# Speech Emotion Recognition Using Support Vector Machine

Yixiong Pan, Peipei Shen and Liping Shen
*Department of Computer Technology*
*Shanghai JiaoTong University, Shanghai, China*
*panyixiong@sjtu.edu.cn, shen@sjtu.edu.cn, 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

Automatic Speech Emotion Recognition is a very active research topic in the Human Computer Interaction (HCI) field and has a wide range of applications. It can be used for in-car board system where information of the mental state of the driver maybe provided to initiate his/her safety. In automatic remote call center, it is used to timely detect customers' dissatisfaction. In E-learning field, identifying students' emotion timely and making appropriate treatment can enhance the quality of teaching. Nowadays, the teachers and students are usually separated in the space and time in E-learning circumstance, which may lead to the lack of emotional exchanges. And the teacher can not adjust his/her teaching method and content according to the students' emotion. For example, when there is an online group discussion, if students are interested in the topic, they will be lively and active, and show their positive emotion. On the contrary, if they get in trouble or are not interested in it, they will show the opposite emotion. If we detect the emotion data, and give helpful feedback to the teacher, it will help the teacher to adjust the teaching plan and improve the learning efficiency.

In recent years, a great deal of research has been done to recognize human emotion using speech information. Many speech databases are built for speech emotion research, such as BDES (Berlin Database of Emotional Speech) which is German Corpus and established by Department of acoustic technology of Berlin Technical University [1](we will introduce it more in Section 2), DES (Danish Emotional Speech) is Danish Corpus and established by Aalborg University, Denmark [2].The data are sentences and words which are located between two silent segments. For example 'Nej'(No), 'Ja'(Yes), 'Kommeddig'(Come with me!). The total amount of data are 500 speech segments (with no silence interruptions), which

are expressed by four professional actors, two male and two female. Speech is expressed in 5 emotional states, such as anger, happiness, neutral, sadness, and surprise.

Many 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 use the Berlin emotional database and SJTU Chinese emotional database built by ourselves to train and test our automatic speech emotion recognition system. Prosody and Spectral features have been widely used in speech emotion recognition. In this paper, we compare the recognition rate using energy, pitch, LPCC, MFCC, and MEDC features and their different combination.

## 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. 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 Chinese speech database.

## 3. Speech Emotion Recognition System

Speech emotion recognition aims to automatically identify the emotional state of a human being from his or her voice. It is based on in-depth analysis of the generation mechanism of speech signal, extracting some features which contain emotional information from the speaker's voice, and taking appropriate pattern recognition methods to identify emotional states. Like typical pattern recognition systems, our speech emotion recognition system contains four main modules: speech input, feature extraction, SVM based classification, and emotion output (Figure 1).

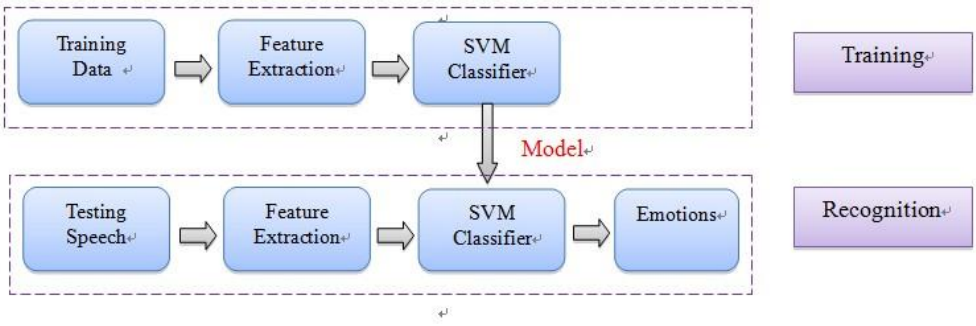


**Figure 1. Speech Emotion Recognition System**

## 4. Feature Extraction

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.

### 4.1. Energy and Related Features

The Energy is the basic and most important feature in speech signal. In order to obtain the statistics of energy feature, we use short-term function to extract the value of energy in each speech frame. Then 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].

### 4.2. Pitch and Related Features

The pitch signal is another important feature in speech emotion recognition. The vibration rate of vocal is called the fundamental frequency F0 or pitch frequency. The pitch signal is also called the glottal wave-form; it has information about emotion, because it depends on the tension of the vocal folds and the sub glottal air pressure, so the mean value of pitch, variance, variation range and the contour is different in seven basic emotional statuses.

### 4.3. 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.

### 4.4. Mel-Frequency Cepstrum Coefficients (MFCC)

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].

MFCC in the low frequency region has a good frequency resolution, and the robustness to noise is also very good, but the high frequency coefficient of accuracy is not satisfactory. In our research, we extract the first 12-order of the MFCC coefficients.
The process of calculating MFCC is shown in Figure2.

