Speech Interactive Emotion Recognition System Based on Random Forest

Susu Yan[1,2], Liang Ye [1,3,4,*], Shuai Han[1,*], Tian Han[2,3], Yue Li[5] and Esko Alasaarela[3]

[1] Department of Information and Communication Engineering, Harbin Institute of Technology, Harbin 150080, China
[2] School of Software and Micro Electronics, Harbin University of Science and Technology, Harbin 150080, China
[3] Health and Wellness Measurement research group, OPEM unit, University of Oulu, Oulu 90014, Finland
[4] Key Laboratory of Police Wireless Digital Communication, Ministry of Public Security, P.R.C., Harbin 150080, China
[5] Electrical Engineering School, Heilongjiang University, Harbin 150080, China
Email:{yeliang@hit.edu.cn, hanshuai@hit.edu.cn}

***Abstract*—In daily life, speech is the main medium of human communication, and interpersonal communication is emotional. People hope that the computer can give a response based on the emotions contained in the voice. In this paper, we build a Wechat program of speech emotion recognition system, which is based on a random forest classifier. Firstly, the system preprocesses the collected speech signals in order to reduce noise. Secondly, 16 acoustic features are extracted from the pre-processed speech signals. The system obtains the emotional features of speech by applying 12 statistical functions to the original acoustic features. The emotional classification of Berlin Speech Emotion Database uses two classifiers: the Random Forest Classifier and the Support Vector Machine. The recognition accuracy of the SVM classifier is 83%. The accuracy of the random forest classifier is 89%. Finally, the random forest classifier is used to build the speech emotion recognition system.**

***Index Terms*—Machine Learning, Speech Emotion Recognition, Random Forest, Wechat Program.**

I. INTRODUCTION

Speech includes two types of information, one is text information, and the other is emotional information. In order to have a harmonious human-computer interaction experience, we hope that the computer can realize automatic emotional recognition of speech signals. In the customer service system, the emotions of customer can be judged by voice. In the educational assistant system, it can improve social emotional ability and academic skills of children [1]. Parents and teachers can deal with problems in time. In the driving system, the system can detect emotions through speech. The system will give an early warning when the driver's mood is extremely anxious or angry. This can reduce the probability of traffic accidents. Automatic speech emotion recognition can be widely used in various fields. In order to improve the recognition accuracy, some researchers apply a variety of classification algorithms to the speech emotion system. Albornoz *et al*. [2] described a hierarchical classifier for identifying a speaker. It had a high recognition accuracy without considering the psychological level information. However, it had significant problems in identifying the gender differences of speakers. Mao *et al*. [3] used a convolutional neural network to classify speech emotions, which effectively utilized the classifier's learning ability for features. However, the classification method required many training sequences, which led to high computational complexity. Few information was considered to establish automatic speech emotion recognition systems.

The main purpose of this study is to build a speech interactive emotion recognition system. This system is based on Wechat program. Features are extracted from the collected speech. Then the extracted features are standardized. Finally, the system gives the classification results through speech emotion interaction interface. In this way, the emotion recognition system of speech interaction is established.

II. METHOD

This paper uses the Berlin Emotional Database. The database contains seven emotions. Each speech in the database gives a corresponding emotional label. In this study, we selected 535 emotional speeches to train and test the model. The following table shows the numbers of speeches of the seven emotions. These seven emotions are labeled from 1 to 7 as shown in Table I.

TABLE I. SEVEN EMOTIONAL LABELS AND NUMBERS OF SPEECH SAMPLES.

Label	1	2	3	4	5	6	7
Emotion	happy	anger	sad	fear	boredom	disgust	neutral
Number	71	128	62	69	81	45	79

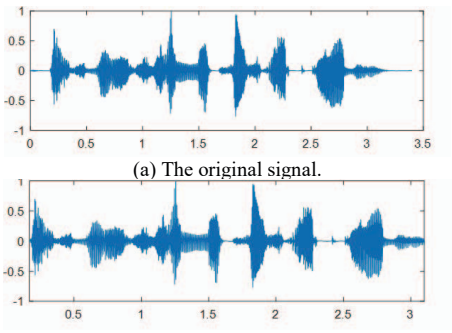
A. Speech Signal Preprocessing

Firstly, Speech signal must be preprocessed. The pre-processing is divided into five steps: sample quantization, pre-emphasis, framing, windowing and endpoint detection.

