Detecting Emotions in Mandarin Speech

Tsang-Long Pao, Yu-Te Chen, Jun-Heng Yeh and Jhih-Jheng Lu

Department of Computer Science and Engineering, Tatung University, Taipei

tlpao@ttu.edu.tw, d8906005@mail.ttu.edu.tw, g9106004@ms2.ttu.edu.tw

Abstract. In this paper, a Mandarin speech based emotion classification method is presented. Five primary human emotions including anger, boredom, happiness, neutral and sadness are investigated. For speech emotion recognition, we select 16 LPC coefficients, 12 LPCC components, 16 LFPC components, 16 PLP coefficients, 20 MFCC components and jitter as the basic features to form the feature vector. Two text-dependent and speaker-independent corpora are employed. The recognizer presented in this paper is based on three recognition techniques: LDA, K-NN, and HMMs. Results show that the selected features are robust and effective in the emotion recognition at the valence degree in both corpora. For the LDA emotion recognition, the highest accuracy of 79.9% is obtained. For the K-NN emotion recognition, the highest accuracy of 84.2% is obtained. And for the HMMs emotion recognition, the highest accuracy of 88.7% is achieved.

1 Introduction

Various opinions of emotions proposed by more than 100 scholars are summarized in a classical article [1]. Research on the cognitive component focuses on understanding the environmental and attended situations that gives rise to emotions; research on the physical components emphasizes the physiological response that co-occurs with an emotion or rapidly follows it. In short, emotions can be considered as communications to oneself and others [1]. Emotions consist of behaviors, physiologic changes and subjective experience as evoked by individual's thoughts, socio-cultures and so on.

Emotions are traditionally classified into two main categories: primary (basic) and secondary (derived) emotions [2]. Primary or basic emotions generally could be experienced by all social mammals (e.g. humans, monkeys, dogs, whales) and have particular manifestations associated with them (e.g. vocal/facial expressions, behavioral tendencies, and physiological patterns). Secondary or derived emotions are the combination or derivation from primary emotions.

Emotional dimensionality is a simplified description of basic properties of emotional states. According to Osgood, Suci and Tannenbaum's theory [3] and subsequent psychological research [4], the computing of emotions is conceptualized as three major dimensions of connotative meaning, arousal, valence and power. In general, the arousal and valence dimensions can be used to distinguish most basic emotions. The emotions location in arousal-valence space is shown in Fig. 1 [3], which results in a representation that is both simple and capable of conforming to wide emotional applications.

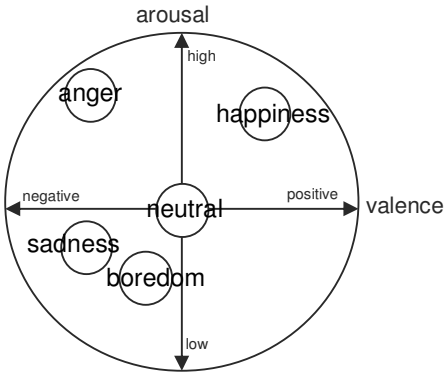

Fig. 1. Graphic representation of the arousal-valence theory of emotions

Table 1. Emotions and speech relations

	Anger	Happiness	Sadness	Fear	Disgust
Speech Rate	Slightly faster	Faster or slower	Slightly slower	Much faster	Very much faster
Pitch Average	Very much higher	Much higher	Slightly lower	Very much higher	Very much lower
Pitch Range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Higher	Higher	Lower	Normal	Lower
Voice Quality	Breathy, chest	Breathy, blaring tone	Resonant	Irregular voicing	Grumble chest tone
Pitch changes	Abrupt on stressed	Smooth, upward inflections	Downward inflections	Normal	Wide, downward terminal inflects
Articulation	Tense	Normal	Slurring	Precise	Normal

There are numerous literatures that indicate emotion on the signs within the psychological tradition and beyond [1-2, 5-13]. The vocal cue is one of the fundamental expressions of emotions [1-2, 5-9, 11, 13]. All mammals can convey emotions by vocal cues. Humans are especially capable of expressing their feelings by crying, laughing, shouting and more subtle characteristics from speech. In ordinary conversation, the emotive cues communicate readily arousal. The communication of valence is believed to be by more subtle cues, intertwined with the content of the speech.

