Volume 6, Issue 10, October 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Language Independent Emotion Recognition in Speech Signals

Sanghamitra Mohanty
Utkal University, Bhubaneswar, Odisha,
India

Abstract: Emotion in speech is a physical expression of the brain reaction and is expressed during a speech process. In the dealings with customers be a purchaser or a patient in the hospital with busy schedule lnguage independent emotion recognition can help in managing the customer in the data analytics activities efficiently. Any speech signals carries the information about the emotions with which it is told or uttered by the speaker. Here is an attempt to recognise the emotions in the piece of speech signal spoken in any language. As features of the speech signals, the Cochleogram Model through the robust Gammatone Frequency Cepstral Coefficients(GFCC) are calculated using the ERB filters and the emotions in the speech signals are recognized successfully. The optimizations technique followed here is the BPNN. The result is satisfactory tough provoking to new chanllenges.

Key words: Emotion Recognition, Cochleogram, Gammatone Filters, GFCC,

I. INTRODUCTION

Speech is a form of expression of one's mind be in any situation. Speech can be in language but emotion is expressed depending upon the situation[1]. It is the expression of mind with respect to any action and thus the conversion of mental staus to different physical behaviour, be it sound or gesture.

Emotion, in everyday speech, is any relatively brief conscious experience characterized by intense mental activity and a high degree of pleasure or displeasure. It is often interwined with mood, temperament, personality, disposition and motivation. During Automatic Speech Recognition (ASR) noise is the major factor which needs attention as with noise the auditory feature extraction is and issue to be handled. For such situation Gammatone based feature extraction is suitable as the Gammatones are modelled with respect to Cochlear Model.

In this piece of work, an attempt has been made to recognize the emotion lying in a piece of speech signal be it of any gendre. With respect to computer analysis a speech signal when it is represented inside the computer digitally. Different speech recognition methods like Filter Bank based, MFCC based, PLP based, and may other methods have been tested earlier by different researchers. We have made a study on the auditory feature based speech recognition system as the method is able to act well in the noisy environment, which is an issue to handle during automatic speech recognition. During this speech recognition we have made a Cochleogram based Gammatone Frequency Cepstrum Coefficient and derived the emotion lying in it.

II. THEORY

Here is an attempt to recognise the emotion in the speech signal applying the psycophysical observations of the auditory periphery and this filter-bank is a standard model of cochlear filtering. This auditory feature parametric analysis is the Gammatone Frequency Cepstral Coefficient (GFCC)[2] as it is modelled on the Cochleogram unlike Spectrogram. It is a Time Frame analysis involving a bank of Gammatone filters[3,4]. The impulse response of a Gammatone filter centered at frequency f is given by

$$g(t) = at^{(n-1)}\ e^{-2\pi bt}\cos(2\pi f_c t + \phi)\quad , t \geq 0 \qquad (1)$$

where f_c is the central frequency of the filter, and ϕ is the phase which is usually set to 0. Constant *a* controls the gain and *n* is the order of the filter which is usually set to be equal to or less than 4. And b is the decay factor which is related to f_c and is given by

$$b=1.019 * 24.7 * (4.37 * f_c /1000 + 1) \qquad (2)$$

where the central frequency of a filter or channel is a measure of a central frequency between the upper and lower cutoff frequencies. It is either the arithmatic mean or geometric mean of the lower cutoff frequency and upper cutoff frequency of a band-pass system or a band-stop system.

A set of Gammatone Filters (GF) with different f_c results in forming a Gammatone filter bank. GF being derived from measured impulse response, has complete amplitude and phase information. This GF is applied to obtain the speech signal chracteristics at different frequency and the resultant is the temporal frequency representation like FFT-based short time spectral analysis. When there is an effort to simulate the human audiory behaviour of the signal, the central frequencies of the filterbank are equally distributed on the Bark Scale.

