Speech emotion recognition using combination of features

Conference Paper · June 2013

DOI: 10.1109/ICICIP.2013.6568131

5 authors, including:

Ning An
Hefei University of Technology
106 PUBLICATIONS **2,017** CITATIONS
SEE PROFILE

Kunxia Wang
Anhui Jianzhu University
15 PUBLICATIONS **339** CITATIONS
SEE PROFILE

Fuji Ren
The University of Tokushima
651 PUBLICATIONS **5,147** CITATIONS
SEE PROFILE

2013 Fourth International Conference on Intelligent Control and Information Processing (ICICIP)
June 9 – 11, 2013, Beijing, China

Speech Emotion Recognition using Combination of Features

Qingli Zhang, Ning An, Kunxia Wang, Fuji Ren and Lian Li

Abstract*— In this paper, we study how speech features' numbers and statistical values impact recognition accuracy of emotions present in speech. With Gaussian Mixture Model (GMM), we identify two effective features, namely Mel Frequency Cepstrum Coefficients (MFCCs) and Auto Correlation Function Coefficients (ACFC) extracted directly from speech signal. Using GMM supervector formed by values of MFCCs, delta MFCCs and ACFC, we conduct experiments with Berlin emotional database considering six previously proposed emotions: anger, disgust, fear, happy, neutral and sad. Our method achieve emotion recognition rate of 74.45%, significantly better than 59.00% achieved previously. To prove the broad applicability of our method, we also conduct experiments considering a different set of emotions: anger, boredom, fear, happy, neutral and sad. Our emotion recognition rate of 75.00% is again better than71.00% of the method of hidden Markov model with MFCC, delta MFCC, cepstral coefficient and speech energy.

I. INTRODUCTION

In Human-Computer Interaction (HCI) domain, speech emotion recognition has drawn attentions from many researchers. Speech signal contains rich information including: 1) speech content; 2) speaker identified [2] and recognized [8] by the voice signal; 3) speaker's emotions. Accurately recognizing the last information, i.e. speaker's emotions can help people in their study, emotional health and other activities. As to study, researchers have used speech emotion recognition techniques to optimize functions of computer-aided learning [15]. As to the emotional health, intelligent pet machines with speech emotion recognition capacity [16] have been introduced to the market and they can interact with people naturally and increase people's enjoyment. To this end, these intelligent pet machines can find a great usage in healthcare for people with need of emotional communication, especially the elderly living alone. They can help track the emotional changes of these people, and provide comforts when no human assistances are available.

In this paper, we first discuss which kind of features to extract from the speech. Next, we use these features to recognize emotions contained in the given speech. Later, we demonstrate that our approach can be used to effectively identify two different sets of six feeling, and the recognition rates only differ about 2%.

Many researchers have studied speech emotion recognition. Bitouk, Verma and Nenkova [10] proposed class-level spectral features for emotion recognition. They divided prosodic and spectral features into different levels of granularity: utterance-level (UL) and class-level (CL). Atassi, Esposito and Smekal [11] showed that high-level features performed well in terms of speech emotion classification, and recognition rates for six emotions are more than 80%. They used a lot of features: the Mel Frequency Cepstrum Coefficients (MFCC), pitch, the Perceptual Linear Predictive (PLP), the subband based cepstral coefficients (SBC), the wavelet decomposition (WADE), the MELBS, the HFBS and the LFBS. Similarly, Emerich and Lupu used the MFCC and the Discrete Wavelet Transform (DWT) in [12] to identify seven kinds of emotions, and achieved 95.42% accuracy rate. Iliou and Anagnostopoulos also achieved very high accuracy (94.3%) in [13]. They used twenty-five features in their work including the pitch, the MFCC, the energy and the formant. The major drawback of the above three works is that the number of features used is so large that their time complexity becomes too high. Other researchers used a large number of features to identify a small number of emotions, and acquired good results. In [14], Pan, Shen and Shen used seven features, which are the energy, the pitch, the linear predictive spectrum coding (LPCC), the MFCC and the mel-energy spectrum dynamic coefficients (MEDC), to identify three emotions. Their method achieved up to 95.1% accuracy rate. Our literature survey shows that most speech emotion recognition methods use spectral and prosodic features. Using different set of features also lead to quite different emotion recognition rate.

