Feature Extraction and Selection in Speech Emotion Recognition

Yixiong Pan[1], Peipei Shen[2], Liping Shen[3]
Department of Computer Technology
Shanghai JiaoTong University
Shanghai, China
[1]panyixiong@sjtu.edu.cn, [2]shen@sjtu.edu.cn, [3]lpshsen@sjtu.edu.cn

***Abstract*—**Speech Emotion Recognition (SER) is a hot research topic in the field of Human Computer Interaction (HCI). In this paper, we recognize three emotional states: happy, sad and neutral. The explored features include: energy, pitch, linear predictive spectrum coding (LPCC), Mel-frequency spectrum coefficients (MFCC), and Mel-energy spectrum dynamic coefficients (MEDC). A German Corpus (Berlin Database of Emotional Speech) and our self-built Chinese emotional databases are used for training the Support Vector Machine (SVM) classifier. Finally results for different combination of the features and on different databases are compared and explained. The overall experimental results reveal that the feature combination of MFCC+MEDC+ Energy has the highest accuracy rate on both Chinese emotional database(91.3%) and Berlin emotional database (95.1%).

Keywords- Speech Emotion; Automatic Emotion Recognition; SVM; Energy; Pitch; LPCC; MFCC;MEDC

1. Introduction

The research of automatic speech emotion recognition, not only can promote the further development of computer technology, but also greatly enhance the efficiency of people's work and study, and help people solve their problems more efficiently, as well as further enrich our lives and improve the quality of life.

Researchers have proposed important speech features which contain emotion information, such as energy, pitch frequency [2], formant frequency [3], Linear Prediction Coefficients (LPC), Linear Prediction Cepstrum Coefficients (LPCC), Mel-Frequency Cepstrum Coefficients (MFCC) and its first derivative [4]. Furthermore, many researchers explored several classification methods, such as Neural Networks (NN) [5], Gaussian Mixture Model (GMM), Hidden Markov model (HMM) [6], Maximum Likelihood Bayesian classifier (MLC), Kernel Regression and K-nearest Neighbors (KNN) and Support vector machines (SVM) [7].

In this paper, we compare the recognition rate using energy, pitch, LPCC, MFCC, and MEDC features and their different combination. The paper is organized as following. Section II describes the database used in the experiments. Section III introduces the automatic speech emotion recognition system. The speech features are presented in this section. Section IV introduces the Support Vector Machine algorithm. The performance of different features combination is shown in section V. Section VI concludes this paper.

2. Speech Database

Two emotional speech databases are used in our experiments: Berlin German Database and SJTU Chinese Database. The Berlin database is widely used in emotional speech recognition [7]. It is easily accessible and well annotated. We use happy, sad and neutral emotion in our recognition system. All of the speech samples are simulated by ten professional native German actors (5 actors and 5 actresses). There are totally about 500 speech samples in this database[1].

Nowadays most databases we use are not Chinese, and there is a lack of Chinese database, which makes it difficult to do the emotion recognition research on Chinese speech. So we design and build our own SJTU Chinese speech database, which includes three kinds of emotional states sadness, happiness and neutral. There are totally about 1500 speech samples in this database. We recorded ten people, (5 men and 5 women students), with everyone reading 150 sentences. As the participants are not professional, so the weakness of this database is that some emotions are not so obvious and accurate, which may lead to wrong classification and influence the final recognition accuracy.

I. SPEECH EMOTION RECOGNITION SYSTEM

Our speech emotion recognition system contains four main modules: speech input, feature extraction, SVM based classification, and emotion output(Figure 1).

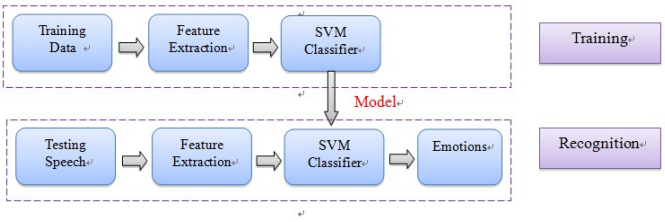

Figure 1. Speech Emotion Recognition System

As the training part to LIBSVM classifier, we need to train a feature model at first with the training voice data. We give each speech sample with the corresponding emotion class label. After that a ".data" file which contains the class label and feature coefficients is put into the LIBSVM [1]classifier, and a ".model" file is gotten. In the recognition part, the input of this system is .wav files, which come from the training database or from the real-time

* Supported by national Natural Science Foundation of China under Grant No. 60873132

speech. The features extracted include: energy, pitch, Linear Prediction Cepstrum Coefficients (LPCC), Mel-Frequency Cepstrum Coefficients (MFCC), and Mel Energy spectrum Dynamic coefficients (MEDC), and some statistical feature calculated. A SVM classifier model is first trained and then used in the speech recognition process. After loading the SVM classifier model, three emotions can be recognized in the recognition stage of the system: happiness, sadness and neutral.

