Emotion Recognition in Mandarin Speech and Its Application in Training of Hearing Impaired

中文語音情緒辨識及其在聽障語言教學系統上之應用

Tsang-Long Pao
Tatung University
tlpao@ttu.edu.tw

Yu-Te Chen
Tatung University
d8906005@mail.ttu.edu.tw

摘要

語言是人類溝通思想、傳遞訊息及表達意願的最基本與最主要的工具，而語言的學習主要仰賴於聽覺系統與發音系統的協調與訓練，對於有聽覺障礙的人們而言，缺少了聽覺的輸入，則其語言能力的發展與訓練幾乎變成不可能。

聽障電腦輔助語言教學系統是一套用來幫助聽障者學習講話的訓練系統，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 79.1%，改採最近群中心法時，正確辨識率可達到 89.1%，藉由此系統的加入，聽障者不僅可以學習到說話說的正確，更可以說話說的「自然」，如同平常人一樣！

關鍵詞：情緒辨識、聽障電腦輔助語言教學系統、梅爾刻度式倒頻譜係數

Abstract

Language is the most basic and main tool for the human to communicate thoughts, convey messages and express aspiration. Language learning mainly relies on the coordination and training of the auditory system and articulatory system. For the hearing-normal person, this kind of learning process is very natural. But for the hearing-impaired person, it becomes almost impossible without the auditory input.

Computer-assisted speech training of hearing impaired is a training system to help the hearing-impaired person to speak. In this paper, we propose a Mandarin speech emotion recognition system. In this system, the features we extracted include 16 LPC coefficients and 20 MFCC (Mel Frequency Cepstrum Coefficients). We adopt two common statistical pattern classification methods, the minimum-distance method and the nearest class mean method, to classify the speech into one of the five basic emotions. In the minimum-distance method, the correct recognition rate is 79.1%. In the nearest class mean method, the correct recognition rate reaches 89.1%. By applying this system, it can assist the hearing-impaired people to learn not only to speak correctly but also to speak “naturally”, just like the hearing-normal people.

Keywords：Emotion recognition, MFCC, Computer-assisted Speech Training of Hearing Impaired

一、Introduction

The feedback and use of the voice is a very natural thing for the hearing-normal people. Ordinary people receive every kind of stimulus and messages via voice, including human voice, music and natural sounds. Through making some special sounds, from simple applause to a complex expression of the language, human can express feelings and convey thoughts. So, it is no doubt that the voice is one of the most important tools of communication. Someday, if we loss this ability, how silence of the surroundings will be?

Hearing-impaired is a generic term including both deaf and hard of hearing which refers to persons with any type or degree of hearing loss that causes difficulty working in a traditional way. It can affect the whole range or only part of the auditory spectrum, which for speech perception, the important region is between 250 and 4,000Hz. The term deaf is used to describe people with profound hearing loss such that they cannot benefit from amplification, while hard of

hearing is used for those with mild to severe hearing loss but who can benefit from amplification.

The survey showed that about 0.08 percentages of children in Taiwan are hearing impaired [1]. These hearing-impaired children are not profoundly deaf and remain some level of hearing. Using these residual hearing and other perception, the hearing-impaired children communicate with other people in different way. Because the difficulties of language learning environment and perceptual ability, hearing-impaired children have many handicaps to learning language.

In Taiwan, there are only one speech therapist among twenty-three hearing impaired people. A hearing impaired person must wait at least 2 to 3 months for one hour treatment at a time. After 2 to 3 months, the previous training result has been forgotten. So in this paper, we present a Mandarin speech based emotion recognition system which can apply in the computer-assisted speech training system for hearing impaired people that they can use at home to learn to pronounce not only correctly but also naturally. The goal is to assist them to learn more speaking skills to communicate effectively in the society.

二、Background

Speech is the most basic and main communication tool in human-to-human interaction. It includes the linguistics information, speaker's tone and emotion. Emotion can make its meaning more complex and the listeners can response differently according to what kind of emotion the speaker transmit. In this section, we will describe some background and review several emotion recognition systems.

(一) Communication Methods

In real life, we can often see that hearing-impaired people converse with others in different ways. These substitute methods of conversation open a new window for hearing-impaired people. People with hearing impairment can communicate using numerous methods of communication, such as:

- Sign language
- Finger spelling
- Lip reading
- Written communication
- Oral communication

In fact, sign language has low popularity among general people. Lip-reading is just a reference because it has some limitations in Mandarin vowels. Writing is not a convenient way. So in many language training, to teach the hearing-impaired person to speak is the ultimate goal.