Fig. 1. Flow chart of preprocessing.

Preprocessing is the basic operation of speech signals. According to the sampling theorem, the speech signal is sampled and quantified. The signal is pre-emphasized after sampling and quantization. The main purpose is to accentuate the high frequency part of the speech signal. Pre-aggravation

can remove the effect of lip radiation. In this paper, we use the first-order FIR high-pass filter to compensate the high-frequency part of the signal. The window function is the Hamming window. The Hamming window can represent the frequency characteristics of speech signals. Finally, the beginning and ending points of speech are detected by the spectral entropy method. The following figure is a comparison of the effect of endpoint detection for a voice in the Berlin Emotional Database.



(a) The original signal.

(b) The signal after endpoint detection.

Fig. 2. Comparison between signals before and after endpoint detection.

B. *Speech Signal Feature Extraction*

In this research, we use "OpenSMILE" software for feature extraction. "OpenSMILE" software is a modular feature extraction tool. It is widely used in signal processing and machine learning. We extract the features of speech signals and calculate the statistics of the function using "OpenSMILE" software. We used 16 original acoustic features shown in Table II. We labeled the corresponding acoustic features from F1 to F16. After extracting the original values frame by frame, the original contour is smoothed.

TABLE II. ORIGINAL FEATURES.

Original feature contour	
F1	RMS energy root-mean-square signal frame energy
F2-F13	Mel-frequency cepstrum coefficients 1-12
F14	Zero-crossing rate of time-domain signals (frame-based)
F15	The voicing probability computed from the ACF
F16	The fundamental frequency computed from the cepstrum

TABLE III. STATISTIC FUNCTION.

Statistic function	
S1	The maximum value of the contour
S2	The minimum value of the contour
S3	Range (max-min)
S4	The absolute position of the maximum value (in frames)
S5	The absolute position of the minimum value (in frames)
S6	The arithmetic mean of the contour
S7	The slope (m) of a linear approximation of the contour
S8	The offset (t) of a linear approximation of the contour
S9	The quadratic error computed as the difference of the linear approximation and the actual contour
S10	The standard deviation of the values in the contour
S11	The skewness (3rd order moment)
S12	The kurtosis (4th order moment)

In this research, the following statistical functions are used as shown in Table III, and they are marked from S1 to S12 in the following content.

Fig. 3 shows the detailed process of feature processing. Firstly, the original feature contours (F1-F16) is smoothed by two methods. Then, 12 statistical functions are applied to the smoothed contour. The 192 features obtained by applying the smoothing method "SMA" are numbered F1-F192. The final features obtained by the smoothing method "de" are numbered F193-F384. A total of 384 features were extracted by smoothing the features using two methods. With this numbering method, we can effectively find out the source of the final feature and facilitate the follow-up work.

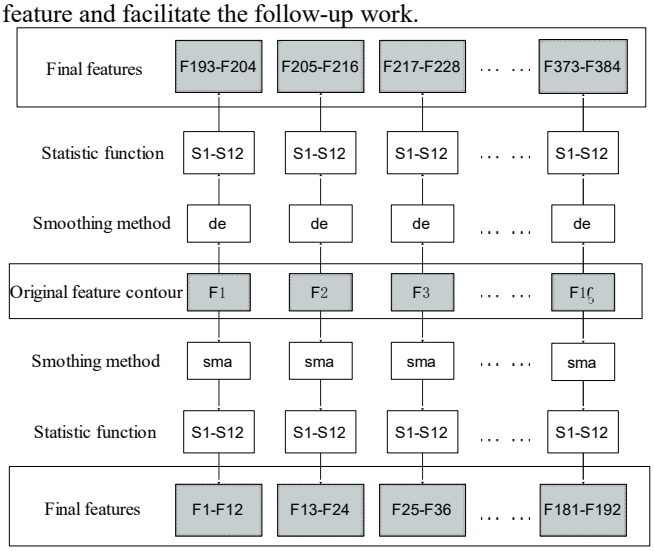


Fig. 3. Detailed feature processing schematic diagram.

In this research, the Berlin Emotional Database is selected for emotion recognition. The Berlin Emotional Database has 535 voices. 384 features are extracted from each speech. The final voice emotion data set consists of 535 rows and 384 columns.

