Computational Linguistics and Chinese Language Processing
Vol. 9, No. 2 , August 2004, pp. 45-62

Multi-Modal Emotion Recognition from Speech and Text

Ze-Jing Chuang[*] and Chung-Hsien Wu[*]

Abstract

This paper presents an approach to emotion recognition from speech signals and textual content. In the analysis of speech signals, thirty-three acoustic features are extracted from the speech input. After Principle Component Analysis (PCA) is performed, 14 principle components are selected for discriminative representation. In this representation, each principle component is the combination of the 33 original acoustic features and forms a feature subspace. Support Vector Machines (SVMs) are adopted to classify the emotional states. In text analysis, all emotional keywords and emotion modification words are manually defined. The emotion intensity levels of emotional keywords and emotion modification words are estimated based on a collected emotion corpus. The final emotional state is determined based on the emotion outputs from the acoustic and textual analyses. Experimental results show that the emotion recognition accuracy of the integrated system is better than that of either of the two individual approaches.

1. Introduction

Human-machine interface technology has been investigated for several decades. Recent research has placed more emphasis on the recognition of nonverbal information, and has especially focused on emotion reaction. Many kinds of physiological characteristics are used to extract emotions, such as voice, facial expressions, hand gestures, body movements, heartbeat and blood pressure. Scientists have found that emotion technology can be an important component in artificial intelligence [Salovey *et al*. 1990], especially for human-human communication. Although human-computer interaction is different from human-human communication, some theories show that human-computer interaction shares

[*] Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, ROC
E-mail: {bala, chwu}@csie.ncku.edu.tw

basic characteristics with human-human interaction [Reeves *et al*. 1996]. In addition, affective information is pervasive in electronic documents, such as digital news reports, economic reports, e-mail, etc. The conclusions reached by researchers with respect to emotion can be extended to other types of subjective information [Subasic *et al*. 2001]. For example, education assistance software should be able to detect the emotions of users and; therefore; choose suitable teaching courses. Moreover, the study of emotions can apply to some assistance systems, such as virtual babysitting systems or virtual psychologist systems.

In recent years, several research works have focused on emotion recognition. Cohn and Katz [Cohn *et al*. 1998] developed a semi-automated method for emotion recognition from faces and voices. Silva [Silva *et al*. 2000] used the HMM structure to recognize emotion from both video and audio sources. Yoshitomi [Yoshitomi *et al*. 2000] combined the hidden Markov model (HMM) and neural networks to extract emotion from speech and facial expressions. Other researchers focused on extracting emotion from speech data only. Fukuda and Kostov [Fukuda *et al*. 1999] applied a wavelet/cepstrum-based software tool to perform emotion recognition from speech. Yu [Yu *et al*. 2001] developed a support vector machine (SVM)-based emotion recognition system. However, few approaches have focused on emotion recognition from textual input. Textual information is another important communication medium and can be retrieved from many sources, such as books, newspapers, web pages, e-mail messages, etc. It is not only the most popular communication medium, but also rich in emotion. With the help of natural language processing techniques, emotions can be extracted from textual input by analyzing punctuation, emotional keywords, syntactic structure, semantic information, etc. In [Chuang *et al*. 2002], the authors developed a semantic network for performing emotion recognition from textual content. That investigation focused on the use of textual information in emotion recognition systems. For example, the identification of emotional keywords in a sentence is very helpful to decide the emotional state of the sentence. A possible application of textual emotion recognition is the on-line chat system. With many on-line chat systems, users are allowed to communicate with each other by typing or speaking. A system can recognize a user's emotion and give an appropriate response.

In this paper, a multi-modal emotion recognition system is constructed to extract emotion information from both speech and text input. The emotion recognition system classifies emotions according to six basic types: happiness, sadness, anger, fear, surprise and disgust. If the emotion intensity value of the currently recognized emotion is lower than a predefined threshold, the emotion output is determined to be neutral. The proposed emotion recognition system can detect emotions from two different types of information: speech and text. To evaluate the acoustic approach, a broadcast drama, including speech signal and textual content, is adopted as the training corpus instead of artificial emotional speech. During feature selection, an initial acoustic feature set that contained 33 features is first analyzed and

extracted. These acoustic features contain several possible characteristics, such as intonation, timbre, acoustics, tempo, and rhythm. We also extract some features to represent special intonations, such as trembling speech, unvoiced speech, and crying speech. Finally, among these diverse features, the most significant features are selected by means of principle component analysis (PCA) to form an acoustic feature vector. The acoustic feature vector is fed to the Support Vector Machines (SVMs) to determine the emotion output according to hyperplanes determined by the training corpus.

For emotion recognition via text, we assume that the emotional reaction of an input sentence is essentially represented by its word appearance. Two primary word types, "emotional keywords" and "emotion modification words," are manually defined and used to extract emotion from the input sentence. All the extracted emotional keywords and emotion modification words have their corresponding "emotion descriptors" and "emotion modification values." For each input sentence, the emotion descriptors are averaged and modified using the emotion modification values to give the current emotion output. Finally, the outputs of the textual and acoustic approaches are combined with the emotion history to give the final emotion output.

