Computational Linguistics and Chinese Language Processing
Vol. 10, No. 3, September 2005, pp. 347-362

Detecting Emotions in Mandarin Speech

Tsang-Long Pao[*], Yu-Te Chen[*], Jun-Heng Yeh[*] and Wen-Yuan Liao[*]

Abstract

The importance of automatically recognizing emotions in human speech has grown with the increasing role of spoken language interfaces in human-computer interaction applications. 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. Combining different feature streams to obtain a more accurate result is a well-known statistical technique. For speech emotion recognition, we combined 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. Two corpora were employed. The recognizer presented in this paper is based on three classification techniques: LDA, K-NN and HMMs. Results show that the selected features are robust and effective for the emotion recognition in the valence and arousal dimensions of the two corpora. Using the HMMs emotion classification method, an average accuracy of 88.7% was achieved.

Keywords: Mandarin, emotion recognition, LPC, LFPC, PLP, MFCC

1. Introduction

Research on understanding and modeling human emotions, a topic that has been predominantly dealt with in the fields of psychology and linguistics, is attracting increasing attention within the engineering community. A major motivation comes from the need to improve both the naturalness and efficiency of spoken language human-machine interfaces. Researching emotions, however, is extremely challenging for several reasons. One of the main difficulties results from the fact that it is difficult to define what emotion means in a precise way. Various explanations of emotions given by scholars are summarized in [Kleinginna *et al.* 1981]. Research on the cognitive component focuses on understanding the environmental and attended situations that give rise to emotions; research on the physical components emphasizes the physiological response that co-occurs with an emotion or rapidly follows it. In short,

[*] Department of Computer Science and Engineering, Tatung University, 40 ChungShan N. Rd., 3rd Sec, Taipei 104, Taiwan, R.O.C, Tel: +886-2-2592-5252 Ext. 2212, Fax: +886-2-2592-5252 Ext. 2288
E-mail: tlpao@ttu.edu.tw; {d8906005, d9306002, d8906004}@ms2.ttu.edu.tw

[Received January 31, 2005; Revised July 15, 2005; Accepted July 19, 2005]

emotions can be considered as communication with oneself and others [Kleinginna *et al.* 1981].

Traditionally, emotions are classified into two main categories: primary (basic) and secondary (derived) emotions [Murray *et al.* 1993]. Primary or basic emotions generally can be experienced by all social mammals (e.g., humans, monkeys, dogs and whales) and have particular manifestations associated with them (e.g., vocal/facial expressions, behavioral tendencies and physiological patterns). Secondary or derived emotions are combinations of or derivations from primary emotions.

Emotional dimensionality is a simplified description of the basic properties of emotional states. According to the theory developed by Osgood, Suci and Tannenbaum [Osgood *et al.* 1957] and in subsequent psychological research [Mehrabian *et al.* 1974], the computing of emotions is conceptualized as three major dimensions of connotative meaning: arousal, valence and power. In 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.

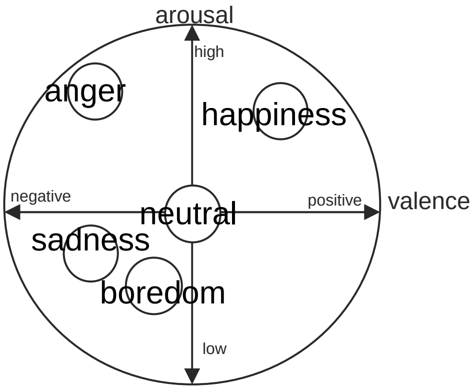

Figure 1. Graphic representation of the arousal-valence dimension of emotions[Osgood et al. 1957]

Numerous previous reports indicated that emotions could be detected by psychological cues [Cowie *et al.* 2000; Ekman 1999; Holzapfel *et al.* 2002; Inanoglu *et al.* 2005; Kleinginna *et al.* 1981; Kwon *et al.* 2003; Murray *et al.* 1993; Nwe *et al.* 2003; Park *et al.* 2002; Park *et al.* 2003; Pasechke *et al.* 2000; Picard 1997; Ververidis *et al.*2004]. Vocal cues are among the fundamental expressions of emotions, on a par with facial expressions [Cowie *et al.* 2000; Ekman 1999; Holzapfel *et al.* 2002; Kleinginna *et al.* 1981; Murray *et al.* 1993; Nwe *et al.*

