22

Emotion Recognition via Continuous Mandarin Speech

Tsang-Long Pao, Jun-Heng Yeh and Yu-Te Chen
Tatung University
Taiwan, R.O.C.

1. Introduction

Emotion plays a significant role in cognitive psychology, behavioural sciences and humanoid robot design. The continuing improvements in speech recognition technology have led to many new and fascinating applications in human-computer interaction, context aware computing and computer mediated communication. A growing number of research studies in emotion recognition via an isolated short sentence are available to shed some light on the implementation of human-computer interface. However, to the best of our knowledge, no work has focused on automatic emotion tracking from continuous Mandarin speech. In this chapter, we will elaborate an emotion recognition method in continuous Mandarin speech, by dividing the utterance into independent segments, each of which contains a single emotional category.

In the growing range of interactive interfaces, the research of emotional voice is still at an early stage, not to mention a paucity of literatures 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. To date, no clear direction has emerged to suggest how such considerations translate into practical interface design. The crux of this problem is that the emotion recognition in continuous speech has not yet been much explored.

From the viewpoint of communication, it is natural for human beings to communicate with others in continuous dialogue. Even though, most proposed methods of emotion recognition 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 chapter attempts to address these areas by processing speech signals rather than interpreting the lexicons of speaking. Moreover, the benefit from the outlook of processing speech signals can also tack the violent change of emotional expression in dialogue. In light of these concerns, this chapter has three purposes: (a) to report on trends in published research in the major journals of emotion recognition; (b) to provide a method in recognition of emotion from continuous Mandarin speech; and (c) to recommend promising research paradigms for recognition of emotion via continuous speech.

This chapter is organized as follows. In section 2, related works are presented. In section 3, the testing corpus is introduced. In section 4, the proposed speech recognition method is

presented in detail. In section 5, the experimental results are shown and commented. The chapter concludes in section 6 showing directions for future research and conclusions.

2. Emotions and Speech

Research on understanding and modelling 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 & Kleinginna, 2005]. 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 immediately follows it. In short, emotions can be considered as communication with oneself and others [Kleinginna & Kleinginna, 2005].
For research related to continuous speech signal, most works are found on the continuous speech recognition. Also, most of the emotion recognition researches are based on short sentences. However, human beings speak continuously. People will change emotions when they are triggered by some incidents in the course of speaking. The short-sentence emotion recognition system may not be able to detect the emotional state correctly because there may have several emotions in a long conversation. One objective of this chapter is to find a proper segmentation algorithm to segment the continuous speech and to develop a method to recognize emotion of each segment correctly so we can track emotion changes of the speaker.

2.1 Emotional Categories

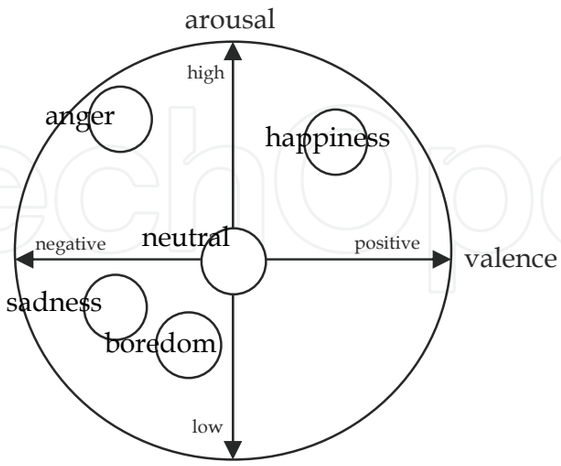

Figure 1. Graphic representation of the arousal-valence dimension of emotions [Osgood et al., 1967]

Traditionally, emotions are classified into two main categories: primary (basic) and secondary (derived) emotions [Murray & Arnott, 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., 1967] and in subsequent psychological research [Mehrabian & Russel, 1974], the computing of emotions is conceptualized as three major dimensions of connotative meaning: arousal, valence and domaniance. 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.

2.2 Speech Features of Emotional Expressions

Speech communication is one of the basic and most essential capabilities possessed by human beings. Emotions play an important role in human-to-human communication and interaction, allowing people to express themselves beyond the verbal domain. Interactions between people not merely transmit through the speech, but also include behaviours, emotion language, heart, and spirit [Sebe et al., 2005]. Detecting emotions in speech is a topic that has been predominantly dealt with in psychology and linguistics. It is attracting the attention of engineering community also. To recognize emotions, we need to know not only what information a user conveys but also how it is being conveyed.

Numerous previous reports indicated that emotions could be detected by psychological cues [Busso & Narayanan, 2007, Cowie et al., 2000, Ekman, 1999, Holzapfel et al., 2002, Inanoglu & Caneel, 2005, Kleinginna & Kleinginna, 1981, Kwon et al., 2003, Murray & Arnott, 1993, Nwe et al., 2003, Park et al., 2002, Park & Sim, 2003, Pasechke & Sendlmeier, 2000, Picard, 1997, Ramamohan, & Dandapa, 2006, Schröder, 2006, Tao et al., 2006, Ververidis et al., 2004]. Vocal cues are among the fundamental expressions of emotions, on a par with facial expressions [Busso & Narayanan, 2007, Cowie et al., 2000, Ekman, 1999, Holzapfel et al., 2002, Kleinginna & Kleinginna, 1981, Murray & Arnott, 1993, Nwe et al., 2003, Park et al., 2002, Park & Sim, 2003, Pasechke & Sendlmeier, 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.

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 & Arnott, 1993], whose results particularized several notable acoustic attributes for detecting primary emotions. Table 1 summarizes the vocal effects most commonly associated with the five primary emotions [Murray & Arnott, 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 modelled for use in recognition. This also implies that the human voice carries abundant information about the emotional states of a speaker.

	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

Table 1. Emotions and speech relations [Murray & Arnott, 1993]

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 (VQ), for performing speaker-dependent emotion recognition. Various speech features, namely, energy, pitch, zero crossing, phonetic rate, linear predictive coding (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 the principal component analysis (PCA) and the support vector machine (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 and 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.