An important research is accomplished by Murray and Arnott [2], whose result particularizes several notable acoustic attributes for detecting primary emotions. Table 1 summarizes the vocal effects most commonly associated with five primary emotions. Classification of emotional states on basis of the prosody and voice quality requires classifying the connection between acoustic features in speech and the emotions. Specifically, we need to find suitable features that can be extracted and models it for use in recognition. This also implies the assumption that voice carries abundant information about emotional states by the speaker.

To estimate a user's emotions by the speech signal, one has to carefully select suitable features. All selected features have to carry information about the transmitted emotion. However, they also need to fit the chosen model by means of classification algorithms. A large number of speech emotion recognition methods adapt prosody and energy related features. For example, Schuller *et al.* chose 20 pitch and energy related features [14]. A speech corpus consisting of acted and spontaneous emotion utterances in German and English language is described in detail. Accuracy in the recognition of 7 discrete emotions (anger, disgust, fear, surprise, joy, neutral, sad) exceeded 77.8%. As a comparison, the similar judgment of human deciders classifying the same corpus at 81.3% recognition rate was reported. Park *et al*. used pitch, formant, intensity, speech speed and energy related features to classify neutral, anger, laugh, and surprise emotions [7]. The recognition rate is about 40% in a 40-sentence corpus. Yacoub *et al*. extracted 37 fundamental frequency, energy and audible duration features to recognize sadness, boredom, happiness, and cold anger emotions in a corpus recorded by eight professional actors [15]. The overall accuracy was only about 50%. But these features successfully separated hot anger from other basic emotions. However, in this experiment, the accuracy obtained from a 15 emotions recognition result is only 8.7%. The accuracy is 63% for male voice and 73.7% for female voice. Tato *et al.* extracted prosodic features, derived from pitch, loudness, duration, and quality features [19] from a 400-utterance database. The most important results achieved are for the speaker-independent case and three clusters (high = anger/happy, neutral, low = sad/bored). The recognition rate is close to 80%. However, the recognition accuracy of five emotions is only 42.6%. Kwon *et al.* selected pitch, log energy, formant, band energies, and Mel frequency spectral coefficients (MFCC) as the base features, and added velocity/acceleration of pitch to form feature streams [12]. The average classification accuracy was 40.8% in a SONY AIBO database. Nwe *et al.* proposed the short time log frequency power coefficients (LFPC) accompanying MFCC as emotion speech features to recognize 6 emotions in a 60-utterance corpus involving 12 speakers [13]. Results show that the proposed system yields an average accuracy of 78%.

According to the experimental results stated previously, the vocal features related prosody and energy that were extracted from time domain seem not stable in distinguishing all primary emotions. Furthermore, the prosodic features between female and male are obviously intrinsic in speech. Simple speech energy feature calculation method is also unconformable to human auricular perception.

In this paper, we make efforts on searching for an effective and robust set of vocal features from Mandarin speech to recognize emotional categories rather than modifying the classifiers. The vocal characteristics of emotions are extracted from a spontaneous Mandarin corpus. In order to surmount the inefficiency of conventional vocal features in recognizing anger/happiness and boredom/sadness valence emotions, we also treat arousal and valence correlated characteristics to categorize emotions in the emotional discrete categories. Several systematic experiments are presented. The characteristic of the extracted features is expected not only facile, but also discriminative.

Table 2. Utterances of Corpus I

Emotion \ Sex	Female	Male	Total
Anger	75	76	151
Boredom	37	46	83
Happiness	56	40	96
Neutral	58	58	116
Sadness	54	58	112
Total	280	278	558

Table 3. Utterances of Corpus II

Emotion \ Sex	Female	Male	Total
Anger	36	72	108
Boredom	72	72	144
Happiness	36	36	72
Neutral	36	36	72
Sadness	72	35	107
Total	252	251	503

The rest of this paper is organized as follows. In Section 2, two testing corpora are addressed. In Section 3, the details of the proposed system are presented. Experiments to assess the performance of the proposed system are described in Section 4 together with analysis of the results of the experiments. The concluding remarks are presented in Section 5.

