International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 3, Issue 10, October 2014 **ISSN 2319 - 4847**

Emotion Recognition in Speech Using Gammatone Cepstral Coefficients

Ekta Garg[1] , Madhu Bahl[2]

[1]Mtech Student, Department of Computer Science, CEC, Landran, Mohali (India)

[2]Assistant Professor , Department of Computer Science CEC, Landran, Mohali(India)

ABSTRACT

The key challenge in speech emotion recognition system is determining emotion from noisy speech, because extraction of emotion gets complicated because of background noise. This complication leads to the mismatching between Training and Testing calculation. Thus, quality of feature extraction algorithm plays an important role. Therefore, to enhance the robustness of the system, gammatone cepstral coefficients (GTCC) is being proposed. This technique is based on gammatone filter bank which attempts to mimic the human auditory system. When GTCC is compared to Mel Frequency Cepstral Coefficients (MFCC), the time-domain GTCC provides better recognition performance under all noise as well as clean conditions. MFCC is a conventional feature extraction technique which does not perform well under noisy conditions. GTCC captures speaker's characteristics and discards irrelevant characteristics. A comparison is done between GTCC and MFCC on the basis of Signal to noise (SNR) and Mean Square Error (MSE) Parameters. Further their performance is evaluated using Feed Forward Back Propagation Neural Network which is a supervised machine learning method. Neural networks are used to make quick and accurate recognizing of emotion. Classification is done using Training phase and testing phase. From the results it is concluded that proposed GTCC algorithm enhance the performance with higher SNR and lower MSE.

Keywords:- Feature extraction, Gammatone Cepstral Coefficients, Emotion Recognition, Back Propagation Neural Network

1. INTRODUCTION

The speech signal is the fastest and the most natural method of communication between humans. The speech emotion recognition system must be able to recognize the user's emotion and act accordingly [1] .Emotions are mostly reflected in voices, on the face, in hand and body gestures. Humans mostly use emotions to express their intentions through the speech. For the computer to be able to interact with humans, it needs to understand the communication skills of humans which are achieved by the ability to understand the emotional state of a person. This requires that the machine should have the sufficient intelligence to recognize human voices which will be achieve through artificial neural networks.

1.1 Basics of Speech emotion Recognition

i) Human Speech Production

Normal human speech is produced when air is exhaled from the lungs, and the oscillating of the vocal folds modulates this air flow to a pulsed air stream, called the glottal pulses. Then this pulsed wave passes through the vocal cords and its frequency content is modified by the resonances of the vocal tract. The vocal folds and vocal tract are two important parts in speech production.

ii) Speech Features of Emotion Detection

An important issue in speech emotion recognition is the extraction of speech features that efficiently characterize the emotional content of speech and at the same time do not depend on the speaker or the lexical content. Speech features can be grouped into three categories: continuous features, qualitative features, spectral features [1].

1.2 Background

The sound event is detected with continuous audio stream recorded by a microphone. The segmented sound event is automatically classified and labeled with a predefined sound name. To carry out the part of identification, the sound signal is parameterized with a set of features. Therefore, feature extraction methods and data compaction process are included.

1.2.1 Pattern Recognition Approach

Pattern recognition is defined as the process where a received pattern or signal is assigned to one of a prescribed number of classes. A neural network performs pattern recognition by first undergoing a training session, during which the network is repeatedly presented a set of input patterns along with the category to which particular pattern belongs[2]. Later, a new pattern is presented to the network that has not been seen before, but belongs to the same population of patterns used to train the network. The network is able to identify the class of that particular pattern because of the information it has extracted from the training data.

1.3 Generic Schema of Speech Emotion Recognition System

The general methodology of audio classification involves extracting discriminatory features from the audio data and feeding them to a pattern classifier [3]. The first phase in a speech emotion recognition system is the creation of an emotional speech database. The second phase is the selection of a feature extraction technique, which extracts some information from the input speech emotional signals .Third phase is the selection of an efficient emotion recognizer, any machine learning algorithm which effectively classifies the different emotional feature vectors to appropriate emotional classes. Figure 1 shows the Structure of speech emotion recognition system.