3. The Testing Corpora

An emotional speech database, Corpus I, was specifically designed and set up for emotion recognition studies. The database includes short sentences portraying the five primary emotions, including 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 speakers (7 females and 5 males) were asked to generate the emotional utterances. The recording format is mono channel pulse-code modulation (PCM) with sampling rate of 44.1 kHz and 16-bit resolution.

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

Table 2. Corpus I

Combining emotions	Combined sentences
Angry-Happy (AH)	35
Angry-Sad (AS)	37
Angry-Bored (AB)	25
Angry-Neutral (AN)	34
Happy-Sad (HS)	27
Happy-Bored (HB)	24
Happy-Neutral (HN)	26
Sad-Bored (SB)	22
Sad-Neutral (SN)	29
Bored-Neutral (BN)	20
Total sentences	279

Table 3. Corpus II

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 2. 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.

The continuous emotional corpus, Corpus II, used in the experiment is obtained by combining the short emotional sentences in the corpus database. There are ten kinds of combination of emotion as shown in Table 3. Each combined utterance is from the same speaker. Every combined utterance consists more than five short sentences. There are 277 combined semntences for the experiments.

Sentences 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.

4. Speech Processing of Emotion Recognition

In this section, we present an emotion tracking system, by dividing the utterance into several independent segments, each of which contains a single emotional category. Figure 2 shows the block diagram of the emotion recognition from continuous Mandarin speech signal.

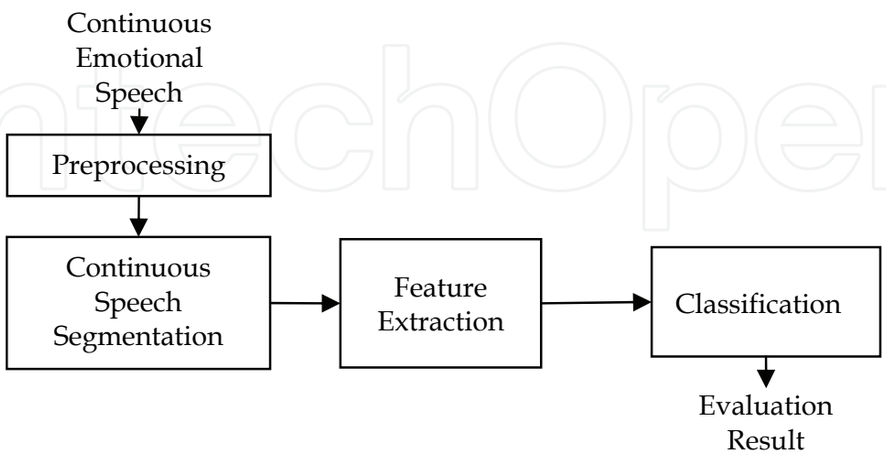

Figure 2. The block diagram of the block diagram of the emotion recognition from continuous mandarin speech signal

4.1 Pre-Processing

The signal from the microphone is an analogy signal. It is essential to transform the analogy signal into digital form so the computer can be used to process the signal. Before the emotion recognition can be done, the input speech signal has to go through some pre-processing. To deal with the discrete-time signal $x(n)$, framing is used to divide the speech signal into sections. In this study, 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 as shown in Eq. (1) 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. Figure 3 depicts the result of frame partition.

$$w(n) = \begin{cases} 0.54 - 0.46\cos(\frac{2n\pi}{N-1}), & 0 \le n \le N-1 \\ 0 & , \text{otherwise.} \end{cases} \quad (1)$$

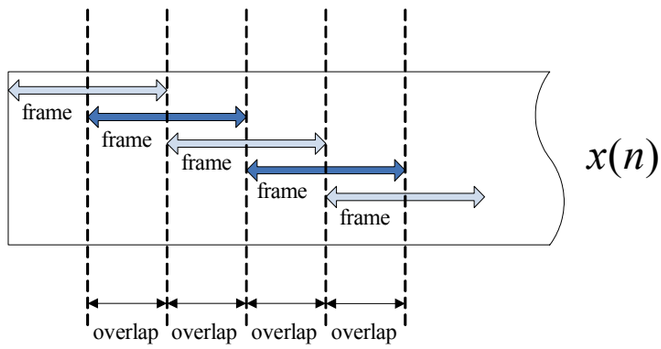

Figure 3. Frame partition of a sequence $x(n)$

4.2 Feature Selection

In order to find a suitable combination of extracted features, we used the regression selection method to determine beneficial features from among more than 100 speech features. 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-Nearest Neighbour (KNN) decision rule (K=3) classifier. Finally, 10 candidates were selected: LPC, linear prediction cepstral coefficient (LPCC), MFCC, Delta-MFCC (dMFCC), Delta-Delta-MFCC (ddMFCC), perceptual linear prediction (PLP), RelAtive SpecTrAl PLP (RastaPLP), LFPC, pitch and formants (F1, F2 and F3).

The feature selection is to reduce the dimensions of feature set. And the forward feature selection (FFS) and backward feature selection (BFS) are used to decrease the computational complexity. FFS and BFS correspond to growing and shrinking feature one at a time, respectively. The FFS starts from an empty set and sequentially adds features, whereas BFS starts from the full set and sequentially removes features. In FFS, the starting set is empty. It then chooses a best single one and adds it to the set. The next step is to choose the second best one. The step repeated until the criteria are full fill. In BFS, the starting set is all the features. It removes the worst one remaining in the set step by step. Figure 4 shows the feature ranking of these 10 speech features by FFS and BFS using KNN.

Without loss of generality, the collected speech samples are split into t data elements $X_1, \cdots, X_t$. The space of all possible data elements is defined as the input space X. The elements of the input space are mapped into points in a feature space F. In our work, a feature space is a real vector space with dimension n, Γ^n. Accordingly, each point f_i in F is represented by an n-dimensional feature vector:

$$f_i = (MFCC_{i1}, \cdots, MFCC_{im}, \cdots, LPCC_{i1}, \cdots, LPCC_{iq}), \tag{2}$$

where m, q are the dimension of MFCC and LPCC respectively, and

$$n = m + q. \tag{3}$$

Finally, MFCC and LPCC are individually obtained from FFS and BFS as the most important features. In the field of speech recognition, LPCC and MFCC are the popular choices as features representing the phonetic content of speech. For each speech frame, 12 LPCC components and 20 MFCC components are used in this study.

