=neutral. People of different genders have different ways of expressing emotions[23], and there age and occupation also have certain influence on speech emotion recognition[24]. Therefore, the training set is clustered into two sub classes according to gender. In two different gender subclasses, 75% of the speech feature data are used to establish the feature model for speech recognition, and 25% of the feature data are used to test the model. In the process of building a random forest, bootstrap method is used for random sampling. Set up 1 000 different sample subsets that can be selected repeatedly. On this basis, each sample subset is trained to generate a decision tree to form a random forest model. Then random forest algorithm is used to classify seven basic emotions[25].

2.3 Simulation and analysis

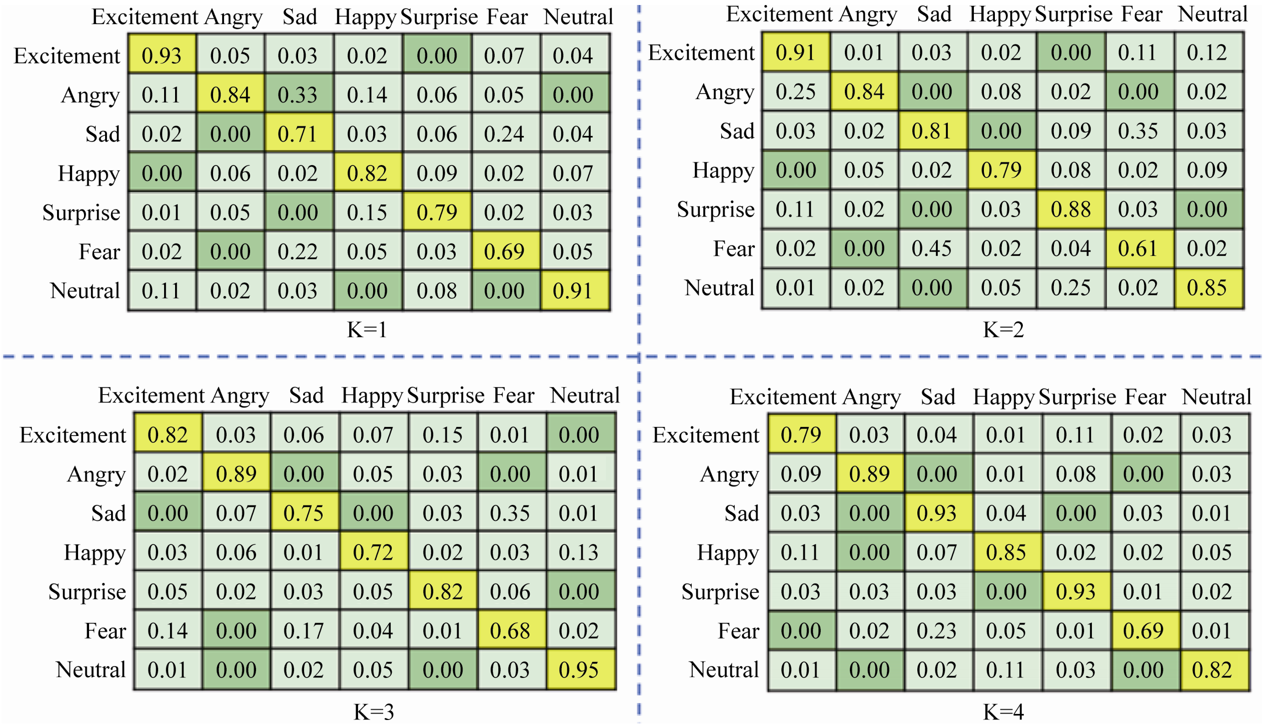
Through the establishment of classification model to verify the accuracy of speech recognition, for example, by establishing convolutional neural network (CNN)[26] model, we can use its algorithm to verify the accuracy of the scheme, and through the establishment of RF[27] model, this kind of verification can also be carried out. Therefore, in order to ensure the accuracy of the data obtained in the simulation experiment, a four fold cross validation method is used to verify the algorithm[28-29]. Suppose that the loop variable is k, and finally the results of four cycle processes are outputted. The comparison of two ser results on IEMO-CAP corpus and rid is shown in Tab. 2.

Tab. 2 Comparison of speech emotion recognition based on IEMO-CAP corpus and RID /%

Index	IEMO-CAP		RID	
	CNN	RF	CNN	RF
K=1	76.78	71.24	81.13	84.56
K=2	75.65	73.12	77.91	82.95
K=3	80.32	76.89	75.93	79.98
K=4	85.13	85.95	76.17	79.11
Average	79.74	76.80	77.78	81.65

Using CNN algorithm[30], in the first case, CNN algorithm includes three layers in the neural network. In the process of experiment, 100 input layer data nodes are set up in the neural network, and seven output layer data nodes are set in IEMO-CAP corpus and RID, respectively representing the test emotions of seven bases. In Fig. 3~4, the confusion matrix obtained by CNN algorithm is shown by cross validation, and ser results of seven different emotions are shown in the figure. Fig. 3 shows that the average recognition rate of IEMO-CAP corpus is 83.25%, and Fig. 4 shows that the average recognition rate of RID is 79.24%.