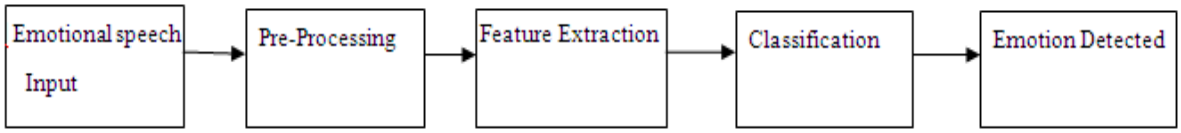

Figure 1: Structure of speech emotion recognition system

1.4 Feature Extraction Using MFCC

MFCC is based on the human peripheral auditory system. The human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency *t* measured in Hz, a subjective pitch is measured on a scale called the 'Mel Scale' [3]. The Mel-Frequency Cepstral Coefficients features is the most commonly used features in speaker recognition.

1.5 Neural Networks

A neural Network is a massively parallel- distributed processor that has a natural tendency for storing experimental knowledge and making it available for use [2]. By adjusting the weight of an artificial neuron we can obtain the output we want for a particular input. The process of adjusting the weights is known as learning [4].

The basic building blocks of the artificial neural network are [5]:

i. **Network architecture-** The manner in which the neurons of a neural network are structured is closely linked with the learning algorithm used to train the network [2].It includes Single layer Feed Forward Networks, Multi layer Feed Forward networks and Recurrent Networks.

ii. Setting the weight- The process of modifying the weights in the connections between network layers with objective of achieving the expected output is called training of network. The internal process the takes place when a network is trained is called learning [5]. It includes three types of learning which are Supervised learning, Unsupervised Learning and Reinforcement learning.

iii. Activation Function- The activation function is used to calculate the output response of a neuron. The sum of the weighted input signal is applied with an activation to obtain the response [5].

1.6 Feed Forward Back-Propagation Network

The network consists of a set of source nodes that constitute the input layer, one or more Hidden layers and an output layer. BPNN is a kind of layered feed-forward network structure which consists of a large number of neurons with nonlinear mapping ability [6]. The input signal propagates through the network in a forward direction, on a layer by layer. These neural networks are called multilayer perceptrons [2]. Error back-propagation learning consists of two passes through the different layers of the network: a forward pass and a backward pass. The purpose of error signal is to adjust the synaptic weights are to make the actual response of the network move closer to the desired response. Error is corrected using the delta rule. The adjustment made to synaptic weight of a neuron is proportional to the product of the error signal and the input signal of the synapse in question [6].Figure 2 shows the architecture of neural networks.

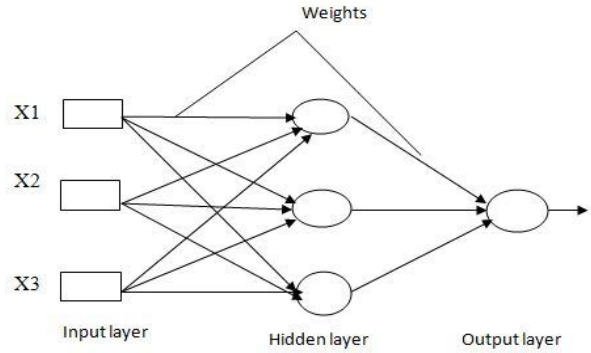

Figure 2: Architecture of 3-layer neural network

1.7 Applications of Speech Emotion Recognition

Speech emotion recognition is useful for applications which require natural man-machine interaction where the response of system depends on the detected emotion of the user. The Applications are as follows [7] :

i. Intelligent Tutoring system
ii. Robots
iii. Computer games
iv. Diagnostic tool for doctors
v. Lie Detection
vi. Call centers
vii. On-board Car systems

2. RELATED WORK

In this paper, the existing technique used for feature extraction is Mel Frequency Cepstral Coefficients (MFCC).The Literature survey related to this work is also discussed here.

