Emotion Recognition from Mandarin Speech

包蒼龍[1]、陳育得[2]、葉俊亨[3]、張原豪[4]

[1]大同大學資訊工程學系副教授兼系主任

[2]大同大學資訊工程學系研究所博士候選人

[3]大同大學資訊工程學系研究所博士班學生

[4]大同大學資訊工程學系研究所碩士班學生

104 台北市中山區中山北路 3 段 40 號

電話：02-2592-5252#3294　Fax：02-2594-1371

E-mail：[1]tlpao@ttu.edu.tw, ([2]d8906005, [3]d9306002, [4]g9206026)@ms2.ttu.edu.tw

摘要

在本論文中，我們對人類的五種基本情緒，包含生氣、厭倦、快樂、平常及悲傷，提出一套中文語音情緒的辨識方法。實驗結果顯示，我們所選擇的語音特徵參數對於辨識語者相關及語者無關的喚醒維度和激發維度情緒種類都是強健且有效的。在 LDA 的分類方式中，語者相關的正確辨識率可達到 85.1%。在 K-NN 的分類方式中，語者相關的正確辨識率可達到 89.2%。在 HMMs 的情緒分類方式中，語者相關的正確辨識率可達到最高的 90.6 %。此外，在語者無關的實驗中，使用 HMMs 的正確辨識率可達到 88.3%。

關鍵字：語音情緒辨識，線性預測編碼，梅爾倒頻譜參數，感知線性預測

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. Results show that the selected features are robust and effective in the emotion recognition not only at the arousal degree but also at the valence degree in the speaker-dependency and speaker-independency. In speaker-dependent experimental results, for the linear discriminate analysis (LDA), the accuracy of 85.1% is obtained. For the k nearest neighbor (K-NN) decision rule, the accuracy of 89.2% is obtained. And For the Hidden Markov models (HMMs), the highest accuracy of 90.6% is achieved. Besides, in the speaker-independent experimental results, the accuracy of 88.3% is obtained with HMMs.

Keywords：Speech Emotion Recognition, LPC, MFCC, PLP

1 Introduction

Many researchers in the area of speech technology during the last decade have worked on different aspects of emotions in speech. In recent years, the interest for automatic detection and interpretation of emotions in speech has grown [1-7]. One of the goals is to recognize immediately the emotional states of a speaker. Such science and technology might improve existing systems, e.g. dialogue/expert systems, but also can be a decisive factor to develop new applications, for instance, aid devices for disabled people. In [5] an "Emotion Recognition Game (ERG)" was developed. The program was designed for a user to compete against the computer or another person to see who can better recognize emotions in recorded utterances. One practical application of the game is to help learning disability people in developing better emotional skills at recognizing emotion in speech. Some clinical diagnoses also depend on detecting vocal signs of emotions, which supplement the subjective impressions of psychiatrists with relevant objective measures. Furthermore, emotion recognition can also be used to make judgments about another person in a more accurate or objective way.

Classification of emotional states on basis of the prosody and voice quality requires classifying acoustic features in speech as connected to certain emotions. However, automatic detection of emotions is considered to be the most difficult part in the emotional recognition and synthesis research. The reason is that those researches have to handle spontaneous speech as input. Specially, we need to find suitable features that the recognition methods can extract and model. This also implies the assumption that voice carries abundant information about emotional states by the speaker. Before the unambiguous agreement of the speech characteristics of emotions is settled, the effective feature extraction is usually considered as a more critical factor in emotional recognition than classifier improvement [3]. Moreover, the practical experience to recognize emotions from Mandarin speech is extraordinarily deficient.

In this paper, we make efforts on searching for an effective and robust set of vocal features from Mandarin speech to recognize five basic emotional categories, rather than modifying the classifiers. The vocal characteristics of emotions are extracted from a spontaneous Mandarin corpus [7].

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

2 EMOTION RECOGNITION METHOD

Two different corpora are involved to validate the robustness and effectiveness of the selected features. Besides, the procedure of the features extraction and the architecture of speech emotion recognition will be discussed in detail.