2003; Park *et al.* 2002; Park *et al.* 2003; Pasechke *et al.* 2000; Ververidis *et al.* 2004]. All mammals can convey emotions by means of vocal cues. Humans are especially capable of expressing their feelings by crying, laughing, shouting and more subtle characteristics of speech.

In this paper, instead of modifying classifiers, we present an effective and robust set of vocal features for recognizing categories of emotions in Mandarin speech. The vocal characteristics of emotions are extracted from a Mandarin corpus. In order to surmount the inefficiency of conventional vocal features, such as pitch contour, loudness, speech rate and duration, for recognizing anger/happiness and boredom/sadness, we also adopt arousal and valence correlated characteristics to categorize emotions in emotional discrete categories. Several systematic experiments are presented. The characteristics of the extracted features are not only facile, but also discriminative.

The rest of this paper is organized as follows. In Section 2, two testing corpora are addressed. In Section 3, the details of the proposed system are presented. Experiments conducted to assess the performance of the proposed system are presented in Section 4 together with analysis of the results of the experiments. Concluding remarks are given in Section 5.

2. The Testing Corpora

An emotional speech database, Corpus I, was specifically designed and set up for emotion classification studies. The database includes short utterances portraying the five primary emotions, namely, anger, boredom, happiness, neutral and sadness. In the course of selecting emotional sentences, two aspects were taken into account. First, the sentences did not have any emotional tendency. Second, the sentences could involve all kinds of emotions. Non-professional speakers were selected to avoid exaggerated expression. Twelve native Mandarin language speakers (7 females and 5 males) were asked to generate the emotional utterances. The recording was done in a quiet environment using a mouthpiece microphone at a sampling rate of 8 kHz.

All of the native speakers were asked to speak each sentence with the five chosen emotions, resulting in 1,200 sentences. We first eliminated sentences that suffered from excessive noise. Then a subjective assessment of the emotion speech corpus by human audiences was carried out. The purpose of the subjective classification was to eliminate ambiguous emotion utterances. Finally, 558 utterances with over 80% human judgment accuracy were selected and are summarized in Table 1. In this study, 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. In order to accommodate the computing time requirement and bandwidth limitation of the practical recognition application, e.g., the call center system [Yacoub *et al.* 2003], a sampling rate of 8 kHz was used. Another corpus, Corpus II, was recorded by Cheng [Cheng 2002]. Two professional Mandarin speakers were employed to generate 503 utterances with five emotions as shown in Table 2. The sampling rate was down-sampled to 8 kHz.

Table 1. Utterances for Corpus I

Sex **Emotion**	**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

Table 2. Utterances for Corpus II

Sex **Emotion**	**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

Utterances can be divided into two sets: one set for training and one set for testing. In this way, several different models, all trained with the training set, can be compared based on the test set. This is the basic form of cross-validation. A better method, which is intended to avoid possible bias introduced by relying on any one particular division into test and train components, is to partition the original set in several different ways and then compute an average score over the different partitions. An extreme variant of this is to split the *p* patterns into a training set of size *p*-1 and a test of size 1, and average the squared error on the left-out pattern over the *p* possible ways of obtaining such a partition. This is called leave-one-out (LOO) cross-validation. The advantage here is that all the data can be used for training; none have to be held back in a separate test set.

3. Emotion Recognition Method

The proposed emotion recognition method has three main stages: feature extraction, feature vector quantization and classification. Base features and their statistics are computed in the feature extraction stage. Feature components are quantized into a feature vector in the feature quantization stage. Classification is done by using various classifiers based on dynamic models or discriminative models.

