Computational Linguistics and Chinese Language Processing
Vol. 12, No. 1, March 2007, pp. 79-90

Emotional Recognition Using a Compensation Transformation in Speech Signal

Cairong Zou* , Yan Zhao+, Li Zhao +, Wenming Zhen+, and

Yongqiang Bao+

Abstract

An effective method based on GMM is proposed in this paper for speech emotional recognition; a compensation transformation is introduced in the recognition stage to reduce the influence of variations in speech characteristics and noise. The extraction of emotional features includes the globe feature, time series structure feature, LPCC, MFCC and PLP. Five human emotions (happiness, angry, surprise, sadness and neutral) are investigated. The result shows that it can increase the recognition ratio more than normal GMM; the method in this paper is effective and robust.

Key words: Speech Emotional Recognition (SER), GMM, Emotion Recognition, Compensation Transformation

1. Introduction

One of the natural goals for research on speech signals is recognizing emotions of humans [Chen 1987; Oppenheim 1976; Cowie 2001]; it has gained growing amounts of interest over the last 20 years. A study conducted by Shirasawa *et al*. showed that SER could be made by ICA and attain an 87% average recognition ratio [Shirasawa 1997; Shirasawa 1999] Many studies have been conducted to investigate neural networks for SER. Chang-Hyun Park tried to recognize sequentially inputted data using DRNN in 2003[Park *et al*. 2003], Muhammad, W. B. obtained about 79% recognition rate using GRNN [Bhatti *et al*. 2004]. Aishah Abdul Razak achieved an average recognition rate of 62.35% using combination MLP [Razak *et al*. 2005]. Fuzzy rules are also introduced into SER such that an 84% rate has been achieved in recognizing anger and sadness [Austermann *et al*. 2005]. A number of studies in SER have

* Foshan University, Foshan, 528000, Guangdong, China

+ Research Center of Learning Science, Southeast University, Nanjing, 210096, China
E-mail: zhaoli@seu.edu.cn

[Received August 11, 2006; Revised September 30, 2006; Accepted October 22, 2006]

also been done with the development of GMM/HMM [Rabiner 1989; Jiang *et al*. 2004; Lin *et al.* 2005]. However, in SER, the variations in speech characteristics, noise and individual differences always influence the recognition results. In addition, the methods above have always handled such problems in the preprocessing stage and have not been able to eliminate the influence effectively. Therefore, a valid solution has still not been proposed. In this paper a compensation transformation is introduced into an algorithm for GMM which operates in the recognition module. The experiments with five emotions (happiness, angry, neutral, surprise and sadness) show that the method in this paper is effective in emotional recognition.

2. Descriptions of Emotion and Selection of Emotion Speech Materials

Usually, emotions are classified into two main categories: basic emotions and derived emotions. Basic emotions, generally, can be found in all mammals. Derived emotions mean derivations from basic emotions. One viewpoint is that the basic emotions are composed by the basic mood. Due to different research backgrounds, different researchers have expressed different definitions of basic emotions. Some of the major definitions [Ortony *et al.* 1990] of the basic emotions are shown in Table 1.

Table 1. Researches about basic emotions definition

Researchers	definitions
Plutchik	Acceptance, joy, anger, anticipation, disgust, fear, sadness, surprise
Ekman/Friesen/ Ellsworth	Anger, disgust, fear, joy, sadness, surprise
James	Fear, grief, love, rage
Izard	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise
Oatley/Johnson -Laird	Anger, disgust, anxiety, happiness, sadness
Panksepp	Expectancy, fear, rage, panic
Weiner/Graham	Happiness, sadness

The common emotion classification which was proposed by Plutchik is shown in Figure 1e. In this paper, the authors only recognize five kinds of emotion.

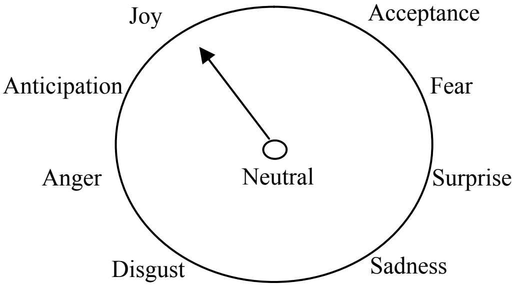

Figure 1. Emotion wheel

This is a relatively conservative view of what emotion is so special attention has been paid to emotional dimension space theory. Three major dimensions (valence, arousal, and control) [Cowie 2001] are used to describe emotions.

a. Valence: The clearest common element of emotional states is that the person is materially influenced by feelings that are valenced, *i.e.*, they are centrally concerned with positive or negative evaluations of people or things or events.

b. Arousal: It has been proven that emotional states involve dispositions to act in certain ways. A basic way of reflecting that theme turns out to be surprisingly useful. States are simply rated in terms of the associated activation level, *i.e.*, the strength of the person's disposition to take some action rather than none.

c. Control: Embodying in the initiative and the degree of control. For instance, contempt and fear are in different ends of the control dimension.