2.1 MFCC Algorithm

After studying the existing technique, the comparison is done between GTCC and MFCC. MFCC algorithm is least capable of capturing speech characteristics under noisy conditions and to improve the performance gains GTCC is proposed.

2.2 Literature Survey

Unluturk et al. (2009) [8] published a paper on "Emotion Recognition using Neural Networks". In this paper, Emotion Recognition Neural Network has been developed to classify the voice signals for emotion recognition. The neural networks are also quick to respond which is a requirement as the emotion should be determined almost instantly. A common problem in existing speech emotion recognition system is determining emotion from noisy speech, which complicates the extraction of the emotion because of the background noise. To extract the emotion signatures inherent to voice signals, the back propagation-learning algorithm is used to design the emotion recognition neural network (ERNN). The results are encouraging and suggest that neural networks are potentially useful for emotion recognition.

Zhao et al. (2013) [9] proposed a paper on "Analyzing noise robustness of MFCC and GFCC features in speaker identification." In this paper, It is suggested how to enhance MFCC robustness, and further improve GFCC robustness by adopting a different time-frequency representation. Comparative Study has been made between MFCC and GTCC. It is concluded that by modifying MFCC extraction, substantial noise robustness improvement is obtained. It is observed that gammatone frequency cepstral coefficients, exhibits superior noise robustness to commonly used Mel-frequency cepstral coefficients.

3. METHODOLOGY

The methodology of work starts with the overview of feature extraction techniques and pattern classification algorithm. GTCC algorithm is used for feature extraction and Back Propagation neural network is used for pattern classification. Algorithms used for the proposed work are:

A. Gammatone Cepstral Coefficients
B. Feed forward backpropagation neural network

A. Gammatone Cepstral Coefficients

Gammatone Cepstral Coefficient is another FFT-based feature extraction technique in speech recognition systems. The technique is based on the Gammatone filter bank, which attempts to model the human auditory system as a series of overlapping band pass filters. Figure 3 shows the block diagram of GTCC.

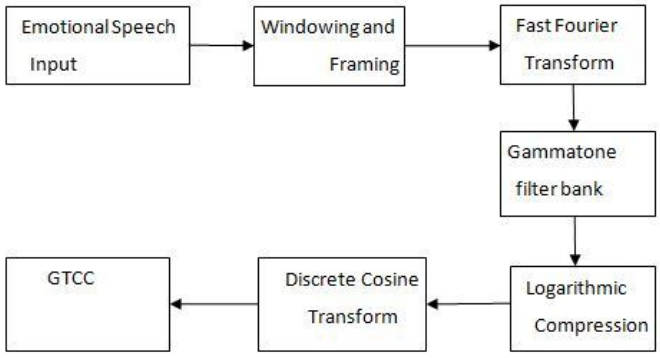

Figure 3: Block diagram of GTCC

Hamming Window

The first step of the algorithm is to subdivide a speech sequence into frames. Speech signal is non-stationary i.e. its statistical characteristics vary with time. Since the glottal system cannot change immediately, speech can be considered to be time-invariant over short segments of time (20-30 ms). Therefore speech signal is split into frames of 20ms. The

windowing function is the Hamming window which aims to reduce the spectral distortion introduced by windowing. Therefore hamming window is the preferred choice.

Fast Fourier Transform

FFT transforms the speech signal from Time domain to frequency domain. To convert each frame of N samples from time domain into frequency domain [13].

Gamma tone Filter Bank

Gammatone filter-bank is a group of filters for the cochlea simulation. The impulse response of a gammatone filter is highly similar to the magnitude charatcteristics of a human auditory filter [10]. The basilar membrane motion can be modeled with gammatone filter-bank. The impulse response of a gammatone filter is the product of a Gamma distribution and a sinusoidal tone whose center frequency is fc,The bandwidth of each filter is described by an Equivalent Rectangular Bandwidth. The ERB is a psychoacoustic measure of the width of the auditory filter at each point along the cochlea.