In our study, we select two features: the MFCC and the Auto Correlation Function Coefficient (ACFC). The MFCC feature has been widely used in the related work because Mel frequency is proposed according to the characteristics of the human auditory and the MFCC included language characters richer than other speech features [4]. In addition to the MFCC, we also choose rarely noticed ACFC. The reason being that the autocorrelation function has several good properties: 1) the autocorrelation function has a holding period which provides a way to estimate signal cycles; 2) the autocorrelation function also has noise immunity when it used for pitch detection. The other property of autocorrelation function is the noise immunity when it is used for pitch detection. Above all, we choose the MFCC and the ACFC in our study.

The rest of the paper is organized as follows: the next section introduces the test model used in the paper: GMM. After describing our feature extraction methods in section III, we present experiment designs and results in section IV.

Manuscript received January 31, 2013.

Qingli Zhang is with Lanzhou University, Lanzhou, Gansu, 730000, P. R. China (e-mail: zhang4722388@163.com).

Ning An is with Hefei University of Technology, Hefei, Anhui, 230000, P. R. China (e-mail: ning.g.an@acm.org). Corresponding Author.

Kunxia Wang is with the Hefei University of Technology, Hefei, Anhui, 230000, P. R. China (kxwang@aiai.edu.cn).

Fuji Ren is with the University of Tokushima, Tokushima, 770-8501, Japan (e-mail: ren@is.tokushima-u.ac.jp）.

Lian Li is with the Hefei University of Technology, Hefei, Anhui, 230000, P. R. China (e-mail: lilian@hfut.edu.cn).

Section V concludes this paper with a summary. Finally, section VI is the acknowledgment.

II. THE GAUSSIAN MIXTURE SPEAKER MODEL

A. The Principle of GMM

Gaussian distribution has many important properties. If we use it to describe the reality data, there are still many limitations. When we put these simple distributions into the form of linear combination, it can better describe the nature of the actual data. Such a model is called hybrid model.

The most commonly used and most popular hybrid model is the Gaussian Mixture Model (GMM). If there is a sufficient number of Gaussian distribution adjusting its expectation， covariance matrix, and the coefficients of the linear combination, it can expressed any continuous distribution. Gaussian Mixture Model is a commonly used statistical models used in speech signal processing.

A Gaussian mixture model consists of M single Gaussian model, and the model is a weighted combination of the M gaussian probability density distribution function (PDF). The model is described as the following equation

$$p(\vec{x}|\lambda)=\sum_{i=1}^{M}\omega_i b_i(\vec{x}) \qquad (1)$$

Where $\vec{x}$ is a D-dimensional random vector, ω_i, i=1……M are the mixture weights, and $b_i(\vec{x})$, i=1……M are the component Gaussian densities. Every component density is a D-variate Gaussian function of the following formula:

$$b_i(\vec{x})=\frac{1}{(2\pi)^{D/2}\left|\sum_i\right|^{1/2}}\exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu_i})'\sum_i^{-1}(\vec{x}-\vec{\mu_i})\right\} \qquad (2)$$

$\vec{\mu_i}$ is the mean vector and $\sum_i$ is the covariance matrix. The mixture weights satisfy the condition that $\sum_{i=1}^{M}\omega_i=1$.

The complete Gaussian mixture density is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by the notation

$$\lambda=\left\{\omega_i,\vec{\mu_i},\sum_i\right\}\ i=1,\ldots\ldots,M \qquad (3)$$

How can we estimate the λ? The paper [1] introduce two methods for us, they are the maximum likelihood (ML) estimation and the expectation-maximization (EM) algorithm.

B. The Usage of GMM

Gaussian mixture model can represent any continuous distribution, so we can analyze the voice signal by GMM. GMM and k-means have some similarities. K-mean's result is that each data point is assigned to one cluster. GMM make the probability of these data points to assign to each cluster. We can use a variety of methods to analyze a voice using the special characteristics of Gaussian mixture model. Finally, we select the maximum probability of the emotion as this voice emotional type.

Article [2] describes the application of GMM in speaker recognition. This paper is the earlier introduction of GMM in speaker recognition applications. The following example gets the PDF after training [3]. Through this example, we can understand the working process of GMM. The experimental results are shown in Figure 1. The histogram is randomly generated data and the chart below is a Gaussian mixture model generation process. $\sum_i \omega_i b_i$ is the final Gaussian mixture model. From the experimental results we can see that the Gaussian mixture model can accurately simulate the distribution of the data.