3.1 Emotion Feature Selection

Determining emotion features is a crucial issue in emotion recognizer design. All selected features have to carry sufficient information about transmitted emotions. However, they also need to fit the chosen model by means of classification algorithms. Important research was done by Murray and Arnott [Murray *et al.* 1993], whose results particularized several notable acoustic attributes for detecting primary emotions. Table 3 summarizes the vocal effects most commonly associated with the five primary emotions [Murray *et al.* 1993]. Classification of emotional states based on prosody and voice quality requires classifying the connections between acoustic features in speech and emotions. Specifically, we need to find suitable features that can be extracted and modeled for use in recognition. This also implies that the human voice carries abundant information about the emotional state of a speaker.

Table 3. Emotions and speech relations [Murray et al. 1993]

	Anger	**Happiness**	**Sadness**	**Fear**	**Disgust**
Speech Rate	Slightly faster	Faster or slower	Slightly slower	Much faster	Very much faster
Pitch Average	Very much higher	Much higher	Slightly lower	Very much higher	Very much lower
Pitch Range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Higher	Higher	Lower	Normal	Lower
Voice Quality	Breathy, chest	Breathy, blaring tone	Resonant	Irregular voicing	Grumble chest tone
Pitch changes	Abrupt on stressed	Smooth, upward inflections	Downward inflections	Normal	Wide, downward terminal inflects
Articulation	Tense	Normal	Slurring	Precise	Normal

A variety of acoustic features have also been explored. For example, Schuller *et al.* chose 20 pitch and energy related features [Schuller *et al.* 2003]. A speech corpus consisting of acted and spontaneous emotion utterances in German and English was described in detail. The accuracy in recognizing 7 discrete emotions (anger, disgust, fear, surprise, joy, neutral and sad) exceeded 77.8%. Park *et al.* used pitch, formant, intensity, speech rate and energy related features to classify neutral, anger, laugh and surprise [Park *et al.*2002]. The recognition rate was about 40% for a 40-sentence corpus. Yacoub *et al.* extracted 37 fundamental frequency, energy and audible duration features for recognizing sadness, boredom, happiness and anger in a corpus recorded by eight professional actors [Yacoub *et al.*2003]. The overall accuracy was only about 50%, but these features successfully separated hot anger from other basic emotions. Tato *et al.* extracted prosodic features, derived from pitch, loudness, duration and quality features [Tato *et al.*2002], from a 400-utterance database. The significant results of emotion recognition were the speaker-independent case and three clusters (high = anger/happy, neutral, low = sad/bored). However, the accuracy in recognizing five emotions was only 42.6%. Kwon *et al.* selected pitch, log energy, formant, band energies and Mel frequency spectral coefficients (MFCC) as base features, and added velocity/acceleration of pitch to form feature streams [Kwon *et al.2003*]. The average classification accuracy achieved was 40.8% in a SONY AIBO database. Nwe *et al.* adopted the short time log frequency power coefficients (LFPC) along with MFCC as emotion speech features to recognize 6 emotions in a 60-utterance corpus produced by 12 speakers [Nwe *et al.*2003]. Results showed that the proposed system yielded an average accuracy of 78%. In [Le *et al.* 2004], the authors proposed a method using MFCC coefficients and a simple but efficient classifying method, Vector Quantization, for performing speaker-dependent emotion recognition. Various speech features, namely, energy, pitch, zero crossing, phonetic rate, LPC and their derivatives, were also tested and combined with MFCC coefficients. The average recognition accuracy achieved was about 70%. In [Chuang *et al.* 2004], Chuang and Wu presented an approach to emotion recognition from speech signals and textual content using PCA and SVM, and achieved 81.49% average accuracy using an extra corpus collected from the same broadcast drama.

According to the experimental results stated above, some simple prosodic features, such as duration, loudness, can not consistently distinguish all primary emotions. Furthermore, the prosodic features of females and males are obviously intrinsic in speech. The simple speech energy feature calculation method is also unconformable to human auricular perception.

