Physiologically-inspired Feature Extraction for Emotion Recognition

Yu Zhou [1] *, Yanqing Sun* [1]*, Junfeng Li* [2]*, Jianping Zhang* [1]*,Yonghong Yan* [1]

[1]Institute of Acoustics, Chinese Academy of Sciences
[2]School of Information Science, Japan Advanced Institute of Science and Technology
zhouyu@hccl.ioa.ac.cn, junfeng@jaist.ac.jp

Abstract

In this paper, we proposed a new feature extraction method for emotion recognition based on the knowledge of the emotion production mechanism in physiology. It was reported by physiacoustist that emotional speech is differently encoded from the normal speech in terms of articulation organs and that emotion information in speech is concentrated in different frequencies caused by the different movements of organs [4]. To apply these findings, in this paper, we first quantified the distribution of speech emotion information along with each frequency band by exploiting the Fisher's F-Ratio and mutual information techniques, and then proposed a non-uniform sub-band processing method which is able to extract and emphasize the emotion features in speech. These extracted features are finally applied to emotional recognition. Experimental results in speech emotion recognition showed that the extracted features using our proposed non-uniform sub-band processing outperform the traditional (MFCC) features, and the average error reduction rate amounts to 16.8% for speech emotion recognition.

Index Terms: speech emotion recognition, feature extraction, non-uniform sub-band.

1. Introduction

Speech emotion recognition has got much attention during the last few years [5]. For speech emotion recognition, one of the most important problems is the analysis and extraction of emotion information, i.e., emotional feature extraction of speech.

There have been plenty of studies on speech emotion recognition. Most of them used prosodic information as their feature parameters [1] [2]. It is commonly thought that the prosodic features of speech contain useful information for discriminating emotions. However the recognition rate is low when using the prosodic information as emotional features [2] [6]. The previous study has shown that the recognition accuracy of 5 emotions (anger, happy, sad, neutral, bored) is 42.6% when the extracted prosodic features are derived from pitch, loudness, duration, and quality features from a 400-utterance database [2]. One reason is that there are less independent components in prosodic features of speech than in phonetic features of speech. For example, 12-16 dimensional MFCC have been used as effective phonetic features for speech recognition. If even a small amount of useful information is kept in phonetic features, the accuracy of emotion recognition can be improved by increasing the number of independent phonetic features [3]. The phonetic feature MFCC, which is often used in speech recognition and speaker identification, has been used as an effective feature for speech emotion recognition [1]. However, the features used for speech emotion recognition should emphasize emotion information while attenuating the speech and speaker information, which is different from speech recognition and speaker identification. For speech recognition, the features should emphasize speech information, while the features used for speaker identification should emphasize speaker information [7]. This difference suggests that MFCC may not meet the requirements for speech emotion recognition. We need to increase the amount of emotion information in phonetic feature. The effect of emotions on the vocal tract shaping during speech production has been investigated [4], and it was found that the tongue tip undergoes an emotion-specific shaping. However, the results have not been utilized in emotion recognition. We used the results to increase emotion information in feature extraction.

In this study, we proposed a new feature extraction method which emphasizes emotion information for speech emotion recognition based on the knowledge of the emotion production mechanism in physiology. we first quantified the distribution of speech emotion information along with each frequency band by exploiting the Fisher's F-Ratio and mutual information techniques, and then proposed a non-uniform sub-band processing method which is able to extract and emphasize the emotion features in speech. These extracted features are finally applied to emotional recognition. At last, Comparisons were made between the proposed feature and MFCC for emotion recognition.

2. Emotional speech articulations

Sungbok Lee investigated speech articulations associated with emotion expression using electromagnetic articulography (EMA) data and vocal tract data [4]. These emotional speech production data were collected using an electromagnetic articulography system as well as a fast magnetic resonance imaging (MRI) technique. The data from this study showed that articulatory maneuvers are largely maintained during emotional speech production when achieving the underlying linguistic contrasts. Mainly, the range and velocity of the tongue tip movements are the primary modulation parameters associated with emotional expression. This is reasonable considering that the primary purpose in speech is to express linguistic information, while articulatory modulation which is relevant to emotion can be considered as a secondary feature. In addition, the vocal cue is one of the fundamental expressions of emotion, i.e., the glottal information contributes a lot to speech emotion expression.

