Speech Emotion Recognition based on Mel-Frequency Cepstral Coefficients and Support Vector Machine

An Hoa Ton-That[#1], Nhan Thi Cao[#2], Hyung-Il Choi[#3]

[#]*School of Media, Soongsil University, Seoul, Korea*

[1]an_tth@yahoo.com

[2]ctnhen@yahoo.com

[3]hic@ssu.ac.kr

Abstract— **Human emotion recognition has been researched much in recent years because of their applications in intelligent communication systems and data-driven animation. Two primary tools for communicating human emotion are the face and the voice. Research on facial expression is highly developed whereas research on speech emotion is much less. Speech emotion recognition has been investigated for past four decades but until now, there are still many challenges and recognition rates are usually low. For recognizing speech emotions, there are many kinds of features and various methods of classification being used. This paper tries to contribute a research on speech emotion recognition based on Mel-frequency cepstral coefficients feature and classification of support vector machine. The research has been experimented on Berlin database of emotional speech (Emo-DB).**

Keywords— **Speech emotion recognition, mfcc, support vector machine**

I. INTRODUCTION

Speech emotion recognition takes aim at identifying the emotional or physical states of humans being from their voices. Emotional state is an important factor in human communication and it provides response information in many intelligent interactive applications [1]. It is also believed that speech emotion recognition can be used to extract useful semantics from speech, and hence, improves the performance of speech recognition systems [2].

There are many challenges in speech emotion recognition. First, it is not clear which speech features are most powerful in distinguishing between emotions. Next is the existence of the different sentences in length or content, differences of speakers, speaking styles or speaking speed. The same utterance may show different emotions. Moreover, each emotion may correspond to the different portions of the spoken utterance or there may be more than one perceived emotion in the same utterance. Therefore it is very difficult to differentiate these portions of utterance. Another problem is that speech emotion is depending on the speakers and their culture and environment. Along with many more other difficult problems for speech emotion recognition, they cause recognition rate of the speech emotion recognition systems quite low.

Following [2], an important issue in speech emotion recognition is the need to determine a set of the key emotions to be classified by a speech emotion classifier. Linguists have defined inventories of the emotional states, most met in our lives. A typical set is given by Schubiger [3] and O'Connor and Arnold [4], which contains 300 emotional states. However, classifying such a large number of emotions is very difficult. So, existing speech emotion recognition researches agree with the 'palette theory', which states that any emotion can be decomposed into primary emotions similar to the way that any color is a combination of some basic colors as the most famous model of Ekman's six basic emotions [5]. These emotions are the most obvious and distinct emotions in our life and they are called the archetypal emotions [6].

The researches on speech emotion recognition commonly deal with two main problems: selecting effective speech emotion features and finding out appropriate model for speech emotion classification. Before the introduction of Mel-Frequency Cepstral Coefficients (MFCCs), Linear Prediction Coefficients (LPCs) and Linear Prediction Cepstral Coefficients (LPCCs) were the main feature for automatic speech recognition (ARS). But now MFCC feature has been used widely. In this work, we present a method of speech emotion recognition based on MFCC feature and classification of support vector machine. The method has implemented on Berlin database (Emo-DB) in German.

The paper is organized as follows: Section II presents Mel-frequency cepstral coefficients and its background. In Section III, the classification of support vector machine is introduced. Section IV learns about the Berlin emotion database. In Section V, the experiments and results are shown. Finally, Section VI, the conclusions are given.

II. MEL-FREQUENCY CEPSTRAL COEFFICIENTS AND ITS BACKGROUND

Mel-Frequency Cepstral Coefficients (MFCCs) were introduced by David and Mermelstein in the 1980's [7], [8] and have been an advanced feature widely used in automatic speech, speaker and speech emotion recognition ever since.

[3]Corresponding author: Hyung-Il Choi

A. Mel scale

As in [9], the Mel scale relates perceived frequency, or pitch, of a pure tone to its actual measured frequency. Humans are much better at discerning small changes in pitch at low frequencies than they are at high frequencies. Incorporating this scale makes our features match more closely what humans hear. The formula for converting from frequency to Mel scale is as (1):

$$\mathrm{M(f)} = 1125 \ln\left(1 + \frac{f}{700}\right) \quad (1)$$

and to invert from Mels back to frequency is as (2):

$$M^{-1}(m) = 700\left(\exp\left(\frac{m}{1125}\right) - 1\right) \quad (2)$$

B. Mel-frequency cepstral coefficients (MFCCs)

As in [10], the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel-scale of frequency.