(二) Speech-Training Methods

There are five main speech-training methods of teaching hearing-impaired person to speak, including:

- Visible speech method: This method directly uses the different distribution of voices to transform the speech to spectral representation changed along the time. Lessons are trained for a long time (220 hours, 800 words) to distinguish different distribution on the spectrogram.
- Oral methods: Learning to pronounce phonemes is the first step and then combines the phonemes to words. The emphasis of this method is not only on correct pronunciation but also speaking training control. This method is analogue to common speech training method.
- Acoustic method: People learn the tips of language learning by using their tactile and visual stimulus to percept, analyze and interpret the pronunciation meaning of sound's vibrations and visual cues.
- Concentric method: Learning the next pronunciation skill only when you have practiced previous one very well.
- Tactile/visual/auditory method: Teaching the hearing impaired person to speak by using the tactile, visual and auditory cues. So this method is a multi-sensory method.

(三) Emotions

What are emotions? Emotions can be considered as communications (messages) to oneself and others [7]. 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.

Traditionally, emotions are classified into two main categories: primary (basic) and secondary (derived) emotions [3, 9]. 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.

(四) Mandarin Sounds

There are 21 initials, 16 finals and 5 tones in Mandarin. These produce about 1340 Mandarin sounds. The initial is a consonant that begins the syllable. A final in Mandarin is a vowel, which may be a simple vowel, or a compound vowel, or a vowel plus a nasal consonant. Some syllables may be without an initial, but no syllable can do without a final. Tone is the variation of pitch within a syllable. Tones are really the most difficult aspect of Chinese at the outset. Chinese uses tones to distinguish words. There are five tones in Mandarin, including four inflected (high-tone, rising tone, falling tone and abrupt glottal stop) and one neutral one (mid-tone). Each word and phrase must be spoken at the right pitch or the meaning is changed and probably will be misunderstood.

(五) Emotion Recognition Review

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 [8] 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%. This study also develops a real-time emotion recognizer using neural networks for call center applications. He achieved approximately 77% classification accuracy in two emotion states, agitation and calm for 8 features chosen by a feature selection method.

In [4] the elicited speech data came from different passages selected because they were effective at evoking specific emotion – fear, anger, happiness, sadness and neutrality. Fourty 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 [2] 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. Best results are obtained using distance measures based classifiers; recognition rates are 70% for neutral, 76% for happy, 83% for sad and 61% for angry utterances. The overall recognition is approximately 70%.

三、Mandarin Emotion Recognition in Speech

Many researchers integrated several different techniques in emotion recognition in speech 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 the way to extract non-textual information to identify emotional state of the utterance.

In our proposed method, the extracted feature is composed of the LPC coefficients and the MFCC 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 [6] 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. 1. After frame blocking, high-pass filtering and windowing, 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.

In addition to these, two statistical pattern recognition techniques, the nearest class mean classification and the minimum distance classification, were used to classify speech samples to

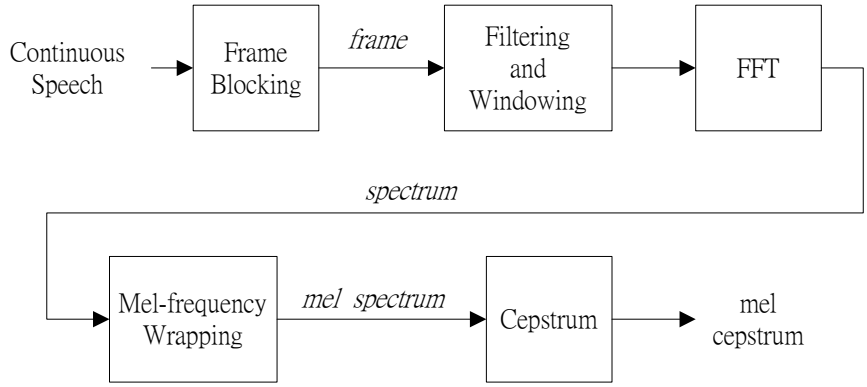

Figure 1: Block Diagram of MFCCs Extraction

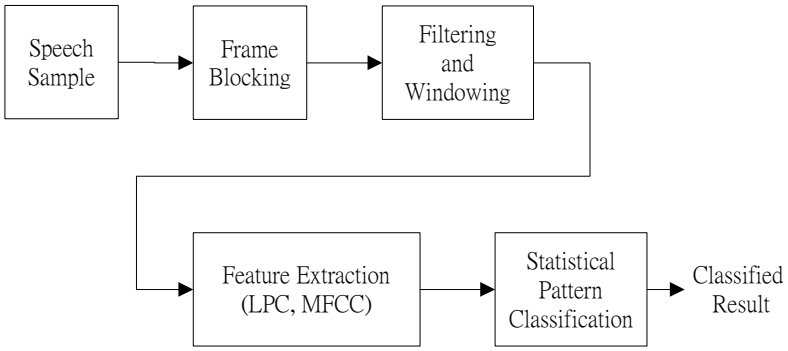

Figure 2: Block Diagram of the Proposed System

one of the 5 emotion classes: anger, boredom, happiness, sadness and neutral, according to their emotional content.