In this paper, two aspects have to be taken into consideration in the selection of emotional materials: 1. the sentence materials can't have any emotional tendency; 2. the materials should relate to five kinds of emotions (happiness, angry, surprise, sadness, and neutral). All recordings were carried out in a large, soundproof room with no echo interference using a high quality microphone, a SONY DAT recorder and a PC164 audio card at a sampling rate of l2KHZ with 16-bit resolution. Six speakers (three male and three female) who are good at acting spoke the sentences with happiness, anger, surprise and sadness, expressing each emotion three times. At the same time, the researchers made the speakers speak each sentence three times in a neutral way. In this way, 2430 sentences for experiments were compiled.

3. Feature Extraction

The emotional features of speech signals are always represented as the change of speech rhythm [Shigenaga 1999; Muraka 1998]. For example, when a man is in a rage, his speech rate, volume and tone will all get higher. Some characteristics of phonemes can also reflect the change of emotions such as formant and the cross section of the vocal tract [Muraka 1998; Zhao *et al*. 2001]. As the emotional information of speech signals is more or less related to the meaning of the sentences, the distributing rules and construction characteristics should be attained by analyzing the relationship between emotional speech and neutral speech to avoid the effect caused by the meaning of the sentences.

The global features used in this paper are duration, mean pitch, maximum pitch, average different rate of pitch, average amplitude power, amplitude power dynamic range, average frequency of formant, average different rate of formant, mean slope of the regression line of the peak value of the formant and the average peak value of formant [Zhao *et al*. 2001; Zhao *et al*. 2000; Zhao *et al*. 2000]. The duration is the continuous time from start to end in each emotional sentence. It includes the silence, because these parts contribute to the emotion. Duration ratio of emotional speech and neutral speech was used as the characteristic parameters for recognition. The frequency of pitch was obtained by calculating cepstrum. Then the pitch-track was gained, and maximum pitch ($F0_{\max}$), average fundamental frequency ($F0$), average different rate of pitch ($F0_{rate}$) of the envelopes of different emotional speech signals can all be extracted from it. $F0_{rate}$ mentioned here, refers to the mean absolute value of the difference between each frame of speech signal 's fundamental frequencies. The authors used the differences in value of the mean pitch, the maximum pitch and the ratio of $F0_{rate}$ between the emotional and neutral speech as the characteristic parameters. In this paper, the average amplitude power (A) and the dynamic range (A_{range}) are to be taken into account. To avoid the influence of the silent and noisy parts of the speech, the authors only took the mean absolute value of the amplitude into account and all the absolute values must above a threshold. The difference of average amplitude power and the dynamic range between the emotional and neutral speech was used for parameters of recognition. Formant is an important parameter that reflects the characteristics of vocal track. Formant was attained as follows [Zhao *et al*. 2001]. At first, LPC method was applied to calculate 14-order coefficients of linear prediction. Then, the coefficients were used to estimate the track's frequency of the formant by analyzing the frequency average ($F1$), frequency-changing rate ($F1_{rate}$) of the first formant, the average and the average slope of recursive lines of the first four formants. The authors use the difference of $F1$, the last two parameters and the ratio of $F1_{rate}$ between the emotional and neutral speech as the characters in each frame.

The structural features of time series for the emotional sentences used in this paper is maximum value of the pitch in each vowel segment, amplitude power of the corresponding frame, maximum value of the amplitude energy in each vowel segment, pitch of the corresponding frame, duration of each vowel segment and mean value and rate of change of the first three formants. For these parameters, the ratio between the emotional and neutral speech was used as the recognition characters.

In addition to the above features, LPCC, PLP, MFCC are also taken into consideration for precise decision. Figure 2 is the module for feature extraction.