Next, random forest algorithm is used to test, as shown in Fig. 5 and Fig. 6.



	Excitement	Angry	Sad	Happy	Surprise	Fear	Neutral
Excitement	0.93	0.05	0.03	0.02	0.00	0.07	0.04
Angry	0.11	0.84	0.33	0.14	0.06	0.05	0.00
Sad	0.02	0.00	0.71	0.03	0.06	0.24	0.04
Happy	0.00	0.06	0.02	0.82	0.09	0.02	0.07
Surprise	0.01	0.05	0.00	0.15	0.79	0.02	0.03
Fear	0.02	0.00	0.22	0.05	0.03	0.69	0.05
Neutral	0.11	0.02	0.03	0.00	0.08	0.00	0.91

K=1

	Excitement	Angry	Sad	Happy	Surprise	Fear	Neutral
Excitement	0.91	0.01	0.03	0.02	0.00	0.11	0.12
Angry	0.25	0.84	0.00	0.08	0.02	0.00	0.02
Sad	0.03	0.02	0.81	0.00	0.09	0.35	0.03
Happy	0.00	0.05	0.02	0.79	0.08	0.02	0.09
Surprise	0.11	0.02	0.00	0.03	0.88	0.03	0.00
Fear	0.02	0.00	0.45	0.02	0.04	0.61	0.02
Neutral	0.01	0.02	0.00	0.05	0.25	0.02	0.85

K=2

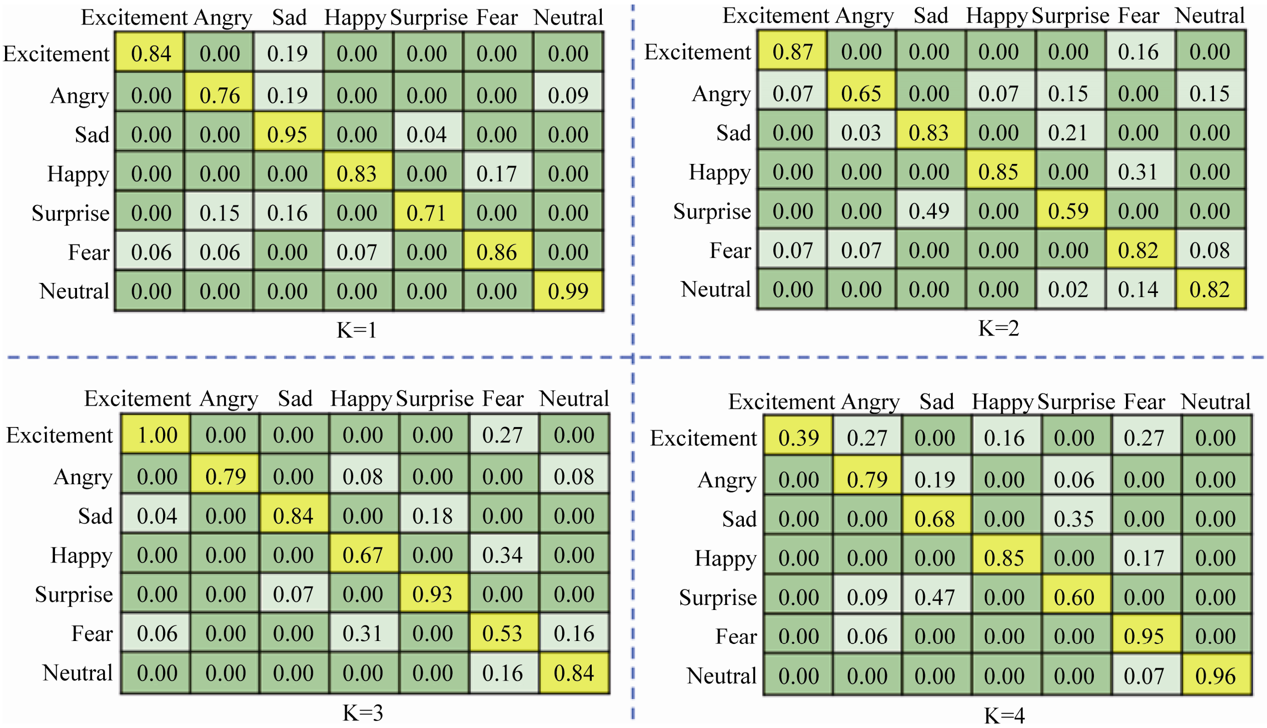
	Excitement	Angry	Sad	Happy	Surprise	Fear	Neutral
Excitement	0.82	0.03	0.06	0.07	0.15	0.01	0.00
Angry	0.02	0.89	0.00	0.05	0.03	0.00	0.01
Sad	0.00	0.07	0.75	0.00	0.03	0.35	0.01
Happy	0.03	0.06	0.01	0.72	0.02	0.03	0.13
Surprise	0.05	0.02	0.03	0.05	0.82	0.06	0.00
Fear	0.14	0.00	0.17	0.04	0.01	0.68	0.02
Neutral	0.01	0.00	0.02	0.05	0.00	0.03	0.95