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:

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

where $\mathbf{x}_{i,j}$ is the jth sample from class i. An unknown sample with feature vector $\mathbf{x}$ is classified as class i if it is closer to the mean vector of class i than to any other class's mean vector. The distance is a Mahalanobis distance calculated as follows:

$$d = (\mathbf{x}-\mathbf{m}_k)\mathbf{C}_k^{-1}(\mathbf{x}-\mathbf{m}_k) \quad k = 1,...,5 \quad (2)$$

$\mathbf{x}$: the unknown sample
$\mathbf{m}_k$: the mean of the kth class
$\mathbf{C}_k$: the covariance matrix of class k

Rather than calculate the distance between the unknown sample and the mean of every classes, the minimum-distance classification estimates the Euclidean distance between the unknown sample and each training sample.

四、Experimental Results

The proposed emotion recognition system was shown in Fig. 2 and was implemented using "MATLAB" software run under a desktop PC platform. The speech corpus recorded includes five basic emotions such as anger, boredom, happiness, sadness and neutral from two Taiwanese, one male and one female. Short sentences were recorded and transformed to "wave" audio format of 8-bit PCM and sampling frequency of 8k Hz. To provide reference data for automatic classification experiments, the obtained speech data were independently tagged by two other human listeners. Only those data that had complete agreement between the taggers were chosen for the experiments reported in this paper. After the database preparation, we obtained 647 utterances with 97 angry, 117 bored,

Table 1: Experimental Results

	Correct Recognition Rate (%) (The minimum-distance method)	Correct Recognition Rate (%) (The nearest class mean method)
Anger	72.1	83.5
Boredom	76.9	88.9
Happiness	80.9	84.4
Neutral	82.2	89.5
Sadness	83.1	98.8
Average	***79.1***	***89.1***

115 happy, 152 neutral, and 166 sad utterances.

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. Suppose we have *N* patterns to train and test the classifier. We divide the set into two subsets, i.e. training set and testing set. The LOO method takes *N*-1 patterns to train the classifier and tests it with the remaining pattern. This procedure is repeated for all the available patterns from 1 to *N*.

The experimental results showed that the correct recognition rate is 79.1% in the minimum distance classification and reaches 89.1% in the nearest class mean classification. All these results are presented in Table 1.

五、Conclusions

In daily lives, we can often see that hearing-impaired people converse with others by sign language, lip reading or writing. These substitute methods of conversation open a new window for hearing-impaired people. However, sign language has low popularity among general people. Lip-reading has some limitations in Mandarin vowels. Writing is not a convenient way. Therefore, to teach the hearing-impaired person to speak is the ultimate goal of language training.

Computer-assisted speech training of hearing impaired is a training system to let the hearing-impaired person to learn at home to pronounce correctly. In this paper, we present a Mandarin speech based emotion recognition system. By applying it, they can learn to pronounce not only correctly but also naturally. In our proposed 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. The correct recognition rate for the minimum distance method is 79.1%. The correct recognition rate for the nearest class mean method reaches 89.1%.

In the future, there are several research directions, including:

- Improve emotion recognition rate: as discussed previously, 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 maybe integrate other features into our system to improve emotion recognition rate.
- Investigate confidence scoring: just like in many singing training system, we can see the singing score on the screen after we have song. We can also apply the score on the computer-assisted speech training system for the hearing-impaired person to tell the trainee how well he speaks.

六、Acknowledgement

The Authors would like to thank the National Science Council (NSC) of the Republic of China (ROC) for financially supporting this research under NSC project NO:
NSC 92-2213-E-036-021

七、References

[1] 鍾玉梅, 聽障兒童之構音治療, 聽語會刊, 8, 41-47, 1992.

[2] Amir, N., “Classifying emotions in speech: a comparison of methods”. Holon Academic Institute of technology, EUROSPEECH 2001, Escandinavia.

[3] Arnott, J. L. and Murray, I., “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, 1993, pp.1097-1108.

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

[5] Huber, R., Prosodische Linguistische Klassifikation von Emotionen. PhD Thesis, 1998.

[6] Biing-Hwang Juang and Lawrence Rabiner, Fundamentals of Speech Recognition, Prentice Hall PTR, 1993.

[7] Kleinginna, P. R. and Kleinginna, A. M., “A categorized list of emotion definitions with suggestions for a consensual definition.” Motivation and Emotion, 5, 345-379. 1981.

[8] Petrushin, V. A., “Emotion Recognition in Speech Signal: Experimental Study, Development and Application.” ICSLP 2000, Beijing.

[9] 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.