The rest of the paper is organized as follows. Section 2 describes the module for recognizing emotions from speech signals. The details of SVM classification model is also provided in this section. Then the textual emotion recognition module and the integration of these two modules are presented in sections 2.3 and 3, respectively. Finally, experimental results obtained using the integrated emotion recognition system are provided in section 5, and some conclusions are drawn in section 6.

2. Acoustic Emotion Recognition Module

Deciding on appropriate acoustic features is a crucial step in emotion recognition. As in similar research, this study adopts the pitch and energy features and their derivatives. In addition, some additional characteristics may be found in emotional speech, such as trembling speech, unvoiced speech, varying speech duration, and hesitation. These features are also extracted in our approach.

2.1 Feature Extraction

A diagram of the acoustic feature extraction approach is shown in Figure 1. In the proposed approach, four basic acoustic features, pitch, energy, formant 1 (F1), and the zero crossing rate (ZCR), are estimated first. Previous research has shown that emotional reactions are strongly related to the pitch and energy of the speech. For example, the pitch of speech associated with anger or happiness is always higher than that associated with sadness or disappointment, and the energy associated with surprise or anger is also greater than that associated with fear.

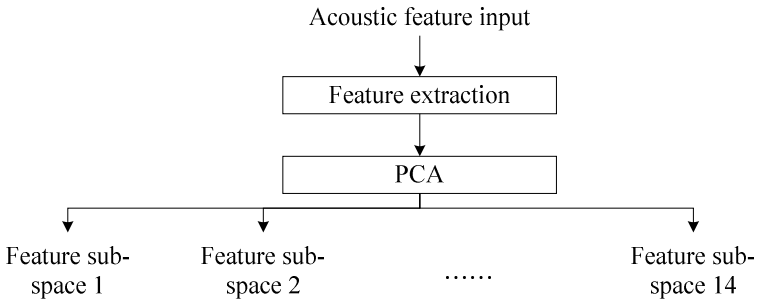

Figure 1. Diagram of the acoustic feature extraction module.

To extract an appropriate feature set, a short-time processing technique is first applied. The contours of the acoustic features are used to represent the time-varying feature characteristics. Each contour can be represented by its mean, slope, and slope difference. The Legendre polynomial [Abramowitz *et al*. 1972] is adopted to represent the contours of these four features.

In feature extraction, we adopt several parameters that are based on pitch and energy. We extract 33 acoustic features in the following 13 categories:

(1) 4^{th}-order Legendre parameters for the pitch contour;

(2) 4^{th}-order Legendre parameters for the energy contour;

(3) 4^{th}-order Legendre parameters for the formant one (F1) contour;

(4) 4^{th}-order Legendre parameters for the zero crossing rate (ZCR) contour;

(5) maximum energy;

(6) maximum smoothed energy;

(7) minimum, median, and standard deviation of the pitch contour;

(8) minimum, median, and standard deviation of the energy contour;

(9) minimum, median, and standard deviation of the smoothed pitch contour;

(10) minimum, median, and standard deviation of the smoothed energy contour;

(11) ratio of the sample number of the upslope to that of the downslope for the pitch contour;

(12) ratio of the sample number of upslope to that of the downslope for the energy contour;

(13) pitch vibration.

The features in categories (1) to (8) are statistical parameters of four basic acoustic features. In order to remove discontinuities from the contour, the pitch and energy features in categories (9) and (10) are smoothed using window method.

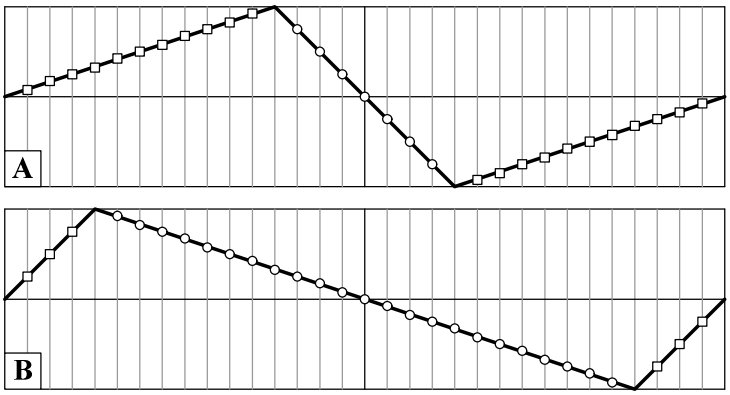

Figure 2. The ratio of up-slope sample number to the down-slope sample number. Two contours with the same wavelength are shown in parts A and B; the square symbols indicate the up-slope sample, and the circle symbols indicate the down-slope sample.

The ratios described in categories (11) and (12) represent not only the slope but also the shape of each vibration in the contour. Figure 2 shows the difference between these parameters. In this figure, each part shows the vibration of a contour. In order to show how the parameters are used, we assume that the length and the amplitude of these two contours are the same. In part A, the length of the upslope contour is longer than that of the downslope contour, while the opposite is shown in part B. The ratio of upslope to downslope is 3.14 (22 upslope samples to 7 downslope samples) in part A and 0.26 (6 upslope samples to 23 downslope samples) in part B.