3 Feature Extraction

The speech signal contains a large number of information which reflects the emotional characteristics. So in the research of speech emotion recognition, the most important thing is that how to extract and select better speech features with which most emotions could be recognized. In recent researches, many common features are extracted, such as speech rate, energy, pitch, formant, and some spectrum features, for example Linear Prediction Coefficients (LPC), Linear Prediction Cepstrum Coefficients (LPCC), Mel-Frequency Cepstrum Coefficients (MFCC) and its first derivative.

A. Energy and related features

The Energy is the basic and most important feature in speech signal. We can obtain the statistics of energy in the whole speech sample by calculating the energy, such as mean value, max value, variance, variation range, contour of energy [2].

B. Pitch and related feature

We calculate the value of pitch frequency in each speech frame, and obtain the statistics of pitch in the whole speech sample. These statistical values reflect the global properties of characteristic parameters. Each Pitch feature vector has the same 19 dimensions as energy.

C. Linear Prediction Cepstrum Coefficients (LPCC)

LPCC embodies the characteristics of particular channel of speech, and the same person with different emotional speech will have different channel characteristics, so we can extract these feature coefficients to identify the emotions contained in speech. The computational method of LPCC is usually a recurrence of computing the linear prediction coefficients (LPC), which is according to the all-pole model.

D. Mel-Frequency Cepstrum Coefficients (MFCC)

MFCC is based on the characteristics of the human ear's hearing, which uses a nonlinear frequency unit to simulate the human auditory system. Mel frequency scale is the most widely used feature of the speech, with a simple calculation, good ability of the distinction, anti-noise and other advantages [11]. In our research, we extract the first 12-order of the MFCC coefficients.

The process of calculating MFCC is shown in Figure2.

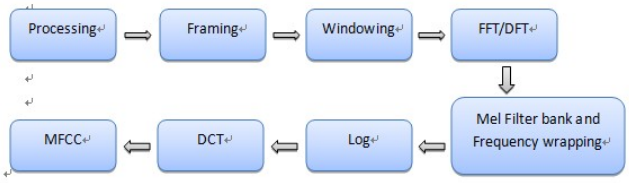

Figure2. Process of Calculating MFCC

E. Mel Energy spectrum Dynamic coefficients (MEDC)

MEDC extraction process is similar with MFCC. The only one difference in extraction process is that the MEDC is taking logarithmic mean of energies after Mel Filterbank and Frequency wrapping, while the MFCC is taking logarithmic after Mel Filterbank and Frequency wrapping. After that, we also compute 1st and 2nd difference about this feature.

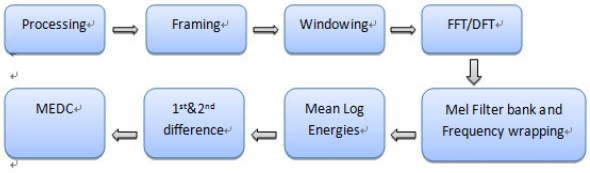

Figure3. MEDC Feature Extraction

4. Experiment and Result

The performance of speech emotion recognition system is influenced by many factors, especially the quality of the speech samples, the features extracted and classification algorithm. This article analyse the system accuracy on the first two aspects with large numbers of tests and experiments.

A. SVM Classification Algorithm

SVM is a simple and efficient computation of machine learning algorithms, and is widely used for pattern recognition and classification problems. Under the conditions of limited training data, it can have a very good classification performance compared to other classifiers [4]. Thus we adopted the support vector machine to classify the speech emotion in this paper.

B. Training Models

Emotion classes sad, happy, neutral are having 62, 71, and 79 speech utterance respectively in Berlin Emotion database. While Our own emotion speech database (SJTU Chinese emotion database) contains 50 speech files for each emotion class. We use both databases, combine different features to build different training models, and analyse their

recognition accuracy. Table1 shows different combination of the features for the experiment.

Training Model	*Combination of Feature Parameters*
Model1	*Energy + Pitch*
Model2	*MFCC + MEDC*
Model3	*MFCC + MEDC + LPCC*
Model4	*MFCC + MEDC + Energy*
Model5	*MFCC + MEDC + Energy + Pitch*

TABLE 1. DIFFERENT COMBINATION OF SPEECH FEATURE PARAMETERS

C. Experimental Results

Table2 shows the models' cross validation rate and recognition rate based on Berlin Emotion database.

