Emotion Recognition in Mandarin Speech

包蒼龍、陳育得

大同大學資訊工程系主任所長兼電算中心主任

大同大學資訊工程系博士候選人

104 台北市中山北路 3 段 40 號

電話：02-2592-5252#3294　Fax：02-2594-1371

E-mail：tlpao@ttu.edu.tw, d8906005@mail.ttu.edu.tw

摘要

在本論文中，我們提出一套中文語音情緒辨識系統，在系統中，我們探討生氣、厭倦、平常、快樂及傷心等五種情緒，所抽取出的特徵參數包含 16 個 LPC 係數及 20 個梅爾刻度式倒頻譜係數，另外，我們採用兩種統計分類方法來做五種不同情緒的分類 ，使用最短距離法時，正確辨識率為 78.3%，改採最近群中心法時，正確辨識率可達到 82.8%。

Abstract

In this paper, a Mandarin speech based emotion classification method is presented. Five basic human emotions including anger, boredom, happiness, neutral and sadness are investigated. The features we extracted include 16 LPC coefficients and 20 MFCC components and the recognizer presented in this paper is based on two statistical pattern recognition techniques, the minimum-distance method and the nearest class mean method. For minimum-distance emotion recognition, an average accuracy of 78.3% is obtained. For the nearest class mean emotion recognition, higher accuracy of 82.8% is achieved.

Keywords：Emotion Recognition, LPC, MFCC

壹、 前言

Speech is the most basic and main communication tool in human-to-human interaction. Emotion can make it's meaning more complex and the listeners can react differently according to what kind of emotion the speaker transmit, e.g., consoling a sad one with soft words.

From the signal processing point of view, speech signal includes the linguistic information, speaker's tone and emotion. There are several applications for automatic machine recognition of the type of emotion expressed in a given speech. For example, it may be desirable to include information of emotions to other party in conventional video teleconferencing and web-based teaching for added effects. In distance teaching, if a student does not understand what the teacher is saying, it may be detected from expression of emotions on his face and in his speech. These responses may have a direct and immediate influence on the teacher who would in

turn try to explain the topic again. The emotion-based feedback is especially important in communicating with young children.

貳、正文

一、 Background

'Emotion in speech' is a topic that has received much attention during the last few years, in the context of speech synthesis as well as in automatic speech recognition (ASR). Speech is the most convenient means of communication between people. In this section, we will describe the background knowledge about emotion recognition in Mandarin speech and review several emotion recognition systems.

(一) Emotions

What are emotions? Emotions can be considered as communications (messages) to oneself and others [1]. They consist of behaviors (e.g., hiding), physiologic changes (e.g., tachycardia) and subjective experience (e.g., "I'm scared") as evoked by thoughts or external events, particularly events that one perceives as important.

Emotions are traditionally classified into two main categories: primary (basic) and secondary (derived) emotions [2, 3]. Primary or basic emotions, including fear, anger, joy, sadness and disgust, are generally those, which are experienced by all social mammals and have particular manifestations associated with them. Secondary or derived emotions, such as pride, gratitude, sorrow, tenderness, irony and surprise, are variations or combinations of primary ones, and may be unique to humans.

(二) Expression of Emotions

There is a large literature on the signs that indicate emotion. Figure 1 shows the different ways to express emotions. The vocal cue is one of the fundamental expressions of emotions, on a par with facial expression. Primates, dolphins, dogs and all the mammals have emotions and can convey them by vocal cues. Humans can express their feelings by crying, laughing, shouting and also by more subtle characteristics of the speech. Besides, emotions can also be expressed by somatic correlates, including heart rate, skin resistivity, temperature, pupillary diameter and muscle activity. They have been widely used to identify emotion-related states, for instance in lie detection. Finally, the face is our emotional signaling system, too. In contrast to speech, emotion recognition through face analysis counts with a good deal of information. Most of the information about facial expression can be extracted from the position of the eyebrows and positions of the corners of the mouth.