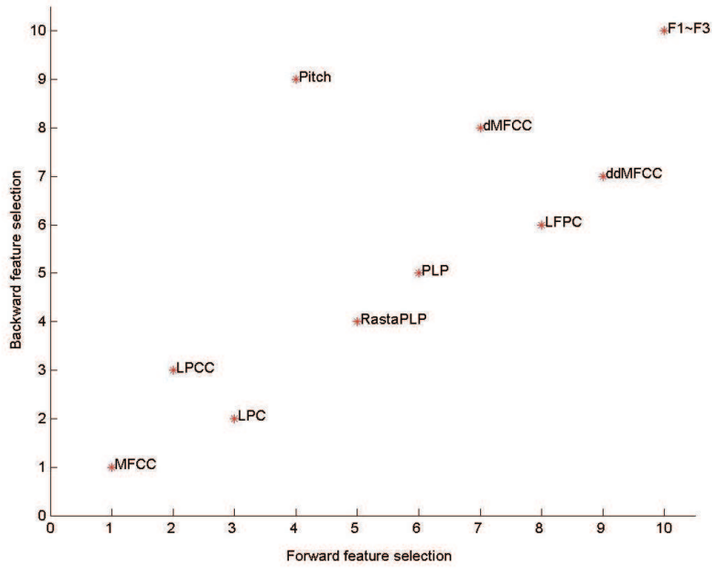

Figure 4. Feature ranking of 10 speech features

4.3 Classifier

A feature map is defined as a function that takes an element in the input space and maps it to a point in the feature space. We use ϕ to define a feature map, that is

$$\phi : X \to F. \tag{4}$$

Being simple, elegant and straightforward, many researchers often adopt KNN as the classifier for their applications. It is an instance-based learning algorithm and classifies unlabeled data based on the similarities with data in the training set. When a new test data x arrives, KNN finds the k neighbours nearest to the unlabeled data from the training space

based on a suitable distance measure. In this study, the Euclidean distance is used. That is, given two samples in the input space x_1 and x_2, the Euclidean distance between them in the feature space is defined as

$$\begin{aligned} & d(x_1, x_2) \\ & = \left|\phi(x_1) - \phi(x_2)\right| \\ & = \left|f_1 - f_2\right| \\ & = \sqrt{(MFCC_{1m} - MFCC_{2m})^2 + \cdots + (LPCC_{1q} - LPCC_{2q})^2}. \end{aligned} \tag{5}$$

Assume that we want to classify the data into one of the l classes and let the k prototypes nearest to x be $N_k(x)$ and $c(y)$ be the class label of y. Then the subset of nearest neighbours within class $j \in \{1, \cdots, l\}$ is

$$N_k^j(x) = \left\{ y \in N_k \left((y) : c(y) = j \right) \right\}. \tag{6}$$

The classification result $j^* \in \{1,\ldots,l\}$ is then defined as a majority vote:

$$j^* = \arg \max_{j=1,\cdots,l} \left| \mathrm{N}_{\mathrm{k,i}}^{\mathrm{j}}(x) \right| . \tag{7}$$

Modified-KNN (M-KNN) is a technique based on the KNN [Pao et al., 2008]. It is based on the comparison of similarity among samples in each class. An unknown sample can be viewed as a point in the n-dimensional feature space, then the k nearest points of the training samples in each class are found by using Euclidean distance as similarity measure. The distance between unknown sample and the ith nearest point in class j is defined as d_j^i.

Then, the classification result $j^* \in \{1, \cdots, l\}$ is obtained by summing up the distance values in each class and picking up the smallest one, that is

$$j^* = \arg \min_{j=1,\cdots,l} \sum_{1}^{k} d_i^j. \tag{8}$$

In this study, we use a weighted D-KNN to improve the performance of M-KNN. M-KNN assigns equal weight to each neighbour. This may cause confusion when there are some irrelevant data in the training set. One obvious refinement to M-KNN is to weight the contribution of each of the k neighbours in each class. The purpose of weighting is to find a vector of real-valued weights that would optimize classification accuracy of the classification or recognition system by assigning lower weights to less relevant features and higher weights to features that provide more reliable information. Let $x_i^j, i = 1, \cdots, z_j$, be the training samples of class j, where z_j is the number of samples belonging to class j. The total number t of training samples is

$$t = \sum_{j=1}^{l} z_j \tag{9}$$

When a test sample x and Euclidean distance measure d_j^i are given, we obtain the k nearest neighbours belonging to class j, $M_j^k(x)$, which is defined as

$$\forall x_i^j \in M_k^j(x), x_p^j \notin M_k^j(x) \Rightarrow d_i^j < d_p^j, \tag{10}$$

$$d_i^j = d\left(x_i^j, x\right), \tag{11}$$

$$d_p^j = d\left(x_p^j, x\right), \tag{12}$$

where the cardinality of the set $\left|M_k^j(x)\right|$ is k. Among the k nearest neighbours in class j, the following relationship is established:

$$d_1^j \le d_2^j \le \cdots \le d_k^j, \tag{13}$$

Let w_i be the weight of the ith nearest samples. From above, we can know that the one have the smallest distance value d_1^j is the most important. Consequently, we set a constraint $w_1 \ge w_2 \ge \cdots \ge w_k$ to conform the idea of weighting. Then, the classification result $j^* \in \{1, \cdots, l\}$is defined as

$$j^* = \arg\ \min_{j=1,\cdots,l} \sum_{i=1}^{k} w_i d_i^j \tag{14}$$

4.4 Segmentation of Emotional Expressions from Continuous Mandarin Speech

In previous studies, the methods to segment continuous speech signal are usually applied in the speech recognition system. To recognize the emotion from continuous emotional speech, the first step is to segment the sentence by finding out the changing points. And emotion of each segment is recognized individually.