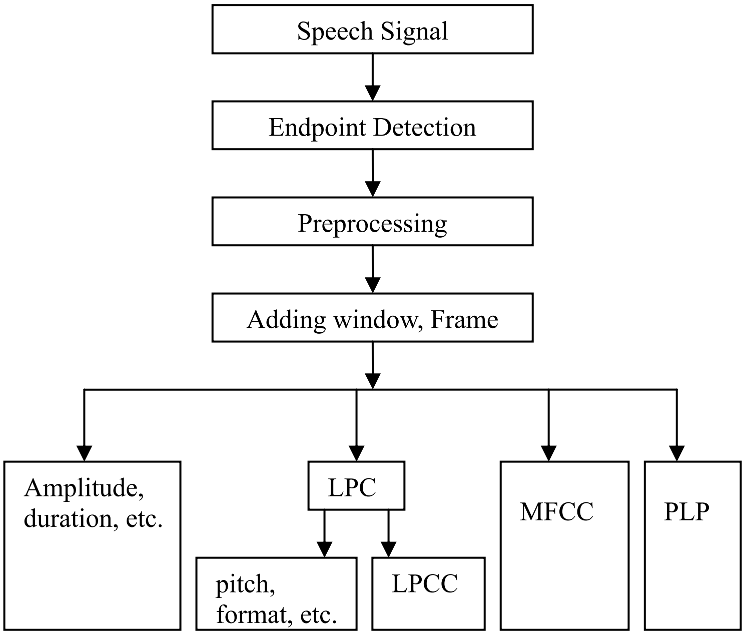

Figure. 2 the module for feature extraction

4. Speech Emotion Recognition based on GMM

GMM can be described as follow:

$$\lambda_i = \{a_i, \mu_i, \Sigma_i\}\ , \tag{1}$$

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} a_i b_i(\vec{x})\ , \sum_{i=1}^{M} a_i = 1\quad , \tag{2}$$

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} \mid \Sigma_i \mid^{1/2}} \cdot \exp\{-\frac{1}{2}(\vec{x} - \mu_i)^t \Sigma_i^{-1} (\vec{x} - \mu_i)\}\ , \tag{3}$$

where $\vec{x}$ is the D-dimensional feature vector, $b_i(\vec{x})$ $(i = 1, 2, ...M)$ is the density function of the member $\vec{x}$, $p(\vec{x} \mid \lambda)$ is the probability density function of $\vec{x}$, and a_i satisfies:

$$\sum_{i=1}^{M} a_i = 1 \quad (i = 1, 2, ...M)\ .$$

The GMM probability function of a speech signal with T frames $X = (\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_T)$ can be denoted as:

$$P(X \mid \lambda) = \prod_{t=1}^{T} p(\vec{x}_t \mid \lambda)\ , \tag{4}$$

or

$$S(X \mid \lambda) = \log P(X \mid \lambda) = \sum_{t=1}^{T} \log p(\vec{x}_t \mid \lambda) . \tag{5}$$

According to the statistical characteristic of likelihood probability (LP) output by Gaussian Mixture Model, the likelihood probability with the best model is generally bigger than that of the other GMM, but due to the existence of variations in speech characteristics and noise, some frames' LP shows a best model that is smaller than that of the others, so the decision may be incorrect. In order to reduce this error recognition rate, some transformation should be introduced to compensate for the likelihood probability, that is, raise the probability with the best model and reduce the probability with the other models. Therefore, a nonlinear compensation transformation is proposed in this paper to solve this problem.

5. Compensation Transformation for GMM

The transformation must satisfy three conditions as follow:

1. The difference of the output probability in different time should be reduced, *i.e.* increase ΔS_1 ;

$$\Delta S_1 = \sum_{\substack{t,k=1 \\ t \neq k}}^{T} \left| \log p(\vec{x}_{tt1} \mid \lambda) - \log p(\vec{x}_k \mid \lambda) \right|$$

2. The difference of the output probability in the same time with different emotion should be increased, *i.e.* increase ΔS_2 ;

$$\Delta S_2 = \sum_{\substack{i,j=1 \\ i \neq j}}^{M} \left| \log p(\vec{x}_t \mid \lambda_i) - \log p(\vec{x}_t \mid \lambda_j) \right|$$