Mel-frequency cepstral coefficients are coefficients that collectively make up a mel-frequency cepstrum (MFC). They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound. MFCCs are commonly used as features in speech recognition systems, such as the systems which can automatically recognize numbers spoken into a telephone. They are also common in speaker recognition, which is the task of recognizing people from their voices [11]. Past several years, MFCCs are widely used as a feature of speech to recognize speech emotions because it models the human perception to speech quite well [12].

III. Support Vector Machine

Support Vector Machine (SVM) proposed by Vladimir N. Vapnik [13], [14] is a powerful, effective and popular machine learning technique for data classification. This method based on statistical learning theory with close foundation of mathematics to ensure that the results are optimal. SVM performs an implicit mapping of data into higher (perhaps infinite) dimensional feature space and then constructs a separating hyperplane with the maximal margin separate data in this higher dimension space. Many applications have confirmed SVM obtaining high results for classifying facial expression [15],[16].

Given a training set of labeled examples $\{(x_i, y_i), i = 1, \ldots, l\}$ where $x_i \in R^n$ and $y_i \in \{-1, 1\}$, a new test example x is classified by the following function:

$$f(x) = sign(\sum_{i=1}^{l} u_i y_i K(x_i, x) + b) \quad (3)$$

where u_i are Lagrange multipliers of a dual optimization problem that describe the separating hyperplane, $K(.\ ,.)$ is a Kernel function, and b is the threshold parameter of the hyperplane. The training sample x_i with $u_i > 0$ is support vectors. $K(x_i, x_j)$ is kernel based on a non-linear mapping Φ that mapped the input data into higher dimensional space and in the form of $\Phi(x_i).\Phi(x_j)$. Some frequently used kernel functions being used in SVM are the linear, polynomial, and Radial Basis Function (RBF) kernels.

SVM makes binary decisions, so the multi-class classification here is accomplished by using the one-against-rest technique, which trains binary classifiers to discriminate one expression from all others, and outputs the class with the largest output of binary classification [17].

In this work, we used the SVM functions based on Radial Basis Functions kernel. In order to choose optimal parameters, we implement grid-search approach as in [18].

IV. Database for The Experiment

Our experiments are applied on Berlin database of emotional speech (Emo-DB) [19]. This database included ten actors (5 females and 5 males) simulated the emotions, producing 10 German utterances (5 short and 5 longer sentences) which could be used in everyday communication and are interpretable in all applied emotions.

The database based on seven basic emotions (German terms in brackets) neutral (Neutral), anger (Ärger), fear (Angst), joy (Freude), sadness (Trauer), disgust (Ekel) and boredom (Langeweile) with following sentences were used:

a01: Der Lappen liegt auf dem Eisschrank. (The cloth is lying on the fridge.)
a02: Das will sie am Mittwoch abgeben. (She will hand it in on Wednesday.)
a04: Heute Abend könnte ich es ihm sagen. (Tonight I could tell him.)
a05: Das schwarze Stück Papier befindet sich da oben neben dem Holzstück. (The black sheet of paper is up there beside the piece of timber.)
a07: In sieben Stunden wird es soweit sein. (In seven hours the time will have come.)
b01: Was sind denn das für Tüten, die da unter dem Tisch stehen? (What are the bags standing there under the table?)
b02: Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter. (They have just carried it upstairs and now they are going down again.)
b03: An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht. (At the weekends I have always gone home now and seen Agnes.)