People may pause for a while when they turn one emotion to the other emotion. In [Lu et al., 2006], two thresholds are defined to preliminarily quantize the energy envelope into three levels instead of binaries. Therefore, it will generate a larger number of potential partition boundaries whenever either threshold is crossed in the energy envelope. This method is applied in this study. The thresholds, T_L and T_U, are defined as

$$T_L = \mu_E - 0.5\sigma_E. \tag{15}$$

$$T_U = \mu_E + 0.5\sigma_E. \tag{16}$$

where μ_E is the mean energy of partition, and σ_E is the standard deviation of the energy of that partition. Then, the points of the energy contour crossing T_L and T_U are checked every two frames. It will place a partition between two points. The partition energy is represented as

$$E_x(m) = \sum_{n=m-N+1}^{m} \left|f_x(n;m)\right|^2 . \tag{17}$$

The feeling of the sound intensity perceived by human ears is not linear but rather logarithmic. Thus, it is better to express the energy function in logarithmic form.

$$E_x(m) = 10 \times \log[\,|\sum_{n=m-N+1}^{m} |f_x(n;m)|^2\,|\,]\,. \tag{18}$$

The μ_E and σ_E are defined as

$$\mu_E = \frac{1}{N_f}\sum_{l=1}^{N_f} E_{x,f}\,, \tag{19}$$

and

$$\sigma_E{}^2 = \frac{1}{N_f}\sum_{l=1}^{N_f}(E_f - \mu_E)^2\,, \tag{20}$$

where *f* is the frame index and N_f is the number of frames.

The mean energy between two intersection points will be calculated. In order to determine the silence, the energy of the partition and T_L are compared after obtaining the intersection points. When the energy of the partition is greater than T_L, the adjacent partitions will be merged. If the merged duration L_i is less than a threshold, the adjacent partitions will be combined again. The threshold *T* is set as the average of the duration

$$T = \frac{1}{N_L}\sum_{i=1}^{N_L} L_i\,, \tag{21}$$

where N_L is the number of the merged duration. After this processing step, the segmented partitions are recognized separately.

4.5 Segmentation with Endpoint Detection

In the 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 energy is calculated according to Equation (18).

There is a zero line on the speech signal. When a zero-crossing occurred, the amplitude is either from the positive to negative or from the negative to positive. The number of zero-crossing of the speech signal in a predetermined time interval, which is counted as the number of times when adjacent sample points have different signs, approximately corresponds to the frequency of the major spectral component. The calculating of the number of zero-crossing in a partition gives the zero-crossing rate. The equation is defined as

$$Z_x(m) = \sum_{n=m-N+1}^{m} \frac{1}{2}\Big|\ \mathrm{sgn}\big[x(n)\big] - \mathrm{sgn}\big[x(n-1)\big]\Big|, \tag{22}$$

where sgn[.] is defined as

$$\mathrm{sgn}[y] = \begin{cases} 1, & y \geq 0 \\ -1, & y < 0 \end{cases}. \tag{23}$$

The absolute value of sgn[x(n)]-sgn[x(n-1)] will be 2 when x(n) and x(n-1) are different in sign, and is 0 otherwise.

The equations for the two energy thresholds and one zero-crossing rate threshold is defined as follows

$$T_L = \mu_E + \alpha_1 \sigma_E, \tag{25}$$

$$T_U = \mu_E + \alpha_2 \sigma_E, \ \alpha_1 < \alpha_2, \tag{26}$$

$$T_Z = \mu_Z + \alpha_3 \sigma_Z. \tag{27}$$

The α_1, α_2, and α_3 are parameters, which are obtained by experiments.

In the sequence of partitions, the first partition with energy greater than T_L is labelled as N_B. If the energies of the next B successive frames are 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, 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 labelled 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 labelled 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 step, the number of the segmented partitions are still too large to process. So, it is necessary to reduce the numbers 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 all the merged segments are calculated individually as the threshold for the corresponding segment. If the length of the frame is less than the threshold, it will be merged with adjacent frames. The threshold is obtained by experiments.

4.6 Emotion Evaluation

The values used in the evaluation are calculated as follows

$$Eva_j = [(\sum_{i=1}^{k} w_i d_j^i)^{-1}]^2. \tag{28}$$

Equation (28) is used to get the scores of the test sample corresponding to each emotion. Five values are obtained with respect to five emotion categories. The five values indicate emotion components of the test sample correspond to each emotional state. The five evaluation values represent the scores for each emotional state. The value of the score is normalized so it is between 0 and 1.

5. Emotion Recognition from Continuous Mandarin Speech

In this section, the method of emotion recognition is presented. And several experimental results are discussed.

5.1 Experimental Results of Segmentation with Silence

In this method, two thresholds are utilized to locate the intersection points between energy contour and thresholds. In this experiment, it is checked every two frames. In Fig. 5, the top figure shows the result of the intersection points between energy contour and the two thresholds. The bottom figure shows the mean energy in every segmented partition. The short horizontal line is the mean energy in every partition, and the long horizontal line is the threshold T_L. Application of the procedure will result a lot of partitions which is not easy to process. Therefore, it needs some post processings. First, the mean energy is calculated in every partition. If the values of the mean energy in adjacent partitions are all smaller or all greater than T_L, these partitions are merged.

Figure 6 is the enlarged plot from the marked rectangle in Fig. 5. Figure 7 shows the merged results. The top one shows the result of the intersection points between energy contour and the two thresholds. The bottom one shows the merged results for partition with similar mean energy. The horizontal line indicates the threshold T_L.

Figure 8 shows the result of segmentation with silence using threshold T_L. The segmentation result is still not good enough, so another threshold is needed to combine small partitions. The threshold T is set to the average of the duration of all the partitions. When the adjacent partitions all have mean energy smaller than T, they are combined together.