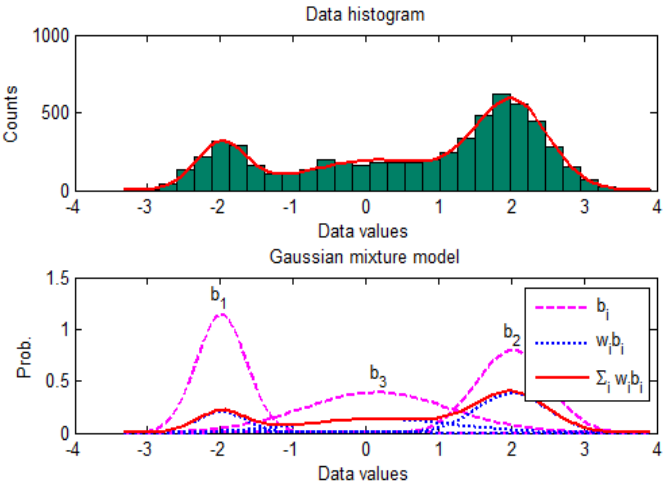

Fig.1. the histogram is randomly generated data and the below line chart is a Gaussian mixture model generation process， b_i means component density, ω_i represents the weights, $\sum_i \omega_i b_i$ is the gaussian probability density distribution function (PDF)

III. FEATURE EXTRACTION

The voice feature extraction occupies an important position in the speech emotion recognition. Because no one can determine which voice characteristic can accurately recognize the emotion of voice. In [4], the author notes that four kinds of problems must be taken into account in feature extraction. He also said the voice characteristics attributed to four categories: continuous features, qualitative features, spectral features, and TEO (Teager energy operator)-based features. In this paper, we will evaluate the emotion of the voice based on the MFCC and Auto Correlation Function Coefficients (ACFC).

A. Mel Frequency Cepstrum Coefficients (MFCC)

In article [5], authors proposed an idea that features based on cepstral analysis such as Mel-frequency Cepstrum Coefficients (MFCC) clearly outperform the linear-based features. The Mel Frequency is based on the characteristics of human auditory. The parameters take into account the feelings of human ear for different frequencies.

The process of solving MFCC is roughly divided into the following steps: Pre-emphasis, Frame blocking, hamming window, Fast Fourier Transform (FFT), Triangular Bandpass

Filters, Discrete cosine transform (DCT) and Log energy. These steps are shown in Figure 2. Finally, we get the feature vectors $\vec{x_1},\vec{x_2},\cdots,\vec{x_t}$.

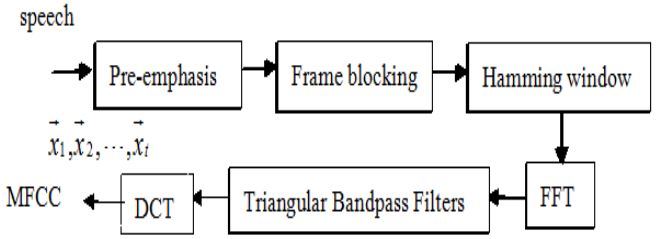

Fig. 2. MFCC Block Diagram

At the Frame Blocking Step, we split the continuous speech signal into N frames. In order to avoid the change of two adjacent frames being too large, we make some overlap region between the frames, and this overlap region contains M sampling points, the value of M is usually about 1/2 or 1/3 of the N. The values used are M=128 and N=256.

Each sound frame multiplied by the Hamming window in order to increase the continuity of the left and right ends of the sound box. We assume that the sound box of the signal is S (n), n=0... N-1, Multiplied by Hamming window, the formula is S'(n) =S (n)*W (n). The form of W (n) is as follows:

$$W(n)=0.54-0.46\cos(\frac{2\pi n}{N-1}),\quad 0\le n\le N-1$$

After being multiplied by the Hamming window, each frame is also necessary to be processed by FFT which can obtain the frequency spectrum energy distribution. The signal change in time domain is often difficult to indicate the characteristics of the signal, so we usually convert the signals into energy distribution in the frequency domain. Different energy distribution can represent the characteristics of different speech.