2 The Testing Corpora

An emotional speech database, Corpus I, is specifically designed and set up for speaker-independent emotion classification studies. The database includes short utterances coveting the five primary emotions, namely anger, boredom, happiness, neutral, and sadness. Non-professional speakers are selected to avoid exaggerated expression. Twelve native Mandarin language speakers (7 females and 5 males) are employed to generate 558 utterances as described in Table 2. The recording is done in a quiet environment using a mouthpiece microphone at 8k Hz sampling rate.

All native speakers are asked to speak each sentence in the chosen five emotions, resulting in 1200 sentences. First, we eliminated the sentences involved excessive nose. Then a subjective assessment of the emotion speech corpus by human audiences was carried out. The purpose of the subjective classification is to eliminate the ambiguous emotion utterances. Finally, 558 utterances were selected over 80% human judgment accuracy rate. In this paper, utterances in Mandarin are used due to an immediate availability of native speakers of the languages. It is easier for the speakers to express emotions in their native language than in a foreign language. In order to accomplish the computing time requisition and bandwidth limitation of the practical recognition application, e.g. the call center system [15], the low sampling rate, 8k Hz, is adopted.

Another corpus, Corpus II, was obtained from [17]. Two professional Mandarin speakers are employed to generate 503 utterances with five emotions as listed in Table 3. The sampling rate is down-sampled to 8k Hz.

3 Emotion Recognition Method

The proposed emotion recognition method has three stages: feature extraction, feature vector quantization and classification. Base features and statistics are computed in feature extraction stage. Feature components are quantized as a feature vector in feature quantization stage. Classification is made by using various classifiers based on dynamic models or discriminative models.

3.1 The Selected Features

Fig. 2 shows the block diagram of feature extraction. In pre-processing procedure, locating the endpoints of the input speech signal is done first. The speech signal is high-pass filtered to emphasize the important higher frequency components. Then the speech frame is partitioned into frames of 256 samples. Each frame is

overlapped with the adjacent frames by 128 samples. The next step is to apply Hamming window to each individual frame to minimize the signal discontinuities at the beginning and end of each frame. Each windowed speech frame is then converted into several types of parametric representation for further analysis and recognition.

Most effective features in speech processing are found in the frequency domain. The speech signal is more consistently and easily analyzed spectrally in the frequency domain than in the time domain. And the common model of speech production corresponds well to separate spectral models for the excitation and the vocal tract. The hearing mechanism appears to pay much more attention to spectral magnitude than to phase or timing aspects. For these reasons, the spectral analysis is used primarily to extract relevant features of the speech signal in this paper.

In base feature extraction procedure, we select 6 features, which are 16 Linear predictive coding (LPC) coefficients, 12 linear prediction cepstral coefficients (LPCC), 16 log frequency power coefficients (LFPC), 16 perceptual linear prediction (PLP) coefficients, 20 Mel-frequency cepstral coefficients (MFCC) and jitter extracted form a frame. LPC provides an accurate and economical representation of the envelope of the short-time power spectrum of speech [18]. For speech emotion recognition, LPCC and MFCC are the popular choices as features representing the phonetic content of speech [19-20]. LFPC is calculated from a log frequency filter bank which can be regarded as a model that follows the varying auditory resolving power of the human ear for various frequencies [13]. The combination of the discrete Fourier transform (DFT) and LPC technique is PLP [21]. PLP analysis is computationally efficient and permits a compact representation. Perturbations in the pitch period are called jitter, such perturbations occur naturally during continuous speech.

