Emotion Recognition Using Voice Based on Emotion-Sensitive Frequency Ranges

Kyung Hak Hyun, Eun Ho Kim, Yoon Keun Kwak

Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
cromno9@kaist.ac.kr, kimeunho@kaist.ac.kr, ykkwak@kaist.ac.kr

Abstract To date, study on emotion recognition has focused on detecting the values of pitch, formant, or cepstrum from the variation of speech according to changing emotions. However, the values of emotional speech features vary by not only emotions but also speakers. Because each speaker has unique frequency characteristics, it is difficult to apply the same manner to different speakers. Therefore, in the present work we considered the personal characteristics of speech. To this end, we analyzed the frequency characteristics for a user and chose the frequency ranges that are sensitive to variation of emotion. From these results, we designed a personal filter bank and extracted emotional speech features using this filter bank. This method showed about 90% recognition rate although there are differences among individuals.

Keywords: emotion, recognition, frequency range, filter

1. Introduction

Emotional speech recognition involves automatically identifying the emotional or physical state of a human being from voice. The importance of emotion recognition for human computer interaction is widely recognized [9]. Although the emotional state does not alter linguistic content, it is an important factor in human communication and it also provides feedback information in many applications. In human-machine interactions, the machine can be made to produce more appropriate responses if the state of

emotion of the person is accurately identified. Other applications of an automatic emotion recognition system include tutoring, alerting, and entertainment [3].

1.1 Previous Works

The first investigations for human emotions were conducted in the mid-1980s using statistical properties of certain acoustic features [2, 13]. In 2000, emotional speech recognition was employed by therapists as a diagnostic tool in medicine [5]. Presently, most researches are focused on finding powerful combinations of classifiers that advance the classification efficiency in real-life applications. As an example, ticket reservations, so-called "SmartKom", can recognize customer frustration and change their response accordingly [1, 12].

A correlation of emotion and frequency was usually used for emotional speech features but the feature characteristic was changed a lot by different speaker. Hence, the results of emotion recognition in speaker independent systems are below 60% [7, 4].

Therefore, in this study, we analyzed the frequency characteristics of speakers and applied a personal filter bank which was proposed in [6].

2 Emotional Speech Database

The Korean database was used in our experiment. The database was recorded in the framework of the G7 project in Korea. In this database emotional sentences from five male speakers and five female speakers were recorded.

2.1 Corpus of Database

Each corpus contains several phonemes. The database contains 45 sentences. The corpora contain short, medium, and long sentences that are context independent. The sentences are chosen upon consideration of the followings:

1. The sentence is able to pronounce in several emotional states.
2. The sentence can express the emotions naturally.
3. The database should contain all phonemes of Korean.
4. The database should contain several dictions such as honorific words.

2.2 Speech Materials

The corpora were recorded over four basic emotions. The recorded emotions are neutrality (N), joy (J), sadness (S), and anger (A).

The database contains speech recorded in three iterations. All speech was recorded four times and the worst record among those was discarded. The aim of this step is to filter clumsy wording and maintain consistency of the speech.

The database contains 5400 sentences. The subjective evaluation tests for all corpora were performed.

2.3 Subjective Evaluation Test

Subjective evaluation tests were made for the database. The subjective evaluation test included 30 listeners. The listeners were engineering students from Yonsei University in Korea. Each listener decided which emotion corresponded to each utterance. The samples were played randomly. 10 listeners made decisions for each utterance. The results of the subjective evaluation test showed, on average, 78.2% accuracy. The confusion matrix of the test is shown in table 1.

Table 1. Human Performance

Recog. (%)	Neutrality	Joy	Sadness	Anger
Neutrality	83.9	3.1	8.9	4.1
Joy	26.6	57.8	3.5	12.0
Sadness	6.4	0.6	92.2	0.8
Anger	15.1	5.4	1.0	78.5
Overall			78.2	

4 Personal Filter Bank Design

4.1 Log Frequency Power Coefficients

Humans perceive audible sound from 20 Hz to 20 kHz. Furthermore, human perception is not linear to physical frequency, and hence there is a different unit for audible sound frequency, that is, Mel [10].

The human auditory system has several sensitive frequency ranges. Under 1,000 Hz, humans feel the pitch is linear to physical frequency; however, when the frequency is increased above 1,000 Hz, the auditory system

becomes insensitive. In other words, as frequency changes become larger, humans can distinguish changes of pitch.

Accordingly, we can regard the human audible system as a filter bank, as described in [11]. Therefore, a log frequency filter bank can be regarded as a model that follows the varying auditory resolving power of the human ear for various frequencies.

In Tin Lay Nwe's study [8], a filter bank is designed to divide speech signals into 12 frequency bands that match the critical perceptual bands of the human ear. The center frequencies and bandwidths of 12 filter banks, which were proposed by Tin Lay Nwe, are described in [8].