K=3

	Excitement	Angry	Sad	Happy	Surprise	Fear	Neutral
Excitement	0.79	0.03	0.04	0.01	0.11	0.02	0.03
Angry	0.09	0.89	0.00	0.01	0.08	0.00	0.03
Sad	0.03	0.00	0.93	0.04	0.00	0.03	0.01
Happy	0.11	0.00	0.07	0.85	0.02	0.02	0.05
Surprise	0.03	0.03	0.03	0.00	0.93	0.01	0.02
Fear	0.00	0.02	0.23	0.05	0.01	0.69	0.01
Neutral	0.01	0.00	0.02	0.11	0.03	0.00	0.82

K=4

Fig. 3 Confusion matrix of recognition results based on CNN (IEMO-CAP corpus)



	Excitement	Angry	Sad	Happy	Surprise	Fear	Neutral
Excitement	0.84	0.00	0.19	0.00	0.00	0.00	0.00
Angry	0.00	0.76	0.19	0.00	0.00	0.00	0.09
Sad	0.00	0.00	0.95	0.00	0.04	0.00	0.00
Happy	0.00	0.00	0.00	0.83	0.00	0.17	0.00
Surprise	0.00	0.15	0.16	0.00	0.71	0.00	0.00
Fear	0.06	0.06	0.00	0.07	0.00	0.86	0.00
Neutral	0.00	0.00	0.00	0.00	0.00	0.00	0.99

K=1

	Excitement	Angry	Sad	Happy	Surprise	Fear	Neutral
Excitement	0.87	0.00	0.00	0.00	0.00	0.16	0.00
Angry	0.07	0.65	0.00	0.07	0.15	0.00	0.15
Sad	0.00	0.03	0.83	0.00	0.21	0.00	0.00
Happy	0.00	0.00	0.00	0.85	0.00	0.31	0.00
Surprise	0.00	0.00	0.49	0.00	0.59	0.00	0.00
Fear	0.07	0.07	0.00	0.00	0.00	0.82	0.08
Neutral	0.00	0.00	0.00	0.00	0.02	0.14	0.82

K=2

	Excitement	Angry	Sad	Happy	Surprise	Fear	Neutral
Excitement	1.00	0.00	0.00	0.00	0.00	0.27	0.00
Angry	0.00	0.79	0.00	0.08	0.00	0.00	0.08
Sad	0.04	0.00	0.84	0.00	0.18	0.00	0.00
Happy	0.00	0.00	0.00	0.67	0.00	0.34	0.00
Surprise	0.00	0.00	0.07	0.00	0.93	0.00	0.00
Fear	0.06	0.00	0.00	0.31	0.00	0.53	0.16
Neutral	0.00	0.00	0.00	0.00	0.00	0.16	0.84