Figure 2 shows a block diagram of the feature extraction process. In the pre-processing procedure, locating the endpoints of the input speech signal is done first. The speech signal is high-pass filtered to emphasize the important high frequency components. Then the speech frame is partitioned into frames consisting of 256 samples each. Each frame overlaps with the adjacent frames by 128 samples. The next step is to apply the Hamming window to each individual frame to minimize the signal discontinuities at the beginning and end of each frame. Each windowed speech frame is then converted into several types of parametric representations for further analysis and recognition.

In order to find a suitable combination of extracted features, we used the regression selection method to determine beneficial features from among more than 200 speech features. Ten candidates were selected: LPC, LPCC, MFCC, Delta-MFCC, Delta-Delta-MFCC, PLP, RastaPLP, LFPC, jitter and shimmer. Then the feature vector of each frame of a sentence from corpus I was calculated. The recognition rate in each step was calculated using the LOO cross-validation method with the K-NN (K=3) classifier.

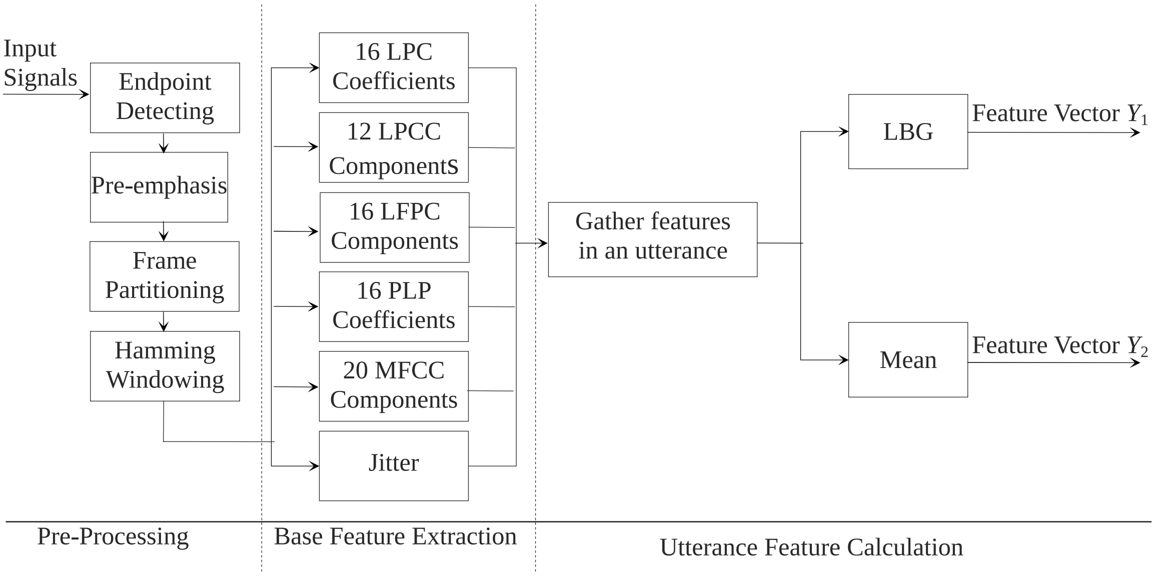

Figure 2. Block diagram of the feature extraction module

Table 4 shows the recognition rate of the first 10 candidates. The highest recognition rate was found to be 83.91% using the forward selection procedure shown in Table 6. In this procedure, the recognition rate grows or declines according to the effectiveness of feature combining. Tables 4-6 list the results of forward selection with 1, 2 and 6 features. Based on these experimental results, we selected six features, which were LPCC, MFCC, LFPC, jitter, PLP and LPC, as a beneficial feature combination for speech emotion recognition.