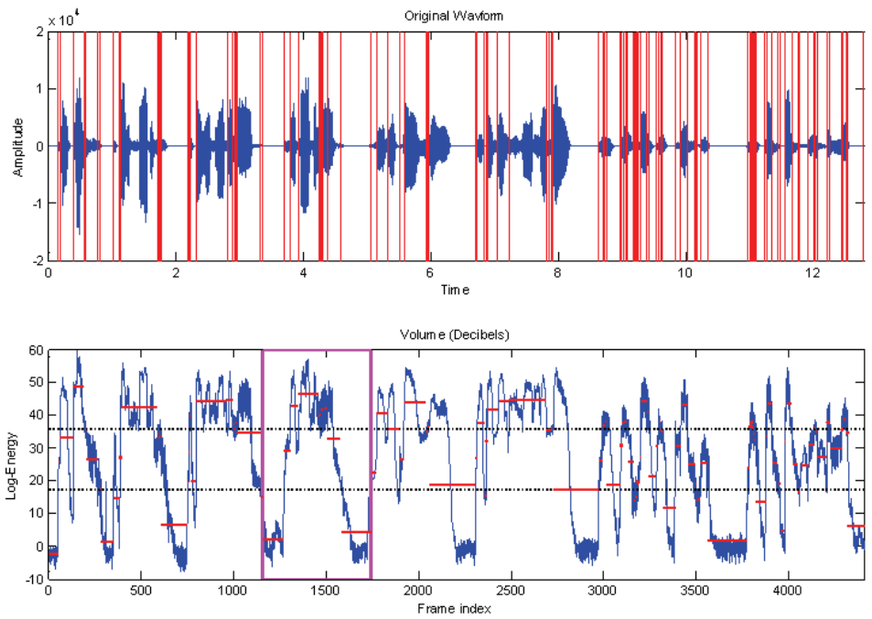

Figure 5. The intersection points between energy contour and two thresholds

Figure 9 shows the segmentation result. The score is shown in Fig. 10. When the mean energy of the segmented partition is greater than T_L, it is regarded as non-silence. Therefore its score is calculated. The short horizontal line is the mean energy for each partition, and the long horizontal line is the threshold T_L. Ten partitions are obtained after the merging

operation. P1 and P2 whose emotional states are anger, P3 and P4 are happiness, P5 and P6 are sadness, and P7 and P8 are neutral. Comparing with the result shown in Fig. 10, the recognition of P6 is wrong.