III. SPEECH SIGNAL PREPROCESSING AND FEATURE EXTRACTION

At present, the widely used classification models include the logistic regression algorithm, the random forest algorithm, the Xgboost algorithm and the support vector machine (SVM) algorithm. We use a variety of classification models to classify speech emotion.

TABLE IV. CLASSIFICATION RESULTS OF MULTIPLE CLASSIFIERS

Classifier	Logistic regression	Support vector machine	Random forest	Xgboost
Accuracy	52%	70%	71%	58%

Table IV shows that the classification results of the support vector machine and the random forest are relatively good. Next, we adjust the parameters of the two algorithms respectively. By comparing the classification results, we choose a good emotion classification algorithm to construct the speech emotion recognition system.

A. Support Vector Machine

SVM can achieve good results in the classification of nonlinear high-dimensional data. Therefore, it is applicable to the problems studied in this research. SVM continuously improves the generalization ability of the model by seeking the method of risk minimization. The optimal segmentation hyperplane is constructed in space to obtain the maximum decision boundary. In this way, the learning machine can get the global optimal solution. The goal of this paper is to find the hyperplane that maximizes the classification interval.

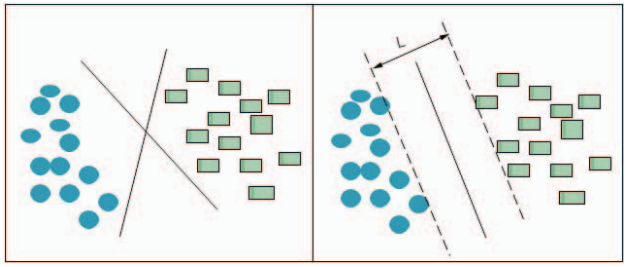


Fig. 4. Diagram of SVM classification model

Blue dots and green squares represent two classes of samples. The solid black line indicates the hyperplane that divides the samples. Fig. 4 represents several possible classification hyperplanes meeting the conditions. The hyperplane with the largest classification interval is selected. The symbol L indicates the maximum classification interval. The parallel dotted lines on both sides represent the closest classification surfaces.

This model is to optimize the objective function. The aim of this model is to find the optimal w and b to maximize the classification interval.

$$\arg\max_{w,b}\{\min(y(w^T x+b))\cdot\frac{1}{\|w\|}\} \quad (1)$$

SVM supports several kernel functions, and in this paper, we use the RBF (Radial Basis Function) kernel function. RBF can map samples to a high-dimensional space. It can be used not only to solve linear inseparable problems, but also to classify multi-class problems. Compared with other kernel functions, RBF has fewer parameters and thus is easier to adjust. If the number of labels in a dataset is n , The number of classifiers is $n(n\text{-}2)/2$ in order to divide these labels. Finally, the label of the sample is determined by counting the votes. The problem of this research is that the classification of seven emotions belongs to multi-classification. In order to satisfy the one-to-one classification scheme, the number of classifiers is 21.

For the RBF kernel function, we need to adjust two parameters, C and *Gamma*. C is used to measure the tolerance of the model to errors. The smaller the value of C, the more likely the model will be under-fitting. Fig. 5 shows the situation of under-fitting. On the contrary, the model with a higher C is prone to over-fitting as shown in Fig. 5. Therefore, an appropriate C value is critical to the performance of the model. *Gamma* is the parameter in the RBF kernel function. *Gamma* determines the number of support vectors in the SVM classification model. If *Gamma* is set too large, the Gauss distribution will be concentrated. There are few support vectors in the model. If *Gamma* is set too small, it will not be able to obtain high accuracy on the training set.

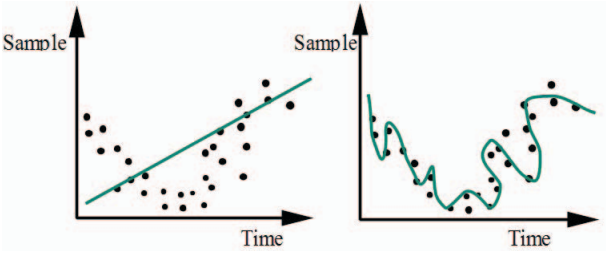


Fig. 5. Diagrams of under-fitting and over-fitting

K-fold cross-validation is usually used for model building and parameter optimization. We use 3-fold cross-validation in this paper. The original data samples are divided into three groups. Two-thirds of the speech samples are used for model training. The testing set consists of the remaining 1/3 speech samples. The testing set is used to evaluate the quality of the model. As shown in Fig. 6, the training set is divided into two parts, one for training the model and the other for validating the model.