Table 4. The recognition rate with single feature

Feature	**Accuracy (%)**
LPCC	68.68
MFCC	68.21
LPC	68.20
PLP	65.59
RastaPLP	65.23
D-MFCC	60.59
LFPC	58.42
Shimmer	53.05
D-D-MFCC	50.18
Jitter	34.77

Table 5. The recognition rate with two feature sets

Feature		Accuracy (%)
LPCC	MFCC	68.97
	D-MFCC	67.38
	LPC	66.52
	PLP	66.52
	LFPC	66.52
	RastaPLP	66.16
	D-D-MFCC	60.06
	Jitter	54.33
	Shimmer	42.86

Table 6. The recognition rate with six feature sets

Feature		Accuracy (%)
LPCC **MFCC** **LFPC** **Jitter** **PLP**	LPC	83.91
	RastaPLP	83.91
	D-MFCC	83.19
	D-D-MFCC	83.19
	Shimmer	79.40

In the base feature extraction procedure, we selected six types of features, which were 16 Linear predictive coding (LPC) coefficients, 12 linear prediction cepstral coefficients (LPCC), 16 log frequency power coefficients (LFPC), 16 perceptual linear prediction (PLP) coefficients, 20 Mel-frequency cepstral coefficients (MFCC) and jitter extracted from each frame. This added up to a feature vector consisting of 81 parameters. LPC provides an accurate and economical representation of the envelope of the short-time power spectrum of speech [Kaiser 2002]. For speech emotion recognition, LPCC and MFCC are popular choices as they represent the phonetic content of speech and convey information about short time energy migration in the frequency domain [Ata 1997; Davis *et al.* 1980]. LFPC is calculated using a log frequency filter bank, which can be regarded as a model that shows the varying auditory resolving power of the human ear for various frequencies [Nwe *et al.*2003]. The combination of the discrete Fourier transform (DFT) and LPC technique is called PLP [Hermansky 1990]. PLP analysis is computationally efficient and permits a compact representation. Perturbations in the pitch period are called jitter. Such perturbations occur naturally during continuous speech.

3.2 Feature Vector Quantization

Each feature vector consists of 81 parameters, which requires intensive computation when classification is performed. To compress the data in order to accelerate the classification process, vector quantization is performed. All the vectors of a frame falling into a particular cluster are coded with the vector representing that cluster. The vector is assigned the codeword c_n^*, according to the best matching codebook cluster. An experiment was conducted with different numbers of centroids obtained using the Linde-Buzo-Gray(LBG) K-means algorithm [Linde *et al.* 1980]. It was found that the effectiveness per centroid diminished significantly when the size exceeded 16. In this study, we took 16 as the number of LBG centroids in all of the experiments. For each utterance with *N* frames, the feature vector Y_1 with 16*81 parameters was then obtained in the form

$$Y_1 = [c_1^* c_2^* ... c_N^*]. \tag{1}$$

Another simple vector quantization method used the mean of the feature parameters corresponding to each frame in one utterance to form a feature vector Y_2 with 81 parameters as follows:

$$Y_2 = [p_1 p_2 ... p_{81}], \tag{2}$$

where p_i is the mean value of the *i*th parameter of all frames.

3.3 Classifiers

Three different classifiers, linear discriminate analysis (LDA), the k-nearest neighbor (K-NN) decision rule, and Hidden Markov models (HMMs), were used to train and test these two testing emotion corpora with the extracted features from Corpus I. In the K-NN decision rule, there are three nearest samples that are closest to the testing sample. In HMMs, the state transition probabilities and output symbol probabilities are uniformly initialized. Our experimental results show that the 4-state discrete ergodic HMM achieved the best performance compared with the left-right structure.

4. Experimental Results

The selected features were quantized using the LBG algorithm to form the vector Y_1 and quantized using the mean method to form vector Y_2. Then the feature vectors were trained and tested with all three classifiers, which were LDA, K-NN and HMMs. All of the experimental results were validated using the LOO cross-validation method.

4.1 The Experimental Results Obtained with the Conventional Prosodic Features

In [Kwon *et al.* 2003], Kwon *et al.* drew a two-dimensional plot of 59 features ranked by means of forward selection and backward elimination. Features near the origin were considered to be more important. By imitating the ranking features method as in [Kwon *et al.* 2003], we could rank the speech features extracted from Corpus I through forward selection and backward elimination as shown in Figure 3. Our experimental results and the Kwon's both show that the pitch and energy related features are the most important components for emotion speech recognition in both Mandarin and English. We selected the first 15 features proposed in [Kwon *et al.* 2003] from Corpus I to examine the efficiency and stability of the conventional emotion speech features. The first 15 features were pitch, log energy, F1, F2, F3, 5 filter bank energies, 2 MFCCs, delta pitch, acceleration of pitch and 2 acceleration MFCCs. Then the feature vector Y_2 and K-NN were used.