Training Model	*Features Combination*	*Cross Validation Rate*	*Recognition Rate*
Model1	*Energy + Pitch*	*66.6667%*	*33.3333%*
Model2	*MFCC + MEDC*	*90.1538%*	*86.6667%*
Model3	*MFCC + MEDC + LPCC*	*72.5275%*	*86.6667%*
Model4	*MFCC + MEDC + Energy*	*95.0549%*	*91.3043%*
Model5	*MFCC + MEDC + Energy + Pitch*	*94.5055%*	*90%*

TABLE 2 THE RECOGNITION RATE AND CROSS VALIDATION BASED ON GERMAN MODEL

Table3 shows the models' cross validation rate and recognition rate based on SJTU Chinese Database.

Training Model	*Features Combination*	*Cross Validation Rate*	*Recognition Rate*
Model2	*MFCC+MEDC*	*88.6168%*	*80.4763%*
Model4	*MFCC + MEDC + Energy*	*95.1852%*	*95.0874%*

TABLE 3 THE RECOGNITION RATE AND CROSS VALIDATION BASED ON MAN MODEL

As is shown at Table2 and Table3, different features combination results in different recognition accuracy rate. To the Berlin Database, the feature combination of Energy and Pitch has the worst recognition rate, which can only recognize one emotional state. That may because these two are simple prosodic features with few numbers of dimensions. The accuracy rate for the feature combination of MFCC and MEDC is higher compared with Model1. It can better recognize three standard emotional states. We also add the LPCC feature, but the performance of the model becomes lower which may result from the feature redundance. The best feature combination is MFCC + MEDC + Energy, for which the cross validation rate can be as high as 95% for nonreal-time recognition. The reason

for this high performance is that it contains prosodic features as well as spectrum features, and the features have excellent emotional characters. For Chinese database, the feature combination of MFCC + MEDC + Energy also shows a good performance there. The cross validation rate is as high as 95%, and the recognition accuracy rate is also around 95%. This combination performs better than that on German database, which means the feature of Energy plays an important role in Chinese speech emotional recognition.

5 Conclusion and Future Works

We can conclude that, different combination of emotional characteristic features can obtain different emotion recognition rate, and the sensitivity of different emotional features in different languages are also different. So we need to adjust our features to different various corpuses.

As can be seen from the experiment, the emotion recognition rate of the system which only uses the spectrum features of speech is slightly higher than that only uses the prosodic features of speech. And the system that uses both spectral and prosodic features is better than that only uses spectrum or prosodic features. Meanwhile, the recognition rate of that use energy, pitch, LPCC MFCC and MEDC features is slightly lower than that only use energy, pitch MFCC and MEDC features. This may be accused by feature redundance.

To extract the more effective features of speech and enhance the emotion recognition accuracy is our future work. More work is needed to improve the system so that it can be better used in real-time speech emotion recognition.

REFERENCES

1 http://www.expressive-speech.net/, Berlin emotional speech database

2 D. Ververidis, C. Kotropoulos, and I. Pitas, "Automatic emotional speech classification", in Proc. 2004 IEEE Int. Conf. Acoustics, Speech and Signal Processing, vol. 1, pp. 593-596, Montreal, May 2004.

3 Xiao, Z., E. Dellandrea, Dou W.,Chen L., "Features extraction and selection for emotional speech classification". 2005 IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS), pp.411-416, Sept 2005.

4 T.-L. Pao, Y.-T. Chen, J.-H. Yeh, P.-J. Li, "Mandarin emotional speech recognition based on SVM and NN", Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), vol. 1, pp. 1096-1100,September 2006.

5 Xia Mao, Lijiang Chen, Liqin Fu, "Multi-level Speech Emotion Recognition Based on HMM and ANN", 2009 WRI World Congress, Computer Science and Information Engineering, pp.225-229, March 2009.

6 B. Schuller, G. Rigoll, M. Lang, "Hidden Markov model-based speech emotion recognition", Proceedings of the IEEE ICASSP Conference on Acoustics,Speech and Signal Processing, vol.2, pp. 1-4, April 2003.

7 Yashpalsing Chavhan, M. L. Dhore, Pallavi Yesaware, "Speech Emotion Recognition Using Support Vector Machine", International Journal of Computer Applications, vol.1, pp.6-9,February 2010.

8 Zhou Y, Sun Y, Zhang J, Yan Y, "Speech Emotion Recognition Using Both Spectral and Prosodic Features", ICIECS 2009.International Conference on Information Engineering and Computer Science, pp.1-4, Dec.2009.

9 An X, Zhang X,"Speech Emotion Recognition Based on LPMCC", Sciencepaper Online.2010.

10 D. Ververidis and C. Kotropoulos, "Emotional Speech Recognition: Resources, features and methods", *Elsevier Speech communication*, vol. 48, no. 9, pp. 1162-1181, September, 2006.

11. Han Y, Wang G, Yang Y, "Speech emotion recognition based on MFCC", Journal of ChongQing University of Posts and Telecommunications(Natural Science Edition),20(5),2008.