ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

EMOTION RECOGNITION FROM MADARIN SPEECH SIGNALS

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

Department of Computer Science and Engineering
Tatung University, Taipei
tlpao@ttu.edu.tw, d8906005@ms2.ttu.edu.tw, d9306002@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. In emotion classification of speech signals, conventional features are statistics of fundamental frequency, loudness, duration and voice quality. However, the recognition accuracy of systems employing these features degrades substantially when more than two valence emotion categories are invoked. 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. A Mandarin corpus recorded by 12 non-professional speakers is employed. The recognizer presented in this paper is based on three recognition techniques: LDA, K-NN, and HMMs. Experimental results show that the selected features are robust and effective for the emotion recognition not only in the arousal dimension but also in the valence dimension.

1. INTRODUCTION

There are numerous literatures that indicate emotion on the signs within the psychophysical and physiological expressions [2-3] [5] [7-11]. The vocal cue is one of the fundamental expressions of emotions. All mammals can convey emotions by it. Humans are especially capable of expressing their feelings by crying, laughing, shouting and subtler characteristics from speech. Classification of emotional states on basis of the prosody and voice quality requires classifying acoustic features in speech as connected to certain emotions. Specially, we need to find suitable features that the methods can extract and model to recognize emotional inflection. This also implies the assumption that speech carries abundant information about emotional states by the speaker. To estimate a user's emotion by the speech signal, one has to carefully select useful features. All selected features have to carry information about the transmitted emotion. However, they also need to fit the chosen classification algorithms. Majority of previous speech emotion recognition methods adapt prosody and energy related features. For example, Schuller *et al.* chose 20 pitch and energy related features [12]. Recognition accuracy of classifying 7 archetypal emotions (anger, disgust, fear, surprise, joy, neutral, sad) exceeded 77.8%. Tato *et al.* extracted prosodic features, derived from pitch, loudness, duration, and quality features [14] from a 400-utterance database. The recognition accuracy of 5 emotions (anger, happy, sad, neutral, bored) is 42.6%. Park *et al*. used pitch, formant, intensity, speech speed and energy related features to classify neutral, anger, laugh, and surprise emotions [10]. The recognition accuracy rate is about 40% in a 40-sentence corpus. Yacoub *et al*. extracted 37 fundamental frequency, energy and audible duration related features to recognize sadness, boredom, happiness, and cold anger emotions in a corpus recorded by eight professional actors [14]. The overall accuracy was only about 50%. 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 [7].

Unfortunately, the practical experience to recognize emotions from Mandarin is extraordinarily deficient. The Mandarin is a tone language. When a stress occurs, the first step of vocal translation is to continue the duration and expand the pitch, then intensity increases finally. If we recognize various emotions through only the speech variation of prosody and intensity, we will confuse with some emotions, as anger and happiness, boredom and sadness, because of the similar magnitude and pitch range. In order to surmount the inefficiency of conventional vocal features in recognizing some valence dimension emotions, we make efforts on searching for an effective and robust set of vocal features from Mandarin speech to recognize emotional categories.

The rest of paper is organized as follows: Section 2 describes the testing Mandarin corpus. Section 3 addresses the Mandarin speech emotion recognition procedure. Section 4 reports the experimental results of classification of five emotion classes. Finally, Section 5 concludes this paper.