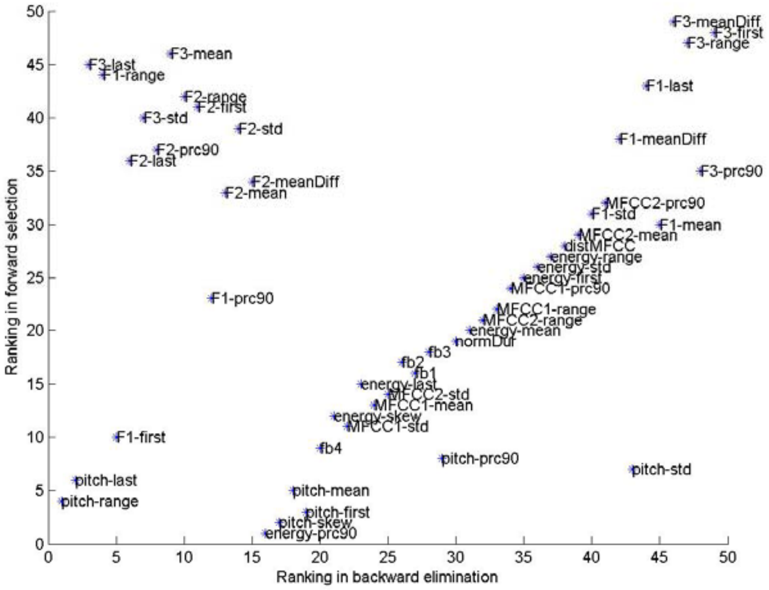

Figure. 3 Ranking of conventional speech features

The confusion matrix that employs conventional emotion speech features is shown in Table 7. The overall average accuracy achieved for the five primary emotions was 53.2%. Similar to most of the previous researches, the pitch and energy related features extracted from the time domain had difficulty distinguishing anger and happiness. The reason is that anger and happiness are close to each other in pitch and energy. Hence, the classifiers often confuse one with the other. This also applies to boredom and sadness.

Table 7. Experimental results obtained using conventional prosodic features

Accuracy (%)	Anger	Boredom	Happiness	Neutral	Sadness
Anger	**59.5**	1.1	32.4	4.4	2.6
Boredom	0	**46.8**	1.1	20.4	31.7
Happiness	32.4	2.5	**58.7**	4.2	2.2
Neutral	9.4	7.7	8.7	**52.1**	22.1
Sadness	1.7	29.4	2.4	17.6	**48.9**

4.2 Experimental Results of Valence Emotions Recognition

The prosodic features related to pitch and energy failed to distinguish the valence emotions. The selected features discussed in Section 3.1 were quantized into feature vector Y_1 and mean feature vector Y_2. The feature vectors from Corpus I were then trained and tested using three different classifiers, the LDA, K-NN and HMMs. All the experimental results were validated using the LOO cross-validation method. According to the experimental results shown in Tables 8 and 9, the three recognizers were undoubtedly able to separate anger and happiness, which most previous emotion speech recognizers usually confuse.

The pairwise emotions, anger and happiness, are considered to be close to each other in the arousal dimension, having similar prosody and amplitude. So do boredom and sadness. The conventional speech emotion recognition method suffers from ineffectiveness and instability in emotion recognition, especially for emotions in the same arousal dimension. On the other hand, using the selected features in the proposed system solves this problem and results in a high recognition rate. The selected features are not only suitable for various classifiers but also effective for speech emotion recognition.

Table 8. 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 9. 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

4.3 Experimental Results for Corpus I and Corpus II

Tables 10 and 11 show the accuracy achieved in classifying the five primary emotions using various classifiers and two feature vector quantization methods applied to Corpus I and II. The various classifiers differ in ability and properties. Hence we achieved various recognition accuracy results with the different classifiers and quantization methods.