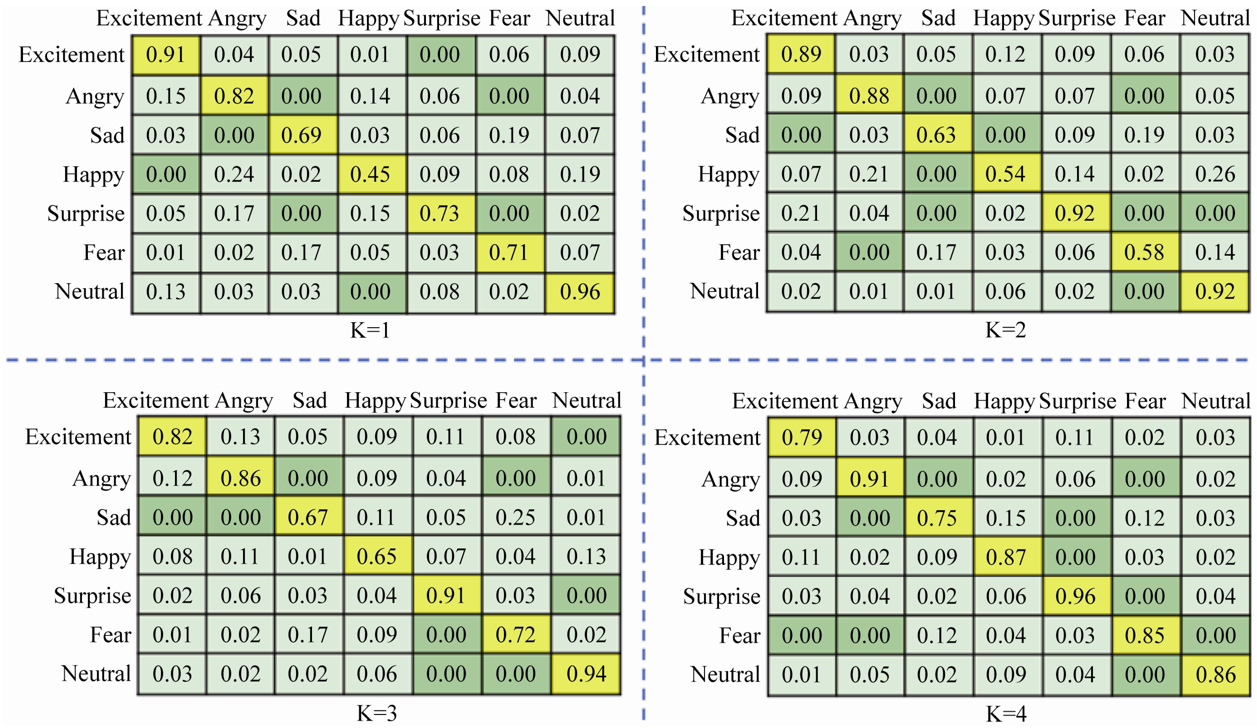
K=3

	Excitement	Angry	Sad	Happy	Surprise	Fear	Neutral
Excitement	0.39	0.27	0.00	0.16	0.00	0.27	0.00
Angry	0.00	0.79	0.19	0.00	0.06	0.00	0.00
Sad	0.00	0.00	0.68	0.00	0.35	0.00	0.00
Happy	0.00	0.00	0.00	0.85	0.00	0.17	0.00
Surprise	0.00	0.09	0.47	0.00	0.60	0.00	0.00
Fear	0.00	0.06	0.00	0.00	0.00	0.95	0.00
Neutral	0.00	0.00	0.00	0.00	0.00	0.07	0.96

K=4

Fig. 4 Confusion matrix (RID) of recognition results using CNN

https://dc-china-simulation.researchcommons.org/journal/vol32/iss12/11
DOI: 10.16182/j.issn1004731x.joss.20-FZ0494E



K=1

	Excitement	Angry	Sad	Happy	Surprise	Fear	Neutral
Excitement	0.91	0.04	0.05	0.01	0.00	0.06	0.09
Angry	0.15	0.82	0.00	0.14	0.06	0.00	0.04
Sad	0.03	0.00	0.69	0.03	0.06	0.19	0.07
Happy	0.00	0.24	0.02	0.45	0.09	0.08	0.19
Surprise	0.05	0.17	0.00	0.15	0.73	0.00	0.02
Fear	0.01	0.02	0.17	0.05	0.03	0.71	0.07
Neutral	0.13	0.03	0.03	0.00	0.08	0.02	0.96

K=2

	Excitement	Angry	Sad	Happy	Surprise	Fear	Neutral
Excitement	0.89	0.03	0.05	0.12	0.09	0.06	0.03
Angry	0.09	0.88	0.00	0.07	0.07	0.00	0.05
Sad	0.00	0.03	0.63	0.00	0.09	0.19	0.03
Happy	0.07	0.21	0.00	0.54	0.14	0.02	0.26
Surprise	0.21	0.04	0.00	0.02	0.92	0.00	0.00
Fear	0.04	0.00	0.17	0.03	0.06	0.58	0.14
Neutral	0.02	0.01	0.01	0.06	0.02	0.00	0.92