3. The relative value of the output probability should not be changed.

Assuming that $\vec{x}$ is a feature vector, λ_0 is the best model corresponded to $\vec{x}$, and λ_1 is the other model that is mismatched. If the transformation is linear:

$$f[p(\vec{x}_t \mid \lambda_i)] = ap(\vec{x}_t \mid \lambda_i) + b$$

$$f[p(\vec{x} \mid \lambda_0)] - f[p(\vec{x} \mid \lambda_1)] = a[p(\vec{x}_t \mid \lambda_0) - p(\vec{x}_t \mid \lambda_1)] , \tag{6}$$

where $a, b = const$. Here set $a > 0$:

$$p(\vec{x} \mid \lambda_0) \geq p(\vec{x} \mid \lambda_1) \Leftrightarrow f[p(\vec{x} \mid \lambda_0)] \geq f[p(\vec{x} \mid \lambda_1)] , \tag{7}$$

$$p(\vec{x} \mid \lambda_0) \leq p(\vec{x} \mid \lambda_1) \Leftrightarrow f[p(\vec{x} \mid \lambda_0)] \leq f[p(\vec{x} \mid \lambda_1)] . \tag{8}$$

From (7) ~ (8), it is obvious that the linear transformation cannot increase or reduce the LP of the output. The compensation could not be linear transformation, so a nonlinear compensation transformation is proposed; the detailed steps are described as follow:

1. Compute the probability of the t-th feature vector, where *N* is the number of the emotions, and T is the number of the frames.

$p(\vec{x}_t \mid \lambda_i)$ $(i = 1, 2, ...N)$, $(t = 1, 2, ...T)$

2. Normalize $p(\vec{x}_t \mid \lambda_i)$.

$$P(\vec{x}_t \mid \lambda_i) = \frac{p(\vec{x}_t \mid \lambda_i)}{\max p(\vec{x}_t \mid \lambda_i)} \tag{9}$$

3. Compute the output LP.

$$S(\vec{x}_t, \lambda_i) = \frac{[P(\vec{x}_t \mid \lambda_i)]^n}{[P(\vec{x}_t \mid \lambda_i)]^n + b}, \tag{10}$$

where $n = 2 \sim 5$, $b > 1$ and b is always set close to 1.

4. Introduce the compensation: compute the average probability with *K* former frames.

$$\overline{S}(\vec{x}_{t-K+1,} \vec{x}_{t-K+2,}, ..., \vec{x}_{t,} \lambda_i) = \frac{1}{K} \sum_{k=1}^{K} S(\vec{x}_{t+k,} \mid \lambda_i) \tag{11}$$

In general, K also has an influence on output probability, here set $K = 2 \sim 5$.

5. Take $\overline{S}(\vec{x}_{t-K+1,} \vec{x}_{t-K+2,}, ..., \vec{x}_{t,} \lambda_i)$ as the compensation for $S(\vec{x}_t \mid \lambda_i)$.

$$S'(\vec{x}_t \mid \lambda_i) = S(\vec{x}_t \mid \lambda_i) + a_{ti}\delta_{ti}[\overline{S}(\vec{x}_{t-K+1}, \vec{x}_{t-K+2}, ..., \vec{x}_t, \lambda_i) - S(\vec{x}_t \mid \lambda_i)], \tag{12}$$

where $a_{ti} \in [0,1)$, $\delta_{ti} = \begin{cases} 1 & \overline{S}(\vec{x}_{t-K+1}, \vec{x}_{t-K+2}, ..., \vec{x}_t, \lambda_i) > S(\vec{x}_t \mid \lambda_i) \\ -1 & otherwise \end{cases}$.

6. Calculate the joint probability for each model.

$$S(X, \lambda_i) = \sum_{t=1}^{T} \log S'(\vec{x}_t \mid \lambda_i) \tag{13}$$

7. Make the decision of which emotion X belongs to. If $S(X, \lambda_j) = \max_i S(X, \lambda_i)$, then *X* belongs to λ_j .

Assuming two emotions: λ_0 , λ_1 and two vectors: $\vec{x}_1, \vec{x}_2$ .Set $T = 2$. The output probability without transformation:

$$S(\vec{x} , \lambda_0) = \ln p(\vec{x}_1 \mid \lambda_0) + \ln[p(\vec{x}_2 \mid \lambda_0)], \tag{14}$$

$$S(\vec{x}, \lambda_1) = \ln p(\vec{x}_1 \mid \lambda_1) + \ln[p(\vec{x}_2 \mid \lambda_1)]. \tag{15}$$