Table 10. Experimental results for five emotion categories in Corpus I

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 11. Experimental results for emotion categories in Corpus II

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

According to the experimental results shown in Tables 10 and 11, the overall accuracy rates achieved for the five primary emotions, namely, anger, boredom, happiness, neutral and sadness, were about the same. In addition, the accuracy rates of the two feature quantization methods were quite close to each other when used under the same conditions. This shows that the set of selected speech features is stable and suitable for recognizing the five primary emotions, using various classifiers with different feature quantization methods. Based on the high recognition accuracy rates achieved for Corpus I and Corpus II, the selected features can be efficiently used to classify the five primary emotions of the arousal and the valence degree simultaneously.

Two different corpora were used to validate the robustness and effectiveness of the selected features. From the experimental results shown in Tables 10 and 11, the overall recognition rates obtained for both corpora are similar.

5. Conclusion

Dealing with the emotions of speaker is one of the challenges for speech processing technologies. Whereas the research on automated recognition of emotions in facial expressions has been quite extensive, that focusing on speech modality, both for automated production and recognition by machines, has been active only in recent years and has mostly focused on English. Possible applications include intelligent speech-based customer information systems, human oriented human-computer interaction GUIs, interactive movies, intelligent toys and games, situated computer-assisted speech training systems and supported medical instruments.

The selection of a feature set is a critical issue for all recognition systems. In the conventional approach to emotion classification of speech signals, the features typically employed are the fundamental frequency, energy contour, duration of silence and voice quality. However, previous proposed recognition methods employing these features perform poorly in recognizing valence emotions. In addition, these features, when applied to different corpora, obtain different recognition results with the same recognizer.

In this study, we combined 16 LPC coefficients, 12 LPCC components, 16 LFPC components, 16 PLP coefficients, 20 MFCC components and jitter as features, and used LDA, K-NN and HMMs as the classifiers. The emotions were classified into five human primary categories: anger, boredom, happiness, neutral and sadness. Two Mandarin corpora, one consisting of 558 emotional utterances made by 12 native speakers and the other consisting of 503 emotional utterances made by 2 professional speakers, were used to train and test the proposed recognition system. Results obtained show that the proposed system yielded top recognition rates of 88.3% for Corpus I and 88.7% for Corpus II.

According to the experimental outcomes, we attained high recognition rates in distinguishing anger/happy and bored/sad emotions, which have similar prosody and amplitude. The proposed method can solve the problem of recognizing valence emotions using a set of extracted features. Moreover, the recognition accuracy results for Corpus I and Corpus II show that the selected speech features are suitable and effective for the speech emotion recognition with different corpora.

Further improvement and expansion may be achieved according to the following suggestions: The set of the most efficient features for emotion recognition is still vague. A possible approach to extracting non-textual information to identify emotional states in speech is to apply all known feature extraction methods. Thus, we may try to incorporate the information of different features into our system to improve the accuracy of emotion recognition. Recognizing emotion translation in real human communication is also a challenge. Thus, it will be worth while to determine the points where emotion transitions occur.

Acknowledgement

The authors are thankful to the National Science Council of Taiwan, R.O.C., for financial supports of this work (grant no. 93-2213-E-036-023 and 92-2213-E-036-021).

References

Ata, B.S., “Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification,” *Journal of the Acoustical Society of America*, 1(55), 1974, pp.1304-1312.

Cheng, P.Y., “Automated Recognition of Emotion in Mandarin,” MD thesis, National Cheng Kung University, 2002.

Chuang, Z.J. and C.H. Wu, “Multi-Modal Emotion Recognition from Speech and Text,” *International Journal of Computational Linguistics and Chinese Language Processing*, 9(2), 2004, pp.1-18.

Cowie, R., E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J.G. Taylor, “Emotion Recognition in Human-Computer Interaction,” *IEEE Signal Processing Magazine*, 18 (1), 2000, pp.32-80.

