The Construction and Testing of a Mandarin Emotional Speech Database

Tsang-Long Pao, Yu-Te Chen, Jhih-Jheng Lu, Jun-Heng Yeh

Department of Computer Science and Engineering, Tatung University, Taipei

E-mail: tlpao@ttu.edu.tw, d890600S@mail.ttu.edu.tw, g9106001@ms2.ttu.edu.tw

Abstract. Researches in speech synthesis and speech analysis are underpinned by the databases they used. The performance of an emotion classifier relies heavily on the qualities of the training and testing data. A good database can make researches in these fields achieving better results. Hearing-impaired people are poor in presenting their emotions in speech. We want develop a computer-assisted speech training system that can help to teach them to present their emotions similar to normal people. In this paper, we present a way to build a Mandarin emotional database, including the process of collecting data, arranging data, clips naming rules, and a listening test. Then we construct a computer-assisted speech training system to help in teaching the hearing-impaired people presenting their emotion in their speech correctly by analyzing the emotion in their speech and those in the database using KNN and M-KNN techniques.

Keywords: Emotional speech database, Emotion evaluation, Emotion Radar chart, M-KNN

1. Introduction

The performance of an emotion classifier relies heavily on the quality of emotional speech data and the similarity of it to real world samples. As mentioned in [1], there are three different categories of emotional speech: acted speech, elicited speech, and spontaneous speech. In this section we will describe the ways to obtain these three kinds of speech data.

In acted speech recording, actors are invited to record utterances, where each utterance needs to be spoken with multiple emotions. The method is adopted by most researches because it can get large amount of data in a short time and the data is undistorted. For general use, we should invite speakers with different age, gender, even with different social or educational background if possible. And if we hope the emotion in the data to be more obvious, we could invite professional actors.

We can also collect the clips that contain utterances with specific emotion in a film. We must avoid the background noise including music, surrounding noise, and other people's voice. This method takes quite a lot of time in viewing the content of films.

In elicited speech recording, the Wizard-of-Oz (WOZ) is used. The WOZ means using a program that interacts with the speaker and drives him into a specific emotion situation and then records his voice. This method needs a good program that can induce the participator to say something in our expected emotion state. So how to design such a program may not be easy.

In spontaneous speech recording, the real-world utterances that express emotions are recorded. Although data got from this method has the best naturalness, it is the most difficult because we need to follow the speaker. When he or she is in some emotion state, his voice is recorded immediately. This method will face many problems. For examples, we must hide our recording device in order to make the speaker without any pressure to present his real emotion. Furthermore, we also cannot assure the environment is quiet. Generally speaking, the method is generally infeasible.

2. Mandarin Emotional Speech Database

In our research, five emotions are investigated: anger, happiness, sadness, boredom, and neutral. We invite 18 males and 16 females to simulate five emotions. A prompting text with 20 different sentences is designed. The length of each sentence is from one word to six words the sentences are meaningful so speakers could easily simulate them with emotions. During the recordings process, speakers are asked to try their best to simulate each

emotion. And speakers can simulate one sentence many times until they are satisfied what they simulated. Finally, we obtained 3,400 emotional speech sentences. After the recording procedure, a listening test is held to evaluate these recorded sentences.

It is very important to use speech with unambiguous emotional content for further analysis. This can be guaranteed by a listening test [2], in which listeners evaluate the emotional content of a recorded sentence. Moreover, we can understand the performance of human in emotion recognition.

We perform the listening test in a three-pass procedure. First, we delete the speech data that is very hard to identify its emotional content. After this process, 1,178 sentences are remained. Then, the remaining sentences are evaluated by three speakers. The sentences with the same agreement are remained. After the stage, 839 sentences are remained. Finally, we invite 10 people whom did not have their speech data in the 839 sentences to take part the final listening test.

The results of the listen test are shown in Fig. 1. We can see the recognition results of the 10 evaluators in the figure and the confusion matrix in Table 1. The results reveal that people are good in recognizing anger (89.56%), sadness (82.76%), and neutral state (83.51%), but are less confident for happiness (73.22%), and boredom (75.16%)

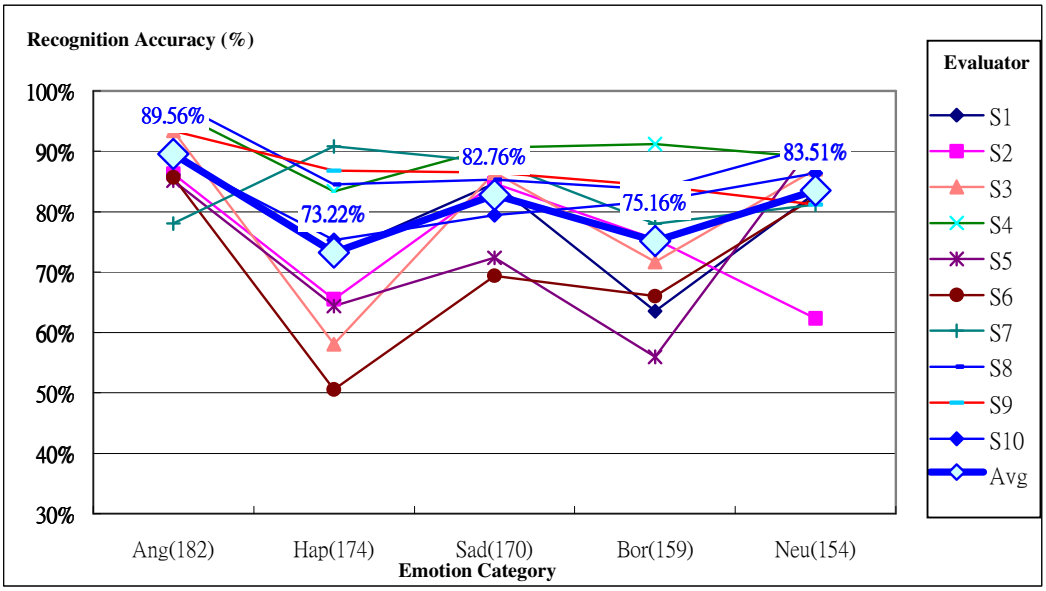

Fig. 1. Recognition results of 10 evaluators.

Table 1: Confusion Matrix of Human Performance.

	Anger	Happiness	Sadness	Boredom	Neutral	None of above
Anger	89.56%	4.29%	0.88%	0.77%	3.52%	0.99%
Happiness	6.67%	73.22%	3.28%	2.36%	13.56%	0.92%
Sadness	2.94%	1.00%	82.76%	9.29%	3.29%	0.71%
Boredom	1.26%	0.44%	8.62%	75.16%	13.65%	0.88%
Neutral	1.69%	0.91%	1.56%	12.27%	83.51%	0.06%

Table 1 shows the human performance confusion matrix. The rows and the columns represent simulated and evaluated categories, respectively. For example, first row says that 89.56% of utterances that were portrayed as angry were evaluated as angry, 4.29% as happy, 0.88% as sad, 0.77% as bored, 3.25% as neutral, and 0.99% if none of above. We can see that the most easily recognizable category is anger (89.56%) and the poorest recognizable category is happiness (73.22%). And we can find that human sometimes are confusing in differentiating anger from happiness, and boredom from neutral.

Table 2 shows the statistics of 10 evaluators for each emotion category. We can see that the variance for anger and sadness are less than for the other emotions. It means that human are better in understanding how to recognize anger and sadness than other emotions.

Figure 2 shows the percentage of remained sentences with different lengths for each emotion. We can see that the shortest sentence (only single word) is least remained in most emotions, especially in neutral. It means that we should avoid too short sentences when we make the prompting text in the future because emotions are hard to be recognized by human if the sentence is too short.

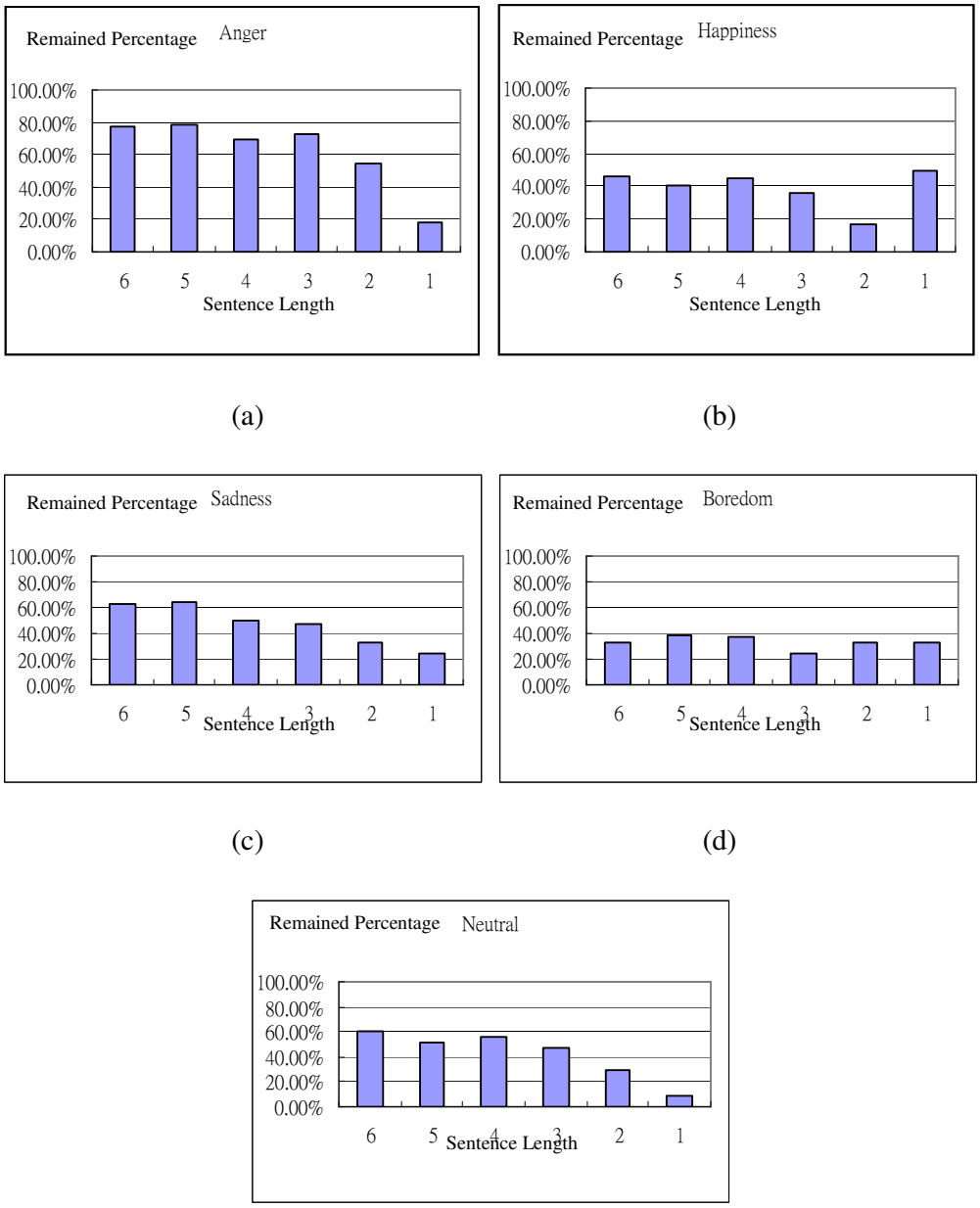

Fig. 2. Percentages of remained sentences with different lengths

Table 2: Evaluator's statistics in each category.

Category	**Mean**	**S.T.D**	**Median**	**Maximum**	**Minimum**
Anger	89.56%	6.11%	89.29%	98.35%	78.02%
Happiness	73.22%	13.36%	74.14%	90.80%	50.57%
Sadness	82.76%	6.92%	85.00%	90.59%	69.41%
Boredom	75.16%	10.87%	76.73%	91.19%	55.97%
Neutral	83.51%	8.37%	84.74%	91.56%	62.34%

For further analysis, we only need the speech data that can be recognized by most human. So we divide speech data into different dataset by their recognition accuracy. We will refer to these data sets as D80, D90, D100, which stand for recognition accuracy of at least 80%, 90%, and 100%, respectively, as listed in Table 3.

Table 3: Datasets and their sizes

Data set	D80	D90	D100
Size (number of sentences)	570	473	283

3. Emotion Recognition and Emotion Evaluation

3.1 Emotion Recognition

Figure 3 shows the block diagram of our emotion recognition system. We calculate the MFCC as the emotional feature from each input data [3]. Then, the speech is classified by pattern classification method (K-NN)[4]. K-NN will find the k neighbors nearest to the new sample from the training space based on some suitable similarity or distance measure methods.

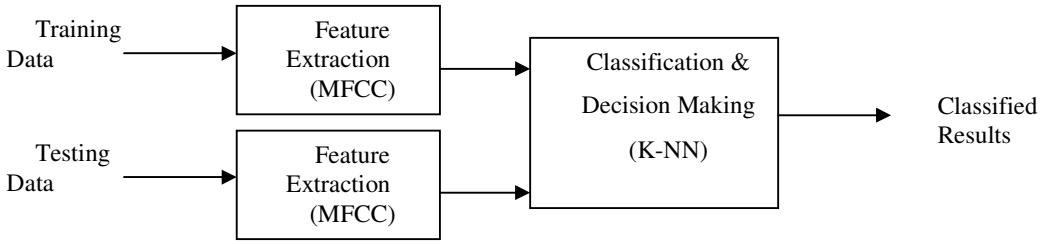

Fig. 3. Block diagram of emotion recognition

Figure 4 shows the recognition accuracy with different values of K of the K-NN classifier where we choose 70% and 30% of the speech data for training and testing, respectively. We notice that while the value of K increased, the recognition accuracy of happiness begins to drop gradually. The higher average recognition accuracy exist in K=1 and K=3. Therefore, we choose K=1 for the K-NN classifier.

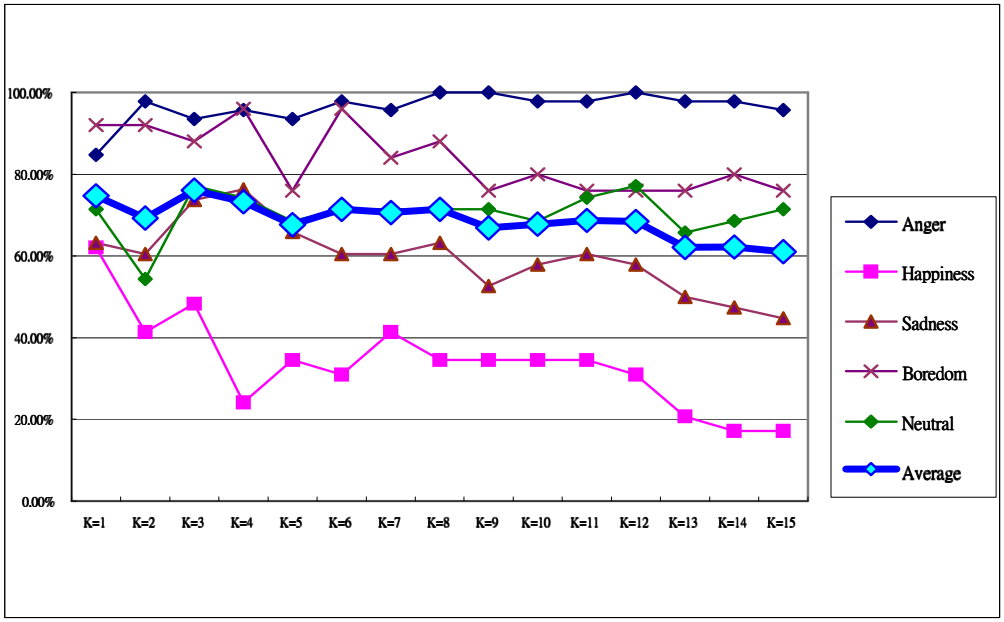

Fig. 4. Recognition rate of K=1 to 15 using KNN classfier.

Table 4 shows the confusion matrix of our recognition system that the value of K is set to one. The rows and the columns represent original and recognized emotion categories, respectively. For example, first row says that 39 sentences that belong to angry were recognized as angry, 6 sentences as happy, 1 sentence as sad, and 0 for the rest. And the recognition accuracy of anger is 84.78%. We can see that our system do better in recognizing anger and boredom. The mean recognition rate is 74.69%.

Table 4 Confusion matrix of our system

	Anger	**Happiness**	**Sadness**	**Boredom**	**Neutral**	**Recognition rate**
Anger	39	6	1	0	0	84.78%
Happiness	7	18	1	0	3	62.07%
Sadness	4	3	24	4	3	63.16%
Boredom	0	0	0	23	2	92.00%
Neutral	1	2	1	6	25	71.43%

3.2 Emotion Evaluation

We use K-NN to classify input testing data, so we modify the K-NN method to be our evaluator in the emotion evaluation stage. We called the method “M-KNN”. Figure 5 shows the block diagram of the evaluation system.

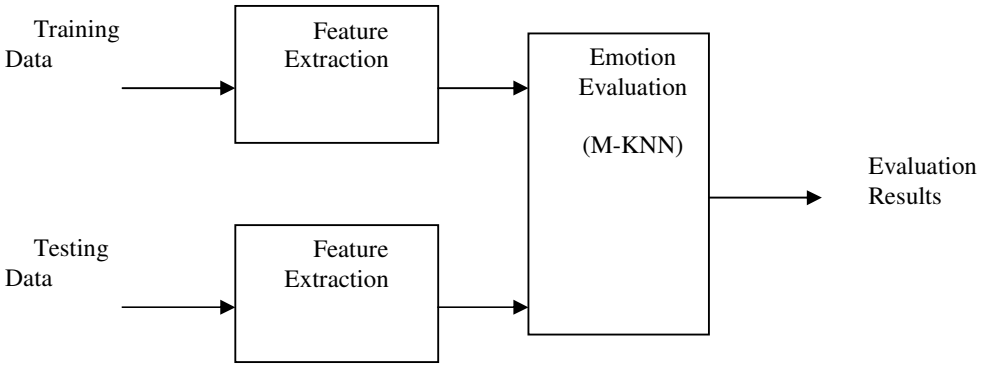

Fig. 5. Block diagram of emotion evaluation

3.3 Emotion Radar Chart

An emotion radar chart is a chart with multi-axes. Each of the axes stands for one category of emotion. In our system, it just looks like a regular pentagon as shown in Fig. 6. We need to measure the distance of a testing data to each category to plot its radar chart. Thus a modified version of KNN (M-KNN) is needed.

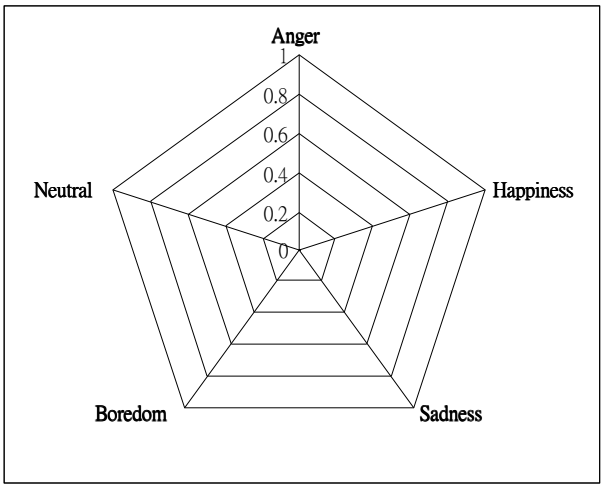

Fig. 6. Emotion radar chart

The M-KNN is based on the KNN technique. It calculates the K-nearest neighbors' distances in each class to the input testing data. We set the value of K to 1 corresponding to the K in emotion recognition. Figure 7 shows the M-1NN method.

After the calculation of M-KNN, we will get five distances from five emotion categories. We take inverse of each distance, and base on these inverses of distances to plot a radar chart. For example, Table 5 list the calculation result of an input testing speech with angry emotion. And Fig. 8 shows the radar chart according to Table 5.

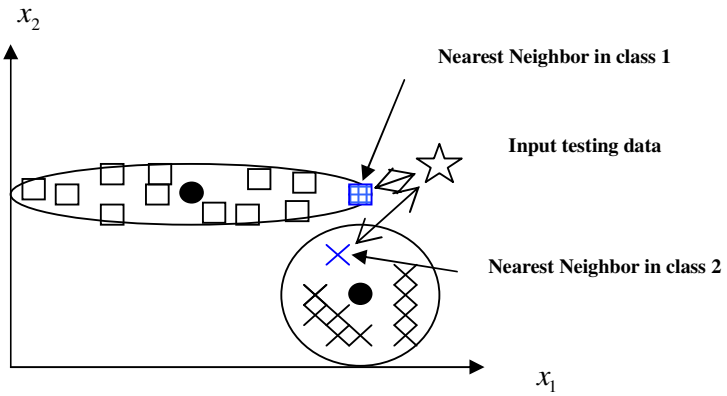

Fig. 7. M-1NN: The computation of the distance to the nearest neighbor in each class

Table 5. Distance measured by M-KNN

Emotion	Anger	Happiness	Sadness	Boredom	Neutral
Distance	12.325	19.31	23.14	27.868	22.83

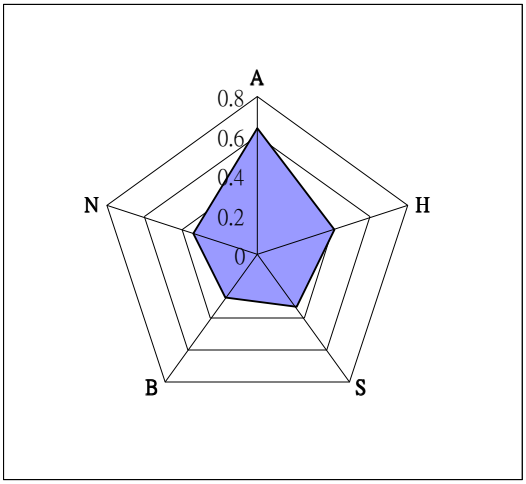

Fig. 8. Emotion radar chart of a speech with anger emotion.

From Fig. 8, we can find that this input data is closed to anger. It means the intensity of anger is greater than the other emotions. An unambiguous emotion should close to one emotion and far away the other emotions similar to the one shown in Fig. 9.

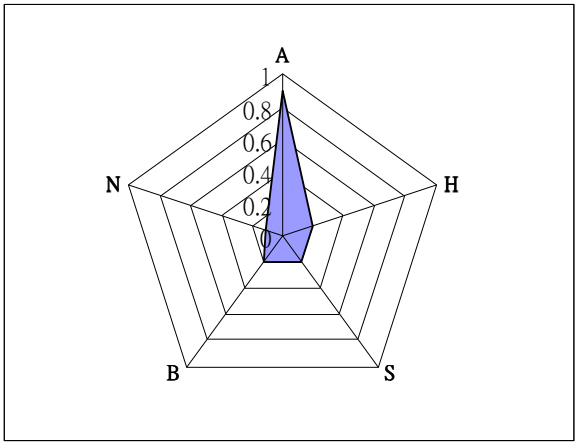

Fig. 9. Emotion radar chart of speech with unambiguous emotion

3.4 System Interface

Figure 10 shows the interface of our system. A user can record his or her voice by pressing the Record button. The user can hear the speech within the selected range by pressing the Play button. Finally, the user can see the evaluation result of his or her speech by pressing the Eva button.

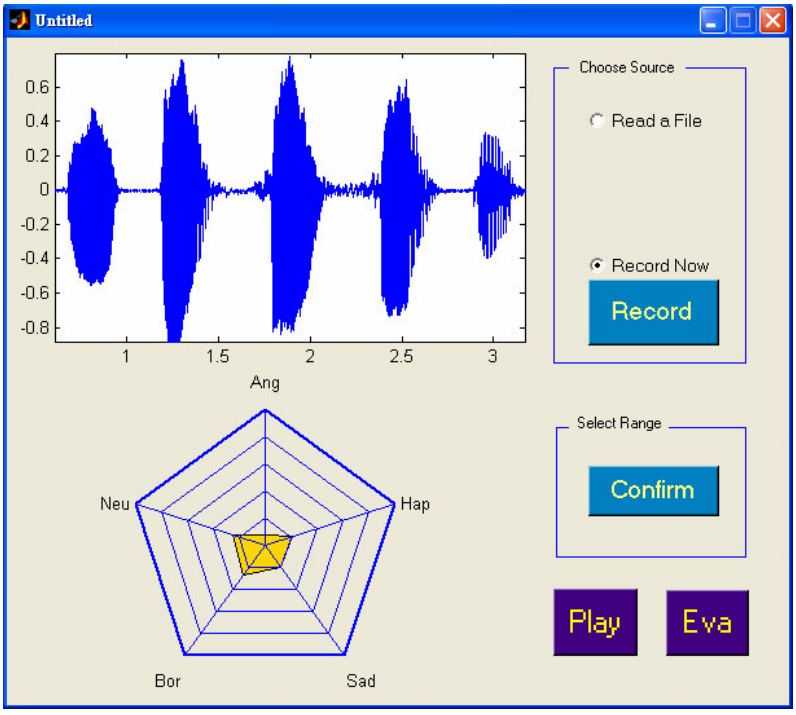

Fig. 10 System interface

When teaching the hearing-impaired people, teacher could ask him or her to say something with certain emotion. Hearing-impaired people can easily understand what emotion is presented by the emotion radar chart. Guiding by the teacher, they can try and improve the naturalness of their emotion expression through the use of the emotion radar chart.

4. Conclusions

In this paper, we build a Mandarin emotional speech database for research in this field. We also propose an emotion recognition and evaluation system. For hearing-impaired people, it could provide an easier way to learn how to speak more naturally.

We will continue to get more speech data into our database, and improve the recognition accuracy of the emotion recognition system. We also want to make the emotion evaluation more effectively. Furthermore, friendlier interface to hearing-impaired people is needed to be designed.

5. Acknowledge

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

6. Reference

[1] Raquel Tato, Rocio Santos, Ralf Kompe, "Emotional Space Improves Emotion Recognition", *Man Machine Interface Lab, Advance Technology Center Struttgart*, Sony International (Europe) GmbH.

[2] Inger Samso Engberg, Anya Varnich Hansen , "Documentation of the Danish Emotional Speech Database", *Department of Communication Technology Institute of Electronic Systems,* Aalborg University, Sep. 1996

[3] Bo-Syong Juang, "Automated Recognition of Emotion in Mandarin", *Department of Engineering Science, National Cheng Kung University Master Thesis*, Jun 2002

[4] Maleq Khan, Qin Ding, William Perrizo, "k-Nearest Neighbor Classification on Spatial Data Streams Using P-Trees", *Computer Science Department, North Dakota State University*.