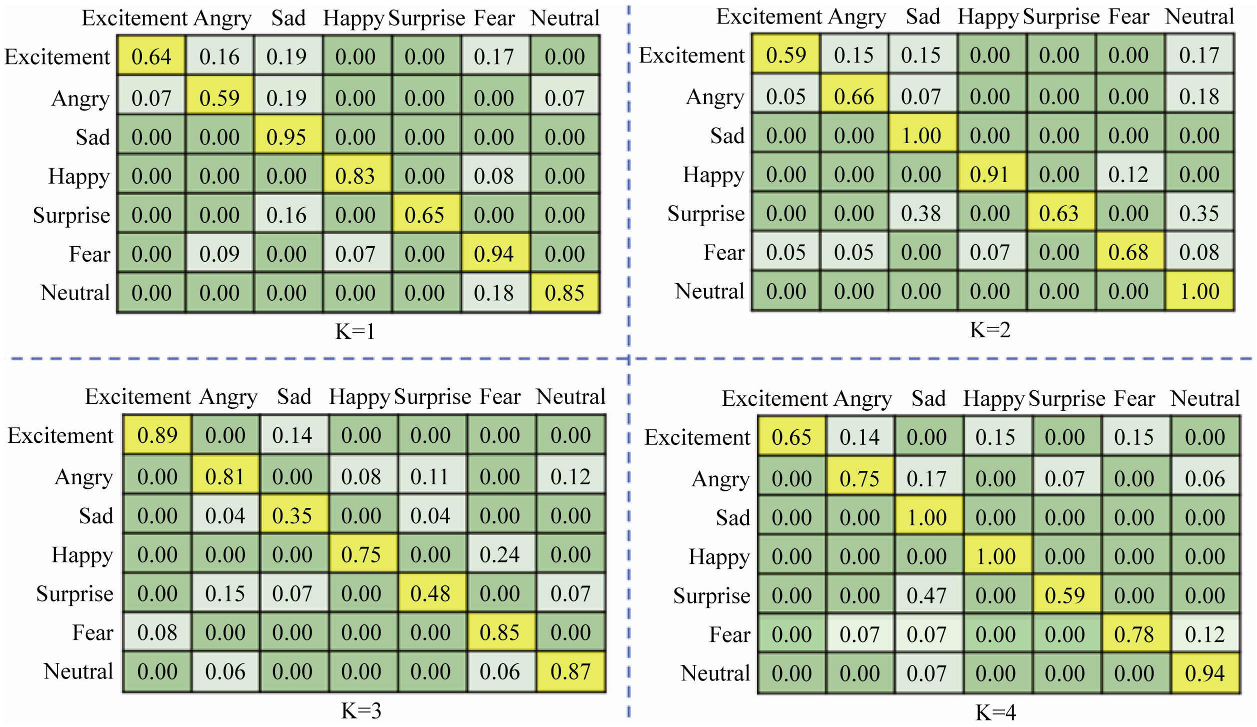
K=3

	Excitement	Angry	Sad	Happy	Surprise	Fear	Neutral
Excitement	0.82	0.13	0.05	0.09	0.11	0.08	0.00
Angry	0.12	0.86	0.00	0.09	0.04	0.00	0.01
Sad	0.00	0.00	0.67	0.11	0.05	0.25	0.01
Happy	0.08	0.11	0.01	0.65	0.07	0.04	0.13
Surprise	0.02	0.06	0.03	0.04	0.91	0.03	0.00
Fear	0.01	0.02	0.17	0.09	0.00	0.72	0.02
Neutral	0.03	0.02	0.02	0.06	0.00	0.00	0.94

K=4

	Excitement	Angry	Sad	Happy	Surprise	Fear	Neutral
Excitement	0.79	0.03	0.04	0.01	0.11	0.02	0.03
Angry	0.09	0.91	0.00	0.02	0.06	0.00	0.02
Sad	0.03	0.00	0.75	0.15	0.00	0.12	0.03
Happy	0.11	0.02	0.09	0.87	0.00	0.03	0.02
Surprise	0.03	0.04	0.02	0.06	0.96	0.00	0.04
Fear	0.00	0.00	0.12	0.04	0.03	0.85	0.00
Neutral	0.01	0.05	0.02	0.09	0.04	0.00	0.86

Fig. 5 Confusion matrix of random forest recognition results (IEMO-CAP corpus)



K=1

	Excitement	Angry	Sad	Happy	Surprise	Fear	Neutral
Excitement	0.64	0.16	0.19	0.00	0.00	0.17	0.00
Angry	0.07	0.59	0.19	0.00	0.00	0.00	0.07
Sad	0.00	0.00	0.95	0.00	0.00	0.00	0.00
Happy	0.00	0.00	0.00	0.83	0.00	0.08	0.00
Surprise	0.00	0.00	0.16	0.00	0.65	0.00	0.00
Fear	0.00	0.09	0.00	0.07	0.00	0.94	0.00
Neutral	0.00	0.00	0.00	0.00	0.00	0.18	0.85

K=2

	Excitement	Angry	Sad	Happy	Surprise	Fear	Neutral
Excitement	0.59	0.15	0.15	0.00	0.00	0.00	0.17
Angry	0.05	0.66	0.07	0.00	0.00	0.00	0.18
Sad	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Happy	0.00	0.00	0.00	0.91	0.00	0.12	0.00
Surprise	0.00	0.00	0.38	0.00	0.63	0.00	0.35
Fear	0.05	0.05	0.00	0.07	0.00	0.68	0.08
Neutral	0.00	0.00	0.00	0.00	0.00	0.00	1.00

K=3

	Excitement	Angry	Sad	Happy	Surprise	Fear	Neutral
Excitement	0.89	0.00	0.14	0.00	0.00	0.00	0.00
Angry	0.00	0.81	0.00	0.08	0.11	0.00	0.12
Sad	0.00	0.04	0.35	0.00	0.04	0.00	0.00
Happy	0.00	0.00	0.00	0.75	0.00	0.24	0.00
Surprise	0.00	0.15	0.07	0.00	0.48	0.00	0.07
Fear	0.08	0.00	0.00	0.00	0.00	0.85	0.00
Neutral	0.00	0.06	0.00	0.00	0.00	0.06	0.87

K=4

	Excitement	Angry	Sad	Happy	Surprise	Fear	Neutral
Excitement	0.65	0.14	0.00	0.15	0.00	0.15	0.00
Angry	0.00	0.75	0.17	0.00	0.07	0.00	0.06
Sad	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Happy	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Surprise	0.00	0.00	0.47	0.00	0.59	0.00	0.00
Fear	0.00	0.07	0.07	0.00	0.00	0.78	0.12
Neutral	0.00	0.00	0.07	0.00	0.00	0.00	0.94

Fig. 6 Confusion matrix (RID) of recognition results using random forest

The results in the Fig. 5 and Fig. 6 are still obtained by cross validation, and the confusion matrix diagram obtained by random forest algorithm is also shown, and ser results of seven different emotions are shown in the figure. The average recognition rate of IEMO-CAP corpus is 76.25%, and that of RID is

82.37%. Through comparison, it is found that the average recognition rate of CNN is relatively high on IEMO-CAP corpus. By analyzing the results of SER, it can be concluded that CNN is a more stable method than random forest.

