Tracking and Visualizing the Changes of Mandarin Emotional Expression

Jun-Heng Yeh, Tsang-Long Pao, Chen-Yu Pai, and Yun-Maw Cheng

Department of Computer Science and Engineering, Tatung University
40 ChungShan North Road, 3rd Section Taipei 104, Taiwan
d9306002@ms.ttu.edu.tw, tlpao@ttu.edu.tw,
g9506017@ms.ttu.edu.tw, kevin@ttu.edu.tw

Abstract. In order to track continuous changes of someone's emotional expressions in Mandarin dialogue, we elaborated an emotion recognition system in continuous Mandarin speech. A new segmentation method was used to estimate the emotional turning points in a dialogue by dividing the utterance into several independent segments, each of which contains a single emotional category. Five basic emotional categories, including anger, happiness, sadness, boredom and neutral, can be recognized in our proposed method. Here, the speech features, MFCC and LPCC, are used in the proposed method. The experimental results show that the new recognition method can provide satisfactory results.

Keywords: emotion recognition, continuous speech, Mandarin.

1 Introduction

Emotion plays a significant role in cognitive psychology, behavioral sciences and human-computer interaction. Humans are especially capable of expressing their feelings by crying, laughing, shouting and subtle characteristics from speech. And classification of emotional states based on the prosody and voice quality requires classifying acoustic features in speech as connected to certain emotional states.

A growing number of research studies in emotion recognition via an isolated short sentence are available to shed some light on the implementations of man-machine interface. However, the short-sentence emotion recognition system may not be able to detect the emotional states correctly because there may have several emotions in a sentence. Even so, human beings speak continuously and people will change emotions when they are triggered by some incidents in the course of speaking. Up to date, few works have focused on automatic emotion tracking of continuous Mandarin speech.

Emotional dimensionality is a simplified description of the basic properties of emotional states. According to the theory developed by Osgood, Suci and Tannenbaum [1] and in subsequent psychological research, the computing of emotions is conceptualized as three major dimensions of connotative meaning: arousal, valence and power. In

D.-S. Huang et al. (Eds.): ICIC 2008, LNCS 5226, pp. 978–984, 2008.

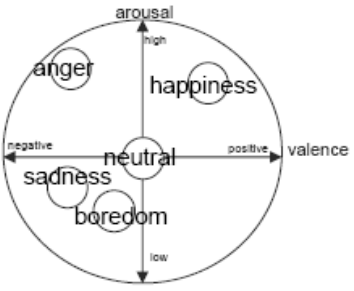

Fig. 1. Graphic representation of the arousal-valence dimension of emotions

general, the arousal and valence dimensions can be used to distinguish most basic emotions. The locations of emotions in the arousal-valence space are shown in Figure 1,which provides a representation that is both simple and capable of conforming to a wide range of emotional applications.

The speech features analysis for emotion recognition [2] has been studied in our previous research to seek for a reliable measurement. These selected speech features are beneficial for us to focus on the emotional segmentation method. Moreover, so far there is relatively little research on the real time emotional speech recognition. Our proposed method tries to implement speech emotion recognition instantly. In our recognition system, we can locate the turning points of emotion changes, and show the emotion categories in a visible interface at the same time.

The rest of the paper is organized as follows: Section 2 describes our features extracted from the speech data and selected from the forward selection method. Section 3 discusses our continuous speech recognition method. Section 4 presents the results of classification experiments. Concluding remarks are given in Section 5.

2 Feature Analysis

In order to find out the optimal combination of all extracted features, we make use of the forward selection to decide the most effective set of features from more than 200 speech features. In the previous research done in our laboratory[3-4], the feature set includes energy, pitch, jitter, shimmer, Formants (F1, F2 and F3), Linear Predictive Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC), Mel-Frequency Cepstral Coefficients (MFCC), first derivative of MFCC (dMFCC), second derivative of MFCC (ddMFCC), Log Frequency Power Coefficients (LFPC), Perceptual Linear Prediction (PLP) and RelAtive SpecTrAl PLP (Rasta-PLP). The forward feature selection (FFS) and backward feature selection (BFS) are then used to decrease the dimensions. Finally, MFCC and LPCC are individually obtained from FFS and BFS as the most important features. For speech recognition, MFCC and LPCC are the popular choices as features representing the phonetic content of speech [5]. For each speech frame, 20 MFCCs and 16 LPCCs are used in this system.