3. Emotional feature extraction

As discussed in section 2, the speech emotion information are encoded in different articulation organs. That is to say, the speech emotion information is encoded in different frequency regions. To extract effective features which are inherent in emotional speech, we need to quantify the contribution of each frequency band for speech emotion recognition, by using two approaches: mutual information and F-Ratio.

3.1. Emotion information measurement based on mutual information

The mutual information of two random variables is a quantity to measure the mutual dependence of the two variables. Supposing an emotional speech feature variable and emotional speech class are X and Y, the mutual information of X and Y can be defined as:

$$I(X;Y) = H(X) + H(Y) - H(X,Y), \tag{1}$$

where $H(X)$ and $H(Y)$ are the marginal entropies, and $H(X,Y)$ is the joint entropy of X and Y, for variable X the entropy is defined as

$$H(X) = -\sum_{x \in X} p(x) \log p(x). \tag{2}$$

Also, the mutual information can be equivalently expressed as

$$I(X;Y) = H(Y) - H(Y|X). \tag{3}$$

We need to estimate the entropy and joint entropy which require estimating the probabilities of the variables. In this study, we used the histogram method to estimate the probabilities for the calculation of entropy and joint entropy. Based on this criterion, we can quantify the dependency between the feature variable at a given frequency and emotional speech class.

3.2. Emotion information measurement based on F-Ratio

To get the contribution of each frequency band to speech emotion, we used linear frequency scale triangle filters to process speech power spectrum. The output of each filter band is weighted integration of the frequency energy around the central frequency of the filter band. We adopted the Fisher's F-Ratio of each frequency band to measure the discriminating power for emotional information in each frequency band. The F-Ratio is defined as:

$$F - Ratio = \frac{\sum_{i=1}^{P}(e_i - e)^2}{\frac{1}{Q}\sum_{i=1}^{P}\sum_{j=1}^{Q}(x_i^j - e_i)^2}, \tag{4}$$

where

$$e_i = \frac{1}{Q}\sum_{j=1}^{Q} x_i^j,$$

$$e = \frac{1}{PQ}\sum_{i=1}^{P}\sum_{j=1}^{Q} x_i^j,$$

and x_i^j is the feature as sub-band spectrum of the jth sample frame of emotional state indexed i with $j = 1, 2, ..., Q$ and $i = 1, 2, ..., P$.

F-Ratio is often used to measure the discriminating power of a feature for pattern recognition. We used it to measure the emotion discriminative score in each of frequency bands.

3.3. Measurement results based on F-Ratio and mutual information

In this study, we used the CASIA Mandarin emotional corpus provided by Chinese-LDC . The corpus is designed and set up for emotion recognition studies. The database contains short utterances from four people, covering five primary emotions, namely anger, happy, surprise, neutral, and sad. Each utterance corresponds to one emotion. For each person, there are 1500 utterances, i.e., 300 utterances for each emotion. Each utterance was recorded at a sampling rate of 16 kHz.

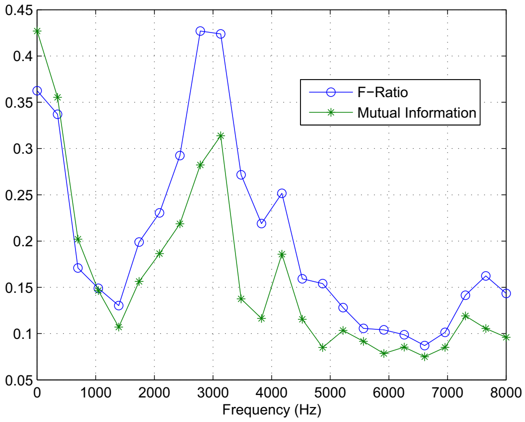

Figure 1: *Emotional speech discriminative score in frequency domain using F-Ratio and mutual information.*

The emotion discriminative ability for each frequency band (totally 24 frequency bands were used) is calculated using F-ratio and mutual information respectively. We used 1600 utterances from the material, which includes utterance from 4 persons. Each person has 100 utterances for each emotion (sad angry happy surprise), and the results are shown in Figure 1.