On IEMO-CAP corpus and RID, CNN and RF are used to compare the average accuracy of seven basic emotions, as shown in Tab. 3. Due to the uneven distribution of samples, the accuracy of fear, happy and sadness in the seven basic emotions is relatively low by using the above two methods. This is because the speech characteristics of anger and fear, excitement and happy are very similar, so it is possible to identify errors in judgment. Moreover, the current recognition method is not suitable for distinguishing the price of emotion due to its lack of prominence in audio. Therefore, there is a high degree of confusion between fear and sadness, between excitement and happy. This is because when expressed in voice signals, fear and sadness are similar, and so are excitement and happy.

Tab. 3　Comparison of basic emotion average accuracy based on IEMO-CAP corpus and RID　　%

Method	IEMO-CAP		RID	
	CNN	RF	CNN	RF
Excitement	77.24	69.25	80.60	98.60
Angry	85.23	86.35	75.70	68.12
Sad	74.60	75.98	61.35	60.67
Happy	78.74	66.45	77.55	85.78
Surprise	88.22	89.12	74.34	67.14
Fear	66.34	70.23	78.89	79.22
Neutral	92.57	89.67	86.23	88.65

In order to further analyze the experimental results, kappa coefficient, sensitivity and specificity can be calculated according to the confusion matrix. The average values of kappa coefficient, sensitivity and specificity are highly consistent. As shown in Tab. 4.

Tab. 4　Comparison of speech emotion coefficient recognition based on IEMO-CAP corpus and RID

Index	IEMO-CAP		RID	
	CNN	RF	CNN	RF
Kappa Coefficient	0.72	0.74	0.61	0.71
Sensibility	0.81	0.83	0.59	0.61
Specificity	0.91	0.96	0.87	0.89
Standard error	1.20	2.31	1.65	1.23

From the error rate results of the two methods, it can be seen from Tab. 4 that CNN and RF and standard errors in the IEMO-CAP corpus are 1.20 and 2.31 respectively. The standard errors of CNN and RF on RID are 1.65 and 1.23 respectively. Therefore, the stability of the method proposed in the experiment is relatively good. In addition, IEMO-CAP is stable in processing some high-dimensional information, and has obvious advantages compared with other methods. Therefore, the data classification takes into account age, gender and occupation.

2.4 Preliminary application test

2.4.1 Experimental environment settings

First of all, through the research and analysis of the system development, a robot system is established that can identify the user's emotion, as shown in Fig. 7. The system includes data transmission equipment, router, PC, workstation capable of emotional computing and emotional robot. Firstly, the system collects the voice signal through Kinect fixed on the emotion robot, and then the collected emotion data is imported by the emotion workstation. At the same time, the data is input into the ser for emotion recognition. Finally, the emotion results after recognition are transmitted to the robot. After the robot gets the results, it makes corresponding output response according to the analysis and understanding. If people are not in the emotional workstation, they can also use PC to remotely control the workstation, so the operability of the whole experimental system is very flexible[31-32].

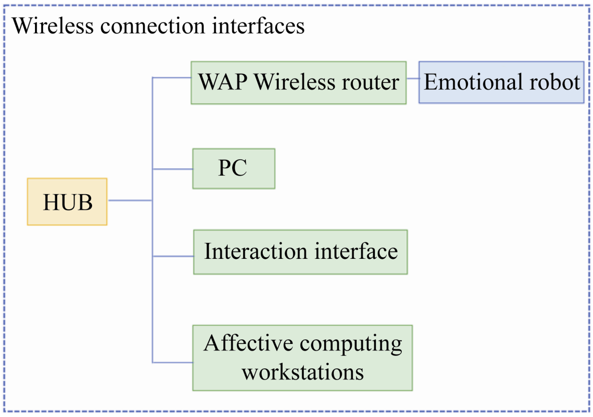

Wireless connection interfaces
WAP Wireless router
Emotional robot
PC
HUB
Interaction interface
Affective computing workstations

Fig. 7 Structure of emotional robot system

2.4.2 Preliminary application test results and analysis

During the experiment, six volunteers of different genders, ages and occupations participate in the whole experiment. The volunteers say the same 200 basic emotional short sentences and finally formed 1 200 voice segments. As the original data of the experiment, the microphone array in Kinect is used to process the original data in the audio stream, that is, multi-channel audio. These original data are input into audio processing software, and then are analyzed and processed by emotion computing workstation. Different feature sets are extracted by using OpenSmile toolbox to get the result of emotion recognition. The results obtained by CNN and RF are shown in Tab. 5.