2.1 Emotional Mandarin Corpora

Two emotion speech corpora are specifically designed and set up for speaker-independent emotion classification studies. The corpora include short utterances covering the five emotions, namely anger, boredom, happiness, neutral, and sadness.

Corpus I was obtained from [7], including short utterances coveting the five primary emotions. Two professional Mandarin speakers are employed to generate 503 utterances with five emotions as listed in Table 1. In addition, non-professional speakers are selected to avoid exaggerated expression in another emotional speech database, Corpus II. 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. First we obtained 1200 emotive utterances. Then a subjective assessment of the emotion corpus by human listeners was carried out. The objective of the subjective classification is to eliminate the ambiguous emotive utterances. In order to accomplish the computing time requisition and bandwidth limitation of the practical recognition application, e.g. the call center system [6], the low sampling rate, 8k Hz, is adopted.

Table 1: Utterances of Corpus I

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

Table 2: Utterances of Corpus II

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

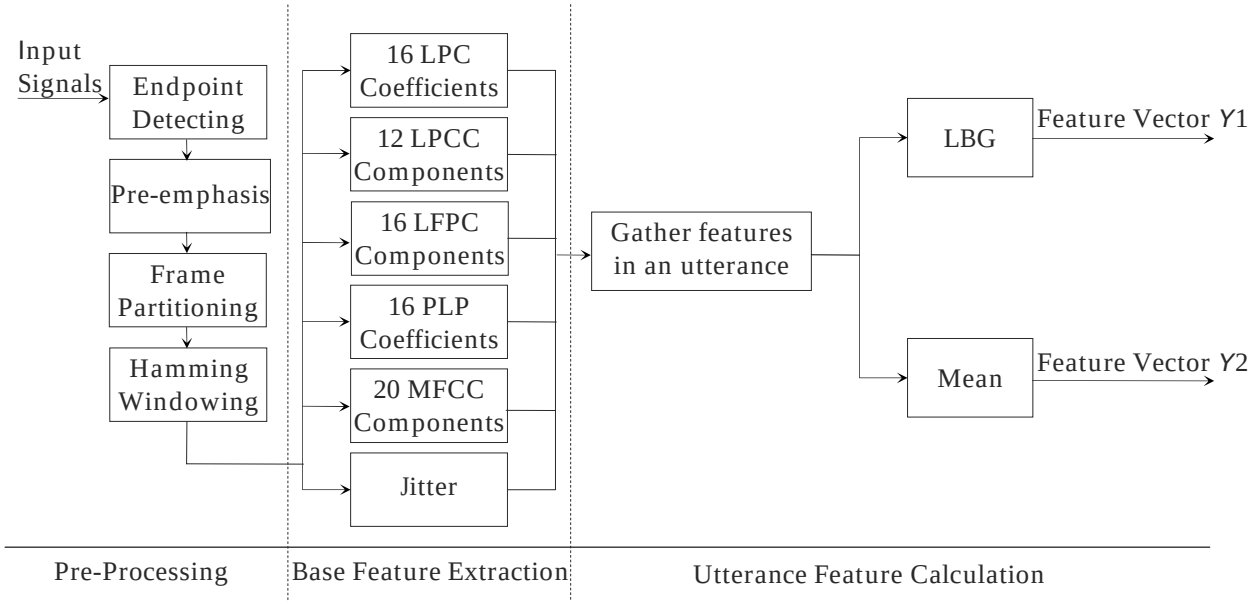

Figure 1. Block diagram of the feature extraction module

2.2 Emotion Recognition Procedure

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. Figure 1 shows the block diagram of feature extraction.

In pre-processing procedure, locating the endpoints of the input speech signal is first done. 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 by 128 samples. The next step is to apply a window function to each individual frame to minimize the signal discontinuities at the beginning and end of each frame. The Hamming window is used. Each windowed speech frame is then converted into some type of parametric representation for further analysis and recognition.

In order to find out a suitable combination of all extracted features, we make use of the forward selection to decide a benefit set of features from more than 200 speech features. Six features, 16 levels Linear predictive coding (LPC) coefficients [4], 12 levels linear prediction cepstral coefficients (LPCC), 16 levels log frequency power coefficients (LFPC), 16 levels perceptual linear prediction (PLP) coefficients, 20 levels Mel-frequency cepstral coefficients (MFCC) and jitter, are extracted from a frame.