Typically the Triangular Bandpass Filters is used. The energy spectrums multiplied by a group of 20 Triangular Bandpass Filters and obtain the log energy from each filter. The 20 Triangular Bandpass Filters is evenly distributed in the Mel frequency. Mel frequency and the average frequency have the following relationship:

$$Mel(f)=2595*\log10(1+f/700)$$

B. Auto Correlation Function Coefficients (ACFC)

Auto Correlation Function is put forward by Ross [6] in 1977. The autocorrelation function of the periodic signal can generate a maximal value when the delay equals to the function cycle. Therefore, by calculating the autocorrelation function of the speech signal we can find the maximum value position and estimate the pitch period of the signal.

Discrete sequence x (n) of digital voice signals have a period as follows:

$$x(n)=x(n+N_p) \qquad (4)$$

The autocorrelation function is a periodic function too:

$$P(k)=\lim_{N\to\infty}\frac{1}{2N+1}\sum_{-\infty}^{\infty}x(m)x(m+k) \qquad (5)$$

Unvoiced signal and its autocorrelation function have no periodicity, and no significant peak. P (k) rapidly decays while k increases. Voiced signal has a quasi-periodicity, and its auto-correlation function P (k) has the same period with k. Autocorrelation function has a peak at the position of integer multiples of the pitch periods. According to this nature, we can judge a speech signal is unvoiced or not, and obtain the pitch period of the voiced. Figure 3 shows the result of a voice handled by the autocorrelation function processing.

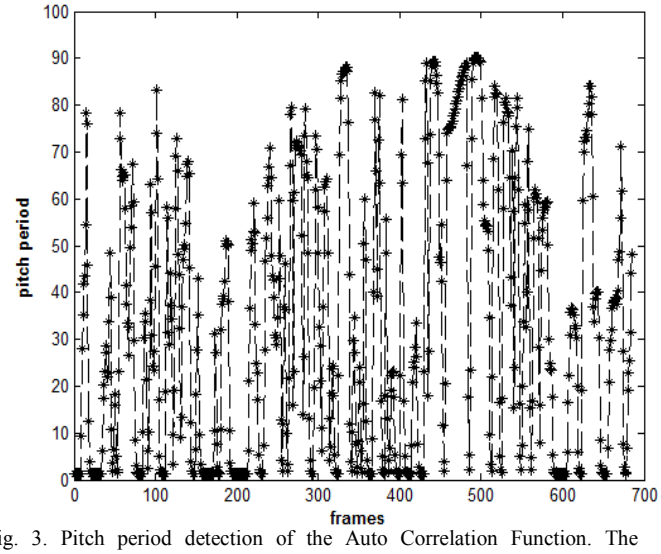

Fig. 3. Pitch period detection of the Auto Correlation Function. The horizontal axis indicates frames and vertical axis represents the pitch period.

IV. EXPERIMENTS AND RESULTS

A. Database

In this work, Berlin database is explored for analyzing the emotion. Berlin database was recorded by F. Burkhardt et al in German language [7]. This database was composed of 800 statements (seven emotions* ten actors* ten sentences + some second versions). Expert listeners selected 10 people (five males and five females) for the establishment of the database. The seven emotions in Berlin database were anger, boredom, disgust, fear, happiness, neutral and sadness. In this paper, we made two experiments. The first experiment was to identify six feelings recognition rate. These six feelings were anger, disgust, fear, happiness, neutral and sadness. The second was to identify the other six emotions recognition rate. The six feelings were anger, boredom, fear, happiness, neutral and sadness.

B. Result

The spectral features included 12 Mel Frequency Cepstral Coefficients (MFCC), 12 delta MFCC and ACFC. In this paper we discuss two factors impacting the emotion recognition rate. The first one is the number of voice features extracted, and the other is the number of the statistical value of the extracted characteristic. In addition, we need to verify the stability of these features for emotion recognition.