By comparison, the variances of the two algorithms have strong stability. It can be seen from Tab. 5 that the average values of kappa coefficient, sensitivity and specificity are highly consistent with the accuracy, indicating the effectiveness of the method. In addition, MDS is used to reduce the dimension of emotional data and extract 523 dimension feature vectors. The dimension value of MDS is 84 dimensions, and the time complexity of the algorithm is 0.001 4 s. Therefore, the calculation time of CNN and RF is 0.009 7 s and 0.004 5 s, respectively. The experimental results are basically consistent with the theoretical analysis, which proves that the accuracy of emotion recognition is in a reasonable range, and the experimental method can be used in other databases.

Tab. 5 Comparative experiment of speech emotion recognition application

Index	CNN	RF
K=1 (%)	72.98	82.34
K=2 (%)	81.78	86.45
K=3 (%)	69.13	75.87
K=4 (%)	78.12	87.33
Average	75.50	82.99
Kappa Coefficient	0.56	0.64
Sensibility	0.54	0.59
Specificity	0.83	0.87

3 Conclusion

The main problems of emotional speech signal recognition in this paper include feature selection and classification method recognition. Through the use of non- personalized speech emotional features, and based on derivative method to extract speech emotional features, and supplement the original personalized voice emotional features, and through experiments, the extraction of general emotion and special emotion is achieved.

Using decision tree and bootstrap method, as well as convolution feature learning to classify speech emotion features, and recognize the classified feature data, the accuracy of recognition is improved by experimental methods. The experiments on IEMO-CAP and rid real scene interaction database show that the proposed ser framework has better performance and is significantly improved compared with the most advanced methods. For the emotional robot, using these indicators in the simulation and application experiment have a certain application prospect.

In future work, we will continue to study the accuracy of emotion robots in recognizing similar emotions. Since emotions need positive facial expressions[31] and body movements and other information to convey[32-33], using visual information will be the best choice for future research[34], and it will really help to identify human emotions[35]. In the next research, we will further improve the accuracy and performance of emotion recognition, and prepare to add intelligent algorithms such as genetic algorithm and guided search method into the human-computer interaction emotional robot system, so the robot can better accurately perceive human emotions, and humans can communicate with robots more smoothly.

Reference:

[1] Flavia Marino. Outcomes of a Robot-Assisted Social-Emotional Understanding Intervention for Young Children with Autism Spectrum Disorders[J]. Journal of Autism and Developmental Disorders (S0162-3257), 2020,50(7): 1973-1987.

[2] Pejman Sajjadi, Laura Hoffmann. A Personality-Based Emotional Model for Embodied Conversational Agents: Effects on Perceived Social Presence and Game Experience of Users[J]. Entertainment Computing (S2875-9521), 2019, 32: 100313.

[3] Robotics Androids. Differences in Spontaneous Interactions of Autistic Children in an Interaction with an Adult and Humanoid Robot[J]. Frontiers in robotics and AI (S1945-8371), 2020, 7: 28.

[4] Man Hao, Wei-Hua Cao. Visual-Audio Emotion Recognition Based on Multi-Task and Ensemble Learning with Multiple Features[J]. Neurocomputing (S0925-2312), 2020, 391: 42-51.

[5] Rao Y, Wu L W, Wang Y M, et al. Research Progress of Affective Computing Technology Based on Semantic Analysis[J]. Journal of Software (S1000-9825), 2018, 29(8): 2397-2426.

[6] Yang M J, Zhang J B. Speech Emotion Recognition Based on Deep Neural Network[J]. Science and Technology and Engineering (S1671-1815), 2019, 19(8): 127-131.

[7] Zhang L, Lü J, Qiang Y, et al. Speech Emotion Recognition Based on Deep Belief Network[J]. Journal of Taiyuan University of technology (S1007-9432), 2019, 50(1): 101-107.

[8] Akçay M B, Oğuz K. Speech Emotion Recognition: Emotional Models, Databases, Features, Preprocessing Methods, Supporting Modalities, and Classifiers[J]. Speech Communication (S0167-6393), 2020, 116: 56-76.

[9] Tang D, Qin B, Liu T. Deep Learning for Sentiment Analysis: Successful Approaches and Future Challenges[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery (S1942-4787), 2015, 5(6): 292-303.

[10] Lu G M, Yuan L, Yang W J, et al. Speech Emotion Recognition Based on Long-Term Memory and Convolution Neural Network[J]. Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition) (S1673-5439), 2018, 38(5): 63-69.

[11] Sun L H, Fu S, Wang F. Decision Tree SVM Model with Fisher Feature Selection for Speech Emotion Recognition[J]. EURASIP Journal on Audio, Speech, and Music Processing (S1687-4714), 2019(1): 1-14.

[12] Li H H. Research on Speech Emotion Recognition Based on Feature Selection and Optimization[D]. Xi'an: Xi'an University of Electronic Science and Technology, 2019.

[13] Koduru A, Valiveti H B, Budati A K. Feature Extraction Algorithms to Improve the Speech Emotion Recognition Rate[J]. International Journal of Speech Technology (S1381-2416), 2020, 23(1): 45-55.

[14] Jiang P X, Fu H L, Tao H W, et al. A Speech Emotion Recognition Method Based on Convolutional Neural Network Feature Representation[J]. Electronic Devices (S1005-9490), 2019, 42(4): 998-1001.