Table 1: Utterances of Mandarin corpus

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

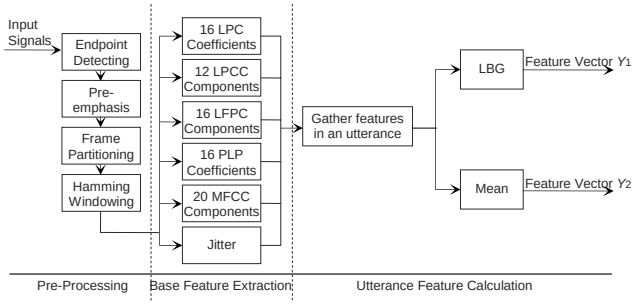

Figure 1: Block diagram of the feature extraction module

2. EMOTIONAL MANDARIN CORPUS

An emotional Mandarin corpus is specifically designed and set up for speaker-independent emotion classification studies. The database includes short utterances covering five primary emotions, namely anger, boredom, happiness, neutral, and sadness. Non-professional speakers are employed to avoid exaggerated expression. A total of 12 native Mandarin language speakers (7 males and 5 females) are involved to generate 1200 utterances. And a subjective assessment by native listeners who didn't participate in recording was carried out. The objective of the subjective classification is to eliminate the ambiguous emotion utterances. Finally, 558 utterances over human judgment accuracy of 80% are selected. Table 1 lists the utterances of each emotional category. The recording is done in a quiet environment with a sampling rate of 8k Hz.

3. RECOGNITION METHOD

The proposed emotion recognition method has three stages: feature extraction, feature quantization and classification. Base features and statistics were computed in feature extraction stage. Feature components were quantized in feature quantization stage. Classification was made by using three different basic classifiers based on dynamic or discriminative models.

In former works [2] [4-5] [7-12] [15], numerous authors compared static and dynamic speech feature sets for the emotion recognition. Due to the classification accuracy and robustness, we select 6 features from more than 200 speech features using forward selection method with K-NN (k nearest neighbor) decision rule.

3.1. Feature Extraction

Figure 1 shows the block diagram of feature extraction module. 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 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.

In base feature extraction procedure, linear predictive coding (LPC) provides an accurate and economical representation of the envelope of the short-time power spectrum of speech [9]. For speech recognition, linear prediction cepstral coefficients (LPCC) [2] and Mel-frequency cepstral coefficients (MFCC) are the popular choices as features representing the phonetic content of speech. Log frequency power coefficients (LFPC) are 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. A combination of discrete Fourier transform (DFT) and LPC techniques is the perceptual linear prediction (PLP) [6]. The PLP analysis is although computationally efficient and permits a compact representation. Perturbations in the pitch period are called jitter, such perturbations occur naturally during continuous speech, but measurements from acoustic waveforms have demonstrated that the perturbations for pathologic and normal speakers differ.

The modulation of spectral components of speech signals, as revealed by the formant structure or speech models, and overall structure of the average spectrum, has been less studied. Such aspects of speech signals, more related to voice quality attributes, have obvious perceptual correlates. 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 as the basic feature parameters.

3.2. Feature Quantization

After the feature extraction stage, a frame of speech samples will be represented as a vector. To further compress the data for presentation to the final stage of the system, vector quantization is performed. The feature

streams were converted into a fixed-length vector for each utterance by computing statistics to represent the streams.

The Linde-Buzo-Gray (LBG) algorithm [13] is carried out to quantify a feature vector Y_1. All vectors of a frame falling into a particular cluster are coded with the vector representing the cluster. For speech recognition using the selected feature parameters, it is found that the benefit per centroid diminishes significantly beyond the size of 16. In another simple vector quantization method, we treat the mean feature parameters corresponding to each frame as a feature vector Y_2.

3.3. Classification

Three different classifiers were selected to train and test the testing corpus with the extracted features. The classifiers we used are linear discriminate analysis (LDA), K-NN decision rule, and Hidden Markov models (HMMs).

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.

Table 2: Experimental results of three clusters recognition

Accuracy (%)	LDA		K-NN		HMMs	
	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
Anger/ Happiness	86.9	88.1	92.1	86.0	87.5	89.4
Neutral	90.4	89.3	87.9	87.2	90.5	90.5
Boredom/ Sadness	89.2	94.5	88.4	92.1	92.4	91.9
Average	88.8	90.6	89.4	88.4	90.1	90.6

Table 3: Experimental results 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 4: Experimental results 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. EXPERIMENTAL RESULTS

The extracted feature parameters 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 experimental results are validated by the leave-one-out (LOO) cross-validation method.