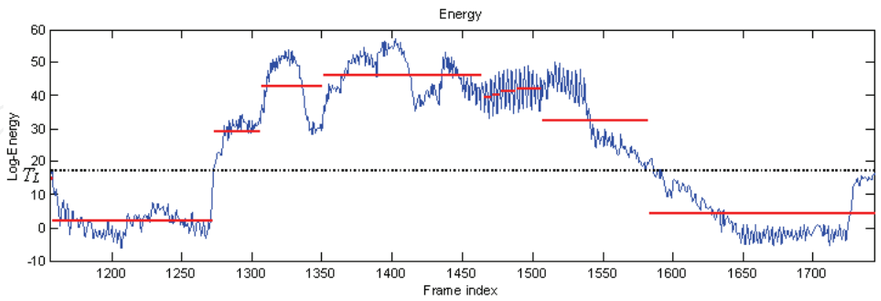

Figure 6. The enlarged plot from the marked rectangle in Fig. 5

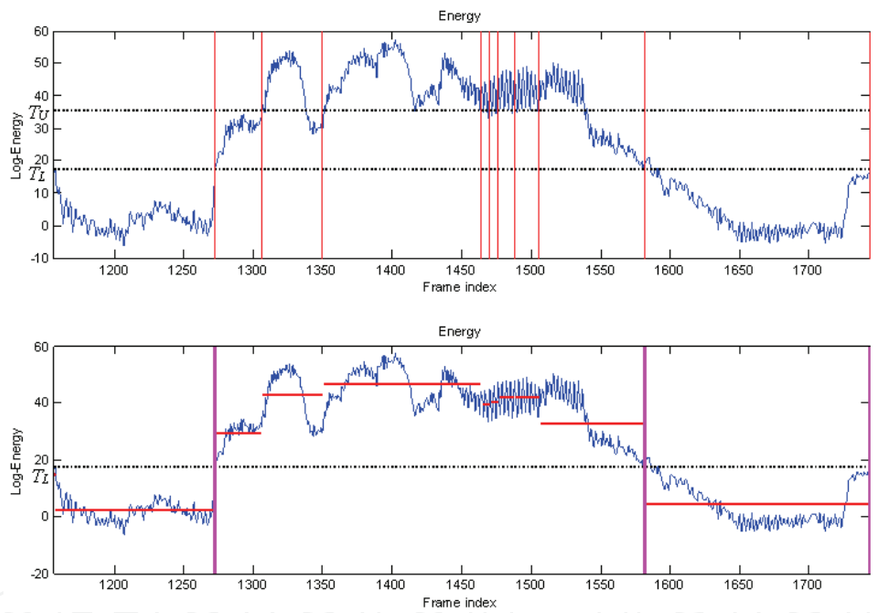

Figure 7. The merged result

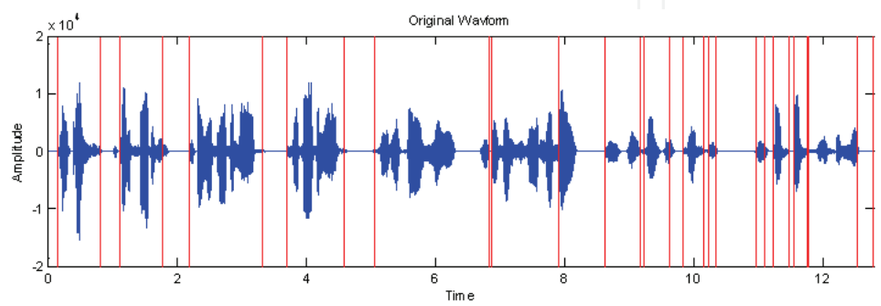

Figure 8. Result of segmentation with silence

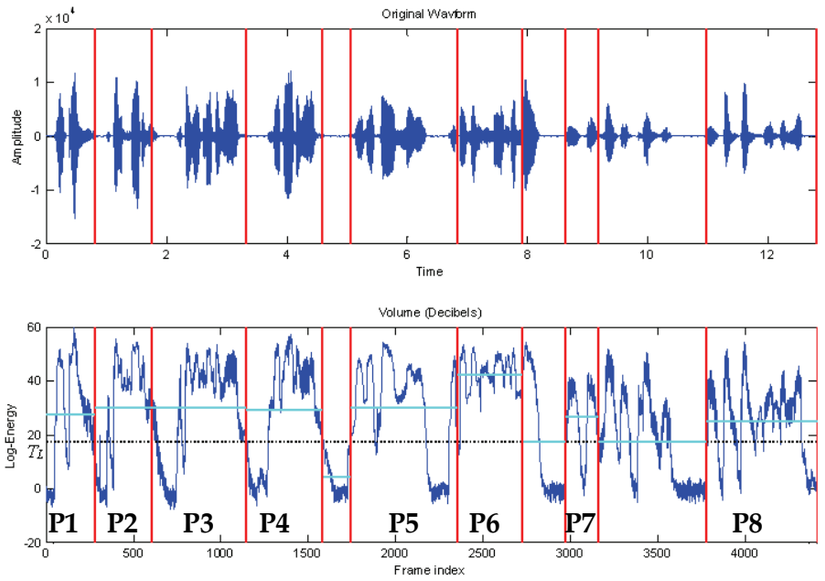

Figure 9. The final segmentation result with segmentation with silence

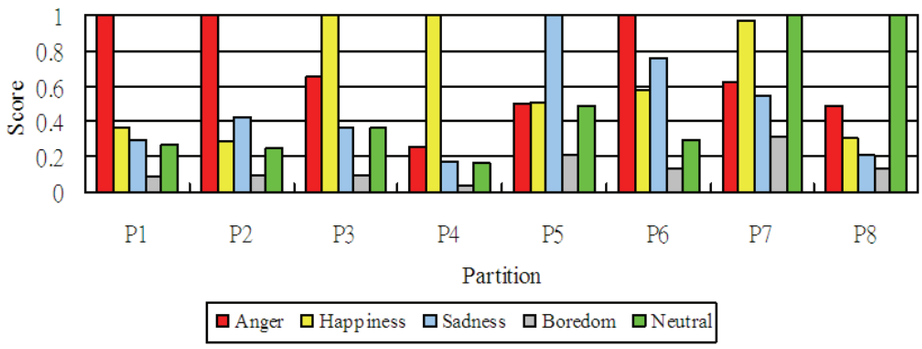

Figure 10. The score bar chart of each partition shown in Fig. 9

5.2 Experimental Results Using Segmentation with Endpoint Detection

From Equations (25)-(27), setting $\alpha_1 = -0.5$, $\alpha_2 = 0.5$, and $\alpha_3 = 0.5$, three thresholds are obtained. Figure 11 shows an example of segmentation with endpoint detection. The thick line is the beginning of the partition and the thin line is the end of the partition. In Fig. 11, we see that there are too many partitions. Thus, we need to merge partitions with similar characteristics together.

In Figure 12, the distances of the beginning and end in two adjacent endpoints are expressed as

$$D_i = B_{i+1} - E_i,\ i = 1,2,...,N \tag{29}$$

where N is the total number of the intervals, B_{i+1} and E_i correspond to the beginning point of partition i+1 and the ending point of partition, respectively. Another threshold is used to merge the partitions. The threshold is calculated as

$$T_D = \overline{D} + \alpha_D D_{STD} \tag{30}$$

where $\overline{D}$ is the average distance between two non-silence partition, and D_{STD} is the standard deviation of the distances. α_D is obtained from experiments and is set to 0.3. If D_i is less than T_D then the two intervals are combined, as shown in Fig. 13.

