Automatic Emotion Recognition of Speech Signal in Mandarin

Sheng Zhang[1], *P.C. Ching*[2], *Fanrang Kong*[1]

[1] Dept. of Precision Machinery and Precision Instrumentation
University of Science & Technology of China, 230027, China
[2] Dept. of Electronic Engineering, the Chinese University of Hong Kong, Hong Kong, China

ABSTRACT

Traditionally, a simultaneous recognition process using the same feature set of a spoken utterance is used to classify the emotional state of the speaker in addition to its content. However, an analysis on the classification performance for every pair of emotions shows that different features have distinctive classification abilities for different emotions. Therefore, we propose an efficient emotion recognition process called cascade bisection (CB-process), which carries out emotion recognition by means of several bisecting steps and applies different feature sets for every step. This process is based on the features' different abilities of classifying emotions. Through this, we can fully utilize the information extracted from features and achieve a better recognition performance. Five discrete emotional states, namely, neutral, anger, fear, joy, and sadness are distinguished from the input Mandarin speech. After extracting the acoustic features that contain information on short-time energy (amplitude), signal amplitude, and pitch, we derive the representation feature set for further use in the CB-process, which achieves better emotion recognition as demonstrated seen from the experimental results.

Index Terms: Mandarin, automatic emotion recognition, cascade bisection process (CB-process)

1. INTRODUCTION

The goal of emotion recognition is to determine the emotional states of a particular speaker from the uttered speech samples. It is a challenging research topic and has received a lot of attention recently. While different approaches have been proposed [1~6] to tackle this problem, there exists no satisfactory solution yet. Kwon [1] adopted a data-driven approach and retried on two different kinds of databases: the text-independent SUSAS database and the speaker-independent AIBO database. Schuller [2] used acoustic features and several classification methods such as linear classifiers, Gaussian Mixture Models, Neural Nets, and Support Vector Machines to differentiate different emotional states. Lee [3], on the other hand, used three different techniques, namely, liner discriminant classifier (LDC), k-nearest neighborhood (k-NN) classifier, and support vector machine (SVM) classifier for classification purposes.

In this paper, we carry out the classification of various emotion behaviors based on a set of comprehensive acoustic information. However, instead of analyzing all the emotions simultaneously as traditional methods have done in the past, our study is focused on analyzing the feature performance for classifying every two of the five emotions. Through this method, more information on emotions is extracted, as it helps build up the Cascade Bisecting process of the decision tree theory. The experimental results show that the proposed Cascade Bisecting process can fully utilize the different information on emotions, as well as the feature sets for classifying them. As there is no convention yet, the focus emotional states are the following four emotional behaviors: anger, fear, joy, and sadness. This set is often supplemented by a neutral state for dissociation from a non-emotional state. This selection also offers a certain degree of international comparability [5].

2. EMOTIONAL SPEECH DATA

Currently, it is difficult to find a common set of emotional speech data, particularly Chinese speech data [7]. We collected the emotional speech signals for experiments in an acoustically isolated room, and we also designed the speech text materials ourselves. The speech data was recorded at a sampling rate of 16 kHz with a 16-bit resolution. There are 20 sentences (uttered three times each) for every emotion. Out of the 20 sentences, there are five common ones, while the other 15 are respectively of an individual emotion. Four males and four females participated in the data collection process.

In order to test the validity of these data, all of the sound clips were played randomly. The listeners (not the speakers) decided how strong the emotion was in each utterance based on their subjective judgment using three levels: very obvious, obvious, and not obvious. Those utterances that were considered to possess different emotional levels were grouped.

3. FEATURES

We computed 28 acoustic [2][4][5][8] correlates including those automatic computing prosodic information of the speech signal. The chosen features comprised utterance-level statistics corresponding to pitch (F0), short-time energy, and short-time amplitude. Frame-based speech signal with duration 20 ms was analyzed for every 10 ms interval using a Hamming window. In addition, we also directly calculated the signal amplitude as signal strength parameters. The following 28 feature coefficients were obtained:

Pitch: (7)

1. – 5. Mean, Median, STD, Range, and Inter-quartile range of pitch.

6. – 7. STD and Inter-quartile range of pitch gradient.

Short-time energy (db): (2)

8. – 9. Mean and STD of energy reversal points' positions.

Short-time amplitude: (11)

10. – 11. STD and Range of amplitude.

12. Mean of fall-time amplitude.

13. – 14. STD and Inter-quartile range of amplitude gradient.

15. STD of the second gradient of amplitude.

16. – 20. Mean, Median, STD, Range, and Inter-quartile range of amplitude reversal points' positions.

Short-time amplitude for Pitch Frames*: (4)

Only the amplitude of pitch frames was measured, those frames without pitch were assumed to have zero amplitude.