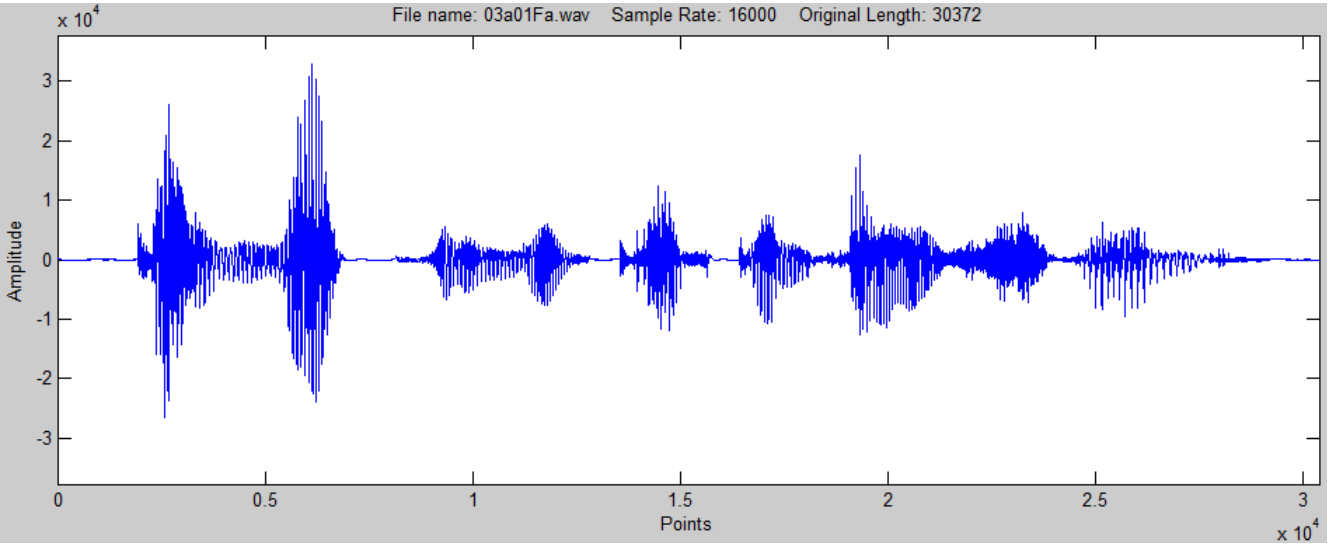

Fig. 1 A sample of emotion speech in Berlin database (Emo-DB)

b09: Ich will das eben wegbringen und dann mit Karl was trinken gehen. (I just want to take this away and then go for a drink with Karl.)

b10: Die wird auf dem Platz sein, wo wir sie immer hinlegen. (It will be in the place where we always put it.)

Audio files are in wav format with sampling rate 16 kHz, 16 bit, mono channel. All the available information of the speech database can be accessed via the internet as [20]. A sample of emotion speech in Emo-DB database for experimenting is shown as in Fig. 1.

V. Experiments and Results

A. Signal pre-process

For experiment in this work, we randomly selected 500 sentences included 7 different emotions of 10 people. The emotional speech files were cut with 1.5 second in length excluded blank or noise parts at beginning of the signals when they were recorded.

The initial speech signal often has a constant component, i.e. a non-zero mean. This is typically due to Direct Current (DC) bias within the recording instruments. For removing DC component, all samples within an utterance are subtracted its mean value as in (4):

$$signal = signal - mean(signal) \quad (4)$$

For compressing the dynamic range of the speech signal's power spectrum by flattening the spectral tilt, the pre-emphasis filtering is applied as in (5):

$$P(z) = 1 - az^{-1} \quad (5)$$

where a is usually set by 0.97.

To enhance the performance in energy-related features, the amplitude normalization is applied to can cancel the inconsistent energy level between signals as in (6):

$$signal = \frac{signal}{\max(abs(signal))} \quad (6)$$

B. Extracting MFCC feature

We set frame window size being 0.03 second, step size being 0.02 second. Max frequency boundary is half of sample rate (8 kHz) and min frequency boundary is 0 Hz. The MFCC filter number is 20 and MFCC cepstrum number is 12.

When applying a spectral analysis such as a Fast Fourier Transform (FFT) to a frame of rectangular windowed signal, the abrupt change at the starting and ending point significantly distorts the original signal in the frequency domain. To alleviate this problem, the rectangular window function is modified, so that the points close to the beginning and the end of each window slowly attenuate to zero. There are many possible window functions. One common type is the generalized window, defined as (7):

$$w(n) = \begin{cases} (1-\propto) - \propto \cos\left[\frac{2\pi n}{N-1}\right] & 0 \le n \le N-1 \\ 0 & n = else \end{cases} \quad (7)$$

where N is the window length. When α = 0.5 the window is called a Hanning window, whereas an α of 0.46 is a Hamming window.

Step 1: To properly frame a signal, the traditional method requires two parameters: window length and step size. Given a speech utterance with window size N and step size M, the utterance is framed by the following steps:

1. Start the first frame from the beginning of the utterance, thus it is centered at the $N^{th}/2$ sample;

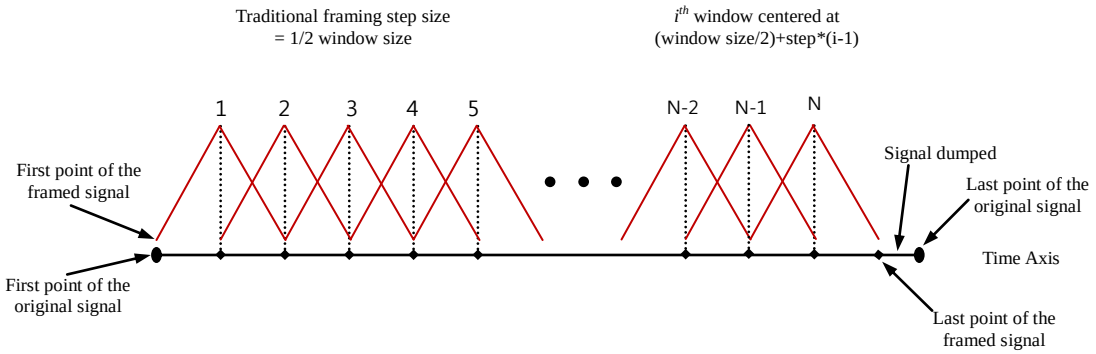

Fig. 2 The framing method

2. Move the frame forward by M points each time until it reaches the end of the utterance. Thus the i^{th} frame is centered at the $(i-1)\times M + N^{th}/2$ sample;
3. Dump the last few sample points if they are not long enough to construct another full frame.

In this case, if the window length N is changed, the total number of frames may change, even when using the same step size M. Note that this change prevents the possibility of combining feature vectors with different frame sizes. Figure 2 shows the framing method.

The next steps are applied to every single frame. First, a set of 12 MFCC coefficients is extracted for each frame. Assume that the time domain signal is called $S(n)$, then once it is framed we have $S_i(n)$ where n ranges over 1to number of samples and i ranges over the number of frame.

Step 2: The periodogram-based power spectral estimate for the speech frame is given by (8):

$$P_i(k) = \frac{1}{N}|S_i(k)|^2 \qquad (8)$$

where $S_i(k)$ is performed by taking the Discrete Fourier Transform of the frame as in (9):

$$S_i(k) = \sum_{n=1}^{N} s_i(n)h(n)e^{-j2\pi kn/N} \quad 1 \le k \le K \qquad (9)$$

where $h(n)$ is an N sample long analysis window (e.g. Hamming window), and K is the length of the Discrete Fourier Transform (DFT).

Step 3: Compute the Mel-spaced filterbank. This is a set of triangular filters that we apply to the periodogram power spectral estimate from step 2.

Step 4: Take the log of each of the energies from step 3. This leaves us with log filterbank energies.

Step 5: Take the Discrete Cosine Transform (DCT) of the log filterbank energies to give cepstral coefficents. Keep coefficients 2-13, discard the rest.

C. Classifying speech emotions

We have chosen about two-third of 500 sentences for training, the remaining sentences for recognition. The experimental process is implemented as Fig. 3 and the experimental results are shown in table 1.