Logarithmic compression

The logarithm is applied to each of the filter output to simulate the human perceived loudness given certain signal intensity and to separate the excitation source produced by the vocal cords and the filter that represents the vocal tract [10].

Discrete Cosine Transform

Since the log-power spectrum is real, Discrete Cosine Transform is applied to the filter outputs which produce highly uncorrelated feature.

B. Feed forward back propagation neural network

The algorithm used for classification of emotion using feature vector are as follows

Procedure:

Step 1- Upload Voice samples for the following emotional categories like sad, happy, neutral, angry, fear.
Step 2-Apply GTCC
Step 3-Generating training data from GTCC
Step 4- Generating test data
Step 5- Initialize target set
Step 6- Generate Neural Network and set Activation Function. Activation function will produce positive numbers over the entire real number range.
Step 7- Initialize the network
Step 8-Define the network parameters. Set default values of number of epochs, goal and so on.
Step 9-Now generate testing data
Step10-Upload files for testing
Step 11-Match the actual output with desired output using error checking
Step 12-Find output of network

The graphical representation of flowchart is shown in figure 4:

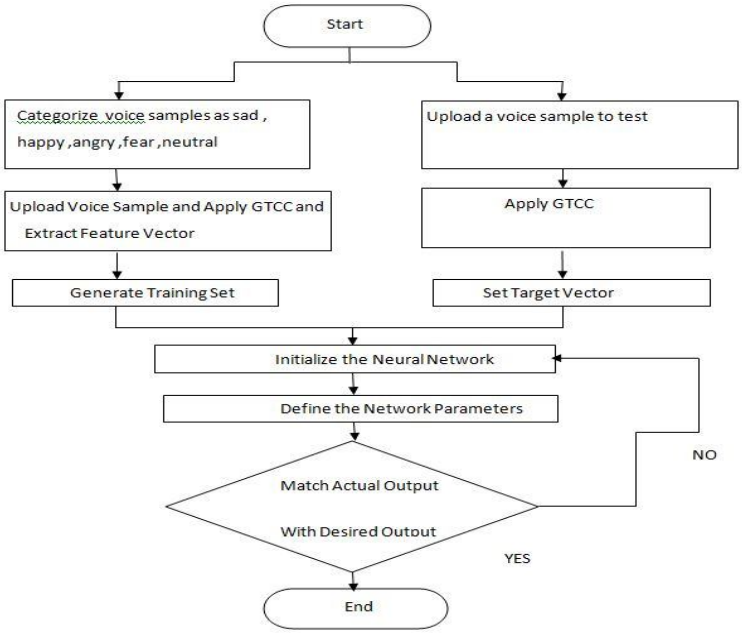

Figure 4: Flowchart of Back Propagation neural network

4. EXPERIMENTAL SET UP AND RESULTS

To see the qualitatively as well as quantitatively performance of the proposed algorithm a set of speech signals of five categories angry, fear, sad, joy and neutral have been trained. Speech files have been saved in .wav format. Figure 5 shows the accuracy of GTCC for recognized emotion using neural networks. Figure 4 (a) shows graphical representation of accuracy percentage of MFCC which is the existing technique .Figure 4(b) shows the highest accuracy percentage of GTCC than MFCC algorithm.

