Speech Emotion Recognition Based on SVM and ANN

Xianxin Ke, Yujiao Zhu, Lei Wen, and Wenzhen Zhang

***Abstract*—Speech emotion recognition mainly includes emotion feature extraction, feature reduction and speech emotion recognition model. This paper chooses valid emotional features and extracts the statistical values of the emotional features. Speech emotion recognition model are constructed respectively based on SVM and ANN and the recognition effect of feature reduction respectively on two types of models are compared. The experimental results show that, based on emotion features which is extracted by CASIA emotion corpus, feature reduction can improve recognition accuracy and the recognition effect of speech recognition model based on SVM is better than ANN.**

***Index Terms*—SVM, ANN, speech emotion recognition, feature reduction.**

I. Introduction

Speech emotion recognition research involves the traditional speech signal processing, pattern recognition, human psychology, artificial intelligence, human-computer interaction and other fields. In human-computer interaction, speech emotion recognition is an important part to determining the emotional state of interaction objects.

The study of speech emotion recognition can be traced back to the early 1980s. Emotion classification can use the acoustic statistics. There are more and more studies of speech emotion recognition with the growing understanding of emotional intelligence. Schuller [1], [2] at Technical University of Munich were conducted a research on speech emotion recognition. A lot of studies on speech emotion reconducted by the Voice and Emotional Group at the University of Southern California and the Emotion Research Laboratory at Universit é de Gen ève. Technical University of Munich developed the openSMILE [3]-[5] speech feature extractor, which can automatically extract speech emotion features in batches, and carried out various classification methods of speech emotion recognition. Tsinghua University [6], CASIA [7], Zhejiang University and Southeast University also achieved outstanding results.

Based on the CASIA Chinese Emotional Corpus, this paper analyzied key technologies in the procession of speech emotion recognition and the effect of feature reduction on the speech emotion recognition. We compared the emotional classification effect of speech recognition model based on SVM and ANN.

II. Speech Emotion Recognition

The premise of speech emotion recognition is the description of emotion. At present, there are two kinds of emotion descriptions: discrete emotion classification system and dimension emotion classification system. Scholars represented by Rene Descartes, Paul Ekman and Silvan Tomkins divide emotions into several basic emotions. The six emotion theory proposed by Paul Ekman is the most widely used, including happy, sad, angry, fear, surprise and disgusted. Scholars represented by Wilhelm Wundt, James Russell and Lisa Feldman Barrett think that the emotional state can be described as a point in multidimensional emotional space. One of the most widely used model is the Arousal-Valence Model.

The speech emotion recognition in this paper is based on the discrete emotion classification system. The structure of speech emotion recognition is shown in Fig. 1, which mainly includes: emotion feature extraction, feature reduction and speech emotion recognition model [8]-[10]. At first, the emotion features of each audio in the emotion corpus are extracted. Secondly, whether the feature dimension reduction is needed or not should be judged. Then speech emotion recognition model is constructed by using feature values.

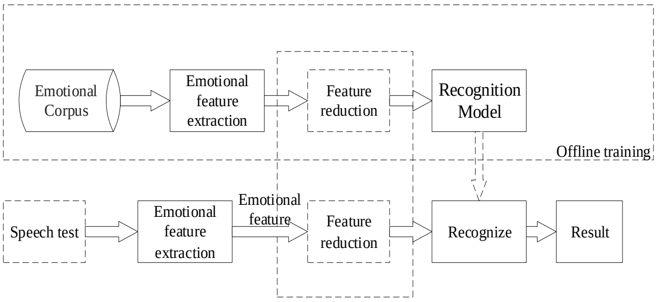

Fig. 1. The Process of speech emotion recognition.

III. Speech Feature Extraction and Reduction

Emotional features in speech signals [11] mainly include prosodic features, spectral-based correlation features and sound quality characteristics. Prosodic features refer to changes of speech speed, volume, pitch and sound wavelength. Common prosodic features include duration, fundamental frequency, energy and so on. Spectral-based correlation features reflect the correlation between the change of channel shape and the actions of speech. The main features of the linear spectrum are LPC, OSALPC, LFPC *et al.* The cepstrum features include LPCC, OSALPCC, MFCC et al. The sound quality features are to measure the speech intelligibility and the difficulty of distinguish. Sound quality features we commonly used are formant frequency, formant bandwidth and glottis parameters.

Manuscript received April 14, 2018; revised June 6, 2018. This work was supported by National Key Scientific Instrument and Equipment Development Projects of China (2012YQ150087), National Natural Science Foundation of China (61273325).