$$\begin{aligned} \ln P(\vec{x}_1 \mid \lambda_0) + \ln P(\vec{x}_2 \mid \lambda_0) > & \ \ln P(\vec{x}_1 \mid \lambda_1) + \ln P(\vec{x}_2 \mid \lambda_1) \\ & \Rightarrow P(\vec{x}_1 \mid \lambda_0) P(\vec{x}_2 \mid \lambda_0) - P(\vec{x}_1 \mid \lambda_1) P(\vec{x}_2 \mid \lambda_1) > 0 \\ & \Rightarrow P(\vec{x}_1 \mid \lambda_0)^n P(\vec{x}_2 \mid \lambda_0)^n - P(\vec{x}_1 \mid \lambda_1)^n P(\vec{x}_2 \mid \lambda_1)^n > 0, \end{aligned} \tag{16}$$

When $S(\bar{x},\lambda_0) > S(\bar{x},\lambda_1)$, $\vec{x}_i$ (i=1, 2) belongs to λ_0 , otherwise belongs to λ_1 . The output probability with transformation:

$$S(x_1 \mid \lambda_0) = \log(\frac{P'(\bar{x}_1 \mid \lambda_0)^n}{P'(\bar{x}_1 \mid \lambda_0)^n + b} + \delta_{1,0}\alpha_{1,0}[\frac{P'(\bar{x}_1 \mid \lambda_0)^n}{P'(\bar{x}_1 \mid \lambda_0)^n + b} - \bar{S}(0,\lambda_0)]) , \tag{17}$$

$$S(x_2 \mid \lambda_0) = \log(\frac{P'(\bar{x}_2 \mid \lambda_0)^n}{P'(\bar{x}_2 \mid \lambda_0)^n + b} + \delta_{2,0}\alpha_{2,0}[\frac{P'(\bar{x}_2 \mid \lambda_0)^n}{P'(\bar{x}_2 \mid \lambda_0)^n + b} - \bar{S}(1,\lambda_0)]) . \tag{18}$$

$S(\bar{x}_1,\lambda_1)$ and $S(\bar{x}_2,\lambda_1)$ are similar to (17)~(18). The decision rule is the same as the one without transformation.

$$\begin{aligned} S(X \mid \lambda_0) - S(X \mid \lambda_1) &= \log(\frac{P'(\bar{x}_1 \mid \lambda_0)^n}{P'(\bar{x}_1 \mid \lambda_0)^n + b} + \delta_{1,0}\alpha_{1,0}[\frac{P'(\bar{x}_1 \mid \lambda_0)^n}{P'(\bar{x}_1 \mid \lambda_0)^n + b} - \bar{S}(0,\lambda_0)]) \\ &+ \log(\frac{P'(\bar{x}_2 \mid \lambda_0)^n}{P'(\bar{x}_2 \mid \lambda_0)^n + b} + \delta_{2,0}\alpha_{2,0}[\frac{P'(\bar{x}_2 \mid \lambda_0)^n}{P'(\bar{x}_2 \mid \lambda_0)^n + b} - \bar{S}(1,\lambda_0)]) \\ &- \log(\frac{P'(\bar{x}_1 \mid \lambda_1)^n}{P'(\bar{x}_1 \mid \lambda_1)^n + b} + \delta_{1,1}\alpha_{1,1}[\frac{P'(\bar{x}_1 \mid \lambda_1)^n}{P'(\bar{x}_1 \mid \lambda_1)^n + b} - \bar{S}(0,\lambda_1)]) \\ &- \log(\frac{P'(\bar{x}_2 \mid \lambda_1)^n}{P'(\bar{x}_2 \mid \lambda_1)^n + b} + \delta_{2,1}\alpha_{2,1}[\frac{P'(\bar{x}_2 \mid \lambda_1)^n}{P'(\bar{x}_2 \mid \lambda_1)^n + b} - \bar{S}(1,\lambda_1)]) , \end{aligned} \tag{19}$$

Set $a_{10} = a_{20} = a_{11} = a_{22} = const = a$, $p_{ti} = p\left(\vec{x}_t \middle| \lambda_i\right)$, $\bar{S}_{ti} = \frac{[P(\vec{x}_{t+1} \mid \lambda_i)]^n}{[P(\vec{x}_{t+1} \mid \lambda_i)]^n + b} - \bar{S}\left(t,\lambda_i\right)$.