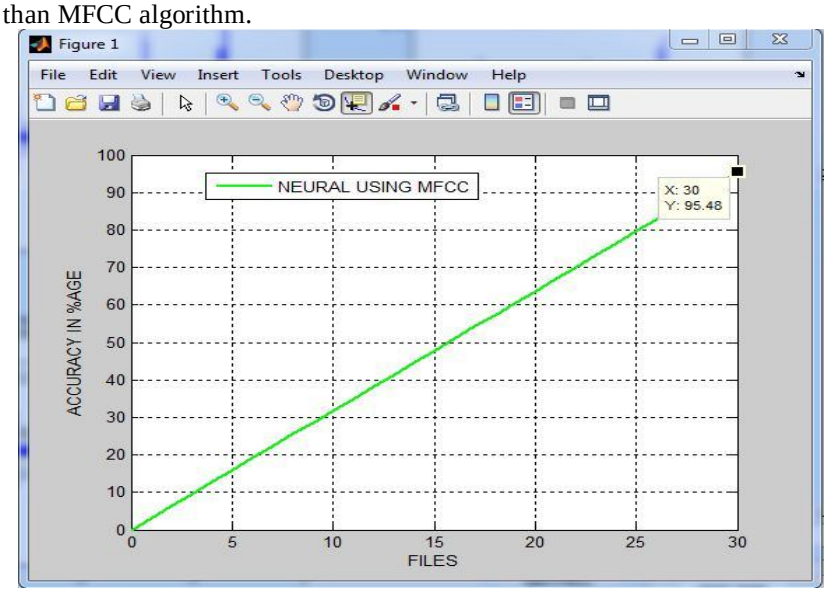

Figure 4(a): Accuracy Percentage of MFCC using Neural

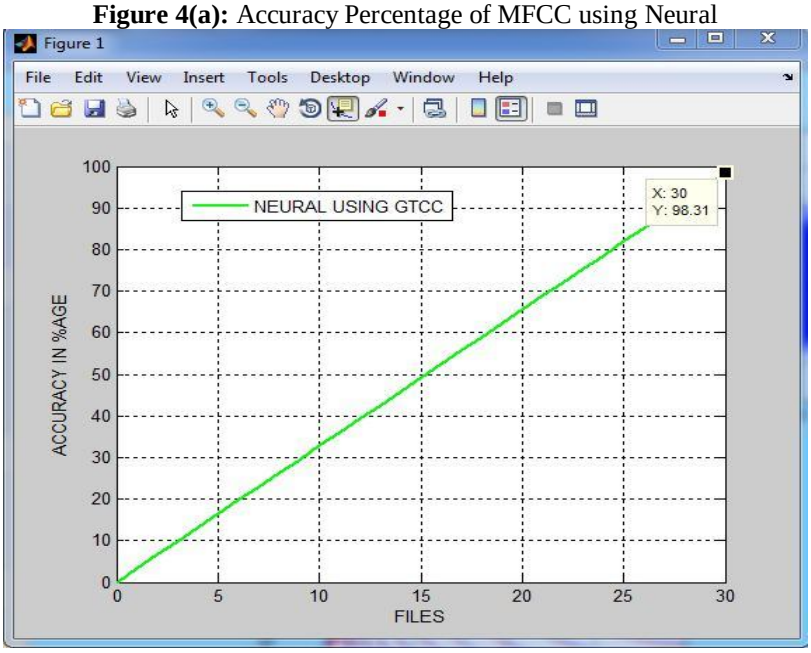

Figure 4(b): Accuracy Percentage of GTCC using Neural Network

The table 1 shows comparison between GTCC and MFCC using Parameter values. The improved result using the proposed technique has been highlighted to show the comparison with existing technique. The Parameters used SNR and MSE for proposed method is a very good improvement as compared to the existing technique. The Table 2(a) shows the accuracy percentage of particular detected emotion from voice sample using MFCC. The Table 2(b) shows the higher accuracy percentage compared to MfCC.

Table 1: Comparison between MFCC and GTCC using Parameter values

Audio Sample Name	Existing value of MSE	Proposed Value of MSE	Existing Value of SNR	Proposed Value Of SNR
1 (3).wav	2.5939	**0.78845**	12.3432	49.2239
2 (2).wav	2.4892	**0.8077**	17.743	46.2306
3 (5).wav	3.2542	**0.3656**	10.2905	42.5442
4 (7).wav	2.4519	**0.9747**	18.3088	44.4392

Table 2(a): Accuracy percentage of MFCC for five emotions and trained with Back propagation

	Angry	Sad	Happy	Fear	Neutral
Angry	95.64	1	0	2	1.36
Sad	2	95.42	2.58	0	0
Fear	0	0	96.03	1	2.97
Happy	1	2	0	95.24	1.76
Neutral	2	2.68	0	0	95.32