Trembling speech can be characterized by means of pitch vibration. For category (13), the pitch vibration is defined and calculated as follows:

$$P_r = \frac{1}{N}\sum_{i=0}^{N-1} d\left[\left(P(i)-\bar{P}\right)\times\left(P(i+1)-\bar{P}\right)\right],\quad d[x]=\begin{cases}1 & x<0\\ 0 & x\geq 0\end{cases}, \tag{1}$$

where $\bar{P}$ is the mean value of the pitch contour.

2.2 Principle Component Analysis

Principal component analysis (PCA) is a standard statistical approach that can be used to extract the main components from a set of parameters. As described in the previous section, an initial set of 33 features is firstly extracted. After PCA is performed, 14 dimensions of principle components are chosen to capture over 90% of the total variance.

Traditionally, the 14 dimensions of principle components are used to perform classification directly. But in our approach, the principle components are used to select a more

detailed subspace. In PCA, each principle component is the linear combination of the original features. If a principle component is selected, the features that have larger combination weights are also selected and form a feature subspace. The combination weights of the original features are represented in the transformation matrix, which is calculated in PCA. By setting the threshold of the combination weights to a value of 0.2, we can select the significant features for each principle component to form a feature set. Therefore, we have 14 feature subspaces.

Table 1 shows an example of feature subspace generation. Suppose that F_1 to F_5 are the original features, that P_1 and P_2 are the selected principle components in PCA, and that the values indicate the combination weights. By selecting the original features according to values that are greater than the threshold of 0.2, we can select $\{F_1, F_3, F_4\}$ as the first feature subspace from P_1 and $\{F_2, F_4\}$ as the second feature subspace from P_2.

Table 1. An example of feature subspace generation. When a threshold value of 0.2 is applied, the generated feature subspaces are {F_1, F_3, F_4} and {F_2, F_4}.

	P_1	P_2
F_1	**0.2**	0.1
F_2	0.15	**0.3**
F_3	**0.2**	0.1
F_4	**0.3**	**0.4**
F_5	0.15	0.1

2.3 Emotion Recognition Using SVM Models

The support vector machine (SVM) [Cristianini *et al*. 2001] has been widely applied in many research areas, such as data mining, pattern recognition, linear regression, and data clustering. Given a set of data belonging to two classes, the basic idea of SVM is to find a hyperplane that can completely distinguish two different classes. The hyperplane is decided by the maximal margin of two classes, and the samples that lie in the margin are called "support vectors." The equation of the hyperplane is described in Eq. (2):

$$D(x) = \sum_{i=1}^{N} a_i y_i (x \cdot x_i) + w_0 \ . \qquad (2)$$

Traditional SVMs can construct a hard decision boundary with no probability output. In this study, SVMs with continuous probability output are proposed. Given the test sample x', the probability that x' belongs to class c is $P(class_c|x')$. This value is estimated based on the following factors:

- the distance between the test input and the hyperplane,

$$R = \frac{D(x')/\|w\|}{1/\|w\|} = D(x') \ ; \tag{3}$$

- the distance from the class centroid to the hyperplane,

$$R' = \frac{R}{D(\bar{x})} = \frac{D(x')}{D(\bar{x})} \ ; \tag{4}$$

where $\bar{x}$ is the centroid of the training data in a class;

- the classification confidence of the class;
  the classification accuracy evaluated based on the training data is used to define the classification confidence of class *c*,

$$P_c = \frac{Number\ of\ sentences\ correctly\ recognized\ as\ class\ c}{Total\ number\ of\ sentences\ in\ class\ c} \ . \tag{5}$$

Finally, the output probability is defined as follows according to the above factors:

$$P(class_c|x') = \frac{P_c}{1+\exp(1-R')} = \frac{P_c}{1+\exp\left(1-\frac{D(x')}{D(\bar{x})}\right)} \ . \tag{6}$$

As described above, the acoustic feature set is divided into 14 feature sub-spaces. For each sub-space, an SVM model is applied to decide on the best class of the speech input. The final output is the combination of these different SVM outputs, and shown as follows:

$$\begin{aligned} P(class_c|x') &= \left(\prod_{i=1}^{S} P_i(class_c|x')\right)^{\frac{1}{S}} \\ &= \left(\prod_{i=1}^{S}\left(\frac{P_c}{1+\exp(1-D(x')/D(\bar{x}))}\right)\right)^{\frac{1}{S}} \ , \end{aligned} \tag{7}$$

where the probability $P_i(class_c|x')$ is the output of SVM in the i-th feature subspace and S (=14) is the number of sub-spaces.

3. Textual Emotion Recognition Module

The most popular method for performing emotion recognition from text is to detect the appearance of emotional keywords. Generally, not only the word level but also the syntactic and semantic levels may contain emotional information. Figure 3 shows a diagram of the textual emotion recognition module. A front-end speech recognizer is first used to convert the speech signal into textual data. To extract the emotional state from the text input, we assume that every input sentence includes one or more emotional keywords and emotion modification words. The emotional keywords provide a basic emotion description of the input sentence, and

the emotion modification words can enhance or suppress the emotional state. Finally, the final emotional state is determined by combining the recognition results from both textual content and speech signal.