First of all, we just took MFCC and delta MFCC characteristics, and the speech emotion recognition rate reached 72.69%. When ACFC characteristics were also taken into account, the accuracy rate was up to 74.01%. So we can

see that multiple characteristics can obviously improve the recognition rate. Table I and table II show the confusion matrices of these two kinds of circumstances. Diagonal represents the correct classification rate, whereas other elements in the row show the miss-classification of the emotions. The average emotion recognition rates for all 6 emotions were observed to be 70.77% and 72.44% for these two situations in order. The results show that multiple features can improve emotion recognition rate. In table I, anger has the highest recognition rate. It may be wrong identification into fear and happiness. The recognition rate of happiness was the lowest and the probability of identified as anger was larger. In table II, neutral has the highest recognition rate. This indicates if we add new features in the test, change the experimental results.

TABLE I
AVERAGE EMOTION CLASSIFICATION USING MFCC AND DELTA MFCC

	Emotion recognition performance (%)					
	anger	disgust	fear	happy	neutral	sad
anger	85.94	0	6.25	7.81	0	0
disgust	4.36	65.22	13.04	8.69	8.69	0
fear	5.71	2.86	65.71	17.14	8.58	0
happy	40	0	11.43	45.71	2.86	0
neutral	0	2.56	7.69	5.13	84.62	0
sad	0	0	16.13	0	6.46	77.41

TABLE II
AVERAGE EMOTION CLASSIFICATION USING MFCC DELTA MFCC AND ACFC

	Emotion recognition performance (%)					
	anger	disgust	fear	happy	neutral	sad
anger	79.69	0	0	18.75	1.56	0
disgust	4.36	60.87	17.39	8.69	0	8.69
fear	8.58	0	74.29	5.71	5.71	5.71
happy	48.57	2.86	8.57	37.14	2.86	0
neutral	0	0	7.69	0	92.31	0
sad	0	0	3.23	0	6.45	90.32

Secondly, we studied the impact on emotion recognition rate which was the statistical value of the extraction features. Suppose that we extract the characteristics of the original data values as x. First we take the mean, median, standard deviation of x and then take the mean, standard deviation conduct the experiment. The obtained experimental results are as follows: when taking three statistical values obtained, the highest recognition rate 74.45%, and when taking two statistical values, the recognition rate is 70.48%. These emotion recognition rates were based on the extracted MFCC delta MFCC and ACFC characteristics.

Table III and table IV show the confusion matrices of emotion classification for the different number of statistical values. The emotion recognition rate of the table III is 70.48% and table IV is 74.45%. Diagonal represent the correct classification rate. The average recognition performance is around 66.08% and 73.31% respectively. It can be seen that from comparison, the number of statistical values has a great impact on emotion recognition rate. From table III and table IV, we can be found that when there are two statistics and three statistics, sad recognition rate is always the highest. The accuracy rate reached 90.32% and 93.48% respectively. Moreover, sad is not easy to be recognized as other feelings. We also found that happiness recognition rate was extremely low; most of the happiness statements were recognized as anger.

TABLE III
AVERAGE EMOTION CLASSIFICATION USING TWO STATISTICAL VALUES

	Emotion recognition performance (%)					
	anger	disgust	fear	happy	neutral	sad
anger	85.94	1.56	0	10.94	1.56	0
disgust	8.69	52.17	17.39	17.39	4.36	0
fear	5.71	0	51.43	25.71	14.29	2.86
happy	40	0	22.86	37.14	0	0
neutral	2.56	0	10.26	5.13	79.49	2.56
sad	0	0	9.68	0	0	90.32

TABLE IV
AVERAGE EMOTION CLASSIFICATION USING THREE STATISTICAL VALUES

	Emotion recognition performance (%)					
	anger	disgust	fear	happy	neutral	sad
anger	79.69	0	1.56	17.19	1.56	0
disgust	0	65.22	17.39	13.04	0	4.35
fear	11.43	0	77.14	2.86	5.71	2.86
happy	42.86	2.86	14.28	37.14	2.86	0
neutral	0	0	10.26	2.56	87.18	0
sad	0	0	3.26	0	3.26	93.48

From above four tables, we can clearly see that the sadness recognition rate is the highest, in other words, when voice emotion is sad, it is not easy to identify compared to other emotions. In addition, neutral is relatively easy to determine. Happiness has the lowest recognition rate of emotions and easily identified as anger. Through all of these experiments, the highest emotion recognition accuracy rate was 74.45%. Article [9], however, only has 59% accuracy rate.