4.1. Results of Three Emotion Clusters Recognition

Due to the failing of separating some emotions at the same valence degree, some previous experiments are designed only to distinguish three emotion clusters which are obviously discriminate at the arousal degree. By contrast, we validate the set of the selected features to distinguish three emotion clusters, anger/happiness, neutral, and boredom/sadness using various classifiers. The experimental results shown in Table 2 are both robust and effective in distinguish three emotion clusters at the arousal degree.

4.2. Results of Anger and Happiness Recognition

The prosodic features as pitch and energy related features are failed to distinguish the valence emotions. The experimental results summarized in Table 3 show that the set of the selected features can distinguish anger and happiness emotions which have similar prosody and magnitude at the valence degree.

By applying the set of the selected emotion speech features, recognizers are undoubted to separate the anger and happiness which most previous speech emotion recognizers are always confused in this emotion cluster. In addition, as shown in Table 3, the high and stable accuracy provides the appropriateness to distinguish the emotions at the valence degree.

4.3. Results of Boredom and Sadness Recognition

Table 4 shows the recognition accuracy of distinguishing boredom and sadness emotion utterances. The experimental results are similar to Table 3. We can observe the robustness and suitability from the stable accuracy of each emotion with different classifiers.

These two emotions, boredom and sadness, are considered to be close to each other at the valence degree with the similar prosody and magnitude. So do anger and happiness. 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.

Table 5: Experimental results of five emotions recognition

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

4.4. Results of Speaker-Independent Recognition

According to the experimental results shown in Table 5, the accuracy over five primary emotions, which are anger, boredom, happiness, neutral and sadness, is approximately equivalent with the same classifier. This shows that the set of the selected speech features is stable and suitable to recognize the five primary emotions. Moreover, the accuracy is higher than most experimental results from the previous surveyed methods. By this high recognition accuracy of the experimental result, the selected features are proofed to be efficient to directly classify five emotions of the arousal and valence degree simultaneously rather than only arousal degree. In addition, the accuracy of two feature quantization methods of LBG and mean is quite close to each other in the same conditions.

5. CONCLUSIONS

In this paper, we proposed an emotion recognizer that makes use of 16 LPC coefficients, 12 LPCC components, 16 LFPC components, 16 PLP coefficients, 20 MFCC components and jitter with LDA, K-NN, HMMs as the classifiers. The emotions are classified into five human basic emotion categories. The category labels used are, the primary emotions of anger, boredom, happiness, neutral and sadness. A corpus consisting of 558 emotional utterances, in which the corpus employed twelve native speakers, are used to train and test in the proposed system. The selected feature parameters of an utterance are quantified as a feature vector using LBG or mean method. Results show that the proposed system yields the best accuracy of 88.3%.

According to experimental outcomes, the proposed method can solve the difficulty of recognizing the valence emotions using the set of selected features. And we attain a high accuracy to distinguish anger/happy or bored/sad emotions that have similar prosody or amplitude. Moreover, the cluster recognition experiments show that the extracted features are outstanding to distinguish the anger/happiness, neutral, and boredom/sadness clusters.

6. ACKNOWLEDGE

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

7. REFERENCES

[1] 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.

[2] 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.

[3] P. Ekman, Handbook of Cognition and Emotion, New York: John Wiley & Sons Ltd, 1999.

[4] H. Hermansky. “Perceptual linear predictive (PLP) analysis of speech,” Journal of the Acoustical Society of America, pp.1738-1752, 1990.

[5] 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.

[6] J.F. Kaiser, Discrete-Time Speech Signal Processing, Prentic Hall PTR, pp.201-209, 2002.

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

[8] 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.

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

[10] 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.

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

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

[13] D. G. Stork, R.O. Duda, P.E. Hart, Pattern Classification, John Wiley & Sons, Inc., 2001.

[14] R.S. Tato, R. Kompe, J.M. Pardo., “Emotional Space Improves Emotion Recognition,” ICSLP, pp. 2029-2032, 2002.

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