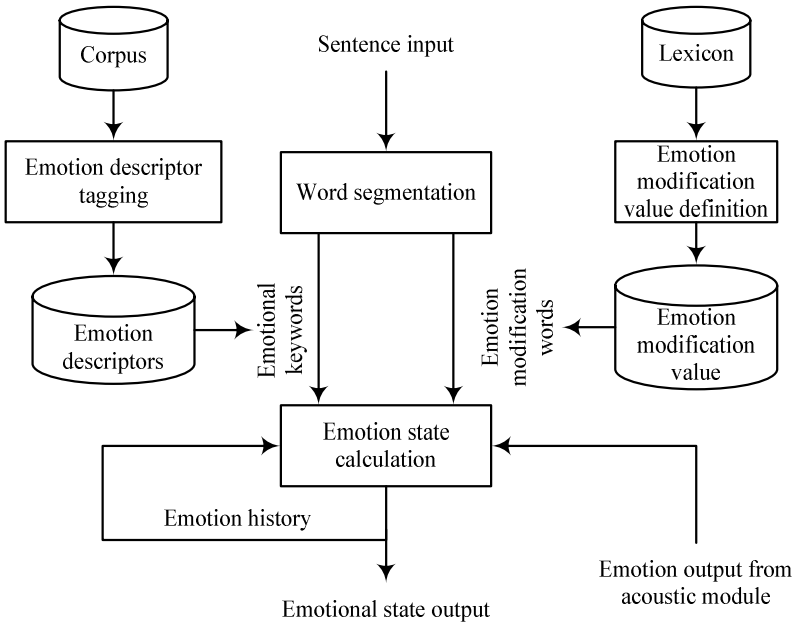

Figure 3. Diagram of textual emotion recognition module.

3.1 Front-end Processor

In order to transform the speech signal into textual data, a keyword spotting system [Wu *et al*. 2001] is applied first. The hidden Markov models (HMM) are adopted to perform keyword spotting, and the Mel-frequency cepstrum coefficients (MFCC) are extracted as the acoustic features. Obviously, the speech recognizer plays a very important role in the textual emotion recognition module. In our approach, since we consider only the emotional keywords and extract their corresponding information using HowNet, a keyword spotting system is adopted to spot the emotional keywords and emotion modification words.

3.2 Emotional Keyword Definition

For each emotional keyword, the corresponding emotion descriptor is manually defined. The emotion descriptor is a set of descriptions of the emotion reactions corresponding to the keywords. Basically, it contains an emotional state label and an intensity value, which ranges

from 0 to 1. The emotional state label can be one of the following six labels: happiness, sadness, anger, fear, surprise, and disgust. The intensity value describes how strongly the keyword belongs to this emotional state. In many cases, however, a word may contain one or more emotional reactions. Accordingly, there may be more than one emotion descriptor for each emotional keyword. For example, two emotional states, sadness and anger, are involved in the keyword "disappointed." However, the keyword "depressed" is annotated with only one emotional state: sadness. After the tagging process is completed, the emotion descriptors of the word "disappointed" are {(2, 0.2), (3, 0.6)}, and the emotion descriptor of the word "depressed" is {(3, 0.6)}. The numbers 2 and 3 in the parentheses indicate the emotional states anger and sadness, respectively. The numbers 0.2 and 0.6 represent the degree of the emotional states. In the following, we describe how the emotional state is calculated. Consider the following input sentence at time *t*:

$\boldsymbol{S_t}$: "We felt very ***disappointed*** and ***depressed*** at the results."

Here, the i^{th} emotional keyword is represented by k_i^t, $1 \le i \le M_t$, and M_t is the number of keywords in sentence S_t. In this example, k_1^t and k_2^t represent the words "***disappointed***" and "***depressed,***" respectively, and the value of M_t is 2. For each emotional keyword k_i^t, the corresponding emotion descriptor is $\left(l_r^{ti}, v_r^{ti}\right)$, $1 \le r \le R_i^t$, where R_i^t represents the number of emotion descriptors of k_i^t. The variable l_r^{ti} is the r^{th} emotional state label, and v_r^{ti} is the r^{th} intensity value of k_i^t. The value of the emotional state label can range from 1 to 6, corresponding to six emotional states: happiness, sadness, anger, fear, surprise, and disgust. In this case, the values of R_1^t and R_2^t are 2 and 1, respectively. For k_1^t, the values of l_1^{t1}, l_2^{t1}, v_1^{t1}, and v_2^{t1} are 2, 3, 0.2, and 0.6, respectively. For k_2^t, the values of l_1^{t2} and v_1^{t2} are 3 and 0.6, respectively.

The emotion descriptors of each emotional keyword are manually defined based on a Chinese lexicon containing 65620 words. In order to eliminate errors due to subjective judgment, all the words are firstly tagged by three people individually and then cross validated by the other two people. For each word, if the results tagged by different people are close, the average of these values will be set as the emotion descriptors of the word. If the three people cannot reach a common consensus, an additional person will be asked to tag the word, and the result will be taken into consideration. Based on experience, only a few words need additional suggestions.