Below, we verify the recognition rate of the other six emotions: anger, boredom, fear, happiness, neutral and sadness. The data of these six feelings are also from the Berlin database [7]. This experiment is similar to the previous experiment, the first step only to take MFCC and delta MFCC. Table V is the confusion matrix of the experiment results. Diagonal represent the correct classification rate, whereas other elements in the row show the miss-classification of the emotions. We took the mean, median, standard deviation of the characteristics. We obtained the accuracy rate of 72.95%. The average emotion recognition rate of all 6 emotions was observed to be 70.88%. Compared with the first experiment, the accuracy rate does not change much. In table V anger has the highest recognition rate and least likely to be recognized as the other feelings. The happiness identification rate is the lowest. It is most likely to be recognized as anger. We can also find that sadness can't easily be identified as the other feelings.

Then, we took into account the Auto Correlation Function Coefficients (ACFC) characteristic, and also took the mean, median, standard deviation of these characteristics. The accuracy was 2.05% higher than last experiment. Table VI is the confusion matrix of the experiment results. Diagonal represent the correct classification rate, whereas other elements in the row show the miss-classification of the emotions. Under such conditions, the average emotion recognition rate was74.22%.

TABLE V
AVERAGE EMOTION CLASSIFICATION USING MFCC DELTA MFCC AND THE MEAN, MEDIAN, STANDARD DEVIATION OF MFCC AND DELTA MFCC

	Emotion recognition performance (%)					
	anger	boredom	fear	happy	neutral	sad
anger	88.89	0	1.59	9.52	0	0
boredom	0	72.5	5	2.5	20	0
fear	5.71	0	74.29	14.29	5.71	0
happy	38.89	0	13.89	44.44	2.78	0
neutral	0	7.69	12.82	5.13	74.36	0
sad	0	9.68	12.90	0	6.45	70.9

From table V and table VI, it can be found that the recognition rate of the happiness is the lowest, and happiness particularly is easily confused with anger. From these confusion matrixes of the two experiments, this relationship can be found between the happiness and anger. This shows that a particular similarity exists between the happiness and anger, and not easy to distinguish. In table VI sadness identification rate is very high. The probability of sadness be identified into other emotion is small. Although the natural recognition rate is the highest, the probability that the natural is identified as the other feelings is relatively large

TABLE VI
AVERAGE EMOTION CLASSIFICATION USING MFCC, DELTA MFCC, ACFC AND THE MEAN, MEDIAN, STANDARD DEVIATION OF MFCC DELTA MFCC AND ACFC

	Emotion recognition performance (%)					
	anger	boredom	fear	happy	neutral	sad
anger	80.95	0	1.59	17.46	0	0
boredom	0	80	0	2.5	15	2.5
fear	11.43	0	80.00	0	5.71	2.86
happy	44.44	0	19.44	33.33	2.79	0
neutral	0	7.69	2.56	2.56	87.19	0
sad	12.9	0	0	0	3.23	83.9

In the above experiment, we only used mean, median and standard deviation of the characteristics. In article [4], the author mentions the statistical value with the mean, median, standard deviation, maximum, minimum and range (max-min). In the following test, we took the mean, median, standard deviation maximum and minimum of the MFCC and delta MFCC. Although the statistical value type increased, the emotion recognition accuracy didn't improve. This phenomenon indicates that the emotion recognition accuracy rate doesn't increase with the increase of the statistic's type. Table VII is the confusion matrix of the experiment results. Diagonal represent the correct classification rate, whereas other elements in the row show miss-classification of the emotions. The average emotion recognition rate was 72.54%. The accuracy rate was lower than the situation which we only took mean, median and standard deviation.. From that table, the happiness recognition rate becomes high and can distinguish between anger and happiness very well. In table VII, anger has the highest recognition rate, the rate reached 88.89%. The anger only be identified into happy not be identified as the other feelings. Neutral may be wrong to think as one of the five kinds of feelings, but the probability is not very high.