3.2 Feature Vector Quantization

To further compress the data for presentation to the final stage of the system, vector quantization is performed. The division into 16 clusters is carried out according to the Linde-Buzo-Gray (LBG) algorithm. The vector f_n is assigned the codeword c_n^* according to the best match codebook cluster z_c using

$$c_n^* = \arg\min_{1 \le c \le C} d(f_n, z_c) \tag{1}$$

For a speech utterance with *N* frames, the feature vector Y_1 with 16 parameters is then obtained as

$$Y_1 = [c_1^* c_2^* ... c_N^*] \tag{2}$$

In another simple vector quantization method, we treat the mean feature parameters corresponding to each frames as a feature vector Y_2. Therefore, another feature vector Y_2 with 81 parameters is then obtained.

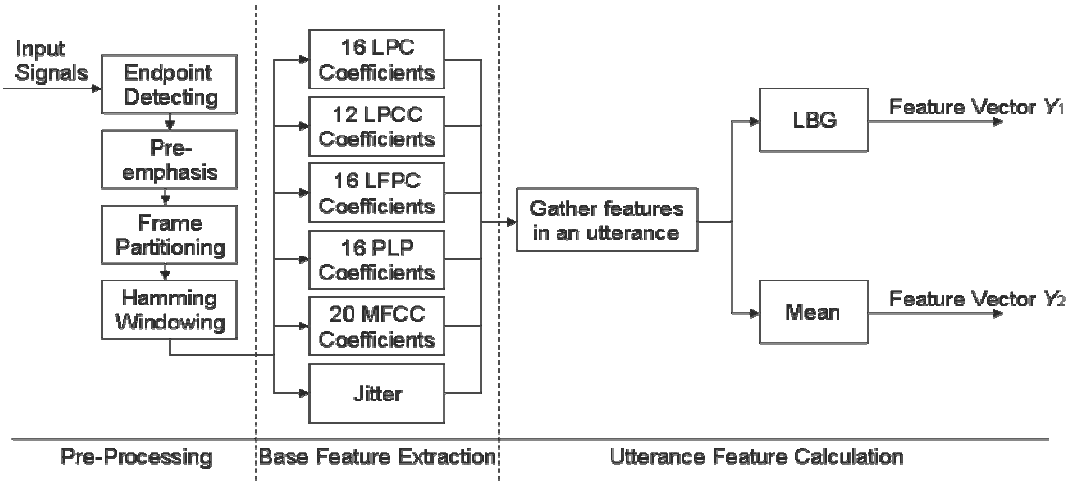

Fig. 2. Block diagram of the feature extraction module

3.3 Classifiers

Three different classifiers, linear discriminate analysis (LDA), k-nearest neighbor (K-NN) decision rule, and Hidden Markov models (HMMs), are selected to train and test these two testing emotion corpora with the extracted features from Corpus I. In K-NN decision rule, there are three nearest samples closest to the testing sample. In HMMs, our experimental studies show that a 4-state discrete ergodic HMM gives the best performance compared with the left-right structure. The state transition probabilities and the output symbol probabilities are uniformly initialized.

4 Experimental Results

The selected features in Section 3.1 will be quantified as the LBG feature vector Y_1 and the mean feature vector Y_2. Then the feature vectors will be trained and tested with three different classifiers, which are LDA, K-NN and HMMs. All these experimental results are validated by the leave-one-out (LOO) cross-validation method.

4.1 The Experimental Results Using the Conventional Prosodic Features

In [9], Kwon *et al.* drawled a two-dimensional plot of 59 features ranked by forward selection and backward elimination. Features near origin are considered to be more important. By imitating the ranking features method as [9], the speech features extracted from Corpus I are ranked by forward selection and backward elimination in Fig. 3. The experimental results of this Mandarin experiment and Kwon's show that the pitch and energy related features are the most important components for the emotion speech recognition in both Mandarin and English. We select the first 15 features proposed by [9] from Corpus I to examine the efficiency and stability of the conventional emotion speech features. The first 15 features are pitch, log energy, F1, F2, F3, 5 filter bank energies, 2 MFCCs, delta pitch, acceleration of pitch, and 2 acceleration MFCCs. Then the feature vector Y_2 and K-NN are used.