The final tagged results for the emotion descriptors are shown in Table 2. A total of 496 words are defined as emotional keywords, and there are some ambiguities. Only 423 of them have unique emotional label definitions, 64 words have 2 emotional label definitions, and 9 words have 3 emotional label definitions. Most of the ambiguities occur in the anger and sadness categories. For example, the word "unhappy" may indicate an angry emotion or a sad emotion, according to the individual's personality and situation.

Table 2. The ambiguity of tagged emotion labels for an emotional keyword.

Number of tagged emotion labels of an emotional keyword			Total
1	**2**	**3**	
423	64	9	496

3.3 Emotion Modification Value

Besides, emotional keywords, emotion modification words also play an important role in emotion recognition. For example, the following three phrases have different emotional states and emotion degrees: "happy," "very happy," and "not happy." The only difference between these three phrases is in the emotion modification words "very" and "not." In order to quantify the emotional effect for different emotion modification words, we define an emotion modification value. According to the previous analysis of emotions [Lang. 1990], all emotion modification words can be classified into two groups: positive emotion modification words and negative emotion modification words. Positive emotion modification words strengthen the current emotional state, while negative emotion modification words reverse the current emotional state. For example, "very happy" is stronger than "happy" because of the use of word "very," but "not happy" may be sad or angry.

The emotion modification value is manually defined for each emotion modification word. It consists of a sign and a number. The sign indicates the positive or negative state of the emotion modification word, and the number indicates the modification strength of the emotion modification word. For example, the emotion modification values of the words or phrases "a little," "very," and "extremely" are +1, +2, and +3, respectively. And the emotion modification values of the words or phrases "not at all," "not," and "never" are -1, -2, and -3, respectively. The degree ranges from 1 to 3. For the example, in previous section, in the case of S_t : *"We felt* ***very*** *disappointed and depressed at the results,"* the emotion modification word is represented by g_j^t, $1 \le j \le N_t$, where N_t is the number of emotion modification words in sentence S_t. The corresponding emotion modification value of g_j^t is represented by u_j^t. In this example, g_1^t represents the word "***very.***" The values of N_t and u_1^t are 1 and +2, respectively.

4. Final Emotional State Determination

The final emotional state is the combination of the three outputs: the emotion recognition result obtained from acoustic features, the emotion descriptors of the emotional keywords, and the emotion modification values of the emotion modification words in this sentence. Given an input sentence S_t at time t, the final emotion reaction obtained from the textual content of sentence S_t is represented by E_t^C, which is a six dimension vector, $E_t^C = \left(e_1^{tC}, e_2^{tC}, e_3^{tC}, e_4^{tC}, e_5^{tC}, e_6^{tC}\right)$.

The six elements in E_t^C represent the relationship between sentence S_t and the six emotional states: happiness, sadness, anger, fear, surprise, and disgust, respectively. Each value is calculated as follows:

$$e_o^{tC} = \frac{1}{3}\left(\prod_{x=1}^{N} u_x^t\right)^{\frac{1}{N}} \left(\frac{\sum_{y=1}^{M}\sum_{z=1}^{R_y^t} S\left(l_z^{ty}, o\right) v_z^{ty}}{\sum_{y=1}^{M}\sum_{z=1}^{R_y^t} S\left(l_z^{ty}, o\right)} \right), 1 \le o \le 6 \quad . \tag{8}$$

The value in the first pair of parentheses is a geometric mean of all the emotion modification values, and the value in the second pair of parentheses is the average of intensity values that belong to emotional state *o*. The function $S\left(l_z^{ty}, o\right)$ is a step function with a value of 1 when $l_z^{ty} = o$ and a value of 0 when $l_z^{ty} \neq o$. The constant 1/3 is used to normalize the emotion intensity value to the range from -1 to 1.

After the emotion reaction from the textual content has been calculated, the final emotion output E_t is the combination of E_t^A and E_t^C,

$$\begin{aligned} &E_t = \left(e_1^t, e_2^t, e_3^t, e_4^t, e_5^t, e_6^t\right) \\ &e_o^t = a\, e_o^{tA} + (1-a)\, e_o^{tC} \quad , \text{ where } a = \max_{1 \le o \le 6}\left(e_o^{tA}\right) \end{aligned} \quad . \tag{9}$$

The emotion output of acoustic module e_i^{tA} ranges from 0 to 1, and the emotion output of textual module e_i^{tC} ranges from -1 to +1.

According to the assumption that the current emotional state is influenced by the previous emotional states, the output of the current emotion vector E_t must be modified by means of its previous emotion vector E_{t-1}. The recursive calculation of the emotion history is defined as follows:

$$E_t' = d\, E_t + (1-d)\, E_{t-1}, \quad t \ge 1, \tag{10}$$

where E_t is the *t*-th emotion vector calculated in as described in the previous section; E_t' indicates that the final output considers the emotion history, and the initial value E_0 is the output without any modification. The combination coefficient δ is empirically set to 0.75.

5. Experimental Results

For the purpose of system evaluation, in order to obtain real emotional states from natural speech signals, we collected the training corpus from broadcast dramas. There were 1085 sentences in 227 dialogues from leading man and 1015 sentences in 213 dialogues from leading woman. The emotional states of these sentences were tagged manually. The emotion tagging results are listed in Table 3.

Table 3. Tagged emotion labels in the testing corpus.

	Number of tagged sentences	
	Male	**Female**
Happiness	126	121
Sadness	121	92
Anger	98	80
Fear	60	58
Surprise	196	172
Disgust	106	113
Neutral	1617	1530

The system was implemented on a personal computer with a Pentium IV CPU and 512 MB of memory. A high-sensitivity microphone was connected to the computer and provided real-time information about speech signals.

5.1 Experiment on Acoustic Feature Extraction with PCA

As described in section II, 33 acoustic features are analyzed using PCA with thresholds of 90% and 0.2, which are the thresholds for deciding on the important principle components and the significant features of each component, respectively. The PCA process also divides the original feature space into 14 feature sub-spaces. The value of the threshold and the number of feature sub-spaces are decided experimentally.

For acoustic feature evaluation, an SVM classification system was constructed for this experiment. The threshold for deciding on the important principle components (R^2) was set to be within a range of from 85% to 100% with a step size of 2%, and the threshold for deciding on the significant features of components (T) was set to be with a range of from −1 to 1 with a step size of 0.1. The experimental results are shown in Figure 4.

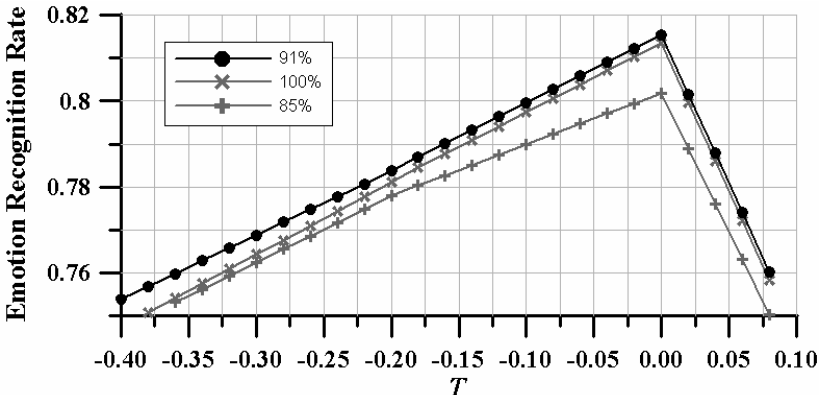

Figure 4. Emotion recognition rates for acoustic features under different PCA thresholds. The black line indicates the results obtained when R^2 = 91%, and the two gray lines indicate the results obtained when R^2 = 85% and 100%.

As shown in Figure 4, the achieved recognition rate was 63.33% when $T = -1$. When R^2 = 91% and T = 0, the achieved recognition rate was 81.55%, the highest rate obtained in all the tests. The results show that after PCA was performed, the orthogonal feature space was extracted from the original feature sets when R^2 = 91% and T = 0, and the emotion recognition rate also increased due to the elimination of dependency.

Based on the results, we could decide on the appropriate number of feature sub-spaces. Figure 5 shows the relation between the number of sub-spaces and R^2. Since the previous experiment indicated that an appropriate value of R^2 was 91%, the appropriate number of sub-spaces was chosen as 14 based on the curve in Figure 5.

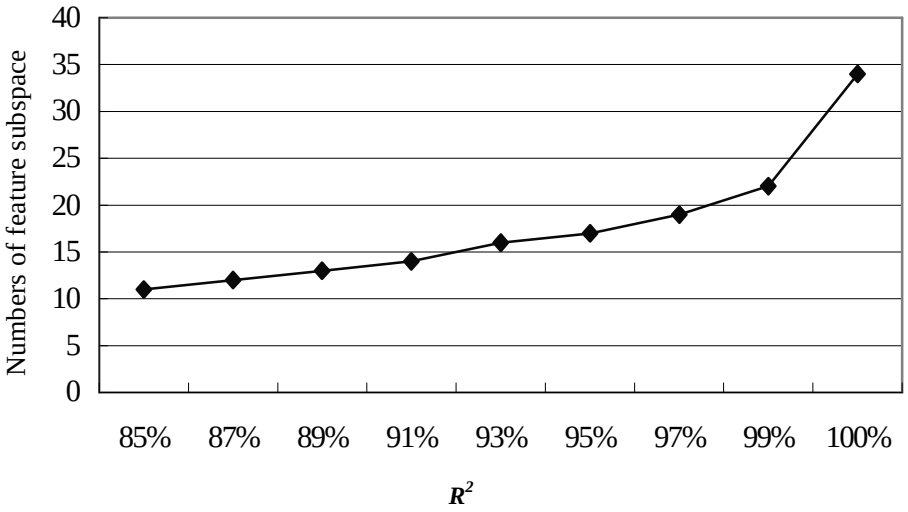

Figure 5. The relationship between R2 and the number of feature sub-spaces.

5.2 Experiment on Keyword Spotting

Since the emotion recognition rate of the textual module depends on the recognition rate of the keyword spotting system, the aim of this experiment was to identify the relationship between the keyword spotting system and textual emotion recognition module. The test corpus is first prepared with all emotional keywords are annotated manually, and then all the emotional keywords in test corpus was selected randomly and fed to the textual emotion recognition module. The emotion recognition rates of the textual module according to varying ratio of the number of emotional keywords are illustrated in Figure 6.

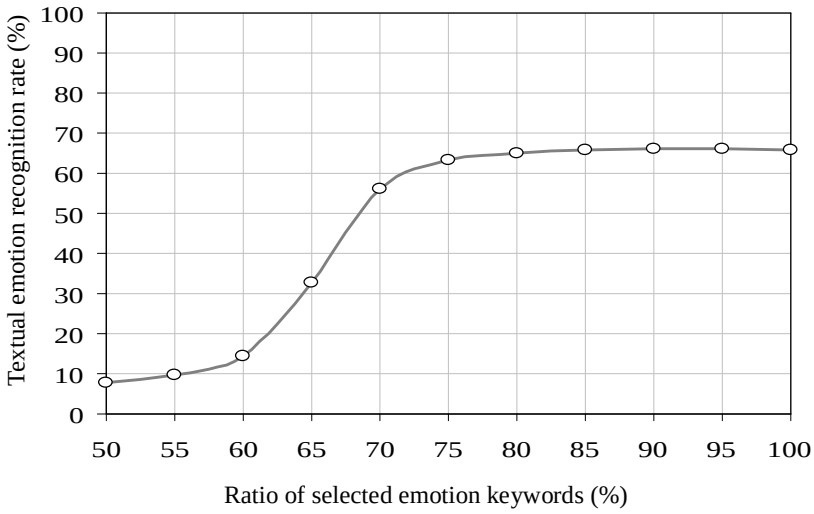

Figure 6. The relationship between the keyword recognition rate and the emotion recognition rate.

As shown in Figure 6, the emotion recognition rate of the textual module did not increase after the ratio of selected keywords reached an accuracy rate of 75%. That means if the keyword recognition rate is higher than 75%, the output of the textual emotion recognition module will reach an upper bound. Since the keyword recognition rate of the system can reach 89.6%, this keyword spotting system is suitable for the textual emotion recognition module.

5.3 Emotion Recognition Results Obtained from Acoustic Information

In this experiment, 14 feature subspaces were adopted. The radial basis function was chosen as the kernel function in the SVM model. Table 4 shows the results obtained by acoustic module. Since the dramatic dialogues were spoken by professional actors, the variation of speech intonation was very large with, therefore, decreased the recognition rate. In addition, the recognition rates for neutral and sadness were a little higher than those for other emotions. Checking the speech corpus, we found that the intonation patterns for neutral and sadness are more stable than those for other emotions. This was the main reason why these experimental results were obtained.

Table 4. Emotion recognition results obtained, based on acoustic information.

	Recognition rate		
	Male	**Female**	***Average***
Happiness	78.85%	71.90%	75.37%
Sadness	85.40%	88.04%	86.72%
Anger	81.52%	75.00%	78.26%
Fear	72.13%	70.18%	71.16%
Surprise	73.55%	62.54%	68.05%
Disgust	76.32%	68.79%	72.56%
Neutral	88.38%	77.53%	82.96%
Average	79.45%	73.43%	76.44%

The acoustic module is based on the assumption that the speech information is too complicated to be classified using only one SVM. Thus, PCA is used to generate the feature subspace. In order to test this assumption, we compared the recognition results for speech input obtained using the classifier with a single SVM and multiple SVMs. Table 5 shows the comparison and confirms the assumption.

Table 5. A comparison of the results obtained using the acoustic module with a single SVM and multiple SVMs.

	Multiple SVM	**Single SVM**
Happiness	75.37%	68.13%
Sadness	86.72%	75.91%
Anger	78.26%	66.57%
Fear	71.16%	60.55%
Surprise	68.05%	55.62%
Disgust	72.56%	64.54%
Neutral	82.96%	70.01%
Total	76.44%	65.90%

5.4 Emotion Recognition Results Obtained from Textual Content

The experimental results obtained by the textual emotion recognition module are listed in Table 6. From these results, we can find that the recognition rate cannot achieve the same level in the case of the acoustic module, i.e., the keyword-based approach cannot achieve satisfactory performance. The reasons of these results are two twofold. Firstly, owing to the complexity of natural language, sentences with the same emotional state may not contain the same emotional keywords. Secondly, as mentioned above, less than 500 words are labeled as emotional keywords from a total of 65620 words. This leads to the low occurrence rate of the occurrence of emotional keywords. But when emotional keywords appear in a sentence, the emotional reaction of the sentence is always strongly related to these keywords. The

keyword-based approach is still helpful for improving performance when integrated with the acoustic module.

Table 6. Emotion recognition results obtained based on textual content.

	Recognition rate		
	Male	**Female**	***Average***
Happiness	66.35%	63.64%	64.99%
Sadness	59.12%	61.96%	60.54%
Anger	76.09%	72.50%	74.29%
Fear	71.03%	65.51%	68.27%
Surprise	66.85%	58.46%	62.66%
Disgust	57.12%	55.34%	56.23%
Neutral	77.98%	64.76%	71.37%
Average	67.79%	63.17%	65.48%

5.5 Emotion Recognition Results Obtained Using the Integrated System

Finally, the experimental results obtained using the integrated system are shown in Table 7. The outside test was performed using an extra corpus collected from the same broadcast drama. There were a total of 200 sentences in 51 dialogues in this corpus. When the integration strategy was used, the performance of the integrated system is better than that any of the individual modules. Compared with the results obtained by the acoustic module, the results obtained with the integrated system were 5.05% higher. In order to understand the results, we verified the test corpus manually and found that when a sentence was recognized as having one emotional state, it usually contained either emotional keywords or no keywords. Only a few sentences contained emotional keywords with opposite the emotional states. Thus when the output of the acoustic module was reliable, the output of the textual module could slightly support the results obtained by the acoustic module. But if the acoustic module could not identify the emotional state of an input sentence, the emotional keywords played an important role in the final calculation.

Table 7. Emotion recognition results obtained using the integrated system.

	Inside	**Outside**
Happiness	84.44%	66.67%
Sadness	82.98%	73.91%
Anger	79.66%	67.65%
Fear	78.24%	62.37%
Surprise	80.33%	69.52%
Disgust	76.51%	70.43%
Neutral	88.24%	76.84%
Average	81.49%	69.63%

6. Conclusion

In this paper, an emotion recognition system with multi-modal input has been proposed. When PCA and the SVM model are applied, the emotional state of a speech input can be classified and fed into the textual emotion recognition module. This approach to recognizing emotions from textual information is based on pre-defined emotion descriptors and emotion modification values. After all the emotion outputs have been integrated, the final emotional state is further smoothed by mean of the previous emotion history. The experimental results show that the multi-modal strategy is a more promising approach to emotion recognition than the single module strategy.

In our study, we investigated a method of textual emotion recognition and also tested the combination of the two emotion recognition approaches. Our method can extract emotions from both speech and textual information without the need for a sophisticated speech recognizer. However, there are still many problems that remain to be solved. For example, in the textual emotion recognition module, syntactic structure information is important for natural language processing but cannot be obtained using HowNet alone. An additional parser may be needed to solve this problem. In the acoustic module, crying and laughing sounds are useful for deciding on the current emotional state but are hard to extract. A sound recognizer may, thus, be useful for improving the emotion recognition performance.

References

Salovey, P. and J. Mayer, "Emotional Intelligence," *Imagination, Cognition and Personality*, vol. 9, no. 3, 1990, pp.185-211.

Reeves, B. and C. Nass, "The Media Equation : How People Treat Computers, Television and New Media Like Real People and Places," *Cambridge Univ. Press*, 1996.

Subasic, P. and A. Huettner, "Affect Analysis of Text Using Fussy Semantic Typing," *IEEE Transactions on Fussy System*, vol. 9, no. 4, 2001, pp.483-496.

Cohn, J.F. and G.S. Katz, "Bimodal Expression of Emotion by Face and Voice," *Proceedings of the sixth ACM international conference on Multimedia: Face/gesture recognition and their applications*, 1998, pp.41-44.

Silva, L. C De and N.P. Chi, "Bimodal Emotion Recognition," *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000, pp.332-335.

Yoshitomi, Y., S.I. Kim, T. Kawano, and T. Kitazoe, "Effect of Sensor Fusion for recognition of Emotional States Using Voice, Face Image and Thermal Image of Face," *Proceedings of the ninth IEEE International Workshop on Robot and Human Interactive Communication*, 2000, pp.173-183.

Fukuda, S. and V. Kostov, "Extracting Emotion from Voice," *Proceedings of IEEE International Workshop on Systems, Man, and Cybernetics*, vol. 4, 1999, pp.299-304.

Yu, F., E. Chang, Y.Q. Xu, and H.Y. Shum, "Emotion Detection from Speech to Enrich Multimedia Content," *Proceedings of IEEE Pacific Rim Conference on Multimedia*, 2001, pp.550-557.

Chuang, Z.J. and C.H. Wu, "Emotion Recognition from Textual Input using an Emotional Semantic Network," *Proceedings of IEEE International Conference on Spoken Language Processing*, 2002, pp.2033-2036.

Abramowitz, M. and I.A. Stegun, "Legendre Functions and Orthogonal Polynomials," in *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, New York: Dover, 1972, pp.331-339.

Cristianini, N. and J. Shawe-Taylor, "An Introduction to Support Vector Machines," *Cambridge University Press*, 2001.

Wu, C.H. and Y.J. Chen, "Multi-Keyword Spotting of Telephone Speech Using Fuzzy Search Algorithm and Keyword-Driven Two-Level CBSM," *Speech communication*, Vol.33, 2001, pp.197-212.

Lang, P.J., M.M. Bradley, and B.N. Cuthbert, "Emotion, atten Lang tion, and the startle reflex," *Psychological Review*, 97, 1990, pp.377-395.