TABLE VII
AVERAGE EMOTION CLASSIFICATION USING MFCC DELTA MFCC AND THE MEAN, MEDIAN, STANDARD DEVIATION, MAXIMUM AND MINIMUM OF MFCC AND DELTA MFCC

	Emotion recognition performance (%)					
	anger	boredom	fear	happy	neutral	sad
anger	88.89	0	0	11.11	0	0
boredom	0	60	7.5	2.5	20	10
fear	11.43	5.71	65.71	8.57	5.71	2.87
happy	33.33	0	5.56	61.11	0	0
neutral	5.13	7.69	5.13	5.13	74.36	2.56
sad	0	19.35	3.23	0	3.23	74.2

All of these experiments show that the speech features for speech emotion recognition accuracy is very stable. When the emotion that we identified changes, the obtained accuracy rates were very close-like. When identified six feelings of anger, disgust, fear, happy, neutral and sad, using MFCC and delta MFCC with their statistics of mean, median and standard deviation, the accuracy rate was 72.69%. When we used the same features and statistics to identify another six kinds of feelings, accuracy rate was 72.95%. We can find that the difference between these two recognition rate is only 0.26%. Only in each test each emotion recognition rate difference is bigger. We find that when the number of types of statistical values changes, some emotion recognition rate increases.

Finally, we studied that how number of Gaussian model impact the emotion recognition rate. This is a difficult problem, because there is no specific theory to illustrate this problem. Taking too much or too little of the Gaussian model will reduce the emotional recognition rate. Figure 4 shows the result of the number of Gaussian components impacts on table I, table II, table III and table IV. There are several obvious discoveries from the figure. First, when mixed combination ranges between 2 and 4, the emotion recognition rate is higher. When mixed combination reaches at 10, the accuracy rate becomes very low. In addition, we also can see that table IV has a higher accuracy rate than others almost all the time. This is because the form of the experimental data for this table contains three characteristics and three statistical values of these characteristics.

V. CONCLUSION

This paper describes a method using Gaussian mixture model with several speech features, including MFCC, delta MFCC and ACFC to identify emotions present in speech. Experiment results demonstrate that our method can reach 74.45% emotion recognition rate. Especially for sadness, it can go as high as 93.48%. We also experiment with two sets of six emotions, and find that emotional recognition rates of our method only differs less than 2% for these two cases. This in some extent show the robustness of our method. While our method shows better emotion recognition rate than previous work, it still has room to improve. We are investigating other methods and features to this goal.

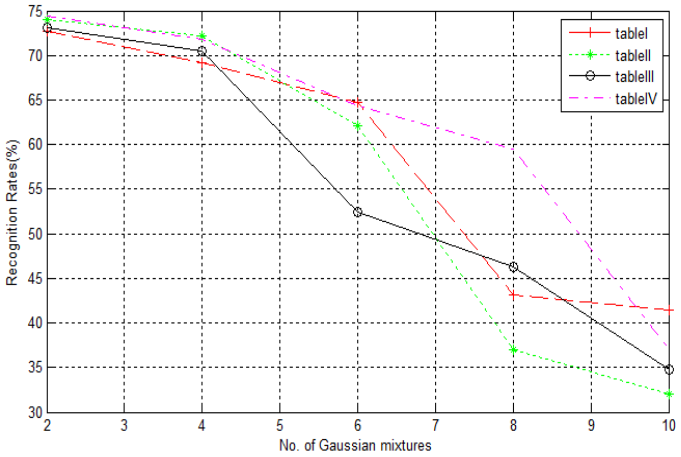

Fig. 4. Number of mixture components

VI. ACKNOWLEDGMENT

This work was supported in part by the National High Technology Research and Development Program of China (863 Program) under grant 2012AA011103, by National Natural Science Foundation of China under Grant no. 61073193, Grant no. 70673030 and Grant no. 90924025, and by Hefei University of Technology under Grant 407-037036 and 2011HGZY0018.

REFERENCES

[1] D. Reynolds, "Gaussian Mixture Models", Encyclopedia of Biometric Recognition, Springer, Feb. 2008.

[2] D. A.Reynolds and R. C. Rose,"Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE transactions on speech and audio processing, vol. 3, pp. 72-83, January 1995.

[3] Jyh-Shing Roger Jang, "Data Clustering and Pattern Recognition", available at the links for on-line courses at the author's homepage at http://mirlab.org/jang.

[4] Moataz El Ayadi, Mohamed S.Kamel, Fakhri Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases", Pattern Recognition, Volume 44, Issue 3, pp. 572-587, March 2011.