The accuracy rate of confusion matrix using conventional emotion speech features is shown in Table 4. The overall average accuracy rate of five primary emotions is 53.2%. As most previous surveyed experimental results and discussion, the pitch and energy related features extracted form the time domain confuse in anger and happiness valence emotions. The reason is that anger and happiness are close to each other in the pitch and energy related speech features; hence the classifiers often confuse one for the other. This also applies to boredom and sadness.

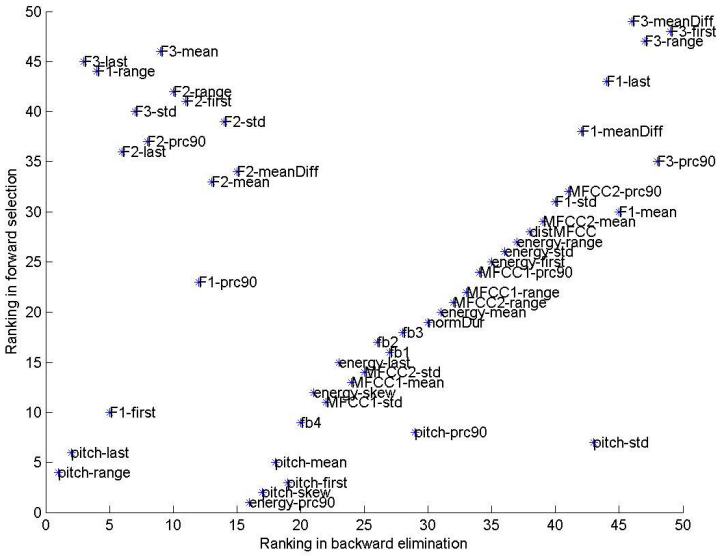

Fig. 3. Conventional emotional speech features ranking

Table 4. The experimental result of conventional prosodic features

Accuracy (%)	Anger	Boredom	Happiness	Neutral	Sadness
Anger	**59.5**	1.1	34.4	4.4	2.6
Boredom	0	**46.8**	1.1	20.4	31.7
Happiness	32.4	2.5	**58.7**	4.2	2.2
Neutral	9.4	7.7	8.7	**52.1**	22.1
Sadness	1.7	29.4	2.4	17.6	**48.9**

Table 5. The experimental result of anger and happiness recognition

Accuracy (%)	LDA		K-NN		HMMs	
	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
Anger	93.1	93.4	93.7	91.6	93.9	92.6
Happiness	87.7	91.2	90.4	92.8	91.2	93.5
Average	90.4	92.3	92.0	92.2	92.5	93.0

Table 6. The experimental result of boredom and sadness recognition

Accuracy (%)	LDA		K-NN		HMMs	
	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
Boredom	89.5	90.5	89.7	92.1	90.5	94.3
Sadness	92.2	87.6	93.5	90.4	93.2	90.9
Average	90.8	89.0	91.6	91.0	91.8	92.6

4.2 Experimental Results of Valence Emotions Recognition

The prosodic features as pitch and energy related speech features are failed to distinguish the valence emotions. The selected features in Section 3.1 will be quantified as the LBG feature vector Y_1 and the mean feature vector Y_2. Then the feature vectors will be trained and tested in Corpus I with three different classifiers, which are LDA, K-NN and HMMs. All the experimental results are validated by the LOO cross-validation method. According to experimental results shown in Table5 and 6, by applying the set of our selected emotion speech features, three recognizers are undoubted to separate the anger and happiness which most previous emotion speech recognizers are always confuse in this emotion cluster. In addition, as shown in Table 5 and 6, the high

and stable accuracy rate of various recognizers with two feature vector quantization methods provides the appropriateness to distinguish the emotions at the valence degree.

These pairwise emotions, anger and happiness, are considered to be close to each other at the valence degree with the similar prosody and amplitude. So do boredom and sadness. Conventional speech emotion recognition method suffers the infectiveness and instability in emotion recognition, especially involving emotions at the same valence degree. On the contrary, the proposed selected features solve the problem and obtain high recognition accuracy. The set of selected features are not only suitable for various classifiers but also effective for the speech emotion recognition.