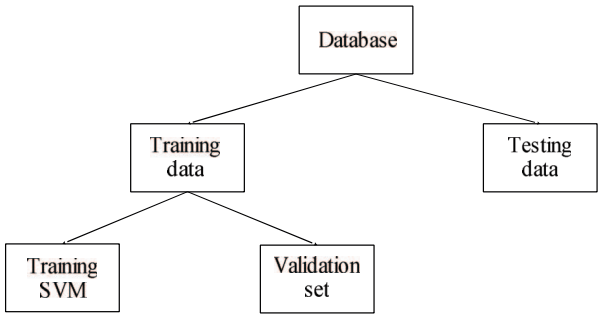


Fig. 6 The data validation process.

The optimized C and *Gamma* values are obtained by parameter optimization, which are 2.06 and 0.002, respectively. Table V shows the recognition accuracies of the seven emotions obtained by the optimized SVM model.

TABLE V. THE RECOGNITION ACCURACIES OF SEVEN EMOTIONS OF THE SVM CLASSIFIER

Emotion	happy	anger	sad	fear	boredom	disgust	neutral
Accuracy	67%	95%	75%	78%	81%	67%	88%

B. Random Forest Algorithm

Random forest (RF) is improved on the basis of decision tree. As shown in Fig. 7, multiple decision trees make up the forest, and each tree in the forest has the same distribution. Moreover, each tree is an independent sample.

The main characteristic of the random forest is double randomness. The first randomness is that sample selection is random. The second randomness is that feature selection is random when training each decision tree model. Because of the double randomness characteristic, the classification algorithm can not only avoid over-fitting, but also has better anti-noise

performance. We optimize the parameters of the random forest classifier. Table 6 shows the recognition accuracies of the seven emotions by the optimized random forest classifier.

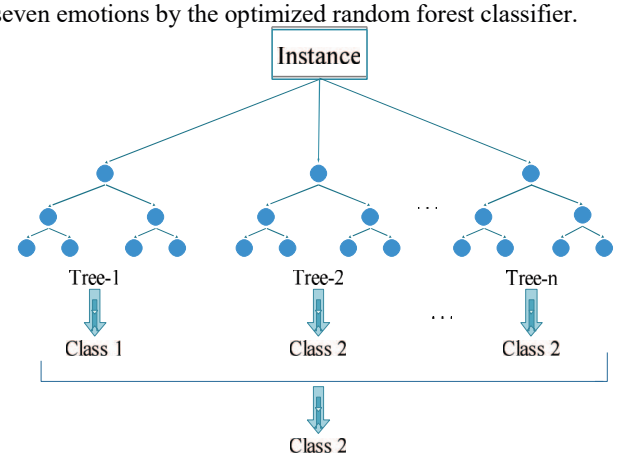


Fig. 7. Simplified graph of random forest classification model.

TABLE VI. THE RECOGNITION ACCURACIES OF SEVEN EMOTIONS OF THE RF CLASSIFIER

Emotion	happy	anger	sad	fear	boredom	digust	neutral
Accuracy	86%	100%	79%	78%	98%	73%	92%

C. Evaluation Index of Classification Model

For the two-category problem, we will divide the samples into positive and negative categories. The positive category is the category studied in the classification problem.

TABLE VII. CONFUSION MATRIX LIST

Confusion Matrix		True value	
		Positive	Negative
Predicted value	Positive	TP	FP
	Negative	FN	TN

The meanings of the symbols in Table VII are given as follows:

TP: The number of samples of the positive class (true value) predicted to be positive.

FP: The number of samples of the negative class (true value) predicted to be positive.

FN: The number of samples of the positive classes (true value) predicted to be negative.

TN: The number of samples of the negative classes (true value) predicted to be negative.