1. $p_{10} = p_{20} = 1$, (16) and (19) can be changed into (20) ~ (21):

$$P_{11}P_{21} < 1 \tag{20}$$

$$\begin{aligned} S(X \mid \lambda_0) - S(X \mid \lambda_1) = &\log[\frac{1}{(1+b)^2} + \frac{a\delta_{1,0}\bar{S}_{00} + a\delta_{2,0}\bar{S}_{10}}{1+b} + a^2\delta_{1,0}\delta_{2,0}\bar{S}_{00}\bar{S}_{10}] \\ &- \log\left[\frac{p_{11}p_{21}}{\left(p_{11}+b\right)\left(p_{21}+b\right)} + \frac{a\delta_{21}\bar{S}_{11}}{\left(p_{11}+b\right)} - \frac{a\delta_{11}\bar{S}_{11}}{\left(p_{21}+b\right)} \quad a^2\delta_{11}\delta_{21}\bar{S}_{11}\bar{S}_{21}\right] > 0 \end{aligned} \tag{21}$$

$$\begin{aligned} &\frac{1}{\left(1+b\right)^2} - \frac{p_{11}p_{21}}{\left(p_{11}+b\right)\left(p_{21}+b\right)} + a\left(\frac{\delta_{10}\bar{S}_{00} + a\delta_{20}\bar{S}_{10}}{1+b} - \frac{\delta_{21}\bar{S}_{11}}{\left(p_{11}+b\right)} - \frac{\delta_{11}\bar{S}_{11}}{\left(p_{21}+b\right)}\right) \\ &\qquad + a^2(\delta_{10}\delta_{20}\bar{S}_{10}\bar{S}_{00} - \delta_{11}\delta_{21}\bar{S}_{11}\bar{S}_{21}) > 0 \end{aligned} \tag{22}$$

s.t.

$$\frac{1}{\left(1+b\right)^2} + \frac{a\delta_{10}\bar{S}_{00} + a\delta_{20}\bar{S}_{10}}{1+b} + a^2\delta_{10}\delta_{20}\bar{S}_{10}\bar{S}_{00} > 0$$

$$\frac{p_{11}p_{21}}{\left(p_{11}+b\right)\left(p_{21}+b\right)}+\frac{a\delta_{21}\bar{S}_{11}}{\left(p_{11}+b\right)}-\frac{a\delta_{11}\bar{S}_{11}}{\left(p_{21}+b\right)}+a^{2}\delta_{11}\delta_{21}\bar{S}_{11}\bar{S}_{21}>0$$

where a is small enough to ignore the influence of the second and the third item in (22).

$$\frac{1}{(1+b)^{2}}-\frac{p_{11}p_{21}}{(p_{11}+b)(p_{21}+b)}=\frac{b^{2}(1-p_{11}p_{21})+bp_{11}(1-p_{21})+bp_{21}(1-p_{11})}{(1+b)^{2}(p_{11}+b)(p_{21}+b)}>0 \qquad (23)$$

Compared to (20), it can be seen that the LP with transformation is increased.

2. $p_{10}=p_{21}=1$, (16) and (19) can be changed into

$$p_{20}-p_{11}>0 \qquad (24)$$

$$S(X\,|\,\lambda_{0})-S(X\,|\,\lambda_{1})=\log\left[\frac{p_{20}}{\left(1+b\right)\left(p_{20}+b\right)}+\frac{a\delta_{20}\bar{S}_{10}}{1+b}+\frac{a\delta_{10}\bar{S}_{00}p_{20}}{p_{20}+b}+a^{2}\delta_{10}\delta_{20}\bar{S}_{10}\bar{S}_{00}\right]$$

$$-\log\left[\frac{p_{11}}{\left(1+b\right)\left(p_{11}+b\right)}+\frac{a\delta_{11}\bar{S}_{01}}{1+b}+\frac{a\delta_{21}\bar{S}_{11}p_{11}}{p_{11}+b}+a^{2}\delta_{11}\delta_{21}\bar{S}_{11}\bar{S}_{21}\right]>0 \qquad (25)$$

$$\frac{p_{20}}{\left(1+b\right)\left(p_{20}+b\right)}-\frac{p_{11}}{\left(1+b\right)\left(p_{11}+b\right)}+a\left(\frac{\delta_{20}\bar{S}_{10}}{1+b}+\frac{a\delta_{10}\bar{S}_{00}p_{20}}{p_{20}+b}-\frac{\delta_{11}\bar{S}_{01}}{1+b}-\frac{\delta_{21}\bar{S}_{11}p_{11}}{p_{11}+b}\right)$$

$$+a^{2}(\delta_{1,0}\delta_{2,0}\,\bar{S}_{00}\,\bar{S}_{10}-\delta_{1,1}\delta_{2,1}\,\bar{S}_{01}\,\bar{S}_{11})>0 \qquad (26)$$

s.t.