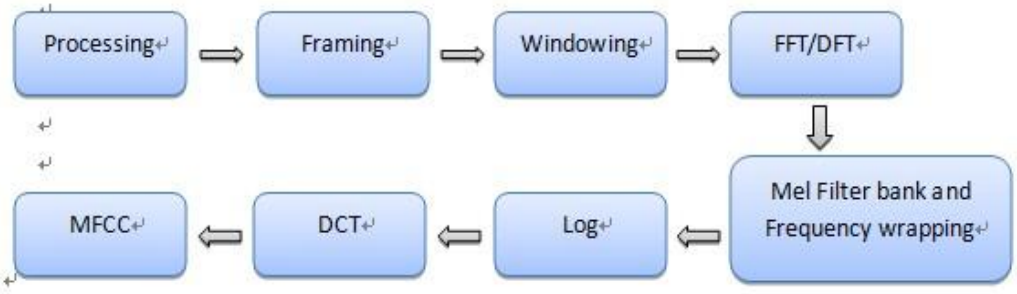


**Figure 2. Process of Calculating MFCC**

### 4.5. 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 Filter bank and Frequency wrapping, while the MFCC is taking logarithmic after Mel Filter bank and Frequency wrapping. After that, we also compute 1st and 2nd difference about this feature.

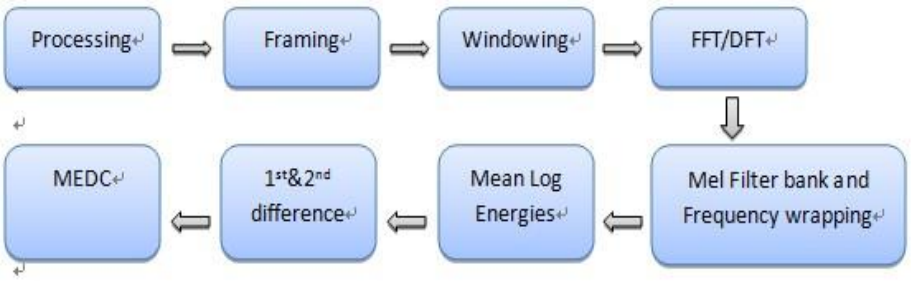


**Figure 3. Process of Calculating MFCC**

## 5. Experiment and Results

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.

### 5.1. SVM Classification Algorithm

Since SVM is a simple and efficient computation of machine learning algorithms, and is widely used for pattern recognition and classification problems, and 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.

### 5.2. Training Models

The Berlin Emotion database contains 406 speech files for five emotion classes. We choose three from it. Emotion classes sad, happy, neutral are having 62, 71, and 79 speech utterance respectively. While our own emotion speech database (SJTU Chinese emotion database) contains 1500 speech files for three emotion classes. There are 500 speech utterances for each emotion class respectively. We use both database, combine different features to build different training models, and analyse their recognition accuracy. Table1 shows different combination of the features for the experiment.

**Table 1. Different Combination of Speech Feature Parameters**

| ***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* |

### 5.3. Experimental Results

We use libsvm tool in Matlab to do the cross validation of models and analyse results. With the experiment, we pick pitch, energy, MFCC, its first-order difference, second-order difference, and MEDC as well as its first-order and second-order difference and their combination to extract features. For each emotion, we divide these speech utterances into two subsets as training subset and testing subset. The number of speech utterances for emotion as

the training subset is 90%, and 10% as the test subset. Table2 shows the models' cross validation rate and recognition rate based on Berlin Emotion database.

**Table 2. The Recognition Rate and Cross Validation Based on German Model**

| *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%* |

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

**Table 3. The Recognition Rate and Cross Validation Based on man Model**

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

As is shown at Table 2 and Table 3, 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.

## 6. 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.

[12] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[13] Lin Y, Wei G, "Speech emotion recognition based on HMM and SVM". Proceedings of 2005 International Conference on Machine Learning and Cybernetics, vol.8, pp. 4898-4901. Agu 2005.

[14] Peipei Shen, Zhou Changjun, Xiong Chen. "Automatic Speech Emotion Recognition using Support Vector Machine," Electronic and Mechanical Engineering and Information Technology (EMEIT), 2011 International Conference on , vol.2, no., pp.621-625, 12-14 Aug. 2011

[15] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2007S09,MASC(Mandarin Affective Speech)

## Authors

**Yixiong Pan**

Now is a graduate student in E-learning Lab at Shanghai JiaoTong University. Research on speech emotion recognition.

**Peipei Shen**
Post graduate student in E-learning Lab at Shanghai JiaoTong University. Research on speech emotion recognition.

**Liping Shen**
An Associate Professor in E-learning Lab at Shanghai JiaoTong University. Research on pervasive learning technology, network computing and speech emotion recognition.