4.2 Filter Bank Analysis

In the analysis of frequency characteristics of a user, we attempted to identify specific frequency ranges that are sensitive to changes of emotion and robust to changes of context. To this end, we compared the rates of emotional speech recognition by changing the filter bank.

The results are presented in figure 1 and figure 2. Both show normalized values on the y-axis in a range between 0 and 1 to verify the effect of each filter. In figure 1, the recognition rates are for speech that was filtered by a one order filter. Hence, we can estimate which filter is most useful with respect to recognizing the emotion. Figure 2 shows the results for speech that was filtered by an eleven order filter bank. Thus, we can identify the worst filter.

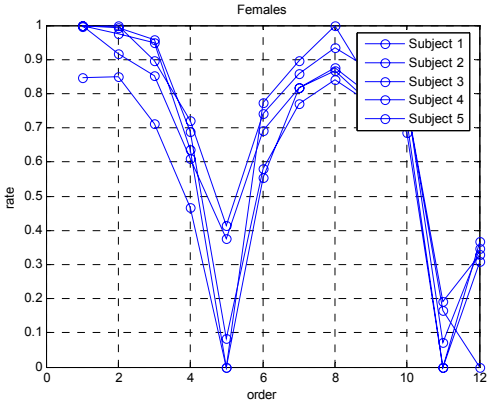

Fig. 1. One filter selection in female subjects

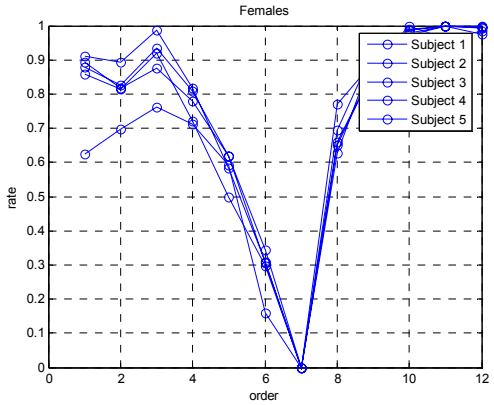

Fig. 2. One filter exception in female subjects

It is possible to select the best filter or abandon the worst filter on the basis of the above results. In this regard, the results demonstrate that accepting or rejecting a specific filter from the filter bank can affect the emotional recognition results.

However, they do not indicate the best set of filters to recognize the emotion, because each filter of the filter bank is not independent of the other filters. According to the study of Rabiner and Juang [11] and a hypothesis of Tin Lay Nwe, each filter shares frequency ranges with its neighbor filters. Therefore, a filter cannot be chosen independently from the filter bank and should be chosen under consideration of the effects of other filters.

4.3 Design of the Personal Filter Bank

4.3.1 Design Method

From the analysis presented above, we find that there are different frequency ranges that can help in the recognition of emotion in speech. Hence, we compared the results obtained using a full order filter bank with those derived using a selected filter set.

The full order filter bank was selected from Tin Lay Nwe's study [8]. For comparison, we assessed several filter sets in order to identify the best set in terms of performance. At this step the full order is twelve and each filter can be selected independently. Thus, it is necessary to compare 4095 cases (2^{12}-1) for one person.

Finally, we examined the results obtained using a Bayesian Classifier and chose the best set for each person. A summary of this procedure is presented in figure 3. In this figure, full order means the classifier uses the full information from filters 1~12, whereas order choicer means the classifier adaptively uses the information among the filters.

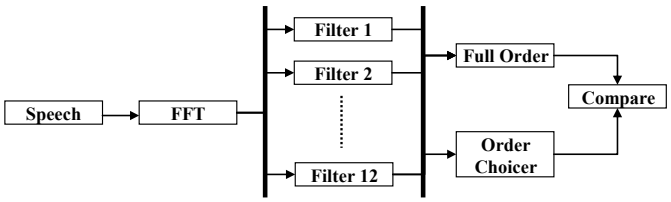

Fig. 3. Flow chart of the adaptive filter bank method

4.3.2 Evaluation

To evaluate the designed filter bank, we compared the results of the designed filter bank with those of the full-order filter bank. We performed 100 iterations for the designed filter bank for cross-validation in order to verify the results statistically. For the cross-validation, the training and test data was randomly chosen from the database; 120 training data were chosen randomly and 300 test data were chosen not only randomly but also differently from the training data for the evaluation.

We found that the designed filter bank method sometimes decreased the recognition rate. However, the degree of decrease was very small. From table 2, however, 80~90% of cases showed an increase in the recognition rate for 100 repetitions of the experiment for each subject.