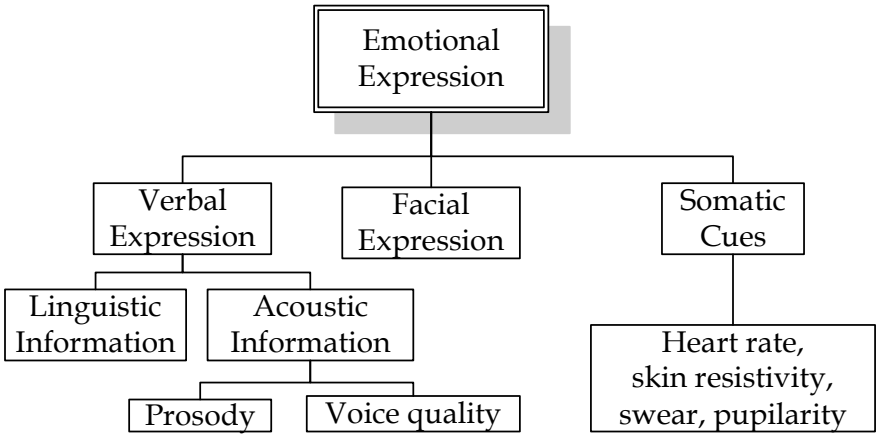

Figure 1: Different channels in the communication of human emotions

(三) Emotion Recognition Review

The first studies that were conducted [4] were not so much trying to get an efficient machine recognition device, but rather were searching for general qualitative acoustic correlates of emotion in speech. The use of acoustic prosodic cues in order to classify *angry* vs. *neutral* speaking style is described in [5]. Twenty speakers were asked to produce 50 neutral and 50 angry utterances and multi-layer perceptrons were trained with these data. Results reach around 90% of accuracy in the simplified tasks of distinguishing *emotional* from *non-emotional* utterances.

Valery A. Petrushin [6] performed an experimental study on vocal emotions and the development of a computer agent for emotion recognition. The study dealt with a corpus of 700 short utterances expressing five emotions: *happiness, anger, sadness, fear* and *normal* (unemotional) state, which were portrayed by thirty subjects. Some statistics of the pitch, the first and second formants, energy and the speaking rate were selected and several types of recognizers were created and compared. The best results were obtained using the ensembles of neural network recognizers. The total recognition accuracy is about 70%.

In [7] the elicited speech data came from different passages selected because they were effective at evoking specific emotion – *fear, anger, happiness, sadness* and *neutrality*. 40 volunteers were recorded. A battery of 32 potentially relevant features, derived from contours tracing the movement of intensity and pitch, were extracted. Two different classifiers were tried and results showed that, for this particular case, discriminant analysis outperformed the neural networks. Using 90% of the data for training, and testing with the remaining 10%, a classification rate of 55% was achieved.

Noam Amir [8] uses a corpus that has been studied extensively, property of Universidad Politécnica of Madrid – Departamento of Ingenieria Electrónica – Group of Technology of Habla, and verifies it through subjective listening tests. The overall recognition is approximately 70%.

二、 Emotion Recognition in Mandarin Speech

Figure 2 shows the block diagram of the proposed system. The input speech is partitioned into frames of 256 samples and a frame is overlapped with the next frame by 128 samples. Then the speech frame is high-pass filtered to emphasize the important higher frequency components. The next step is to window each individual frame as 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. Finally, the input speech is recognized by statistical pattern classification methods.

Many researches integrated several different techniques in emotion recognition in speech

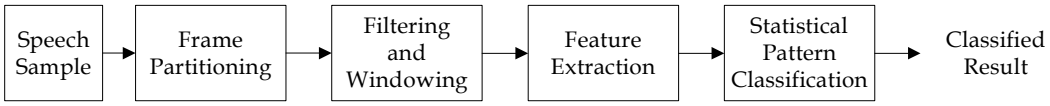

Figure 2: Block Diagram of the Proposed System

to improve performance of emotional speech recognizer. According to the results of emotion recognition research that has been done today, the aspect of features that is most suitable for emotion recognition in speech is still in research stage. A possible approach is to apply various different and known feature extraction methods to investigate how far we can go by using only these information.

There are two kinds of features used in the proposed system. They are MFCC (Mel Frequency Cepstrum Coefficients) and LPC coefficients as they convey information of short time portions that describe the power of speech signal and are used successfully in many automatic speech recognition (ASR) systems. The LPC coefficients can be calculated by either autocorrelation method or covariance method [9] and the order of linear prediction used is 16.

The MFCC is a widely used form of cepstrum in automatic speech recognition system as they convey information of short time energy migration in frequency domain. We expect that these information can help to determine the emotional content of the Mandarin speech. The procedure to obtain MFCC features is shown in Fig. 3. After frame partitioning, high-pass filtering and windowing described previously, the next step is the Fast Fourier Transform, which converts each frame of 256 samples from the time domain into the frequency domain. Then, a set of 20 Mel scaled filter banks which has frequency span between 200 Hz to 4k Hz is applied to the FFT power spectrum. In the final step, we convert the log mel spectrum back to time domain to obtain the MFCC coefficients.

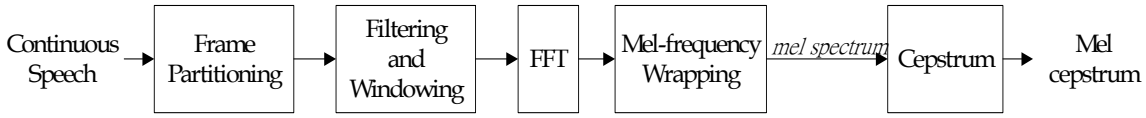

Figure 3: Block Diagram of the MFCCs Extraction

The final step in Fig. 2 is the classification. The classifier uses the features extracted from the feature extraction step to assign the input speech sample to one of the five emotion classes: anger, boredom, happiness, sadness and neutral. There are two statistical pattern classification methods we used in this research.