From Figure 1, we can see that the discriminative information is distributed non-uniformly in the frequency domain. And the peaks and valleys of the two curves are located in similar frequency regions. Based on the common characteristics of the two curves in figure 1, one can see that the discriminative information is mainly concentrated in two regions, the region of less than 300Hz and around 3000 Hz. As discussed in section 2, the range and velocity of the tongue tip movements are the primary modulation parameters associated with emotional expression, and the glottal information contributes a lot to emotion expression. The movement of the tip of the tongue concerns the third formant of vowels, which is around 3000 Hz [9]. So we believed that the region of less than 300Hz is concerned with glottal information, i.e., the fundamental frequency. The location of around 3000 Hz is concerned with tongue tip movement.

As discussed in section 2, articulatory maneuvers are largely maintained during emotional speech production when achieving the underlying linguistic contrasts. To keep linguistic messages, the discriminative information of emotion should be low, that is to say, for vowels, there should be less discriminative information at the first formant F1 and the second formant F2, which is consistent with Figure 1. This statistical result confirms the result in section 2 that emotion information is not distributed uniformly in each frequency band.

Besides the consistent characteristics of the two curves, the main difference is that in the curve estimated with mutual information, the first large peak region in the fundamental frequency region is comparatively higher than peaks in other frequency regions. In fact, the fundamental frequency should not be emphasized that much for one can change the fundamental frequency with conscious efforts. In addition, to measure the dependency between frequency and emotional speech characteristics for speech emotion recognition, it is easier to use F-

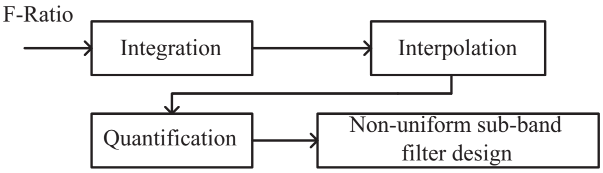

Figure 2: *Design of non-uniform sub-band filter.*

Ratio than mutual information. So in this study, we use F-Ratio to measure the dependency between frequency and emotional speech characteristics.

3.4. Non-uniform sub-band processing

To reflect the importance of each frequency region, we adopted non-uniform sub-band processing to extract the emotional feature for speech emotion recognition.

In order to utilize the contributions of these frequency bands with different amount of emotion information, we increased frequency resolution in these frequency regions with high F-Ratio values, while decreased frequency resolution in frequency regions with low F-Ratio values, and kept total band number invariable. In the design of the sub-band filters, the bandwidth of each sub-band is inversely proportional to the F-Ratio of each frequency band. By this processing, the resolution of the spectral structure around frequency regions with high F-Ratio can be highlighted. Figure 2 shows the algorithm for design of the non-uniform sub-band filters.

The idea of the design is based on calculating bandwidth according to inverse of F-Ratio. As it is hard to implement directly, a transformation method is introduced by calculating distribution function of band numbers according to F-Ratio. 24 was chosen as the number of critical bands. F-Ratio was calculated on each band. After that a cumulative summation operation was applied. By interpolation it could be mapped to the FFT domain. After employing quantification to total channel number, distribution of band number in the full range could be obtained. The bandwidth for each band could be calculated directly from the above distribution, which is shown in figure 3 and can be used to design the non-uniform sub-band filters (please note that the ranges of curves in figure 3 have been regulated, convenient for comparison, not the real value.). The curve of inverse F-Ratio in each frequency band is also shown in Figure 3. We can see that the curve of inverse F-Ratio has almost the same variation trend to bandwidth. Therefore the feature extracted with the non-uniform filters emphasizes emotion information.

3.5. Feature extraction

The proposed non-uniform frequency cepstral coefficient is referred to as NUFCC, whose processing diagram is shown in Figure 4. In the feature extracting processing of NUFCC, first, a voice activity detector (VAD) is used to cut off the silences of the speech, but the pause periods within speech sentences which contain emotional information should be kept. Then 512 point FFT is performed for each frame in which a hamming window with 25 ms frame length and 10 ms shift is employed. The designed non-uniform sub-band filters are used to integrate each frequency band to get power spectrum. After applying Logarithm, the discrete cosine transform is adopted to get 12 order cepstral coefficient vectors plus energy, which are the features used for speech emotion modeling.