Table 2. Adaptive filter orders and the improvement rate

Subject	Adaptive filter order	Improvement (%)
Female 1	3,4,5,7,9,10,11,12	97
Female 2	2,5,6,7,8,9,10,11,12	93
Female 3	4,6,7,9,11,12	98
Female 4	1,4,5,6,8,10,11	94
Female 5	3,5,6,7,8,9,10,11,12	84

Finally, the recognition rates of all subjects are presented in figure 4. In addition, we present a comparison with the principal component analysis (PCA) results. Hence, we find that the proposed method is superior to the PCA method, as the latter can only find the principal component, which is

good for representation, whereas the proposed method finds the frequency ranges that are sensitive to change of emotion.

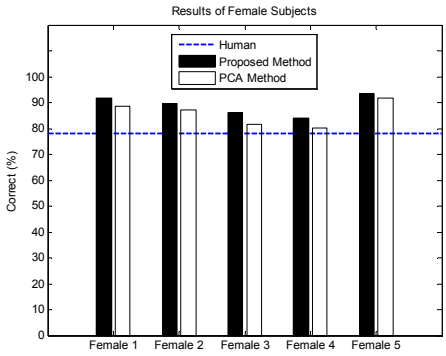

Fig. 4. Result of Female Subjects

4.3.3 Discussion

Although proposed method is available with learning data for each subject and collecting user's speech data with emotion is not easy, this works contributed that each person has different frequency characteristics in emotional speech and some frequency ranges is not important to classify emotions. In proposed method, we analyzed each subject's frequency characteristics by filter bank method and designed personal filter bank for emotion recognition. As presented in figure 4, proposed method shows better result than PCA method and human performance. In simulation, we assumed clean environment because all speech data were loaded from database, however, the noise problem is serious in real application. Proposed method generally uses high frequency regions which are easily distorted by noise. Therefore noise reduction process will be needed before applying proposed method to attain the desired results.

5 Conclusion

In this work, we consider the correlation of speech frequency with emotion and develop a personal filter bank for emotion recognition. We verified the superiority of the proposed method through comparison with PCA results. The strength of the proposed approach lies in selecting frequency ranges that are sensitive to the change of emotion. Although the improve-

ment varied on a case by case and person to person basis, a roughly 2~5% overall improvement in recognition rate was obtained.

References

1. Ang J, Dhillon R, Krupski A, Shriberg E, Stolcke A (2002) Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In: Hansen JHL, Pellom B (eds) International Conference on Spoken Language Processing, ISCA Archive, pp 2037-2040
2. Bezooijen Rv (1984) The Characteristics and Recognizability of Vocal Expression of Emotions. Walter de Gruyter, Inc., The Netherlands
3. Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, Taylor JG (2001) Emotion recognition in human-computer interaction. In: Chang S-F, Schneiderman S, (eds) IEEE Signal Processing Magazine, IEEE Signal Processing Society, pp 32-80
4. Esau N, Kleinjohann B, Kleinjohann L, Stichling D (2003) MEXI: Machine with Emotionally eXtended Intelligence. In: Abraham A, Köppen M, Franke K (eds) Hybrid Intelligent systems, Design and Application, IOS Press, The Netherlands pp 961-970
5. France DJ, Shivavi RG, Silverman S, Silverman M, Wilkes M (2000) Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans Biomed Eng 7:829-837
6. Hyun KH, Kim EH, Kwak YK (2006) Emotion Recognition Using Frequency Ranges Sensitive to Emotion In: Proceeding of 3rd International Conference on Autonomous Robots and Agents, pp 119-124
7. McGilloway S, Cowie R, Douglas-Cowie E, Gielen S, Westerdijk M, Stroeve S (2000) Approaching automatic recognition of emotion from voice: A rough benchmark. In: Proceeding of the ISCA Workshop on Speech and Emotion, pp 207-212
8. Nwe TL, Foo SW, De Silva LC (2003) Speech emotion recognition using hidden markov model. Speech Communication, 41:603-623.
9. Pantic M, Rothkrantz L (2003) Toward an affect-sensitive multimodal human-computer interaction. In: Trew RJ, Calder J (eds) Proceeding of the IEEE, IEEE, pp 1370-1390
10. Quatieri TF (2002) Discrete-Time Speech Signal Processing Principles and Practice, Prentice Hall, New Jersey.
11. Rabiner LR, Juang BH (1993) Fundamentals of Speech Recognition, Prentice Hall, Englewood Cliffs, New Jersey
12. Schiel F, Steininger S, Turk U (2002) The Smartkom multimodal copus at BAS. In: The 3rd International Conference on Language Resources and Evaluation, pp 35-41
13. Tolkmitt FJ, Scherer KR (1986) Effect of experimentally induced stress on vocal parameters. J Exp Psychol: Hum Percept Perform 12:302-313