Four performance evaluation indexes derived from the confusion matrix are given as follows.

$$accuracy = \frac{TP + TN}{P + N} \quad (2) \qquad precision = \frac{TP}{TP + FP} \quad (3)$$

$$recall = \frac{TP}{TP + FN} \quad (4) \qquad F_1 = 2 * \frac{precision \times recall}{precision + recall} \quad (5)$$

Table VIII shows the values of four performance evaluation indexes of SVM. The average values of the four indicators are all above 80%, so SVM is effective in classifying the speech emotion data set.

TABLE VIII. PERFORMANCE EVALUATION MATRIX OF SVM CLASSIFIER

Label	1	2	3	4	5	6	7	Mean
Precision	82%	82%	86%	82%	71%	83%	88%	82%
Recall	67%	95%	75%	78%	81%	67%	88%	81%
F_1	74%	88%	80%	80%	76%	74%	88%	81%
Accuracy				83%				

Table IX shows the confusion matrix of the RF classifier. The values on the diagonal of the confusion matrix represent the correctly classified speech samples.

TABLE IX. CONFUSION MATRIX OF RF CLASSIFIER

Confusion matrix		True Emotional Label						
		1	2	3	4	5	6	7
Predicted Emotional Labels	1	18	0	0	0	0	0	0
	2	1	42	0	0	0	2	2
	3	0	0	17	0	2	1	0
	4	2	0	0	20	0	2	0
	5	0	0	7	1	25	0	2
	6	0	0	0	1	0	10	0
	7	0	0	0	1	0	0	23

We can conclude that most samples are correctly classified. As can be seen from the confusion matrix, the largest number of false predictions is label 3 (sad), which is predicted to be label 5 (boredom). The emotion of label 2 (anger) has the least number of prediction errors.

We also obtain the RF classifier performance evaluation matrix. Table X shows the values of four performance evaluation indexes of the RF classifier. The average values of the four indicators are all above 89%. Obviously, the RF classifier is more effective than SVM. Therefore, we apply the random forest classification algorithm to the speech interaction emotion recognition system.

TABLE X. PERFORMANCE EVALUATION MATRIX OF RF CLASSIFIER

Label	1	2	3	4	5	6	7	Mean
Precision	100%	89%	86%	82%	81%	92%	96%	89%
Recall	67%	100%	79%	78%	96%	73%	92%	89%
F_1	74%	94%	83%	80%	88%	81%	94%	89%
Accuracy				89%				

IV. SPEECH EMOTION RECOGNITION SYSTEM

We establish the speech interactive emotion recognition system based on random forest. Fig. 8 shows the basic framework of speech emotion recognition system.

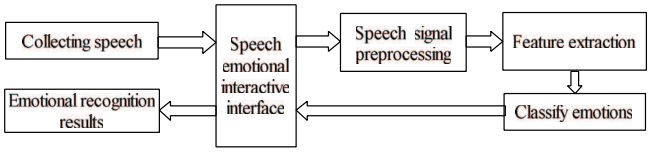
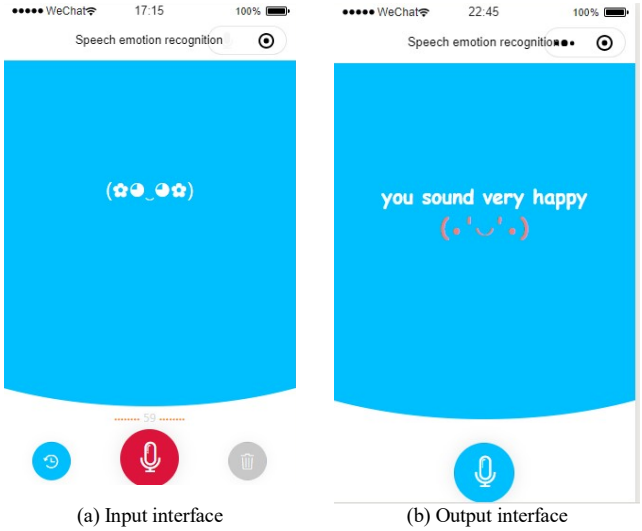


Fig. 8. The basic framework of the speech emotion recognition system

Firstly, we collect speech signals through speech acquisition interface. Then, end-point detection is performed on the speech data. Acoustic features related to emotional states are extracted by openSMILE software, and applied to statistical functions.