$$\frac{p_{20}}{\left(1+b\right)\left(p_{20}+b\right)}+\frac{a\delta_{20}\bar{S}_{10}}{1+b}+\frac{a\delta_{10}\bar{S}_{00}p_{20}}{p_{20}+b}+a^{2}\delta_{10}\delta_{20}\bar{S}_{10}\bar{S}_{00}>0$$

$$\frac{p_{11}}{\left(1+b\right)\left(p_{11}+b\right)}+\frac{a\delta_{11}\bar{S}_{01}}{1+b}+\frac{a\delta_{21}\bar{S}_{11}p_{11}}{p_{11}+b}+a^{2}\delta_{11}\delta_{21}\bar{S}_{11}\bar{S}_{21}]>0$$

The first and the second item in (26)

$$\frac{b}{(1+b)(p_{11}+b)(p_{20}+b)}(p_{20}-p_{11})\,. \qquad (27)$$

Compared to (24), (27) has little effect in increasing or reducing probability, except according to the convention: If $P(\vec{x}_{1}\,|\,\lambda_{0})>P(\vec{x}_{2}\,|\,\lambda_{0})$, then $P(\vec{x}_{1}\,|\,\lambda_{1})<P(\vec{x}_{2}\,|\,\lambda_{1})$. So $\delta_{20}=1$, $\delta_{21}=-1$, the first and third items in (26) are positive, the second item is far smaller than the first one. Even if the second and the fourth items were negative, the output probability with the best modal would still be bigger than the one with other modals. S_{10} is always bigger than S_{01}, and a is small enough to ignore the fourth item. When the LP of $\vec{x}_{1}$ with λ_{0} and LP of $\vec{x}_{2}$ with λ_{1} is big, the compensation transformation can enlarge the distance between these

two probabilities.

3. $p_{11} = p_{20} = 1$, the analysis is similar to Derivation 2.

6. Experiment Results

In this paper, six people (three male and three female) have taken part in a recording test. They read 27 sentences using five kinds of emotion (happiness, angry, neutral, surprise and sadness), every sentence was read three times, and 2430 sentences were taken as the experiment materials.

GMM with compensation and GMM without compensation are compared first. In the first experiment, globe features and structural features of the time series were utilized. The result is shown in Table 2. In the second experiment, 12 LPCC, 12 MFCC, 16 PLP were utilized. The result is listed in Table 3. Set $K = n = 3$, $a_{ii} \equiv const = 0.01$

Table 2. the result of the experiments between compensated and uncompensated emotion recognition (globe features and structural features %)

Emotion	Uncompensated GMM	Compensated GMM
Anger	77.6	86.2
Sadness	84.5	99.8
Happiness	73.4	80.0
Surprise	75.8	79.3
Neutral	71.6	77.1

Table 3. the result of the experiments between compensated and uncompensated emotion recognition (LPCC, MFCC, PLP %)

Emotion	Uncompensated GMM	Compensated GMM
Anger	76.3	84.2
Sadness	82.1	97.8
Happiness	79.6	88.3
Surprise	77.8	82.1
Neutral	80.4	87.0

The experiments indicate that the compensation transformation can improve the recognition rate effectively. Angry recognition rate increased 8.2%, sadness recognition rate increased 15.5%, and happiness recognition rate increased 8.5%, surprise recognition rate increased 4%, and neutral recognition rate increased 6%. The selection of K, n, a_{ti} also can improve recognition rate. Here, the authors only selected a set of parameters to explain the effectiveness and robustness of the method. Due to the compensation for GMM, the probability of the output has been stabilized and ΔS_2 has been increased.

Table 4 shows another experiment which compared three methods: KNN, NN [7] and compensated GMM (CGMM).

Table 4. KNN, NN, Compensated GMM (%)

Emotion	KNN	NN	CGMM
Anger	76.0	82.3	86.2
Sadness	82.3	86.0	99.8
Happiness	70.5	71.4	80.0
Surprise	72.2	64.0	79.3
Neutral	78.9	70.6	77.1

Compared to KNN, the recognition rate of anger using CGMM increased 10.2%, sadness increased 17.5%, happiness increased 7.5%, and surprise increased 7.1%, while neutral decreased 1.7%. This decrease doesn't effect the improvement of the whole recognition rate. Compared to NN, the average recognition rate also has been increased about 9.7% using CGMM. The results indicate that CGMM also can improve some other methods to a certain degree.