21. – 23. Mean, STD, and Range of short-time amplitude*

24. STD of short-time amplitude* reversal points' positions.

Signal strength: (4)

We calculated the range of each adjacent maximum and minimum point of signal.

25. – 28. Mean, Median, STD, and Maximum of signal amplitude.

4. FEATURE ANALYSIS FOR EVERY TWO EMOTIONS

To obtain more information on the difference between every two of the five emotions, we analyzed their different features' classification performances for every two of the five emotions. The K-means clustering method is applied to get the average successful clustering rate for every feature. By sorting the features from best to worse, the 10 top features are extracted as shown in the following. The data used here include two males and two females making up a total of 880 sentences within the 'very obvious level.' This set of data is also to be used for training.

Neutral vs. Anger:

Short-time amplitude and signal strength features excel others to distinguish between neutral and anger emotion. They all show higher values in anger emotion than in neutral emotion. The best one is feature 15, and its ASCR (average successful clustering rate) is 86.58%.

Neutral vs. Fear:

The reversal points' positions related features excel others to distinguish between neutral and fear emotion. They all show lower values in fear emotion than in neutral emotion. The best one is feature 20, and its ASCR is 75.72%.

Neutral vs. Joy:

Short-time amplitude and pitch features excel others in this case. The feature values in joy emotion are all higher than in neutral emotion. The best feature is No. 14, and its ASCR is 79.72%.

Neutral vs. Sadness:

It is difficult to distinguish between these two emotions. Short-time amplitude, signal amplitude, and pitch features are all shown in the 10 top features. The best feature is No. 12, and its ASCR is only 68.83%.

Anger vs. Fear:

Short-time amplitude and signal amplitude features excel others to distinguish this case. The feature values in anger emotion are all higher than those in fear emotion. The best feature is No. 28, and its ASCR is 77.00%.

Anger vs. Joy:

It is even more difficult to distinguish between these two emotions than to distinguish between Neutral and Sadness. Short-time amplitude and pitch features excel others in this case. The best feature's ASCR is only 63.00%. Moreover, feature No. 15 is the best one in the Neutral and Anger case; however, the effect is very bad.

Anger vs. Sadness:

It is comparatively easy to distinguish between these two emotions. Short-time amplitude and signal amplitude features excel others in this case as in the Neutral vs. Anger case. The best feature is No. 13, and its ASCR is 90.00%.

Fear vs. Joy:

It is also difficult to distinguish between these two emotions. Signal amplitude and Reversal points' positions related features excel others in this case. The best feature is No. 28 as in the Anger vs. Fear case, and its ASCR is only 69.25%.

Fear vs. Sadness:

It is comparatively easy to distinguish between these two emotions because of the Reversal points' positions related features. The best feature is No. 18, and its ASCR is 89.00%.

Joy vs. Sadness:

Short-time amplitude and pitch features excel others in this case. The feature values in joy emotion are all higher than in sadness emotion as in the Neutral vs. Joy case. The best feature is also No. 14, and its ASCR is 82.00%.

Based on the analysis above, the features have different abilities to distinguish between two different emotions. Short-time amplitude features excel others to distinguish 'Neutral and Sadness' vs. 'Anger and Joy.' Signal amplitude features excel others in the case of Anger vs. Fear, and Anger vs. 'Neutral and Sadness.' Pitch features have good ability in the case of Joy vs. 'Neutral and Sadness,' while reversal points' positions related features excel others in the case of Fear vs. 'Neutral and Sadness.'

5. THE CASCADE-BISECTION PROCESS

Based on the analysis in Part 4, different features show variant abilities to distinguish between two different emotions. If this character is fully utilized, better classification results can be achieved than by using just one feature set for all emotions and classifying them simultaneously. This leads to the cascade-bisection process (CB-process) of the decision tree theory [9], which carries out the classification by several bisecting steps and different feature sets applied for every step. There are two important issues for building an effective decision tree here: (1) how to build the tree structure by the node-splitting rule (classification steps), and (2) how to get the decision rules at each node (feature set for every step)

5.1 The node-splitting rule

In our problem, the node-splitting rule is based on the relationship of emotions. *Table 1* is the average successful

clustering rate of the 10 top features between every two of the five emotions.

	Anger	Fear	Joy	Sadness
Neutral	81.02	73.03	73.07	**66.37**
Anger	—	70.35	**60.77**	86.42
Fear		—	64.27	84.72
Joy			—	79.00

Table 1: The average successful clustering rate (%) of the 10 best features between every two of the five emotions

We can see that the recognition rate between anger and joy is the smallest- 60.77%, which means that anger and joy are closer in some way than other emotions. Similarly in the other three, neutral and sadness are closer (66.37%). The sum of the recognition rate between fear and group 'anger and joy' is 70.35% and 64.27%, while between fear and group 'neutral and sadness,' it is 73.03% and 84.72%. Therefore, fear is closer to group 'anger and joy' than 'neutral and sadness.' We now arrive at the tree structure, as shown in *Figure 1*.