Davis, S. and P. Mermelstein, “Comparison of Parametric Representations of Monosyllabic Word Recognition in Continuously Spoken Sentences,” *IEEE Transactions on Acoustics*, 28(4), 1980, pp.357-366.

Ekman, P., *Handbook of Cognition and Emotion*, John Wiley & Sons, New York, 1999.

Hermansky, H., “Perceptual Linear Predictive (PLP) Analysis of Speech,” *Journal of the Acoustical Society of America*, 87(4), 1990, pp.1738-1752.

Holzapfel, H., C. Fügen, M. Denecke and A. Waibel, “Integrating Emotional Cues into a Framework for Dialogue Management,” In *Proceedings of International Conference on Multimodal Interfaces*, 2002, Pennsylvania, USA, pp.141-148.

Inanoglu, Z. and R. Caneel, “Emotive Alert: HMM-Based Emotion Detection in Voicemail Messages,” In *Proceedings of Intelligent User Interfaces*, 2005, San Diego, USA, pp.251-253.

Kaiser, J.F., *Discrete-Time Speech Signal Processing*, Prentice Hall, New Jersey, 2002.

Kleinginna , P.R. and A.M. Kleinginna, “A Categorized List of Emotion Definitions with Suggestions for a Consensual Definition,” *Motivation and Emotion*, 5(4), 1981, pp.345-379.

Kwon, O.W., K. Chan, J. Hao and T.W. Lee , “Emotion Recognition by Speech Signals,” In *Proceedings of Eurospeech*, 2003, Geneva, Switzerland, pp.125-128.

Le, X.H., G. Quenot and E. Castelli, “Recognizing Emotions for the Audio-Visual Document Indexing,” In *Proceedings of the Ninth IEEE International Symposium on Computers and Communications*, 2004, Alexandria, Egypt, pp.580-584.

Linde, Y., A. Buzo and R.M. Gray, “An Algorithm for Vector Quantizer Design,” *IEEE Transactions on Communications*, 28(1), 1980, pp.84-95.

Mehrabian, A. and J. Russel, *An Approach to Environmental Psychology*, The MIT Press, Cambridge, 1974.

Murray, I. 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*, 93(2), 1993, pp.1097-1108.

Nwe, T.L., S.W. Foo and L.C. De-Silva, “Speech Emotion Recognition Using Hidden Markov Models,” *Speech Communication*, 41(4), 2003, pp.603-623.

Osgood, C.E., J.G. Suci and P.H. Tannenbaum, *The Measurement of Meaning*, The University of Illinois Press, Urbana, 1957.

Park, C.H., 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*, 2(2), 2002, pp.122-126.

Park, C.D. and K.B. Sim, “Emotion Recognition and Acoustic Analysis from Speech Signal,” In *Proceedings of International Joint Conference on Neural Networks*, 2003, Portland, USA, pp.2594-2598.

Pasechke, A. and W.F. Sendlmeier, “Prosodic Characteristics of Emotional Speech: Measurements of Fundamental Frequency Movements,” In *Proceedings of ISCA Workshop on Speech and Emotion*, 2000, Northern Ireland, pp.75-80.

Picard, R.W., *Affective Computing*, The MIT Press, Cambridge, 1997.

Schuller, B., G. Rigoll, and M. Lang, “Hidden Markov Model-based Speech Emotion Recognition,” In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2003, Hong Kong, China, pp.401-405.

Tato, R.S., R. Kompe and J.M. Pardo., “Emotional Space Improves Emotion Recognition,” In *Proceedings of International Conference on Spoken Language Processing*, 2002, Colorado, USA, pp.2029-2032.

Ververidis, D., C. Kotropoulos and I. Pitas, “Automatic Emotional Speech Classification,” In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing,* 2004, Montreal, Canada, pp.593-596.

Yacoub, S., S. Simske, X. Lin and J. Burns, “Recognition of Emotions in Interactive Voice Response Systems,” In *Proceedings of Eurospeech*, 2003, Geneva, Switzerland, pp.729-732.