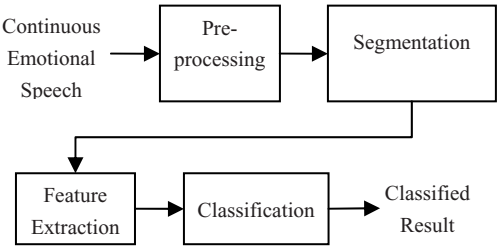

Fig. 2. Block diagram of the emotion recognition system

3 Continuous Emotion Recognition System

Figure 2 shows the block diagram of the emotion recognition system. The analog speech signal is recorded through a microphone and converted to discrete-time signal by sampling and quantization. Framing is used to divide the speech signal into pieces. In this paper, the frame size is set to 256 samples and frame overlapping period is 50% of the frame size. The continuous speech is then divided into suitable length segments. Before further processing, the individual frames are windowed with Hamming window to reduce the discontinuity of both ends of a frame. Finally, these partitions are recognized individually after feature extraction. In previous studies, the methods of continuous segmentation are usually applied in the speech recognition system. To recognize the emotion from continuous emotional speech, the main step is to segment the whole utterance by finding out the turning points of emotional changes. Then, emotion of each segment is recognized in partition basis. The segmentation method used is described as follows. In real-time processing, it is important for the system to be able to detect the endpoints of an utterance so that an assessment can be constructed immediately. There exist some noises in the beginning and the end of the sound. The purpose of endpoint detection is to find the start and the end of meaningful partitions. A simple method to obtain endpoints is to calculate the energy contour and zero-crossing rate contour. The equations for the two energy thresholds and one zero-crossing rate threshold is defined as follows,

$$T_L = \mu_E + \alpha_1 \sigma_E . \tag{1}$$

$$T_U = \mu_E + \alpha_2 \sigma_E , \quad \alpha_1 < \alpha_2 \tag{2}$$

$$T_Z = \mu_Z + \alpha_3 \sigma_Z . \tag{3}$$

where μ_E and σ_E are the corresponding mean and standard deviation of the energy, μ_Z and σ_Z are the corresponding mean and standard deviation of zero crossing rate. The α_1, α_2 and α_3 are parameters, which are obtained by experiments.

Along the sequence of partitions, the first partition with energy greater than T_L is labeled as N_B. If the energies of the next B successive frames are greater than T_L and the

energy of the frame next to B frames is greater than T_L, N_B may be regarded as the beginning of a sound. On the other hand, if the energy of one of the B frames is less than T_L or the energy of the frame next to B frames is less than T_U, it is not the beginning of the sound. In this case N_B will be neglected. After locating the N_B, the next step is to check the zero-crossing rate of all the B frames to see if their zero-crossing rate is greater than T_Z. Now the frame is regarded as the true beginning of the sound, and is labeled as N_S. The frame after N_S with energy greater than T_L means that the sound exists. The first frame after N_S with energy less than T_L is the end of the sound, and is labeled as N_E. As a result, the region of the sound is from N_B to N_E or from N_S to N_E.

After the endpoint detection processing, the number of the segmented partitions may be too large to process. So, it is necessary to reduce the number of partitions. The way is to combine adjacent partitions if they have similar characteristics or too short in length. The mean of the lengths between the beginning and end of the adjacent endpoints are calculated. If the length of the frame is less than the threshold, it will be merged with adjacent frames. The threshold is obtained from experiments. In classification, to let the computer knows what the input signal means, a group of features extracted from frequency components and acoustic models is collected as the basis for the recognition engine. As mentioned in Section 2, the 20 MFCCs and 16 LPCCs features are used to recognize the emotional state of the input utterance by our proposed weighted D-KNN classifier [3].