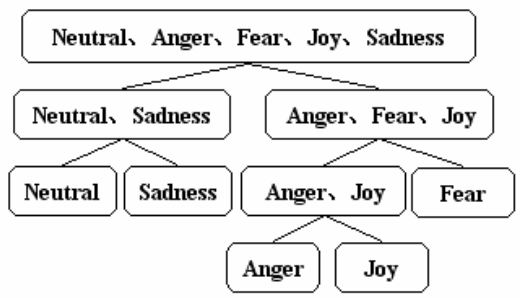

Figure 1: tree structure

5.2 The decision rule at each node

Notice that the decision tree is applied to our problem with the aim of fully utilizing the features. Thus, different features are chosen for each node based on the feature performance of classifying every two of the five emotions described in Part 4. First, we calculated the correlation of the features and grouped those with a correlation larger than 0.9. At most, one feature in the same group is selected to be part of the feature set. According to the results of the average successful clustering rate and still the varying classification performance of the features for different emotions, we got the feature sets at all nodes.

Table 2 shows the feature sets used at each node. As described in Part 4, pitch, short-time amplitude, and the reversal points' positions related features composed the feature set to distinguish between 'N+S vs. A+J+F.' Signal amplitude related features cannot be selected into the feature set because Fear is similar to 'Neutral and Sadness' using these features. The reversal points' positions related features are used to distinguish fear from 'Neutral and Sadness,' but they are not used in other three nodes. Even if we used the same kind of features at the 'A+J vs. F,' the 'N vs. S,' and the 'A vs. J' nodes, the feature sets would be quite different. In this way, the features are fully utilized during the classification process.

	The Feature Set
N+S vs. A+J+F	3+4+6+7+8+11+12+18
A+J vs. F	2+6+8+10+13+21+25+28
N vs. S	1+5+10+21+25
A vs. J	1+3+4+5+7+13+15+23+25

*In this paper, N- Neutral, A-Anger, F-Fear, J-Joy, S-Sadness

Table 2: Feature sets used in each node

6. EXPERIMENTS

Based on the proposed decision tree structure and the decision rule at each node, we build up the cascade bisection process (CB-process). In the CB-process, three steps at most are required in order to classify a particular emotion. For example, X is the feature vector of the speech signal for recognition. In the first step, the features in X are selected and ranked as the feature set of 'N+S vs. A+J+F' in *Table 2*, and come into the classification model of 'N+S vs. A+J+F' which had been trained. If the classification result is 'A+J+F,' X is classified into 'A+J' or 'F' in the second step, and the feature set of 'A+J vs. F' in *Table 2* is used. If the result is 'F,' then we get the final result. Otherwise, we continue the steps to classify X into 'A' or 'J' and work until we obtain the final result. The process of the other classification branch from 'N+S' is similar.

	Recognition Rate (%)					
	N	A	F	J	S	Average
T-KNN	78	35	42	40	20	43.00
CB-process*	73	40	37	40	45	47.00
CB-process	77	40	55	25	50	49.40

Table 3: RR (Recognition Rate) compared with different methods using test 1 data

	Recognition Rate (%)					
	N	A	F	J	S	Average
T-KNN	78	29	39	44	56	49.20
CB-process*	73	32	35	47	64	50.02
CB-process	77	42	66	41	61	57.40

Table 4: RR compared with different methods using test 2 data

The recognition rate of the five emotions in the CB-process is compared with the traditional method (T-KNN) and the CB-process* (a quasi CB-process which uses the same separation order with the CB-process, but at the same time uses the same feature set for every step), as shown in *Table 3* and *Table 4*. This is to show the better performance brought about by the CB-process because of the idea of cascade bisection and the utilization of different feature sets. Here, the classification method is the K-nearest neighbor decision rule (K-NN, K=7). We composed different kinds of testing data. One is the testing data (test1) from another set of male and female speech data in the 'very obvious' level, which makes up a total of 280 sentences. Another is the testing data (test2) from three males and three females, which include two different persons with training data in the obvious level for a total of 440 sentences.

When we use the traditional method to classify five emotions, the classification model has five classes, and each class corresponds to one emotion. Then the recognizing vector will be classified into one class (one emotion) by the majority rule. In this process, only one feature set (1+4+6+7+13+25+26 derived using the same method as shown in Part 5.2) is used for classification. We can see that the average recognition rates for the CB-process are better than those for the traditional method and the CB-process*, and the average recognition rates for the CB-process* are better than that for the traditional method. The process of the CB-process* is built based on the relationship of five emotions and the analysis between two of the five emotions as shown in Part 5.1. Hence, the average recognition rates for the CB-process* are better than those for the traditional method, even though both used the same feature set. As for the CB-process, it used different feature sets in different steps. Hence, the average recognition rates are better than the CB-process* only because of the full utilization of the feature set in each step.