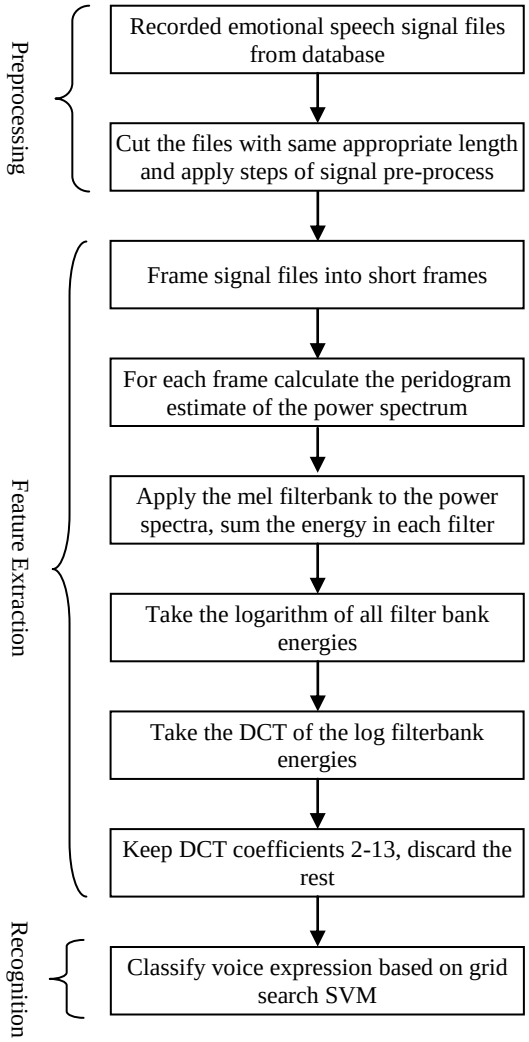

Fig. 3 Experimental process

TABLE I
CONFUSION MATRIX OF BERLIN EMO-DB DATABASE

	Anger (%)	Boredom (%)	Disgust (%)	Fear (%)	Joy (%)	Neutral (%)	Sadness (%)
Anger	**92.68**	2.44	0.00	2.44	2.44	0.00	0.00
Boredom	3.85	**73.08**	0.00	3.85	0.00	15.38	3.85
Disgust	14.29	14.29	**42.86**	14.29	0.00	0.00	14.29
Fear	15.00	5.00	0.00	**55.00**	5.00	10.00	10.00
Joy	54.55	0.00	0.00	0.00	**45.45**	0.00	0.00
Neutral	0.00	36.00	0.00	4.00	0.00	**56.00**	4.00
Sadness	0.00	15.79	0.00	0.00	0.00	0.00	**84.21**
						Average:	**64.18**

VI. CONCLUSIONS

The paper presents a method of speech emotion recognition based on mel-frequency cepstral coefficients and the classification of support vector machine. There are three ways to improve the results of speech emotion recognition. First is changing parameters of MFCC feature to find the better result. Second is associating some appropriate features for speech emotion recognition. The final is finding a better classifier for speech emotion recognition.

ACKNOWLEDGMENT

This research was supported by the Seoul R&BD program (SS110013).

We would like to thank Professor F. Burkhardt et al. for the use of Berlin Emo-DB database,

REFERENCES

[1] Dimitrios Ververidis and Constantine Kotropoulos, *Emotional speech recognition: Resources, features, and methods*, Speech Communication 48, pp. 1162-1181, Elsevier Ltd., 2006

[2] Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray, *Survey on speech emotion recognition: Features, classification schemes, and databases*, Pattern Recognition 48, pp. 572-587, Elsevier Ltd., 2011.

[3] M. Schubiger, *English intonation: its form and function*, Niemeyer, Tubingen, Germany, 1958.

[4] J. O'Connor and G. Arnold, *Intonation of Colloquial English*, second ed., Longman, London, UK, 1973.

[5] P. Ekman, *Are there basic emotions*, Psychol. Rev. 99 (3), pp. 550–553, 1992.

[6] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, S. Kollias, W. Fellenz and J. Taylor, *Emotion recognition in human–computer interaction*, IEEE Signal Process. Mag. 18, pp. 32–80, 2001.

[7] P. Mermelstein, *Distance measures for speech recognition, psychological and instrumental*, in Pattern Recognition and Artificial Intelligence, C. H. Chen, Ed., pp. 374–388. Academic, New York, 1976.

[8] S.B. Davis and P. Mermelstein, *Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences*, in IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4), pp. 357–366, 1980.

[9] Stanley Smith Stevens, John Volkman and Edwin B. Newman, *A scale for the measurement of the psychological magnitude pitch*, Journal of the Acoustical Society of America 8 (3), pp. 185–190, 1937.

[10] Fang Zheng, Guoliang Zhang and Zhanjiang Song, *Comparison of Different Implementations of MFCC*, Journal on Computer Science & Technology, 16(6): 582–589, 2001.

[11] T. Ganchev, N. Fakotakis and G. Kokkinakis, *Comparative evaluation of various MFCC implementations on the speaker verification task*, in 10th International Conference on Speech and Computer, Vol. 1, pp. 191–194, 2005.

[12] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[13] Corinna Cortes and Vladimir Vapnik, *Support-Vector Networks*, Machine Learning, 20, pp. 273-297, 1995.

[14] Vladimir N. Vapnik, *An Overview of Statistical Learning Theory*, IEEE Transactions on Neural Network, Vol. 10, pp. 988-999, 1999.

[15] Chen-Chiung Hsieh and Meng-Kai Jiang, *A Facial Expression Classification System based on Active Shape Model and Support Vector Machine*, IEEE International Symposium on Computer Science and Society, pp. 311–314, 2011.

[16] Irene Kotsia and Ioannis Pitas, *Facial expression recognition in image sequences using geometric deformation features and support vector machines*. IEEE Trans on Image Process, vol. 16, no. 1, pp. 172-187, 2007.

[17] Caifeng Shan, Shaogang Gong and Peter W. McOwan, *Facial expression recognition based on Local Binary Patterns: A comprehensive study*, Image and Vision Computing 27, pp. 803-816, 2009.

[18] Chih-Wei Hsu, Chih-Chung Chang and Chih-Jen Lin, *A Practical Guide to Support Vector Classification*, Tech. Rep., Taipei, 2003.

[19] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier and B. Weiss, *A database of German emotional speech*, http://pascal.kgw.tu-berlin.de/home/publications/p1078.pdf

[20] http://www.expressive-speech.net/emoDB/.