The nearest class mean classification is a simple classification method that assigns an unknown sample to a class according to the distance between the sample and each class's mean. The class mean, or centroid, is calculated as follows:

$$m_i = \frac{1}{n}\sum_{j=1}^{n_i}\mathbf{x}_{i,j} \quad (2)$$

where $\mathbf{x}_{i,j}$ is the jth sample from class i. An unknown sample with feature vector $\boldsymbol{x}$ is classified as class i if it is closer to the mean vector of class i than to any other class's mean vector. Rather than calculate the distance between the unknown sample and the mean of every classes, the minimum-distance classification estimates the distance between the unknown sample and each training sample.

三、 Experimental Results

The speech corpus used in the experiments was recorded from two Chinese female, in 8-bit PCM with a sampling frequency of 8k Hz. To provide reference data for automatic classification experiments, only those data that had complete agreement between two other taggers were chosen for the experiments reported in this paper. Finally, we obtained 415 utterances with 76 angry, 88 bored, 75 happy, 89 neutral, and 87 sad utterances.

The Mandarin emotion recognition system was implemented using "MATLAB" software run under a desktop PC platform. The correct recognition rate was evaluated using leave-one-out (LOO) cross-validation which is a method to estimate the predictive accuracy of the classifier. The experimental results showed that the correct recognition rate is 78.3% in the minimum distance classification and reaches 82.8% in the nearest class mean classification. All these results are presented in Table 1 and Table 2.

Table 1: Experimental results (with minimum-distance method)

	Anger	Boredom	Happiness	Neutral	Sadness	Correct Recognition Rate (%)
Anger	57	5	5	3	6	75.1
Boredom	3	72	3	8	2	81.8
Happiness	6	4	56	5	4	74.6
Neutral	4	10	2	69	4	77.5
Sadness	4	3	4	5	71	81.6

Table 2: Experimental results (with nearest class mean method)

	Anger	Boredom	Happiness	Neutral	Sadness	Correct Recognition Rate (%)
Anger	53	5	8	6	4	69.7
Boredom	1	69	7	8	3	78.4
Happiness	1	2	67	5	0	89.3
Neutral	1	5	2	77	4	86.5
Sadness	1	3	2	3	78	89.6

參、結論

We express our emotions in three main ways: the words that we use, facial expression and intonation of the voice. Whereas research about automated recognition of emotions in facial expressions is now very rich, research dealing with the speech modality, both for automated production and recognition by machines, has only been active for very few years and is almost for English.

In this paper, we proposed a Mandarin emotion recognition system. Sixteen LPC and twenty MFCC coefficients are selected as the feature to identify the emotional state of the speaker. Five basic emotions of anger, boredom, happiness, neutral and sadness are classified using two common statistical pattern classification methods, the minimum-distance method and the nearest class mean method. For minimum-distance emotion recognition, an average accuracy of 78.3% is obtained. For the nearest class mean emotion recognition, higher accuracy of 82.8% is achieved.

肆、誌謝

NSC 92-2213-E-036-021

伍、參考文獻

(1). Kleinginna, P. R. and Kleinginna, A. M., A categorized list of emotion definitions with suggestions for a consensual definition, Motivation and Emotion, 5, pp.345-379 (1981).

(2). Murray, I. and Arnott, J. L., Towards the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion, in Journal of the Acoustic Society of America, pp.1097-1108 (1993).

(3). Stibbard, R. M., Vocal Expression of Emotions in Non-laboratory Speech: An Investigation of the Reading/Leeds Emotion in Speech Project Annotation Data, Unpublished PhD Thesis. University of Reading, UK. (2001).

(4). Williams, U.; Stevens K. N., Emotion and Speech: some acoustical correlates, 1972

(5). Huber, R., Prosodische Linguistische Klassifikation von Emotionen. PhD Thesis, (1998).

(6). Petrushin, V. A., Emotion Recognition in Speech Signal: Experimental Study, Development and Application, ICSLP 2000, Beijing (2000).

(7). McGilloway, S.; Cowie, R.; Doulas-Cowie, E.; Gielen, S.; Westerdijk, M.; Stroeve S.: Approaching Automatic Recognition of Emotion from Voice: A Rough Benchmark. (2000).

(8). Amir, N., Classifying emotions in speech: a comparison of methods, Holon Academic Institute of technology, EUROSPEECH 2001, Escandinavia. (2001).

(9). Lawrence Rabiner and Biing-Hwang Juang, Fundamentals of Speech Recognition, Prentice Hall PTR, (1993).