[15] Miao Y Q, Zou W, Liu T L, et al. Speech Emotion Recognition Based on Parameter Migration and Convolution Recurrent Neural Network[J]. Computer Engineering and Application (S1002-8331), 2019, 55(10): 135-140, 198.

[16] Zeng R H, Zhang S Q. Speech Emotion Recognition Method Based on Improved Convolutional Neural Network[J]. Journal of Applied Sciences (S0255-8297), 2018, 36(5): 837-844.

[17] Tian L L, Zou J Z, Zhang J, et al. Emotion Recognition of EEG Signals Based on Improved Convolutional Neural Network[J]. Computer Engineering and Application (S1002-8331), 2019, 55(22): 99-105.

[18] G Viswanatha Reddy, C V R Dharma Savarni, Snehasis Mukherjee. Facial Expression Recognition in the Wild, by Fusion of Deep Learnt and Hand-Crafted Features[J]. Cognitive Systems Research (S1389-0417) ,2020, 62: 23-34.

[19] Hou P W, Wang Y J. Continuous Emotion Recognition and Tracking Algorithm Based on Random Forest[J]. Science and Technology and Engineering (S1671-1815), 2018, 18(22): 77-83.

[20] Li G L, Tie Y, Qi L. Multi Feature Speech Emotion Recognition Based on Stochastic Forest Classification Optimization[J]. Microelectronics and Computer (S1000-7180), 2019, 36(1): 70-73.

[21] Chen P, Zhang J H, Wen Z Z, et al. EEG Emotion Recognition Based on Nuclear Spectrum Regression and Random Forest[J]. Journal of East China University of science and Technology (Natural Science Edition) (S1006-3080), 2018, 44(5): 744-751.

[22] Zhong Q, Feng Y Q, Wang W. A Comparative Study of Speech Emotion Recognition Based on Cross Language Corpus[J]. Journal of Nanjing University (Natural Science) (S0469-5097), 2019, 55(5): 765-773.

[23] Liu Chang, Zhang Yike, Zhang Pengyuan, et al. Neural Network Language Model Based on Improved Topic Distribution Characteristics[J]. Acta electronics and information (S1009-5896), 2018, 40(1): 219-225.

[24] Jiang Xiaoqing, Tian Lan, Cui Guohui. Prosodic Feature Analysis and Emotion Recognition of Multilingual Emotional Speech[J]. Acta acoustics Sinica (S0371-0025), 2006(3): 217-221.

[25] Zhao J F, Mao X, Chen L J. Speech Emotion Recognition Using Deep 1D & 2D CNN LSTM Networks[J]. Biomedical Signal Processing and Control (S1746-8094), 2019, 47: 312-323.

[26] Bo Hongjian, Ma Lin, Kong Xianghao, et al. Research on Dimension Reduction Method of Speech Emotion Features Based on Convolutional Neural Network Learning[J]. High tech communication (S1002-0470), 2017, 27(S2): 889-898.

[27] Man Hao, Wei-Hua Cao. Visual-Audio Emotion Recognition Based on Multi-Task and Ensemble Learning with Multiple Features[J]. Pattern Analysis and Applications (S0925-2312), 2020, 391: 42-51.

[28] Sun B H, Wang H, Fang B F, et al. Task Allocation of Emotional Robot Pursuit Based on Self-Organizing Algorithm[J]. Robot (S1002-0446), 2017, 39(5): 680-687.

[29] Ren Hao, ye Liang, Li Yue, et al. Speech Emotion Recognition Algorithm Based on Multilevel SVM Classification[J]. Computer application research (S1001-3695), 2017, 34(6): 1682-1684.

[30] Han Wenjing, Li Haifeng, Ruan Huabin, et al. Review of Speech Emotion Recognition Research[J]. Acta software Sinica (S1000-9825), 2014,25(1): 37-50.

[31] Li Aijun, Shao Pengfei, Dang Jianwu. Cross Cultural Multimodal Perception of Emotional Expression[J]. Journal of Tsinghua University (NATURAL SCIENCE EDITION) (S1000-0054), 2009, 49(S1): 1393-1401.

[32] Zhang Jian, Qu Dan, Li Zhen. Cyclic Neural Network Language Model Based on Word Vector Feature[J]. Pattern recognition and artificial intelligence (S1003-6059), 2015, 28(4): 299-305.

[33] Zhang Ziheng, Sun Ying, Yao Hui. Research on Nonlinear Characteristics of Emotional Speech Based on Chaotic Characteristics[J]. Microelectronics and computer (S0469-5097), 2017, 34(4): 65-68, 75.

[34] Liang Ruiyu, Zhao Li, Tao Huawei, et al. Speech Emotion Recognition Algorithm Based on Selective Attention Mechanism[J]. Acta acoustics (S0371-0025), 2016, 41(4): 537-544.

[35] Ke X X, Yun Y W, Qiu B R. Emotional Robot Expression Research with Different Personalities[J]. Journal of Intelligent Systems (S1673-4785), 2017, 12(4): 468-474.