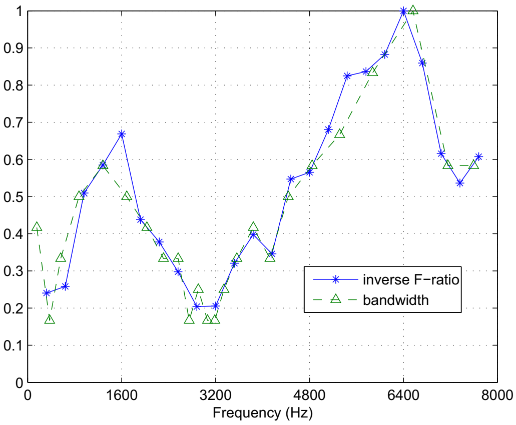

Figure 3: *Bandwidth according to inverse F-ratio.*

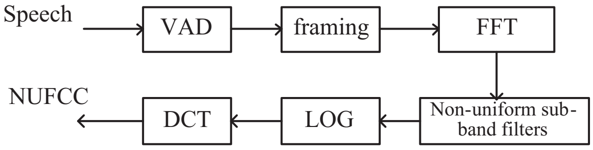

Figure 4: *Feature extraction diagram.*

For comparison, all feature sets are extracted using the framework of Figure 4, when the filter bands are Mel frequency sub-band filters, the features are MFCC. And the features are referred to as NUFCC when using the proposed non-uniform sub-band filters as filter bands. since the proposed feature emphasizes emotion information, it is believed that the proposed feature extraction method will improve the emotion recognition performance. In the following sections, we test the features which are extracted using the non-uniform filter bands by performing experiment.

4. Emotion recognition experiment

4.1. Experiment

We conducted speech emotion recognition experiment on the CASIA Mandarin emotional corpus provided by Chinese-LDC. We have described the corpus in detail in Section 3. 200 utterances from each emotion of each person were used for training, and the other utterances were used for evaluation.

As discussed in section 1, the range and velocity of the tongue tip movement are one of the primary factors that affect emotional speech expression [4]. To reflect the range and velocity of tongue tip movements, we extracted both the static and dynamic information of the emotional speech. 39-dimensional acoustic feature vector MFCC and NUFCC are used for speech emotion recognition separately, which include 12 MFCC or NUFCC, the normalized power as well as their first and second order derivatives. 512 mixture GMM were trained to model each emotion based on the above 39-dimensional features.

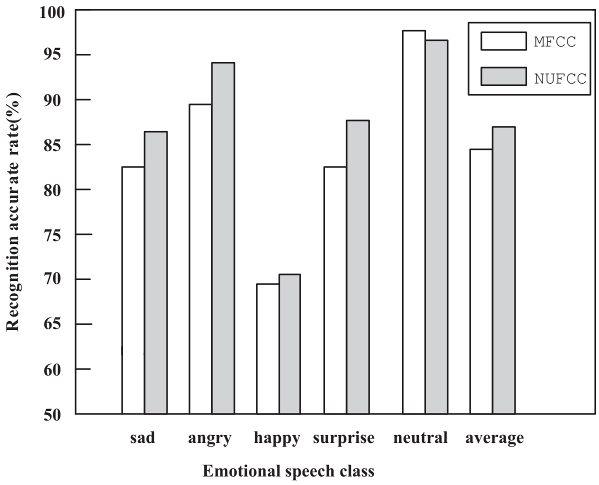

Figure 5: *Recognition accurate rate.*

4.2. Results

The contrast recognition results for each emotion are plotted in figure 5 for the two feature sets, and the error reduction rate for NUFCC over MFCC is shown in table 1.

According to the experimental results shown in figure 5 and table 1, we can see that the proposed 24-band non-uniform frequency feature (NUFCC) performs better than MFCC for speech emotion recognition. The recognition accurate rate increased for sad angry happy surprise when using the proposed feature, with an error reduction rate of 16.8% on average.