7. Conclusion and Future Works

In this paper, a method based on GMM with compensation transformation is proposed. In speech emotion recognition, the variations in speech characteristics and noise always influence the recognition results. The common method to solve this problem is conventional preprocessing. As the method in this paper deals with this problem in the recognition stage, the likelihood probability of the output with different models has been increased or decreased to reduce these influences. According to a simple analysis, this compensation transformation can reduce this impact effectively, and the examination results also proved it has better emotion recognition rates. However, the recognition rate of happiness and surprise is still not ideal, and the test materials are too few to further experiments. In further research, the authors will extend the experiment sentences first, then do some studies, such as adding more types of noise and the consideration of gender.

Reference

Austermann, A., N. Esau, L. Kleinjohann, and B. Kleinjohann, "Fuzzy Emotion Recognition in Natural Speech Dialogue," *IEEE International Workshop on Robots and Human Interactive Communication*, 2005, pp. 317-322.

Bhatti1, M. W., Y. Wang, and L. Guan, "A Neural Network Approach for Human Emotion Recognition in Speech," *IEEE Circuits and System, Proceedings of the 2004 International Symposium* ISAS, 2004, vol. 2, pp. 181-184.

Chen, Y.-B., "Automatic Segmentation of Chinese Continuous Speech," In *Proceedings of IEEE Asian Electronics Conference*, 1987, pp. 163-168, Hong Kong, (1987, 09, 1-4).

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), 2001, pp. 32-80.

Jiang, D.-N., and L.-H. Cai, "Speech Emotion Classification with the Combination of Statistic Features and Temporal Features," *IEEE International Conference on Multimedia and Expro (ICME)*, June 2004, vol.3, pp. 1967-1970.

Lin, Y. L., and G. Wei, "Speech Emotion Recognition Based on HMM and SVM," In *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics*, Guangzhou, August 2005, vol. 8, pp.18-21.

Muraka, S.,"Emotional Constituents in Text and Emotional Components in Speech," Ph. D. Theis, *Kyoto: Kyoto Institute of Technology*, Japan, 1998.

Zhao, L., X. Qian, C. Zou, and Z. Wu, "A study on emotional recognition in speech signal," *Journal of Software,* 12(7), 2001, pp. 1050-1055 (in Chinese).

Oppenheim, A. V., C.E. Kopec, and J.M.Tribolet, "Speech Analysis by Homomorphic Prediction," *IEEE Trans*., Vol. ASSP-24, pp. 327-332, 1976.

Ortony, A., and T. J. Turner, "What's Basic About Basic Emotions?" *Psychological Review,* 1990, vol. 97, pp. 315-331.

Park, C.-H., and K.-B. Sim, "Emotion Recognition And Acoustic Analysis From Speech Signal," *IEEE Neural Networks, Proceedings of the International Joint Conference.* vol. 4, 2003 July, pp. 2594-2598.

Rabiner, L. R., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Processing of the IEEE,* 1989, 77(2), pp. 257 - 286.

Razak, A.A., R. Komiya, and M. I. Z. Abidin, "Comparison Between Fuzzy and NN Method for Speech Emotion Recognition," *Third International Conference of Information Technology and Applications, 2005, ICITA 2005.* vol. 1, 4-7 July 2005, pp. 297-302.

Shigenaga, M., "Features of Emotionally Uttered Speech Revealed by Discriminant Analysis(VI)," *The preprint of the acoustical society of Japan*, 2-p-18 (1999.9) (in Japan).

Shirasawa, T., and T. Yamamura, "Discriminating Emotion Intended in Speech," *The Preprint of the Acoustical Society of Japan,* HIP: 96-38(1997) (in Japanese).

Zhao, L., X. Qian, C. Zou, and Z. Wu, "A study on emotional feature analysis and recognition in speech signal," *Journal of China Institute of Communications,* 21(10), 2000, pp. 18-25.

Zhao, L., X. Qian, C. Zou, and Z. Wu, "A study on emotional feature extract in speech signal," *Data Collection and Process,* 15(1), 2000, pp. 120-123.