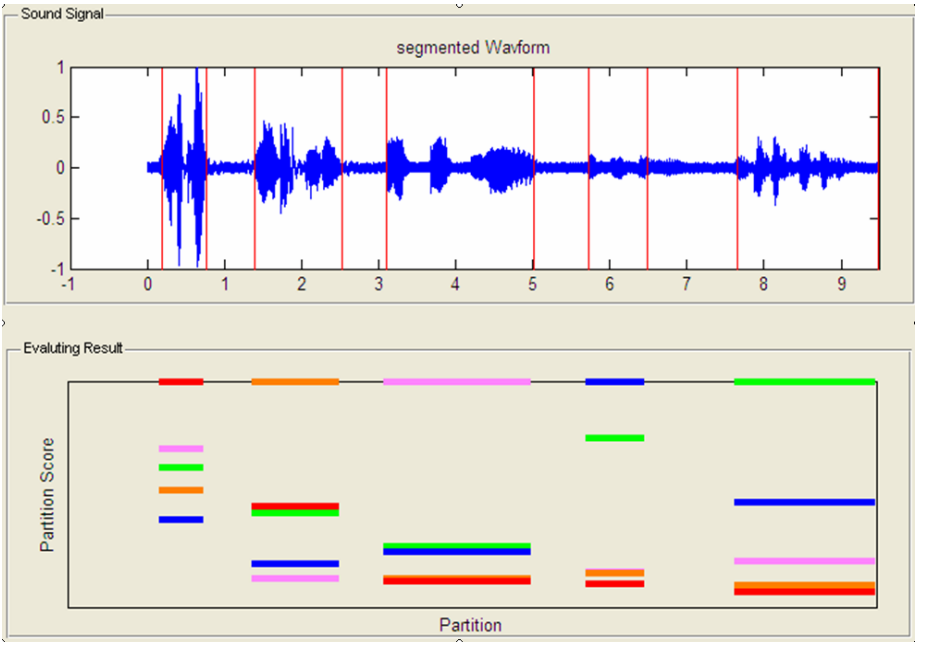

Fig. 3. The interface of the emotion recognition system

From the experimental results done in [6], the Fibonacci weighting function performs best among various weighting schemes. Therefore, we adopted Fibonacci series as the weighting function in this paper.

4 Experimental Results

The objective of our research is to find the turning points of emotional expressions in continuous Mandarin speech. The database in [2] was used. In this paper, utterances in Mandarin were used due to the immediate availability of native speakers of the language. It is easier for speakers to express emotions in their native language than in a foreign language. Twelve native Mandarin language speakers (7 females and 5 males) were asked to generate the emotional utterances. From the corpus, 558 utterances with over 80% human judgment accuracy were selected [2]. The recording format is mono channel PCM with sampling rate of 44.1 kHz and 16-bit resolution. Then we used the features by forward selection mentioned in Section 3 in our proposed recognition method. The testing sentence is a sentence that was combined with emotional short sentences randomly chosen from our database.