The recognition accurate rate for neutral speech decreased when using NUFCC as feature, since the speech emotion models are trained using the features extracted with non-uniform sub-band. This result suggests that the proposed feature extraction method could improve speech emotion recognition, while MFCC performs better in neutral speech recognition.

Table 1: *Error reduction rate (ERR).*

emotion	ERR(%)
sad	21.9
angry	42.9
happy	2.64
surprise	29.5
neutral	-43.5
average	16.8

4.3. Discussions

In MFCC feature representation, the Mel frequency scale is used to get a low resolution in high frequency region and a high resolution in low frequency region. While for emotional speech, the discriminative information of emotion is distributed non-uniformly in the frequency domain. The frequency region around the fundamental frequency and 3000 Hz should have high resolution. Therefore, compared with MFCC, NUFCC is more suitable for emotion feature extraction. Emotion recognition experiment showed that the features extracted using the proposed non-uniform sub-band processing improved speech emotion recognition performance prominently.

5. Conclusion

In this paper, we proposed a new feature extraction method for emotion recognition based on the knowledge of the emotion production mechanism in physiology. Based on the results reported by physiacoustist, we first quantified the distribution of speech emotion information along with each frequency band by exploiting the Fisher's F-Ratio and mutual information techniques, and then proposed a non-uniform sub-band processing method which is able to extract and emphasize the emotion features in speech. These extracted features were finally applied to emotional recognition. Experimental results in speech emotion recognition showed that the extracted features using our proposed non-uniform sub-band processing outperform the traditional (MFCC) features, and an average error reduction rate of 16.8% for speech emotion recognition was achieved.

To further improve the emotion recognition performance, it is also necessary to use other novel types of features, like articulatory feature, which has been proven to be an effective feature for speech recognition and speaker identification. Such investigations are goals for our future work.

6. Acknowledgement

This work is partially supported by The National High Technology Research and Development Program of China (863 program, 2006AA010102), National Science & Technology Pillar Program (2008BAI50B00), MOST (973 program, 2004CB318106), National Natural Science Foundation of China (10874203, 60875014, 60535030), The China-Japan (NSFC-JSPS) Bilateral Joint Projects and the SCOPE (071705001) of Ministry of Internal Affairs and Communications (MIC), Japan.

7. References

[1] O.W. Kwon, K. Chan, J. Hao, T.W. Lee., Emotion Recognition by Speech Signals, in proceedings of 8th European conference on speech communication and technology (Eurospeech-2003), pp.125-128 Geneva, Switzerland.

[2] R.S. Tato, R. Kompe, J.M. Pardo., Emotional Space Improves Emotion Recognition, ICSLP, pp. 2029-2032,2002.

[3] Nobuo Sato and Yasunari Obuchi. Emotion Recognition using Mel-Frequency Cepstral Coefficients, Information and Media Technologies, Vol. 2, No. 3, pp.835-848, (2007).

[4] S.Lee, E. Bresch, and S. Narayanan, An exploratory study of emotional speech production using functional data analysis techniques, in Proc.7th Int. Seminar Speech Production, Ubatuba,Brazil, Dec. 2006, pp. 11-17.

[5] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, Emotion recognition in human-computer interaction, IEEE Signal Processing Magazine, vol. 18, pp. 32C80, January 2001.

[6] C.H. Park, K.S.Heo, D.W.Lee, Y.H.Joo and K.B.Sim, Emotion Recognition based on Frequency Analysis of Speech Signal, International Joumal of Fuzzy Logic and Intelligent Systems, pp. 122-126, 2002.

[7] Xugang Lu, Jianwu Dang, Physiological Feature Extraction for Text Independent Speaker Identification Using Non-uniform Subband Processing, ICASSP 2007, pp. IV-461-IV-464.

[8] Paliwal K. K. (1992) Dimensionality Reduction of The Enhanced Feature Set for the HMM-Based Speech Recognizer. Digital Signal Processing. Vol.2. 157-173.

[9] http://hyperphysics.phy-astr.gsu.edu/hbase/music/vowel.html.