To further compress the data for presentation to the final stage of the system, vector quantization is performed. Division into 16 clusters is carried out according to the Linde-Buzo-Gray (LBG) algorithm. The feature vector Y_1 with 16 parameters is then

obtained. 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 obtained.

3 EMOTION RECOGNITION RESULTS

The Corpus I and Corpus II will be trained and tested with three different classifiers, which are LDA, K-NN decision rule, and HMMs. All the experimental results are validated by the Leave-One-Out (LOO) cross-validation method.

3.1 The Experimental Results of Speaker-Independency

Table 3 and 4 show the accuracy of all emotions classified by various classifiers and facture vector quantization methods. The different classifiers have different ability and property. Thus, we have different recognition accuracy in each classifier or quantization method. According to the results shown in Table 3 and 4, the accuracy over five primary emotions, which are anger, boredom, happiness, neutral and sadness, is approximately equivalent with the same classifier. By this high recognition rate of the experimental results, the selected features are proven to be efficient in classifying five emotions in the arousal and valence degrees. In addition, the accuracy of two feature quantization methods of LBG and mean is quite close to each other in the same conditions of two different corpora. This shows that the set of the selected speech features is stable and suitable to recognize the five primary emotions in different corpora.

3.2 The Experimental Results of Speaker-Dependency

Table 5 and 6 show the recognition accuracy of a female and a male from Corpus I. The accuracy is better than the speaker-independent case because of the interrelationship of resemble speech features from the same person. The high accuracy of speech emotion recognition provides another application in not only speaker-independent but also speaker-dependent situation.

4 CONCLUSION

In this paper, the proposed emotion recognizer makes use of 16 LPC coefficients, 12 LPCC coefficients, 16 LFPC coefficients, 16 PLP coefficients, 20 MFCC components and jitter with LDA, K-NN, HMMs as the classifiers. According to experimental outcomes, the proposed method can solve the difficulty of recognizing five primary human emotions using the set of selected features in both speaker-independent and speaker-dependent corpora.

Table 3: Experimental results of Corpus I

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

Table 4: Experimental result of Corpus II

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 5: Experimental results of a female

Accuracy (%)	LDA		K-NN		HMMs	
	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
Anger	86.5	83.5	90.5	89.7	91.2	88.5
Boredom	82.3	86.2	86.7	84.9	88.8	88.3
Happiness	84.9	82.3	86.4	87.6	90.6	89.5
Neutral	82.6	86.4	89.8	87.1	87.9	86.7
Sadness	85.4	87.3	92.7	93.1	94.8	92.3
Average	84.3	85.1	89.2	88.4	90.6	89.0

Table 6: Experimental results of a male

Accuracy (%)	LDA		K-NN		HMMs	
	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
Anger	84.2	89.8	83.9	88.4	92.5	90.4
Boredom	82.5	79.5	80.5	82.3	84.0	82.5
Happiness	87.2	88.3	86.7	88.5	89.9	92.3
Neutral	80.4	82.4	81.0	83.5	83.2	86.5
Sadness	81.5	76.5	84.6	92.8	92.4	94.3
Average	83.1	83.3	83.3	87.1	90.4	89.2

5 ACKNOWLEDGE

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

6 REFERENCES

[1] 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).

[2] 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).

[3] O.W. Kwon, K. Chan, J. Hao, T.W. Lee , “Emotion Recognition by Speech Signals,” Eurospeech, pp.125-128 (2003).

[4] C.D. Park and K.B. Sim, “Emotion Recognition and Acoustic Analysis from Speech Signal,” Proceedings of IJCNN, pp. 2594-2597 (2003).

[5] V.A. Petrushin, “Emotion Recognition in Speech Signal: Experimental Study, Development and Application,” ICSLP, pp.222-225 (2000).

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

[7] 張柏雄, “中文語音情緒之自動辨識,” the master thesis of Engineering Science department, National Cheng Kung University (2002).