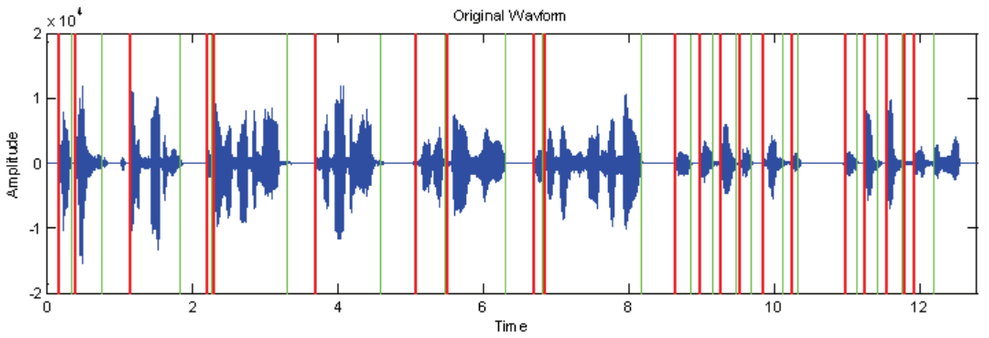

Figure 11. The initial result of endpoint detection

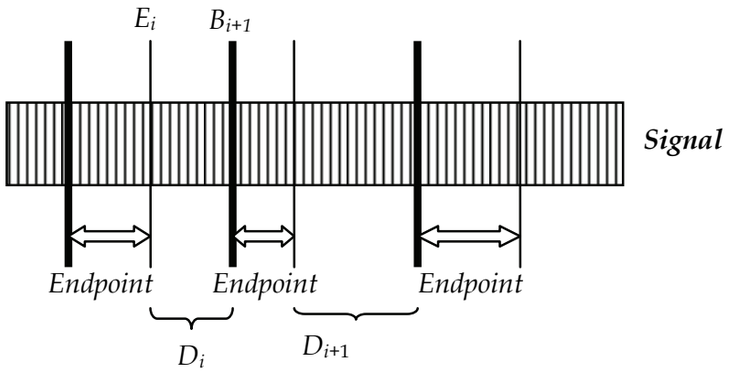

Figure 12. Distance between two adjacent partitions

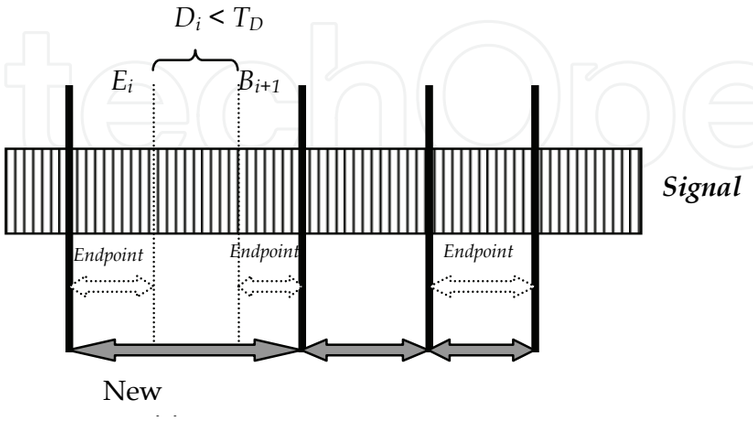

Figure 13. Two non-silence partitions are merged when the distance between them is less then T_D

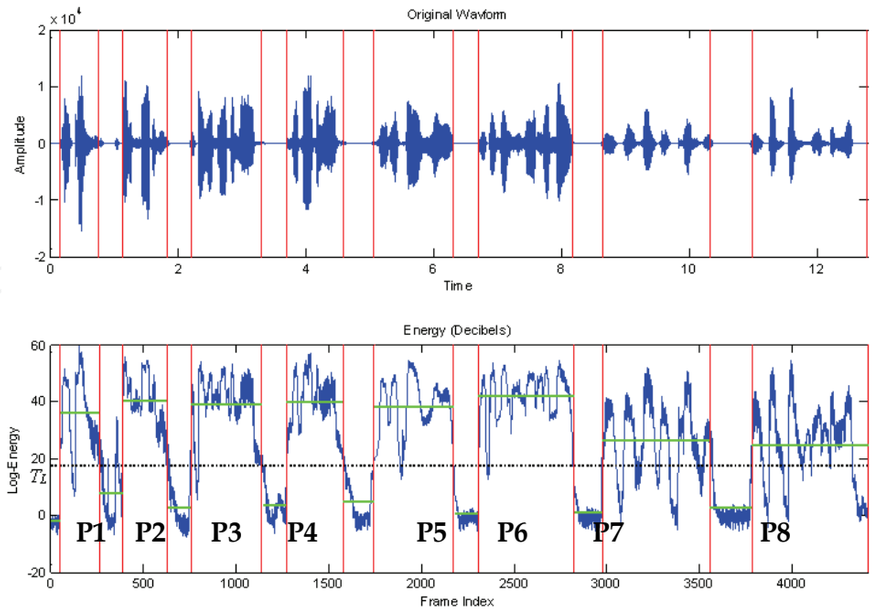

Figure 14. Partition result after merging using the threshold T_D

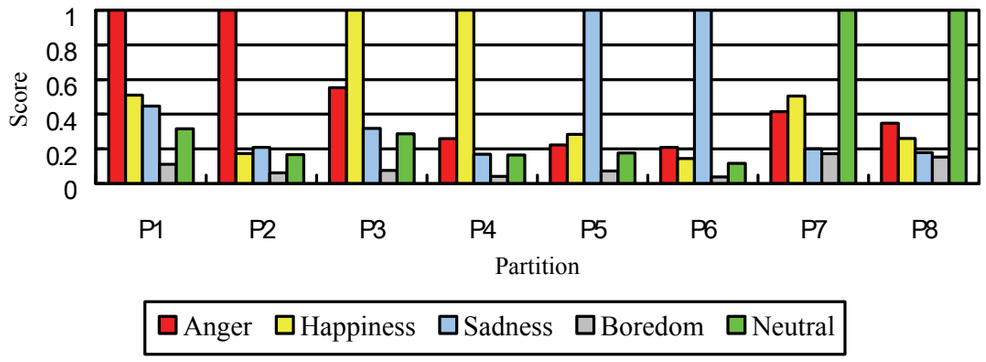

Figure 15. The score bar chart of each partition shown in Fig. 14

The result of the application of threshold T_D is shown in Fig. 14. In the bottom figure, the short horizontal line is the mean energy of each partition, and the long horizontal line is the threshold T_L. When the mean of the segmented partition is greater than T_L, it is regarded as non-silence. Eight partitions are obtained, whose original emotions are 2 anger short sentences, 2 happiness short sentences, 2 sadness short sentences, and 2 neutral short sentences, respectively. The score for each partition is shown in the Fig. 15. The classified result matches with the original emotion in the corpus. Since segmentation with endpoint will obtain better results than the other two segmentation methods, we adopt it as the segmentation method in this study.

5.3 Recognition Accuracy of Continuous Speech

Table 4 shows the recognition accuracy for all the sentences listed in Table 3. The recognition accuracy is calculated from dividing the total correct recognition by the total

number of short corpora in the sentences in each category. Finally, the overall recognition accuracy is 83%.

	AH	AS	AB	AN	HS	HB	HN	SB	SN	BN	Average
Accuracy	0.91	0.89	0.72	0.84	0.89	0.80	0.80	0.86	0.78	0.81	**0.83**

Table 4. The recognition accuracy of Corpus II

6. Conclusions and Future Works

In recent years, emotion recognition is used in more and more applications. In this study, the emotion recognition from continuous speech is realized. Emotion recognition used in real world can be expected soon. The application in the call-centre can help the customer service personnel to better serve the customer. Endpoint detection is used to segment the continuous speech. The feature sets are 12 LPCCs and 20 MFCCs. The classifier is D-KNN with Fibonacci series weighting. The recognition accuracy of 83% is obtained.

It is not easy to obtain the emotional corpus, especially continuous emotional corpus. In the future, it is necessary to get more emotional speech corpora, and hope to collect them in various forms, such as recording in a noisy environment or not so perfect sound quality. In real life, speeches do not always be recorded in a quiet environment or with high quality devices. The emotion recognition from continuous speech can be used in business, such as call centre. If the real-time system is going to be realized, the performance of the program must be improved. In other words, the programming language need to be changed to that can be used not only in personal computers but also in other devices, such as personal digital assistants (PDAs) or mobile phones.