Xianxin Ke, Yujiao Zhu, Lei Wen, and Wenzhen Zhang are with School of Mechatronic Engineering and Automation, Shanghai University, China (e-mail: xxke@staff.shu.edu.cn, 1449803020@qq.com, 405969292@qq.com, 1299896265@qq.com).

doi: 10.18178/ijmlc.2018.8.3.687

In this paper, the acoustic parameters are extracted from the preprocessed speech signal, which includs the root mean square of energy, zero crossing rate, fundamental frequency, sounding probability, MFCC, the first, second and third formant frequency and bandwidth. Dynamic parameters of speech signals are extracted from first-order difference of those corresponding feature values. The statistical characteristics of acoustic parameters and dynamic parameters are adopted as the eigenvectors of speech emotion recognition. The features as shown in Table I, with a total of 382-dimensional features:

TABLE I: EMOTIONAL FEATURES

acoustic parameters	statistical features
the root mean square of energy, zero crossing rate, pronunciation probability, fundamental frequency, 12th MFCC and corresponding first-order difference	range, maximum, minimum, maximum frame, minimum frame, mean, slope and offset of linear approximation, standard deviation, Skewness, kurtosis
first, second, third formant frequency and bandwidth	range, maximum, average, slope and offset of linear approximation

Feature reduction can reduce the dimensionality of speech emotional features. It is one of the key technologies in pattern recognition. There are two ways to reduce the dimension of features: feature selection and feature extraction. Feature selection does not change the original eigenvalue but select a valid subset of features through removing the irrelevant or redundant features from the original set. Feature extraction maps high-dimensional features to low-dimensional space.

Common algorithms include Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). PCA looks for a set of orthonormal base as the main component and uses the principal component to describe the original sample. PCA simplifies the original sample data, which can eliminate the correlation between the sample data and remove the interference of noise and excess features. The application of PCA do not have parameter restrictions, which make PCA to have a very wide range of applications. This paper chooses PCA to reduce the dimension of features.

IV. SPEECH EMOTION RECOGNITION MODEL

The recognition model is a key part of speech emotion recognition. CASIA Chinese Emotional Corpus is one of discrete speech emotional corpus. For discrete emotional recognition, emotional recognition can be modeled into a pattern classification problem. In this paper, the classifiers are constructed respectively based on SVM and ANN.

A. SVM-Based Speech Emotion Recognition

Support Vector Machine (SVM) is a dichotomous model. However, CASIA Chinese Emotional Corpus have 6 emotion types. For multi-classification problems, there are generally three ways to construct a classifier, include One-to-many algorithm, one-to-one algorithm and hierarchical support vector machine.

One-to-many algorithm is: one type is classified as a positive set and the remaining categories are classified as negative set. The samples of negative set are much larger than positive set in this method. The training results will be offset, and the practicality is not high.

One-to-one algorithm is: for the k-type problem, *k* (*k*-1)/2 classifiers will be constructed between any two classes. To classify an unknown data, the data will be judged through the *k* (*k*-1)/2 classifiers and the result is the highest number of occurrences. LIBSVM [12] toolkit developed by Professor Lin Zhiren adopts this classification method.

Hierarchical Support Vector Machine [13]-[15] is: At first, all types are divided into two sub-categories. Then the sub-category will be divided into two sub-categories, until a separate category is achieved. Dividing sub-categories can be based on the degree of confusion between classes.

Mix_{ij} is the confusion degree between the G_i type the G_j type, which is the average between the probability of misjudging the *i* type data as the *j* type data and the probability of misjudging the *j* type data as the *i* type data value.

$$Mix_{ij} = \frac{(P(r=i \mid x \in G_j) + P(r=j \mid x \in G_i))}{2} \qquad (1)$$

The higher confusion value, the harder to differentiate between type *i* and type *j*. When determining the first level of classifier, if the confusion value between two types is bigger than 0.1, the two types will be divided into a sub-category. If the confusion values between one type and other types are always less than 0.02, the type will be a sub-category.

When confusion values between one type and other types are between 0.02 and 0.1, the classification of this type cannot be directly judged. By calculating the total confusion values between this type and a sub-category, the situation with the highest confusion value will be selected.

$$T(a,B) = \sum_{i=1}^{i=m} Mix_{ai} \qquad (2)$$

Tips: *a* represents a type; *B* is a definite sub-category and sub-category *B* includes *m* emotional categories; *T* (*a*, *B*) is the total confusion between type *a* and sub-category *B*.

B. ANN-Based Speech Emotion Recognition

Artificial Neural Network (ANN) [16] is a kind of pattern matching algorithm, which is usually used to solve classification and regression problems. In this paper, neural network model with two hidden layers is trained by error Back Propagation (BP) algorithm to realize speech emotion recognition.

The input of neural network are speech emotional feature values. The activation function of hidden layers neuron is the linear correction unit ReLU function. When a large gradient flows through the ReLU neurons, the weights become smaller due to the large gradient and gradually to 0, so that the neurons have the activating effect no longer. We avoid the problem as much as possible by setting the learning rate to 0.001.

The output categories of speech emotion recognition are changed into vector by one-hot encoding. For n-type problems, a single category output is transformed into an n-dimensional vector output, with each dimension in the output vector representing a category. In the training samples, the category corresponding dimension is set 1, other dimensions is set 0. The cross entropy is used to calculate the distance between the actual prediction vector and the target vector. BP algorithm is used to reduce the cross entropy. The activation function of the output neurons is the Softmax function so that the sum of each unit output is equal to 1. The category corresponding to the highest probability in the output neuron is the final recognition result.

V. Speech Emotion Recognition Experiment

The CASIA Chinese Emotional Corpus, recoded by institute of automation Chinese academy of sciences, consists of 9,600 audios including 6 kinds of emotion (neutral, angry, fear, happy, sad and surprise). The corpus is made up of 300 identical text and 100 different texts. This paper uses 1200 audios from the CASIA Chinese Emotional Corpus. And 20 audio samples are randomly selected from each emotion to make up a total of 120 test data, the remaining 1080 audios are used as training data.

A. Speech Emotion Recognition Experiment Based on SVM

Speech emotion recognition classifier is constructed by one-to-one method and based on different kernel functions. The recognition accuracy is shown in Table II. The accuracy without feature reduction is low.

TABLE II: SVM Classifier Recognition Rate without Feature Reduction

kernel	LINEAR	POLY	POLY	POLY
functions		(2)	(3)	(4)
accuracy	43.33	46.67	47.5	45.83
（%）				

We use PCA to reduce the dimension of the feature values. Four dimensional gradients are set: 85%, 90%, 95% and 96%, and the corresponding feature dimensions respectively are 78,104,144 and 157. And we use this to make the comparison of linear, polynomial, RBF and SIGMOID kernel functions based on the classification results. The results are shown in the Table III.

TABLE III: Different Dimensions and Different Kernel Function Classification Accuracy Rate (%)

dimensions	78	104	144	157
LINEAR	65	63.3333	71.6667	69.1667
POLY *d*=2	68.3333	70	69.1667	70.8333
POLY *d*=3	75.8333	75.8333	76.6667	76.6667
POLY *d*=4	70	68.3333	65	62.5
RBF			47.5	
SIGMOID			23.3333	

The results show that the recognition effect of classifier trained by 95% contribution, 144-dimensional is the best. By comparing the recognition rates of different kernel functions under the 95% contribution, we find that the classification results of RBF and SIGMOID kernel functions are not as good as the linear and polynomial kernel functions in classifying high dimensional features.

From the experimental results, we can see that we get the best accuracy which is 95% and 96% when the third polynomial is applied. The lower the dimension, the faster the calculation. Therefore, the dimension corresponding to the 95% contribution is selected and the polynomial kernel function is 3 Poly.

In this paper, we construct hierarchical support vector machine to compare the effect of different SVM multi-classification strategies on recognition effect. The confusion degree of each emotion category is calculated, and the hierarchy support vector machine model is constructed by comparing confusion values between emotional types. Emotion recognition results is shown in Table IV when the dimension is 144 and the degree of Poly kernel function is 3.

TABLE IV: Emotion Status Identification Results Statistics

	neutral	happy	surprise	fear	sad	angry	accuracy(%)
neutral	18	2					90
happy	2	13	1	2		2	65
surprised	1	2	14	2		1	70
fear			2	15	3		75
sad				5	15		75
angry			1		2	17	85

TABLE V: Confusion between the Various Categories

	neutral	happy	surprise	fear	sad
happy	0.1				
surprise	0.025	0.075			
fear	0	0.05	0.1		
sad	0	0	0	0.2	
angry	0	0.05	0.05	0	0.05

According to the Table IV, we calculated the degree of confusion between the various emotional types. The results are shown in Table V.

The table shows that: the confusion between happy and neutral is 0.1, which are not easy to distinguish; a confusion between surprise and fear, fear and sad are 0.1 and 0.2 respectively, and the three categories are indistinguishable. Therefore, when the first-level classifier is created, classify happy and peaceful as sub-category A, and surprise, fear and sadness as sub-category B.

The degree of confusion between angry and A is T (angry, A) = 0.05, while T (angry, B) = 0.167. The degree of confusion between angry and B is higher. So angry is divided into B group. In group B, referring to the degree of confusion among the categories of emotions, angry as a group, surprise, fear and sad as another group. Hierarchical support vector machine structure is shown in Fig. 2.

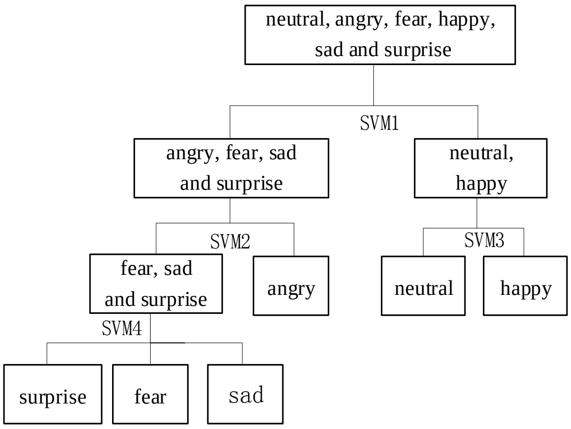

Fig. 2. Hierarchical support vector machine model.

Through the classification of test data, the recognition rate of each classifier and overall recognition rate are shown in Table VI.

TABLE VI: CLASSIFIERS AND LEVELS SVM RECOGNITION RATE

Classifiers	1	2	3	4	Hierarchical SVM
Accuracy (%)	85.83	90	85	73.33	66.67

Through analyzing all the experiments, we can draw the conclusion that based on CASIA Chinese Emotional Corpus, the multi-classification SVM model established by the one-to-one method and 3 poly kernel function is the best one in high-dimensional features.

B. Speech Emotion Recognition Experiment Based on ANN

The neural network-based classifier uses the same dataset as SVM-based. We take 120 sets of data as a test set, 1080 sets of data as a training set. 0.05% of the training set as a verification set.

In this paper, the different number of neurons in first and second neural networks are tested by grid search method. The number of neurons in the first layer are set to 300, 600, 900 and 1200 respectively. The number of neurons in the second layer are set to 32, 50, 100 and 200, respectively. The experimental results show that without feature dimension reduction, when first layer neurons are 900 and the second layer neurons are 50, we get a pretty good accuracy of 45.83%.

In the case of dimensionality reduction, the input data selects the 144 features dimension corresponding to the 95% contribution. Due to the decrease of input data dimension, the number of neurons in the first layer are set to 300, 600 and 900 respectively, while the number of neurons in the second layer are set to 32, 50 and 100, respectively. Among them, when the first layer neurons are 600 and the second layer neurons are 32, the structure of speech emotion recognition network is reasonable, and the accuracy rate is also the highest. The accuracy of test samples is 75%. With the first layer have 600 neurons, the experimental results are as shown in the Table VII.

From the experimental results of the feature-reduced dimension data, we know the second neurons in the range of 32 to 100 have little effect on the final test results. Considering the computational complexity, the number of neurons in the first layer is 600 and the number of neurons in the second layer is 32. The structure of the speech emotion recognition network is reasonable, and the accuracy is the highest.

TABLE VII: ANN EXPERIMENTAL RESULTS

The first layer	600	600	600
The second layer	32	50	100
Training accuracy（%）	99.21	98.97	99.40
Verifying accuracy（%）	74.07	79.63	73.15
Test accuracy（%）	75	73.33	74.17

C. Speech Emotion Recognition Results Comparison

This paper establishes two speech emotion recognition classifiers based on SVM and ANN respectively, and compares the effect of dimensionality reduction of the original data on the two classifiers respectively. The comparison results are shown in Fig. 3. The speech emotion classifier based on PCA and ANN has obvious improvement in accuracy.

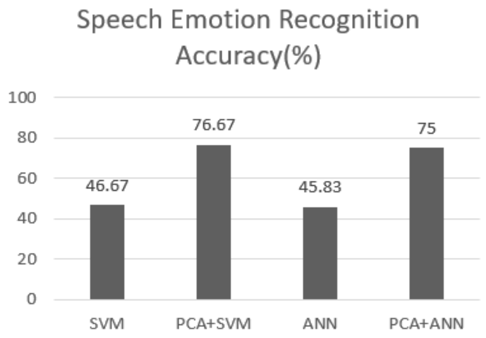

Fig. 3. Classifier comparison results.

From the experimental results, we can see that, the features after PCA are in favor of the improvement of classifier performance. And the speech emotion recognition model based on SVM is slightly better than the ANN-based classifier.

VI. CONCLUSION

Based on the CASIA Chinese Emotional Corpus, the speech emotional features are analyzed and the statistical values of speech emotional features are extracted. Based on SVM and ANN respectively, two speech emotion classification models are constructed. In the two classification models, the effect of feature dimension reduction on accuracy is analyzed, and the two classification methods are compared. The experimental results show that the feature dimension reduction is helpful to improve the classification performance. In this paper, the performance of SVM in speech emotion recognition is slightly better than ANN.

In the future research, we need to expand the corpus to carry out research and analyze the distinction of speech emotion features and the effect of different feature reduction methods on speech emotion recognition to improve the accuracy of speech emotion recognition.

ACKNOWLEDGMENT

This research is supported by National Key Scientific Instrument and Equipment Development Projects of China (2012YQ150087), National Natural Science Foundation of China (61273325).

REFERENCES

[1] B. Schuller, G. Rigoll, M. Lang, "Hidden Markov model-based speech emotion recognition," in *Proc. 2003 IEEE International Conference on Acoustics, Speech, & Signal Processing*, 2003, pp. 401-404.

[2] S. Björn and R. Gerhard, "Timing levels in segment-based speech emotion recognition," in *Proc. Interspeech 2006 and 9th International Conference on Spoken Language Processing*, 2006, vol. 4, no. 1818-1821.

[3] F. Eyben, R. S. Klaus, W. S. Bjorn *et al*., "The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190-202, 2016.

[4] F. Eyben, G. L. Salomão, J. Sundberg *et al*., "Emotion in the singing voice — a deeperlook at acoustic features in the light ofautomatic classification," *Eurasip Journal on Audio, Speech, and Music Processing*, vol. 19, 2015.

[5] F. Eyben, B. Huber, E. Marchi, D. Schuller, and B. Schuller, "Real-time Robust Recognition of Speakers' Emotions and Characteristics on Mobile Platforms," in *Proc. International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015, pp. 778-780.

[6] D. N. Jiang and L. H. Cai, "Speech emotion recognition using acoustic features," *J Tsinghua Univ (Sci& Tech)*, pp. 86-89, 2006.

[7] L. X. Huang, L. Xin, L. Y. Zhao, and J. H. Tao, "Bimodal emotion recognition based on adaptive weights," *J Tsinghua Univ (Sci& Tech)*, vol. 48, pp. 715-719, 2008.

[8] W. Q. Zheng, J. S. Yu, and Y. X. Zou, "An experimental study of speech emotion recognition based on deep convolutional neural networks," in *Proc. 2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015, pp. 827-831.

[9] P. Song, W. M. Zheng *et al*., "Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization," *Speech Communication*, vol. 83, pp. 34-41, 2016.

[10] E. Reda and A. Masato, "Optimizing fuzzy inference systems for improving speech emotion recognition," *Advances in Intelligent Systems and Computing*, vol. 533, pp. 85-95, 2017.

[11] W.-J. Han, H.-F. Li, H.-B. Ruan, L. Ma, "Review on speech emotion recognition," *Ruan Jian Xue Bao/Journal of Software*, vol. 25, no. 1, pp. 37-50, 2014.

[12] C.-C. Chang, C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[13] S. Besbes and Z. Lachiri, "Multi-class SVM for stressed speech recognition," in *Proc. 2016 2nd International Conference on Advanced Technologies for Signal and Image Processing*, pp. 782-787, 2016.

[14] M. Elleuch and R. Mokni, "Offline arabic handwritten recognition system with dropout applied in deep networks based-SVMs," in *Proc. 2016 International Joint Conference on Neural Networks*, 2016, pp. 3241-3248.

[15] X. J. Wang, *Speech Emotion Recognition Research Based on Feature Selection*, Jiangsu University, 2007

[16] Z. H. Zhou, *Maching Learning*, Beijing: Tsinghua University Press, 2016.

Ke Xianxin was born in Hefei, Anhui Province, China in 1973. In 2015, he received a Ph.D. in mechanical and electronic engineering from Shanghai University. Since 2005, she has been a teacher, associate professor and tutor for graduate students at Shanghai University. He was a short-term student at the University of Manchester, UK from 2014.3-2014.4. He was a domestic visiting scholar at the Ministry of Education in Shanghai Jiaotong University from 2012 to 2013. He was a postdoctoral fellow in the Department of Mechanical and Aeronautical Engineering at George Washington University, USA. His main research direction is AI Robotics.

He has presided over two national-level projects such as the National Natural Science Foundation of China, and the State-level major scientific instrument development special secondary subproject. As a major researcher or moderator, he participated in projects including the National 863 Program, the National Natural Science Foundation of China, the US DARPA Research Project, the Science and Technology Research Project of the Shanghai Science and Technology Commission, and the Shanghai Municipal Education Commission.