This Gammatone filters, which is otherwise known as Cochleogram is a time frequency representation of the input signal which is mimicking the components of a cochlea, the sensitive part of the human auditory system. Here the frequency is downsampled into frequency bands with Equivalent Rectangular Bandwidth (ERB)[5] scale, given by

$$ERB(f_c) = 6.23\ f_c^{\ 2} + 93.39.\ f_c + 28.52 \tag{3}$$

This is due to Moore and Glassberg in 1983 who latter in 1990 gave another linear equation

$$ERB(f_c) = 24.7.(4.37.\ f_c\ + 1) \tag{4}$$

Where f_c is in kHz and ERB(fc) is in Hz.

ERB(f_c) is a measure used in psychoacoustics, which gives an approximation to the bandwidths of the filters in human hearing, using the unrealistic but convenient simplification of modelling the filters as rectangular band-pass filters. After the Gammatone Filter bank is defined, it is applied to the raw speech signal to generate the respective cochleogram, which represents transformed raw speech signal in the time and frequency domain. The advantage of using a cochleogram over spectrogram is that the features of a cochleogram is based on ERB scale with finner resolution at low frequency than the Mel-scale used in spectrogram. Besides it allows more number of coeficients in comparision to MFCC. The Mel filter-bank for a power specrum is with 257 coefficients[6] while the GFCC is with 512 coefficients.

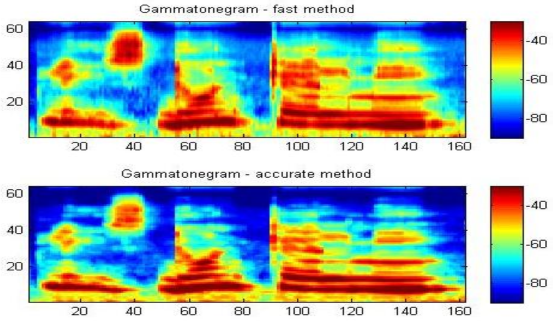


Fig. 1 Cochleogram for the speech signal sad3_mono.wav

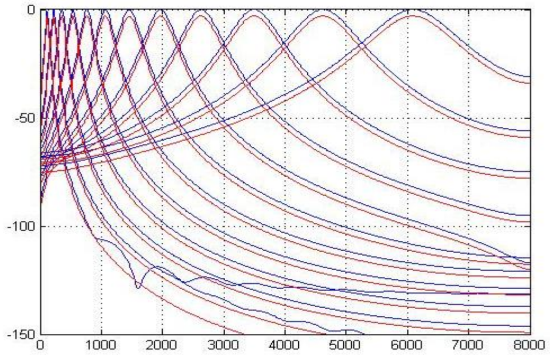
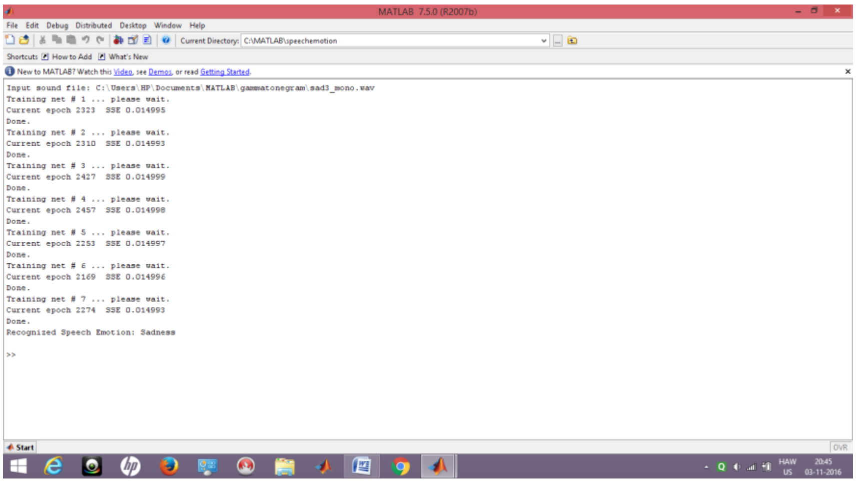


Fig. 2 Gammatone Frequency Cepstral Coefficients for the speech signal sad3_mono.wav.