7. References

Busso C. & Narayanan S.S. (2007). Between Speech and Facial Gestures in Emotional Utterances: A Single Subject Study, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 8, pp. 2331-2347, ISSN: 1558-7916.

Chang, B.H. (2002). *Automated Recognition of Emotion in Mandarin*, Master thesis, National Cheng Kung University.

Cowie, R. E.; Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. & Taylor, J. (2001). Emotion Recognition in Human-Computer Interaction, *IEEE Signal Processing Magazine*, Vol. 18, No. 1, pp. 32-80, ISSN: 1053-5888.

Chuang, Z.J. & Wu, C.H. (2004). Multi-Modal Emotion Recognition from Speech and Text, *International Journal of Computational Linguistics and Chinese Language Processing*, pp. 1-18, Vol. 9, No. 2, ISSN: 0349-1021.

Ekman, P. (1999), *Handbook of Cognition and Emotion*, John Wiley & Sons, ISBN-10: 0471978361, New York, USA.

Inanoglu, Z. & Caneel, R. (2005). Emotive Alert: HMM-Based Emotion Detection in Voicemail Messages, *Proceedings of Intelligent User Interfaces*, pp. 251-253, January 2005, San Diego, USA.

Kleinginna Jr., P.R. & Kleinginna, A.M. (2005). A Categorized List of Emotion Definitions with Suggestions for a Consensual Definition, *Motivation and Emotion*, Vol. 5, No. 4, pp. 345-379, ISSN: 0146-7239.

Kwon, O.W.; Chan, K., Hao, J. & Lee, T.W. (2003). Emotion Recognition by Speech Signals, *Proceedings of Eurospeech*, pp.125-128, September 2003, Geneva, Switzerland.

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

Lu, L.; Liu, D.H. & Zhang, J. (2006). Automatic Mood Detection and Tracking of Music Audio Signals, IEEE Transactions on Audio, Speech and Language Processing, pp. 5-18, Vol. 14, ISSN: 1558-7916.

Mehrabian, A. & Russel, J. (1974). *An Approach to Environmental Psychology*, the MIT Press, ISBN-10: 0-262-63071-0, Cambridge, USA.

Murray, I. &. Arnott, J.L. (1993). Towards the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion. *Journal of the Acoustic Society of America*, Vol. 93, No. 2, pp. 1097-1108, ISSN: 0001-4966.

Nwe, T.L.; Foo, S.W. & De-Silva, L.C. (2003). Speech Emotion Recognition Using Hidden Markov Models, *Speech Communication*, Vol. 41, No. 4, pp. 603-623, ISSN: 0167-6393.

Osgood, C.E.; Suci, J.G. & Tannenbaum, P.H. (1967). *The Measurement of Meaning*, the University of Illinois Press, ISBN-10: 978-0252745393, Urbana, USA.

Pao, T.L.; Chen, Y.T. & Yeh, J.H. (2008). Emotion Recognition and Evaluation from Mandarin Speech Signals, International Journal of Innovative Computing, Information and Control (IJICIC), pp. 0-07-107, Vol. 4, No. 7, ISSN: 1349-4198.

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

Park, C.H.; Heo, K.S., Lee, D.W., Joo, Y.H. & Sim, K.B. (2002). Emotion Recognition based on Frequency Analysis of Speech Signal, *International Journal of Fuzzy Logic and Intelligent Systems*, Vol. 2, No.2, pp. 122-126, ISSN: 1064-1246.

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

Picard, R.W. (1997). *Affective Computing*, the MIT Press, ISBN-10: 0-262-16170-2, Cambridge, USA.

Ramamohan, S. & Dandapa, S. (2006). Sinusoidal Model-Based Analysis and Classification of Stressed Speech, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, Issue. 3, pp. 737-746, ISSN: 1558-7916.

Sebe N.; Cohen, I., Gevers, T. & Huang, T.S. (2005). Multimodal Approaches for Emotion Recognition: A Survey, *Proceedings of the International Society for Optical Engineering (SPIE)*, pp. 56-67, Vol. 5670, February 2005. San Jose, CA.

Schröder, M. (2006). Expressing Degree of Activation in Synthetic Speech, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, Issue. 4, pp. 1128-1136, ISSN: 1558-7916.

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

Tato, R.S.; Kompe, R. & Pardo, J.M. (2002). Emotional Space Improves Emotion Recognition, *Proceedings of International Conference on Spoken Language Processing*, pp. 2029-2032, September 2002, Colorado, USA.

Tao J.; Kang, Y. & Li, A. (2006). Prosody Conversion From Neutral Speech to Emotional Speech, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, Issue. 4, pp. 737-746, ISSN: 1558-7916.

Ververidis, D.; Kotropoulos, C. & Pitas, I. (2004). Automatic Emotional Speech Classification, *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, pp. 593-596, May 2004, Montreal, Canada.

Yacoub, S.; Simske, S., Lin, X. & Burns, J. (2003). Recognition of Emotions in Interactive Voice Response Systems, *Proceedings of Eurospeech,* pp. 729-732, September 2003, Geneva, Switzerland.

Advances in Human Computer Interaction

Edited by Shane Pinder

ISBN 978-953-7619-15-2
Hard cover, 600 pages
Publisher InTech
Published online 01, October, 2008
Published in print edition October, 2008

In these 34 chapters, we survey the broad disciplines that loosely inhabit the study and practice of human-computer interaction. Our authors are passionate advocates of innovative applications, novel approaches, and modern advances in this exciting and developing field. It is our wish that the reader consider not only what our authors have written and the experimentation they have described, but also the examples they have set.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Tsang-Long Pao, Jun-Heng Yeh and Yu-Te Chen (2008). Emotion Recognition via Continuous Mandarin Speech, Advances in Human Computer Interaction, Shane Pinder (Ed.), ISBN: 978-953-7619-15-2, InTech, Available from: http://www.intechopen.com/books/advances_in_human_computer_interaction/emotion_recognition_via_continuous_mandarin_speech

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen