Emotion Recognition and Evaluation of Mandarin Speech Using Weighted D-KNN Classification

Tsang-Long Pao, Yu-Te Chen, Jun-Heng Yeh, Yuan-Hao Chang
Department of Computer Science and Engineering, Tatung University, Taipei
tlpao@ttu.edu.tw, d8906005@mail.ttu.edu.tw, g9206026@ms2.ttu.edu.tw

Abstract. In this paper, we proposed a weighted discrete K-nearest neighbor (weighted D-KNN) classification algorithm for detecting and evaluating emotion from Mandarin speech. In the experiments of the emotion recognition, Mandarin emotional speech database used contains five basic emotions, including anger, happiness, sadness, boredom and neutral, and the extracted acoustic features are Mel-Frequency Cepstral Coefficients (MFCC) and Linear Prediction Cepstral Coefficients (LPCC). The results reveal that the highest recognition rate is 79.55% obtained with weighted D-KNN optimized based on Fibonacci series. Besides, we design an emotion radar chart which can present the intensity of each emotion in our emotion evaluation system. Based on our emotion evaluation system, we implement a computer-assisted speech training system for training the hearing-impaired people to speak more naturally.

1 Introduction

Recognizing emotions from speech has gained increased attention recently, because automatic emotion recognition can help people to develop and design many applications about human-machine communication. In emotion recognition, collecting corpus and selecting the suitable features and classification algorithms are the two most difficult problems.

Language is the most basic and main tool for the human to communicate thoughts, convey messages and express aspiration. For the hearing-normal people, the process of speak learning is very natural. But for the hearing-impaired people, it becomes almost impossible due to no auditory input. Fortunately, the hearing-impaired people are not profoundly deaf and remain some level of hearing. Using these residual hearing and other perception, the hearing-impaired people can still communicate with other people. In real life, we can often see that hearing-impaired people converse with others by sign language, lip reading or writing. In fact, sign language has low popularity among general people. Lip-reading is just a reference because it has some limitations in Mandarin vowels. Writing is not a convenient way. So in many language training, to teach the hearing-impaired people to speak is the ultimate goal.

For this reason, we want to design a computer-assisted emotional speech training system. By using this system, it can assist the hearing-impaired people to learn not only to speak correctly but also to speak naturally, just like the hearing-normal people. Besides, we also use the visual feedback in our system. Just like in many singing training system, we can see the singing score on the screen after singer has sung. Speech therapist can have no trouble to teach hearing-impaired people to speak with emotions when they communicate with people. This mechanism can let the hearing-impaired people better understand their speaking state and make the whole system more complete.

The emotional state of a speaker can be identified from the facial expression [1] [2] [3], speech [4] [5] [6], body language, perhaps brainwaves, and other biological features of the speaker. A combination of these features may be the way to achieve high accuracy of recognition. But they all are not unconditionally prerequisites necessary for extraction of an emotion.

In this paper, a system is proposed to classify and evaluate the emotions, including anger, happiness, sadness, boredom and neutral, from Mandarin speech. Several early research works in this area are reviewed as follows.

ASSESS [4] is a system that makes use of a few landmarks – peaks and troughs in the profiles of fundamental frequency, intensity and boundaries of pauses and fricative bursts in identifying four archetypal emotions. Using discriminant analysis to separate samples that belong to different categories, a classification rate of 55% was achieved.

In [5], over 1000 utterances emotional speeches, incorporating happiness, sadness, anger and fear from different speakers were classified by human subjects and by computer. Human subjects were asked to recognize the emotion from utterances of one speaker in random order. It was found that human's classification error rate was 18%. For automatic classification by computer, pitch information was extracted from the utterances. Several pattern recognition techniques were used and a miss-classification rate of 20.5% was achieved.

Nicholson et al. [6] analysed the speech of radio actors involving eight different emotions. The emotions chosen were joy, teasing, fear, sadness, disgust, anger, surprise and neutral. In the study, which was limited to emotion recognition of phonetically balanced words, both prosodic features and phonetic features were investigated. Prosodic features used were speech power and fundamental frequency while phonetic features adopted were Linear Prediction Coefficients (LPC) and the Delta LPC parameters. A neural network was used as the classifier. The best accuracy achieved in classification of the eight emotions was 50%.

Machine recognition of emotions using audiovisual information was conducted by Chan [7]. Six basic emotions, happiness, sadness, anger, dislike, surprise and fear, were classified using audio and video model separately. The recognition rate for audio alone is about 75% and video alone is about 70%. For audio processing, statistics of pitch, energy and the derivatives are extracted. Nearest mean criterion approach was adopted for classification. Joint audiovisual information of facial expression and emotive speech were used. The correct recognition rate is 97%.

For the system proposed in this paper, 20 MFCC components and 16 LPCC components were selected as the features to identify the emotional state of the speaker. Subsequently, a weighted D-KNN modified from K-Nearest Neighbor decision rule is adopted as classifier.

2 System Architecture

Figure 1 shows the block diagram of the proposed emotion recognition and evaluation system. The process of emotion recognition is the same as most previous studies. Differently, the evaluation of emotional speech is a research that only a few people focus on. Therefore, we will emphasize the evaluation of emotional speech in our research. Of course the recognition of emotional speech is also the core of our research we were interested in.

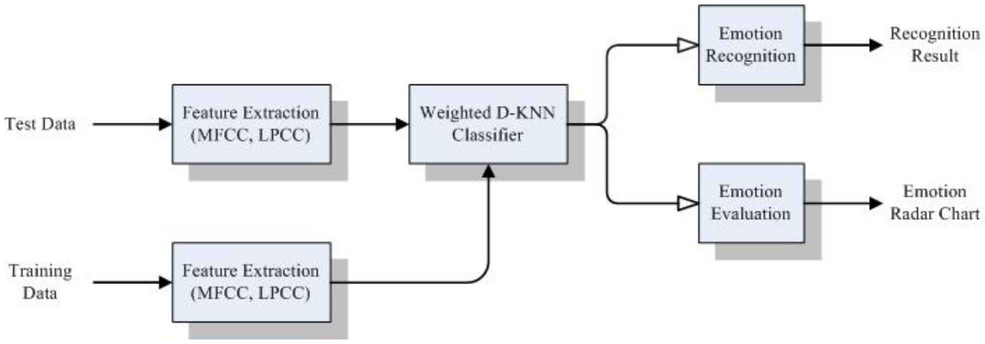


Figure 1: System architecture block diagram

2.1 Emotional Speech Database

We invite 18 males and 16 females to express five emotions, anger, happiness, sadness, boredom and neutral, in their speech. A prompting text with 20 different sentences is designed. The sentences are meaningful so speakers could easily simulate them with emotions. Finally, we obtained 3,400 emotional speech sentences.

After the three-pass listening test procedure, 839 sentences are remained and evaluated by 10 people whom did not have their speech data in the 839 sentences to take part for the final listening test [8]. Table 1 shows the human performance confusion matrix. The rows and the columns represent simulated and evaluated categories, respectively. We can see that the most easily recognizable category is anger and the poorest is happiness. And we can find that human sometimes are confusing in differentiating anger from happiness, and boredom from neutral.

Table 1: Confusion matrix of human performance (%)

	Angry	**Happy**	**Sad**	**Bored**	**Neutral**	**Others**
Angry	89.56	4.29	0.88	0.77	3.52	0.99
Happy	6.67	73.22	3.28	2.36	13.56	0.92
Sad	2.94	1.00	82.76	9.29	3.29	0.71
Bored	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 2: Datasets size

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

For further analysis, we only need the speech data that can be recognized by most people. 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 2.

In this research, the D80 dataset containing 570 utterances was used. Table 3 shows the distribution of sentences among the five emotion categories for the data set.

Table 3: Distribution of 570 sentences

Emotion Category	Number of Sentence
Angry	151
Happy	96
Sad	124
Bored	83
Neutral	116

2.2 Feature Extraction

A critical problem of all recognition systems is the selection of the feature set to use. In our previous experiment, we investigated the following feature set. Formants (F1, F2 and F3), Linear Predictive Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC), Mel-Frequency Cepstral Coefficients (MFCC), first derivative of MFCC (dMFCC), second derivative of MFCC (ddMFCC), Log Frequency Power Coefficients (LFPC), Perceptual Linear Prediction (PLP) and RelAtive SpecTrAl PLP (Rasta-PLP). Due to the highly redundant information, a forward feature selection (FFS) or backward feature selection (BFS) should be carried out to extract only the most representative features.

In FFS, LPCC is the most representative feature while in BFS, it is the MFCC. Finally, we combine MFCC and LPCC as the feature set and is used in our emotion recognition system.

2.3 Classifiers

The simplest classification algorithm is K-Nearest Neighbor (KNN) which is based on the assumption that samples residing closer in the instance space have same class values. Thus, while classifying an unclassified sample, the effects of the k nearest neighbors of the sample were considered. It yields accurate results in most of the cases. However, the classification seems unfair only determine with a point. The k-nearest neighbor classification takes k nearest samples of the testing sample to make a decision.

When a new sample data *x* arrives, KNN finds the k neighbors nearest to the unlabeled data from the training space based on some distance measure. In our case, the Euclidean distance is used. Now let the *k* prototypes nearest to *x* be $N_k(x)$ and *c(z)* be the class label of *z*. Then the subset of nearest neighbors within class $j \in \{ 1 , \ldots , \text{number of classes } l \}$ is

$$N_k^j(x) = \{ y \in N_k(x) : c(y) = j \} \qquad (1)$$

Finally, the classification result $j^* \in \{1,...,l\}$ is defined as a majority vote:

$$j^* = \arg\max_{j=1,...,l} \left| N_k^j(x) \right| \tag{2}$$

Modified-KNN is a technique based on the KNN [8]. When a new sample data *x* arrives, M-KNN finds the k neighbors nearest to the unlabeled data in each class from the training space and calculates their distances d^i. Now let the *k* prototypes nearest to *x* be $N_{k,i}(x)$ which is defined as

$$N_{k,i}(x) = \arg\ \min_{j=1,...,l}\ d_j^i\ ,\quad i = 1,...,k \tag{3}$$

The following steps are similar to KNN method in making a decision from majority vote.

In this paper, we propose a weighted D-KNN which is a combination of weighting scheme and M-KNN to improve the performance of M-KNN. The purpose of weighting is to find a vector of real-valued weights that would optimize classification accuracy of some classification or recognition system by assigning low weights to less relevant features and higher weights to features that are more important.

2.4 Emotion Evaluation

Emotion evaluation is used to evaluate emotion expression of a sentence. In this paper, the evaluation method we used is based on weighted D-KNN classification. When we take a test data to evaluate, we can obtain five values by the M-KNN classifier. The five values are the distance to the sets of emotion categories respectively. Then each emotional evaluation value can be obtained from each distance set.

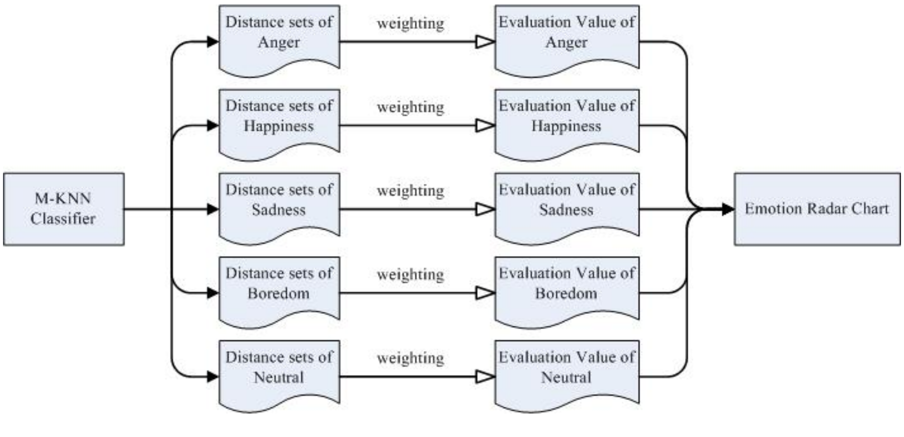


Figure 3: Block diagram of emotion evaluation

Figure 3 shows a block diagram of emotion evaluation. After the calculation of weighted D-KNN, we will obtain five evaluation values from five emotion categories. Moreover, each evaluation value of emotion can be plotted in Emotion Radar Chart that is described in detail in section 3.3.

3 Experimental Results

The weighted D-KNN classification is used in our experiment. All experiments used the MATLAB software and all results are based on the Leave-One-Out cross-validation which is a method to estimate the predictive accuracy of the classifier [9]. The extracted acoustic features were MFCC and LPCC.

3.1 Experimental Results of Weighted D-KNN Classifier

In the beginning of the experiment, we try to assign different weighting series to the calculation. These series are often used in lots of previous studies that were not limited in the field of signal processing. In addition, the constraint $w_1 \geq w_2 \geq \cdots \geq w_k$ were enforced in the weighting series lookup process. Three different series, from 10 to 1, the power of 2 and Fibonacci series, were chosen as our weighting series.

In KNN based classification, larger weighting values in the series are more important. So, in the case of Fibonacci series, such assumption is not groundless in our experiments. In our assumption, we want to assign the series that the importance of a certain value in the series is equal to the importance of the sum of two values behind that value.

Table 4: Comparison of weighted D-KNN using different weighting schemes

Weighting Scheme	Accuracy (%)
$w_i = k - i + 1$ **(k→1)**	75.39
$w_i = 2^{k-i}$ **(The power of 2)**	78.86
$w_i = w_{i+1} + w_{i+2}$, $w_k = w_{k-1} = 1$ **(Fibonacci series)**	79.31

The experimental results of the weighted D-KNN with different weighting series are summarizes in Table 4. Their corresponding confusion matrices are given in Table 5 to Table 7. The results show that different weighting scheme have different ability and property. The best accuracy is obtained with Fibonacci series scheme, and the best recognition rate is 79.31%.

Table 5: Confusion matrix of weighted D-KNN (*k*=10, weighting: 10→1)

Accuracy (%)	Angry	Happy	Sad	Bored	Neutral
Angry	88.74	3.97	2.65	0	4.64
Happy	22.92	54.17	6.25	0	16.67
Sad	4.03	1.61	79.03	6.45	8.87
Bored	0	0	9.64	84.34	6.02
Neutral	0.86	4.31	12.93	11.21	70.69

Table 6: Confusion matrix of weighted D-KNN (*k*=10, weighting: the power of 2)

Accuracy (%)	Angry	Happy	Sad	Bored	Neutral
Angry	90.07	5.29	1.99	0	2.65
Happy	19.79	61.46	4.17	0	14.58
Sad	3.23	2.42	82.26	2.42	9.68
Bored	0	2.41	6.02	85.54	6.02
Neutral	0.86	6.03	7.76	10.35	75.00

Table 7: Confusion matrix of weighted D-KNN (*k*=10, weighting: Fibonacci series)

Accuracy (%)	Angry	Happy	Sad	Bored	Neutral
Angry	90.73	4.64	1.32	0	3.31
Happy	18.75	62.50	3.13	0	15.63
Sad	4.03	2.42	82.26	2.42	8.87
Bored	0	1.20	8.43	84.33	6.02
Neutral	0.86	5.17	6.90	10.35	76.72

3.2 Weighting Optimization

Furthermore, we try to optimize the weighting series based on weighting value we used in last subsection. In addition, weighting series are also follow the constraint $w_1 \geq w_2 \geq \cdots \geq w_k$. In the experiment, we modified one weighting value and kept others fixed. The modification was done from right to left or from left to right, in the process of searching the optimum weighting values.

Table 8: Recognition accuracy of different optimum weighting series

Weighting Scheme	Accuracy (%)	
	From Left to Right	From Right to Left
k→1	78.44	77.13
The power of 2	79.07	79.31
Fibonacci series	79.52	79.55

In the next experiment, we try to optimize the weighting series that were used in section 3.1 in accordance with the directions from left to right and from right to left respectively. Table 8 shows the recognition accuracy of each optimized series, and we can find that weighted D-KNN classifier with optimum weighting series yields better results than without optimum weighting series: 3.05% improvement for weighting scheme in k→1, 0.45% improvement for the power of 2, and 0.24% improvement for Fibonacci series. The best recognition accuracy of 79.55% is obtained with weighted D-KNN optimized based on Fibonacci series. The corresponding confusion matrix is given in Table 9.

Table 9: Confusion matrix (optimized weighting from right to left with Fibonacci series)

Accuracy (%)	Angry	Happy	Sad	Bored	Neutral
Angry	90.73	4.64	1.32	0	3.31
Happy	18.75	62.50	3.13	0	15.63
Sad	4.032	2.42	82.26	2.42	8.87
Bored	0	1.20	7.23	85.54	6.02
Neutral	0.86	5.17	6.90	10.35	76.72

3.3 Emotion Radar Chart

An emotion radar chart is a multi-axes plot. Each of the axes stands for one emotion category. In our system, emotion radar chart just look like a regular pentagon as shown in Fig. 4. Figure 4 is an Emotion Radar Chart plotted using the data from Table 10 and 11. We can find that this input data is closed to angry emotion, and anger intensity of the speech is greater than the other emotions.

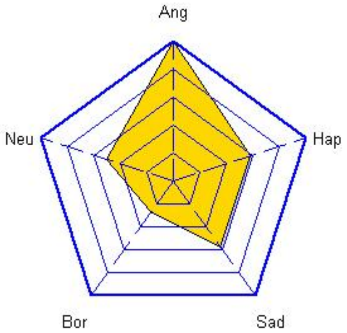


Figure 4: Emotion Radar Chart of test data with angry emotion

Table 10 shows the fifty distance values, 10 neighbors from each emotion class nearest to the input test data which is an angry speech. For example, first row shows the first 10 distances from input test data to training data of angry emotion. Here we call the value of the first row the distance set of Anger, and detailed description and operation is described in section 2.4. We can see clearly that the minimum distance in each round is almost the distance from input test data to the training data of angry emotion. Table 11 shows the calculation result of each distance set obtained by weighted D-KNN classification.

Table 10: Distance measured by M-KNN with *k* =10

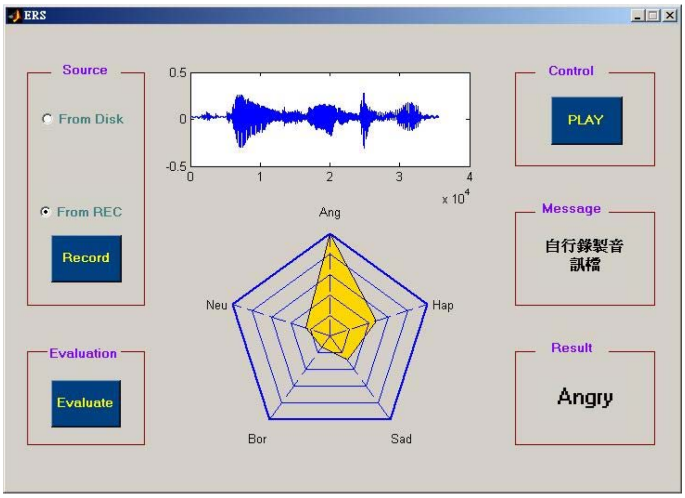
Round	1	2	3	4	5	6	7	8	9	10
Angry	8.17	9.62	9.64	10.23	11.44	11.53	11.62	12.58	12.66	12.67
Happy	11.26	11.72	13.16	13.80	14.65	11.53	11.62	12.58	12.66	12.67
Sad	11.34	12.21	12.83	13.06	13.21	15.24	15.91	16.14	16.17	16.64
Bored	16.40	19.04	19.06	19.20	19.29	19.67	19.85	20.02	20.17	20.26
Neutral	11.96	12.40	14.55	15.12	15.57	15.72	15.74	15.87	15.95	16.09

Table 11: Evaluation value obtained by weighted D-KNN (Normalized with the maximum)

Emotion	**Anger**	**Happiness**	**Sadness**	**Boredom**	**Neutral**
Evaluation Value	1.0000	0.6032	0.5768	0.2699	0.5048

3.4 System Interface

Figure 5 is the user interface of our system. First, the source of the test speech has to be chosen from Source block. Test speech can get from disk or from recording. Second, after choosing the source, the Evaluation button in the Evaluation block can be pressed to plot the emotion radar chart on the lower graph. Finally, the Message frame will show the current state or error message, and Result block shows the recognition result of emotion of the test speech.



○ From Disk

● From REC

Message: 自行錄製音訊檔

Result: Angry

Figure 5: System interface (Evaluate test data from recording)

4 Conclusions

In this paper, we optimized the weights in weighted D-KNN to improve the recognition rate in our emotion recognition system. That is, we tried to modify slightly the weights in weighted D-KNN, and the accuracy of emotion recognition increased. The highest recognition rate of 79.55% is obtained with weighted D-KNN optimized based on Fibonacci series.

We also propose an emotion recognition and evaluation system. We regard the system as a computer-assisted emotional speech training system. For hearing-impaired people, it could provide an easier way to learn how to speak with emotion more naturally or help speech therapist to guide hearing-impaired people to express correct emotion in speech.

In the future, it is necessary to collect more acted or spontaneous speech sentences. Furthermore, it might be useful to measure the confidence of the decision after performing classification. Based on confidence threshold, classification result might be classified as reliable or not. Moreover, we also want to make the emotion evaluation more effectively, and a more user friendly interface of system for hearing-impaired people needs to be designed. Besides, how to optimize the weights in weighted D-KNN to improve the recognition rate in emotion recognition system is still a challenge work.

5 Acknowledge

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

References

[1] P. Ekman, *Darwin and Facial Expressions*, Academic, New York, 1973.

[2] M. Davis and H.College, *Recognition of Facial Expression*, Arno Press, New York, 1975.

[3] K. Scherer and P. Ekman, *Approaches to Emotion*, Lawrence Erlbaum Associates, Mahwah, NJ, 1984.

[4] S. McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk, S. Stroeve, “Approaching Automatic Recognition of Emotion from Voice: A Rough Benchmark,” *ISCA Workshop on Speech and Emotion*, Belfast, 2000.

[5] F. Dellaert, T. Polzin and A. Waibel, “Recognizing Emotion in Speech,” *Fourth International Conference on Spoken Language Processing*, Vol. 3, 1996, pp. 1970-1973.

[6] J. Nicholson, K. Takahashi, R. Nakatsu, “Emotion Recognition in Speech Using Neural Networks,” *6th International Conference on Neural Information Processing*, ICONIP ’99, Vol. 2, 1999, pp. 495-501.

[7] L. S. Chan, H. Tao, T.S. Huang, T. Miyasato, and R. Nakatsu, “Emotion Recognition from Audiovisual Information,” *IEEE Second Workshop on Multimedia Signal Processing*, 1998, pp. 83-88.

[8] Tsang-Long Pao, Yu-Te Chen, Jhih-Jheng Lu and Jun-Heng Yeh, “The Construction and Testing of a Mandarin Emotional Speech Database,” *Proceeding of ROCLING XVI*, Sep. 2004, pp. 355-363.

[9] Tsang-Long Pao, Yu-Te Chen, Jun-Heng Yeh and Jhih-Jheng Lu, “Detecting Emotions in Mandarin Speech,” *Proceeding of ROCLING XVI*, Sep. 2004, pp. 365-373.