4.3 Experimental Results of Corpus I and Corpus II

Table 7 and 8 show the accuracy of five primary emotions classified by various classifiers with two feature vector quantified methods in Corpus I and II. The different classifiers have different ability and property, and then we have the different recognition rates in each classifier or quantization method.

According to the experimental results shown in Table 7 and 8, the accuracy overall five primary emotions, which are anger, boredom, happiness, neutral and sadness, is approximately equivalent with the same classifier. In addition, the accuracy of two feature quantization methods of LBG and mean is quite close to each other in the same conditions. This shows that the set of the selected speech features is stable and suitable to recognize the five primary emotions in various classifiers with different feature quantization methods. By this high recognition rate of the experimental results in Corpus I and II, the selected features are proofed to be efficient to directly classify five primary emotions of arousal and valence degree simultaneously rather than only arousal degree.

Table 7. Experimental result of five emotion classes in Corpus I

Accuracy (%)	**LDA**		**K-NN**		**HMMs**	
	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
Anger	81.5	80.4	82.3	84.8	86.4	86.7
Boredom	80.3	79.8	84.9	82.3	89.1	88.4
Happiness	76.5	72.3	79.5	82.1	82.3	83.6
Neutral	78.4	80.5	80.4	81.2	84.5	90.5
Sadness	82.5	81.3	91.2	89.1	92.4	92.3
Average	79.8	78.8	83.6	83.9	86.9	88.3

Table 8. Experimental result of five emotion classes in Corpus II

Accuracy (%)	**LDA**		**K-NN**		**HMMs**	
	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
Anger	82.4	76.2	83.2	84.5	90.2	91.4
Boredom	78.9	80.2	81.5	80.9	84.3	86.7
Happiness	81.4	77.8	86.4	82.5	87.5	88.1
Neutral	76.5	79.8	84.1	83.2	90.3	86.0
Sadness	80.3	76.5	86.0	87.5	89.5	91.5
Average	79.9	78.1	84.2	83.7	88.3	88.7

Two different corpora are involved to validate the robustness and effectiveness of the selected features that the conventional speech emotion recognition method is difficult to accomplish. As the relative experimental results shown in Table 7 and 8, the overall recognition rates of both corpora are similar. The proposed selected features solve the thorny problem and obtain a high accuracy recognition rate. The set of selected features are not only suitable for various classifiers but also effective for the recognition outperform in different corpora.

5 Conclusion

In conventional emotion classification of speech signals, the popular features employed are fundamental frequency, energy contour, duration of silence and voice quality. However, some recognizers employing these

features confuse in the recognition of the valence emotions. In addition, these features employed in different corpora reveal the instable recognition results of the same recognizer.

In this paper, we use 16 LPC coefficients, 12 LPCC components, 16 LFPC components, 16 PLP coefficients, 20 MFCC components and jitter as featuers, and LDA, K-NN, HMMs as the classifiers. Presentation of the selected feature parameters is quantified as a feature vector using LBG and mean methods. The emotions are classified into five human primary categories. The emotional category labels used are anger, boredom, happiness, neutral and sadness. Two Mandarin corpora, one consisting of 558 emotional utterances employed 12 native speakers and the other consisting of 503 emotional utterances employed 2 professional speakers, are used to train and test in the proposed recognition system. Results show that the proposed system yields the best accuracy of 88.3% for Corpus I and 88.7% for Corpus II to classify five emotions.

According to experimental outcomes, we attain a high accuracy rate to distinguish anger/happy or bored/sad emotions that have similar prosody and amplitude. The proposed method can solve the difficult of recognizing the valence emotions using the set of extracted features. Moreover, the recognition accuracy of the experimental results of Corpus I and II shows that the selected speech features are suitable and effective in different corpora for the speech emotion recognition.

Further improvements and expansions may be achieved by using one or more of the following suggestions:

A possible approach to extract non-textual information to identify emotional state in speech is to apply various different and known feature extraction methods. So we may integrate other features into our system to improve emotion recognition accuracy. Besides, recognizing the emotion translation in real human communication is an arduous challenge in this field. We will try to find out the point where the emotion transition occurs

6 Acknowledge

A part of this research is sponsored by NSC 93-2213-E-036-023.