III. EXPERIMENT

In this experiment we have used emotional speech database gerenated from the persons of our laboaratory, 10 persons uttering 10 sentences, with same meaning, in three langauges namely English, Odia and Hindi. The experiment is done using 24 channels, frequency lowered upto 50Hz, summation window 0.025 Secs, hop between successive windows is 0.01 secs. The ERB filter is made first and then the Filter Bank. Appling FFT the Gammatone filters are generated. The coefficients thus generated are matched with the trained dataset using Feed Forward Back Propagation Neural Network method. First training of the database is done with epoch 5000. Then the test dataset are experimented with the epoch being dynamically decided by the program. It accommodates mono as well as stereo recordings and recognizes the emotion lying in it. Result for sad emotion is shown below.


MATLAB 7.5.0 (R2007b)
File
Edit
Debug
Distributed
Desktop
Window
Help
Current Directory: C:\MATLAB\speechemotion
Shortcuts
How to Add
What's New
New to MATLAB? Watch this Video, see Demos, or read Getting Started.
Start

```
Input sound file: C:\Users\HP\Documents\MATLAB\gammatonegram\sad3_mono.wav
Training net # 1 ... please wait.
Current epoch 2323  SSE 0.014995
Done.
Training net # 2 ... please wait.
Current epoch 2310  SSE 0.014993
Done.
Training net # 3 ... please wait.
Current epoch 2427  SSE 0.014999
Done.
Training net # 4 ... please wait.
Current epoch 2457  SSE 0.014998
Done.
Training net # 5 ... please wait.
Current epoch 2253  SSE 0.014997
Done.
Training net # 6 ... please wait.
Current epoch 2169  SSE 0.014996
Done.
Training net # 7 ... please wait.
Current epoch 2274  SSE 0.014993
Done.
Recognized Speech Emotion: Sadness

>>
```

Fig. 3 Output of the emotion recognition for tere_mono.wav.

Fig. 1, Fig 2. And Fig. 3 represent the Cochleogram, GFCC and result of recognized emotion respectively for the sppech signal sad3_mono.wav. The experiment is setup with 7 layers of training network. The test is for six types of Emotions namely (1) Anger, (2) Boredom, (3) Fear (5) Joy (6) Sadness along with Normal or Neutral status are taken into consideration during the experiment. In every case it runs till the SSE value saturates with value << 1. 80% database are used for training and 20% are used for testing. The result observed are 80% for mono recordings and 60% for stereo recosrdings.

IV. CONCLUSION

The experiment is performed with MATLAB. Back Propagation Neural Network optimization technique is used for the optimization of the results during emotion recognition. More robust speech database is under developeme to make the recognition system sturdy. This piece of work has a great potential in the field of Data Analytics to analyse the emotion of the users those are not in a situation to spend time in answering the queries as per the normal Data Analytics procedure.

REFERENCES

[1] Maganti H. K. And Matassoni M. “Auditory processing-based features for improving speech recognition in adverse acoustic conditions”, EURASHIP Journal of Audio, Speech and Music Processing, (21), 2014.

[2] Darling A. M. “Properties and Implementation of Gammatone Filter: A Tutorial”, 1991.

[3] Yang Shao, Z. Jin, D. Wang and S. Srinivasan, “An Auitory Based Features For Robust Speech Recognition”, ICASSP 2009.

[4] A. Tjandra, S. Sakhi, G. Neubig, T. Toda, M. Adrian and S. Nakamura, “Combination of Two-Dimensional Cochleogram and Spectrogram Features For Deep Learning-Based ASR”, ICASSP 2015.

[5] B. C. J. Moore and B. R. Glasberg, “Suggested formulae for calculating auditory-filter bandwidths and extraction patterns” Journal of Acoustical Society of America Vol. 74 1983.

[6] Honig F. , Stemmer G. Hacker C. And Brugnara F. “ Revising Perceptual Linear Prediction (PLP)”, INTERSPEECH 2005.