The final feature is input into the classification model, and the classifier gives the emotional category of speech.

The speaker collects speech through the speech input interface of the Wechat program. Fig. 9(a) shows the speech input interface. The speaker begins to collect speech by pressing the acquisition button of the input interface. Fig. 9(b) shows the interface of speech acquisition. The system can collect speech signals less than 60 seconds.



(a) Input interface (b) Output interface

Fig. 9. Speech interface of the speech interactive emotion recognition system

The speaker releases the button to complete the signal acquisition. At the same time, the collected speech samples are sent to the background. The speech signal collected by Wechat program is in MP3 format. However, openSMILE software cannot process speech signals of MP3 format. We need to convert MP3 format to wav format. After the transformation of speech signal format, the system extracts and chooses speech signal features, and inputs them into the random forest classification model. Background will return a JSON string. One kind of the JSON string is error, used to indicate possible errors, and the other is the category of emotions. The emotional recognition result of the speech is presented to the user through the interactive platform.

We show the recognition results of the speech emotion recognition model by using the WeChat program. Compared with the traditional client software framework, WeChat program is easy to use, easy to get, and easy to design. Therefore, WeChat program is more suitable for the establishment of speech interaction emotion recognition system.

V. Discussion and Conclusion

Emotional recognition system for voice interaction is an important part of the ideal goal of human-computer interaction. In order to realize emotional recognition of speech, we select 535 voice data from the Berlin Voice Emotion Database to train the model. We extract a total of 384 features. The final voice emotion data set consists of 535 rows and 384 columns. After standardizing the data, we use logistic regression, support vector machine, random forest and Xgboost classification algorithms to classify emotions. We use cross validation to evaluate the model. By comparing the recognition accuracies, we choose support vector machine and random forest classification algorithms which outperform the other two. The recognition accuracies of SVM and RF reach 83% and 89%, respectively by means of model optimization. The evaluation indexes of classification models include accuracy, precision, recall and F_1 score. We find that the RF classifier is better than the SVM classifier by comparing the four indicators. We use the WeChat program to build the speech interactive emotion recognition platform. The user collects the speech signals through the speech acquisition interface. Finally, the classification results are presented through the speech emotion interaction interface.

Acknowledgment

This work was supported by the National Key R&D Program of China (No.2018YFC0807101). The authors would like to thank all the people who participated in the project.

References

[1] Z. A. Khan, and W. Sohn, "Abnormal human activity recognition system based on R-transform and kernel discriminant technique for elderly home care," IEEE Trans. Consumer Electronics, vol.57, No.4, pp. 1843-1850, 2011.

[2] M. Basten, H. Tiemeier, and Althoff RRs, "The stability of problem behavior across the preschool years: an empirical approach in the general population," Abnorm Child Psychol, vol.44, no.2, pp. 393-404, 2016.

[3] E. Albornoz, H. Milone, and L. Rufiner, "Spoken emotion recognition using hierarchical classifiers," Computer Speech & Language, vol.25, no.3, pp. 556-570, 2011.

[4] Q. M., "Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks," IEEE Transactions on Multimedia, vol.16, no. 8, pp. 2203-2213, 2014.

[5] Mencattini, and Aet, "Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure," Knowledge-Based Systems, vol.63, no.2, pp. 68-81, 2014.

[6] H. Palo, N. Mohanty, "Wavelet based feature combination for recognition of emotions," Ain Shams Engineering Journal, vol.9, no.4, pp. 1799-1806, 2018.

[7] S. Shahnawazuddin. "Studying the role of pitch-adaptive spectral estimation and speaking-rate normalization in automatic speech recognition," Digital Signal Processing, vol.79, no.3, pp. 142-151, 2018.

[8] T. Han, J.C. Zhang, Z. Zhu, *et al*, "Emotion recognition and school violence detection from children speech," EURASIP Journal on Wireless Communications and Networking, vol.2018, no.1, pp. 235, 2018.

[9] J. Lorenzo-Trueba, "Emotion transplantation through adaptation in HMM-based speech synthesis," Computer Speech & Language, vol.34, no.1, pp. 292-307, 2015.

[10] Z. Liu, "Speech emotion recognition based on feature selection and extreme learning machine decision tree," Neurocomputing, vol.273, no.8, pp. 271-280, 2018.