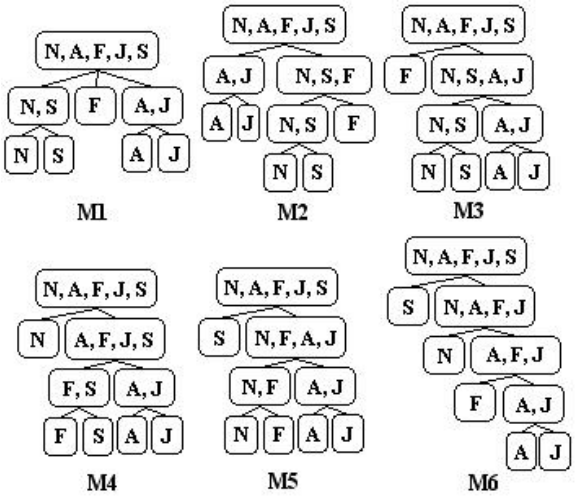

Figure 2: Different separated orders

In *Figure 2*, we give six different kinds of cascade section processes and their experimental results in *Table 5*. M1* is not a cascade 'bisection' process. However it first separates the five emotions into three groups: 'N, S', 'F', and 'A, J' using only one feature set. We use different feature sets in different steps with the CB-process result for the purpose of making a comparison. The CB-process has better result because of its sufficient utilization of the feature sets (in the CB-process, we need to use two steps to get three groups, and each step has its own feature set), although it has more steps than M1*. As for the other processes (M2 ~ M6), they meet the characters of the cascade bisection process, while the separation orders used are different from the optimal one. In order to eliminate the effect of the feature sets, we used here the same feature set in different steps and then compare it with the CB-process* result. The experiment's results show that all of their recognition rates are not good. We see that the separation order is crucial for the recognition results, and that good results can only be obtained based on the research of the recognition rates between two of the five emotions.

	Average Recognition Rate (%)					
	M1*	M2	M3	M4	M5	M6
Test1 data	48	46	45	43	41	41
Test2 data	53	49	49	49	47	48

Table 5: Average RR in different separated orders

7. CONCLUSION

This paper proposed the Cascade Bisection process of the decision tree theory for automatic emotion classification by speech signals. For every two of the five emotions, the features show different types of classification performance. This information is noted and applied when building the proposed process that can fully exploit the classification ability of the emotions' features. The experiments with emotional speech data showed that the CB-process gave a better emotion recognition rate not only by means of the cascading bisection structure, but also by different feature sets used at every step.

The Cascade Bisection process is a promising tool for emotion recognition in speech. Hence, this research direction should be further explored. With regard to future work, more focus will be given to the feature sets used in different steps. In this paper, we used the K-NN as a classification method. However, in actuality, different classification methods may be used in the cascade bisection process in different steps.

8. REFERENCES

[1] Oh-Wook Kwon, etc, "Emotion Recognition by Speech Signals", in Proceedings of EUROSPEECH-2003, pp125-128.

[2] Bjorn Schuller, etc, "Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine – Belief Network Architecture", in Proceedings of ICASSP '04. Volume 1, 17-21 May 2004 Page(s): I - 577-80 vol.1

[3] Chul Min Lee, etc, "Classifying Emotions in Human-Machine Spoken Dialogs", in Proceedings of ICME '02. Volume 1, 26-29 Aug. 2002: pp737 - 740

[4] Oudeyer Pierre-Yves, "The production and recognition of emotions in speech: features and algorithms", Int. J. Human-Computer Studies 59 (2003) 157-183

[5] Valery A. Petrushin, "Emotion Recognition in Speech Signal: Experimental Study, Development, and Application", In Proceedings of ICSLP-2000, vol.2, pp222-225

[6] Tsang-Long Pao, etc, "Detecting Emotions in Mandarin Speech", Computational Linguistics and Chinese Language Processing, Vol. 10, No. 3, Sep. 2005, pp347-362

[7] N. Campbell, "Databases of Emotional Speech," In: Proc. of ISCA Workshop on Speech and Emotion, 2000.

[8] Yuan, J.; etc, "The Acoustic Realization of Anger, Fear, Joy and Sadness in Chinese," Proceedings of ICSLP 2002, Denver, Colorado, pp. 2025-2028.

[9] Sankar K. Pal, Amita Pal, "Pattern Recognition-From Classical to Modern Approaches", World Scientific Publishing, 2001, pp169 – 184