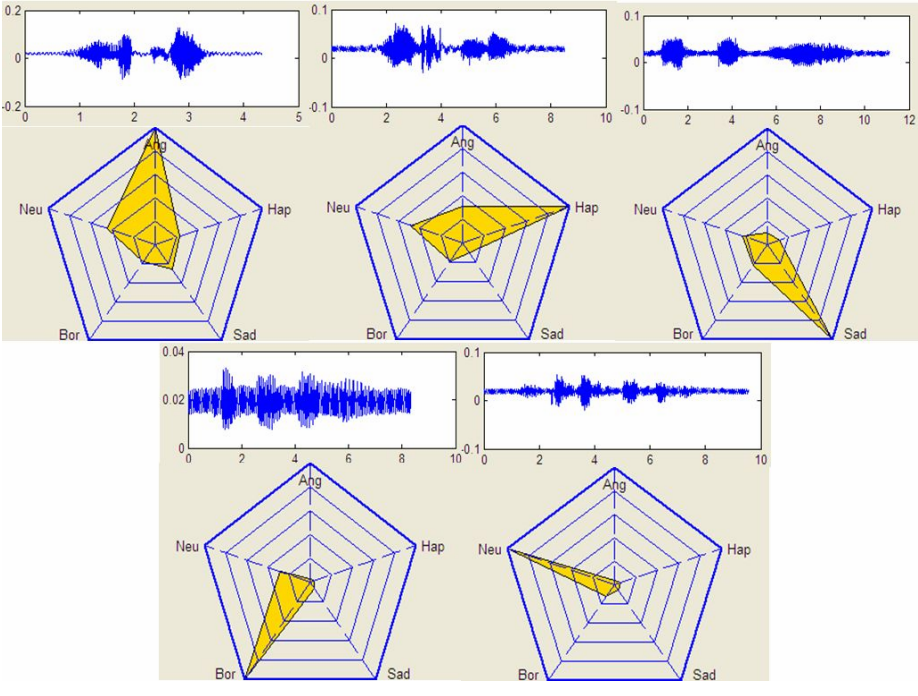

Fig. 4. Radar chart of the emotion recognition system

The interface of our proposed recognition system is shown in Figure 3. The five colors, red, orange, purple, blue and green represent angry, happiness, sadness, boredom and neutral, respectively. In Figure 3, the x-axes of both boxes (Source Signal and

Evaluating Result) are concurrent in time. The Source Signal box shows the results of emotional speech segmentation. In our proposed method, the unit of segmented sentences is an emotional category rather than a word. And the recognition rate achieves 82%. The Evaluating Result shows the relative intensity in each segmented emotional category. Users can easily recognize what is the major or minor emotional expression of a test corpus from the colors.

Each segmented sentences from the concatenated sentences can also be validated individually to estimate the emotional intensity as shown in Figure 4. In this figure, a concatenated sentence with angry, happiness, sadness, boredom and neutral emotions was validated. In our proposed recognition system, each recognized emotional category can be visualized as a "radar chart" to analyze more complicated emotional expressions. It can also present the relative intensity of emotional categories in a segmented sentence as shown in the Evaluating Result box of Figure 3.

In this emotion recognition system, the input speech will be recognized and mapped to the emotional expressions continuously.

5 Conclusion

It is natural for human beings to communicate with others in continuous dialogue. Even though, most proposed emotion recognition methods via voice can only be provided with a fragmented sentence (i.e. a manual and deliberate cutting sentence). To ensure the practicability, the purpose of this paper attempts to address these areas concerned on processing speech signals rather than interpreting the lexicons of speech. Moreover, the benefit from the processing can also tack the change of emotional expression in a signal sentence. The experimental results show that the proposed recognition method can be practically implemented and provide satisfactory results. In our recognition system, we can find out the turning points of emotional changes, and show the emotional categories in a visible interface instantly.

In the growing range of interactive interfaces, the research of emotional voice is still at an early stage, not to mention a paucity of literature on real applications. The crucial difficulty of this subject is how to blend the knowledge of interdisciplinary, especially in speech processing, applied psychology and human-computer interface. There are lots of applications that can benefit from the emotion speech recognition. First, in the hearing-impaired learning system, we can teach the hearing-impaired people to speak more naturally. Second, in the call-center application, it is possible to pick up the angry customer immediately. Third, in the slogan validator, the system may provide a convenient testing tool for marketers and researchers to evaluate a brand slogan that gives consumers the intended impression of the product.

References

1. Osgood, C.E., Suci, J.G., Tannenbaum, P.H.: The Measurement of Meaning. The University of Illinois Press, Urbana (1957)
2. Pao, T.L., Chen, Y.T., Yeh, J.H.: Emotion Recognition from Mandarin Speech Signals. Chinese Spoken Language Processing. In: International Symposium on Digital Object Identifier, pp. 301–304 (2004)

3. Vidrascu, L., Devillers, L.: Annotation and Detection of Blended Emotions in Real Human-Human Dialogs Recorded in a Call Center. In: IEEE International Conference on Multimedia and Expo., pp. 944–947 (2005)
4. Cheng, Y.M., Kuo, Y.S., Yeh, J.H., Chen, Y.T., Pao, T.L., Chien, S.C.: Using Recognition of Emotions in Speech to Better Understand Brand Slogan. In: Proceedings of the International Workshop on Multimedia Signal Processing (2006)
5. Kondoz, A.M.: Digital Speech: Coding for Low Bit Rate Communication Systems. John Wiley & Sons, Chichester (1994)
6. Chang, Y.H.: Emotion Recognition and Evaluation of Mandarin Speech Using Weighted D-KNN Classification. Master Thesis, Tatung University, Taipei (2005)