Table 2(b): Accuracy percentage of GTCC for five emotions and trained with Back propagation

	Angry	Sad	Happy	Fear	Neutral
Angry	98.74	1	0.26	0	0
Sad	0	98.81	0	1	0.81
Happy	0	0	98.7	1.30	0
Fear	0	0.98	0	99.02	0
Neutral	1	0	0	0.79	98.21

5. CONCLUSION

A speech emotion recognition system consists of feature extraction algorithm and machine learning algorithm. It is observed that emotion recognition in speech using GTCC algorithm is superior to the other existing techniques. A comparison study has been made between the existing work and the proposed work on basis of Signal to Noise Ratio and Mean Square Error. Therefore, the proposed GTCC algorithm enhances the performance with higher SNR and lower MSE. Also, the experimental results show that GTCC outperforms well from existing technique in terms of classification recognition accuracy percentage. Therefore, it is concluded that GTCC has potential to substitute MFCC in the field of speech emotional recognition.

REFERENCES

[1] El Ayadi, Moataz, Mohamed S. Kamel, and Fakhri Karray. "Survey on speech emotion recognition: Features,classification schemes, and databases." ,ELSEVIER, Pattern Recognition 44, no. 3 (2011): 572-587,2010.
[2] Haykin, Simon, and Neural Network. "A comprehensive foundation." Neural Networks 2, no. 2004 ,2004.
[3] Tiwari, Vibha. "MFCC and its applications in speaker recognition." International Journal on Emerging Technologies 1, no. 1 ,pp : 19-22,2010.
[4] Firoz, S. A., S. A. Raji, and A. P. Babu. "Automatic Emotion Recognition from Speech Using Artificial Neural Networks with Gender-Dependent Databases." In Advances in Computing, Control, & Telecommunication Technologies, IEEe. ACT'09. International Conference on, pp. 162-164, 2009.
[5] Sivanandam, S. N., and S. N. Deepa. Introduction to neural networks using Matlab 6.0. Tata McGraw-Hill Education, 2006.
[6] Wang, Sheguo, Xuxiong Ling, Fuliang Zhang, and Jianing Tong. "Speech emotion recognition based on principal component analysis and back propagation neural network." In Measuring Technology and Mechatronics Automation (ICMTMA),IEEE, International Conference on, vol. 3, pp. 437-440., 2010. Iliou, Theodoros, and Christos-Nikolaos Anagnostopoulos. "Classification on speech emotion recognition-a comparative study." International Journal on Advances in Life Sciences 2, no. 1 and 2 ,pp: 18-28,2010.
[7] Ramakrishnan, S. "Recognition of emotion from speech: a review." Speech Enhancement, Modeling and Recognition-Algorithms and Applications. InTech ,2012.
[8] Unluturk, Mehmet S., Kaya Oguz, and Coskun Atay. "Emotion recognition using neural networks." In Proceedings of the 10th WSEAS International Conference on NEURAL NETWORKS, Prague, Czech Republic, pp. 82-85. 2009.
[9] Zhao, Xiaojia, and DeLiang Wang. "Analyzing noise robustness of MFCC and GFCC features in speaker identification." In Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on, pp. 7204-7208. , 2013.

[10] Liu, Jia-Ming, Mingyu You, Guo-Zheng Li, Zheng Wang, Xianghuai Xu, Zhongmin Qiu, Wenjia Xie, Chao An, and Sili Chen. "Cough signal recognition with Gammatone Cepstral Coefficients." In Signal and Information Processing (ChinaSIP), IEEE China Summit & International Conference on, pp. 160-164. 2013.

[11] Iliou, Theodoros, and Christos-Nikolaos Anagnostopoulos. "Classification on speech emotion recognition-a comparative study." International Journal on Advances in Life Sciences 2, no. 1 and 2 : 18-28,2010.

[12] Cheng, Octavian, Waleed Abdulla, and Zoran Salcic. "Performance evaluation of front-end processing for speech recognition systems." School of Engineering Report. The University of Auckland, Electrical and Computer Engineering ,2005.