References

[1] P.R. Kleinginna and A.M. Kleinginna, “A Categorized List of Emotion Definitions with Suggestions for a Consensual Definition,” Motivation and Emotion, pp. 345-379, 1981.

[2] I. Murray and J.L. Arnott, “Towards the Simulation of emotion in Synthetic Speech: A review of the Literature on Human Vocal Emotion,” *Journal of the Acoustic Society of America*, pp. 1097-1108, 1993.

[3] C.E. Osgood, J.G. Suci and P.H. Tannenbaum, *The Measurement of Meaning*, University of Illinois Press, pp. 31-75, 1957.

[4] A. Mehrabian and J. Russel, *An Approach to Environmental Psychology*, Cambridge MA: MIT Press, pp. 192-203, 1974.

[5] A. Pasechke and W.F. Sendlmeier, “Prosodic Characteristics of Emotional Speech: Measurements of Fundamental Frequency Movements,” In SpeechEmotion-2000, pp.75-80, 2000.

[6] C.D. Park and K.B. Sim, “Emotion Recognition and Acoustic Analysis from Speech Signal,” Proceedings of IJCNN, pp. 2594-259, 2003.

[7] C.H. Park, K.S.Heo, D.W.Lee, Y.H.Joo and K.B.Sim, ”Emotion Recognition based on Frequency Analysis of Speech Signal,” International Journal of Fuzzy Logic and Intelligent Systems, pp. 122-126, 2002.

[8] H. Holzapfel, C. Fügen, M. Denecke and A. Waibel, “Integrating Emotional Cues into a Framework for Dialogue Management,” Proceedings de International Conference on Multimodal Interfaces, pp.141-148, 2002.

[9] O.W. Kwon, K. Chan, J. Hao, T.W. Lee , “Emotion Recognition by Speech Signals,” Eurospeech, pp.125-128, 2003. [10][13] P. Ekman, *Handbook of Cognition and Emotion*, New York: John Wiley & Sons Ltd, 1999.

[11] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J.G. Taylor, “Emotion Recognition in Human-Computer Interaction,” IEEE Signal Proc. Mag., 18(1), pp. 32-80, 2000.

[12] R.W. Picard, *Affective Computing*, MIT Press, Cambridge, pp. 178-192, 1997.

[13] T.L. Nwe, S.W. Foo and L.C. De Silva, “Speech Emotion Recognition Using Hidden Markov Models,” *Speech Communication*, pp. 603-623, 2003.

[14] B. Schuller, G. Rigoll, and M. Lang, “Hidden Markov Model-based Speech Emotion Recognition,” Proceedings of IEEE-ICASSP, pp. 401-405, 2003.

[15] S. Yacoub, S. Simske, X. Lin, J. Burns, “Recognition of Emotions in Interactive Voice Response Systems,” Eurospeech, HPL-2003-136, 2003.

[16] R.S. Tato, R. Kompe, J.M. Pardo., “Emotional Space Improves Emotion Recognition,” ICSLP, pp. 2029-2032, 2002.
[17] 張柏雄, “中文語音情緒之自動辨識,” master thesis of Engineering Science department, National Cheng Kung University, 2002.
[18] J.F. Kaiser, *Discrete-Time Speech Signal Processing*, pp.11-99, Prentic Hall PTR, 2002.
[19] B.S. Ata, “Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification,” *Journal of the Acoustical Society of America*, pp.1304-1312, 1974.
[20] S.B. Davis and P. Mermelstein, “Comparison of Parametric Representations of Monosyllabic Word Recognition in Continuously Spoken Sentences,” IEEE Transactions on Acoustics, Speech, and Signal Processing, pp. 357-366, 1980.
[21] H. Hermansky. “Perceptual linear predictive (PLP) analysis of speech,” *Journal of the Acoustical Society of America*, pp.1738-1752, 1990.