[5] S. Bou-Ghazale, J. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress", IEEE Trans. Speech Audio Process, 8 (4) (2000) 429-442

[6] M. J. Ross, H. L. Shaffer, A. Cohen, et al, "Average magnitude difference function pitch extractor [J]", IEEE Trans. Acoustics, Speech and Signal Processing, 22(5):353-362,1974.

[7] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss,"A database of german emotional speech," in Proc Interspeech 2005, Lisbon, Portugal, pp. 1517-1520, 2005.

[8] J. Naik, "Speaker verification: A tutorial," IEEE Commun. Mag., vol. 28, pp. 42-48, Jan 1990.

[9] S. G. Koolagudi, R. Reddy and K. S. Rao, "Emotion Recognition from Speech Signal using Epoch Parameters", IEEE International Conference on Signal Processing and Communications (SPCOM), pp. 1-5, 2010.

[10] D. Bitouk, R. Verma and A. Nenkova, "Class-level spectral features for emotion recognition", Speech Communication 52 (2010) pp. 613-625.

[11] H. Atassi, A. Esposito and Z. Smekal, "Analysis of High-level Features for Vocal Emotion Recognition", International conference on telecommunications and signal processing (TSP), pp. 361-366, 2011.

[12] S. Emerich and E. Lupu, "Improving Speech Emotion Recognition using Frequency and Time Domain Acoustic Features", Signal processing and applied mathematics for electronics and communications, SPAMEC, pp. 85-88, 2011.

[13] Theodoros Iliou, Christos-Nikolaos Anagnostopoulos, "Classification on Speech Emotion Recognition – A Comparative Study", International Journal on Advances in Life Sciences, vol. 2 no 1 & 2, ISSN. 1942-2660, pp. 18-28, year 2010.

[14] Y. X. Pan, P. P. Shen and L. P. Shen, "Speech Emotion Recognition Using Support Vector Machine", International Journal of Smart Home, vol. 6, No 2, pp. 101-108, April, 2012.

[15] L. P. Shen, M. J. Wang and R. M. Shen, "Affective e-learning: Using "emotional" data to improve learning in pervasive learning environment", Educational Technology & Society, 12 (2), 176-189, 2009.

[16] Y. M. Huang, G. B. Zhang and X. L. Xu, "speech emotion recognition research based on Wavelet Neural Network for robot pet", Emerging intelligent computing technology and application. With aspects of artificial intelligence, vol. 5755, pp. 993-1000, 2009.

[17] H. Hu, M. X. Xu and W. Wu, "GMM supervector based SVM with spectral features for speech emotion recognition", IEEE international conference on acoustics, speech and signal processing, vol.4, pp. IV-413-IV-416, 2007.

[18] B. Schuller, G. Rigoll and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture", IEEE international conference on acoustics, speech, and signal processing, vol.1, pp. I-577-80, 2004.

[19] L. J. Chen, X. Mao and Y. L. Xue, "Speech emotion recognition: Features and classification models", Digital signal processing, vol.22, pp. 1154-1160, 2012.

[20] S. Mansour, G. Davood and A. Farhad, "Using DTW neural-based MFCC warping to improve emotional speech recognition", Neural computing & applications, vol. 21, pp. 1765-1773, Oct 2012.

[21] S. Yun, C. D. Yoo, "Loss-scaled large-margin Gaussian Mixture Models for speech emotion classification", IEEE transactions on audio speech and language processing, vol. 20, pp. 585-598, Feb 2012.

[22] S. Bartlomiej and R. K. Krzysztof, "Fundamental frequency extraction in speech emotion recognition", 5th international conference on multimedia communications, services and security, MCSS, vol. 287, pp. 292-303, 2012.

[23] M. Kockmann, L. Burget and J. H. Cernocky, "Application of speaker-and language identification state-of-the-art techniques for emotion recognition", speech communication, vol.53, pp.1172-1185, Nov-Dec 2011.

[24] A. Bihar Kandali, A. Routray and T. Kumar Basu, "Emotion recognition from Assamese speeches using MFCC features and GMM classifier", TENCON 2008-2008 IEEE Region 10 conference, pp. 1-5, 2008.

[25] M. M. H. El Ayadi, M. S. Kamel, F. Karray, "Speech emotion recognition using Gaussian Mixture Vector Autoregressive models", IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 4, pp. 957-960, April 2007.