Volume 4, Issue 11, November 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Feature Extraction techniques for Classification of Emotions in Speech Signals

Gurpreet Kaur*
Research Scholar
RIMT Mandi Gobindgarh, India

Mr Abhilash Sharma
Asst Prof. CSE Department
RIMT Mandi Gobindgarh, India

Abstract— Automatic speech emotion recognition is a process of recognizing emotions in speech. This has wide applications in the area of phsycatrics help and in robotics'he human computer interaction the challenging area of research. Any effective HCI system has two sections Training and testing. The techniques used in the system are feature extraction and classification. This paper focuses on the brief introduction of the GFCC feature extraction, optimization algorithm and the back propagation neural network for the classification of the emotions in speech.

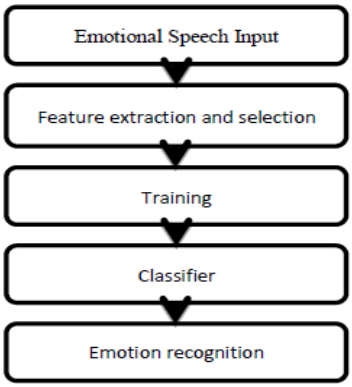
Keywords— emotions in speech, emotion recognition, Back propagation, Gammatone Frequency Cepstral Coefficients, GFCC, Bacterial Forging optimization.

I. INTRODUCTION

Speaker recognition refers to recognize the person from their speech. The speech signal contains the message being spoken, the emotional state of the speaker and the information of the speaker. So the speech signal can be used for the recognition of the speaker and the emotional state of the speaker .Emotion recognition in speech extracts the speech features in each utterance. It is the process of automatically recognizing who is speaking and in which emotional state the words are spoken on the basis of features present in the speech signal. [1].Speaker recognition can be text independent and text dependent .The context dependent has more accuracy. Speaker recognition is helpful in the areas such as voice dialling, banking by telephone, telephone shopping, database access services, information services, and security control for confidential information areas voice mail and remote access to computers [2].The detection of emotions in speech is gaining attention in wide range of application like mostly used to develop wide range of application like application for call centre and learning, gaming software, security applications and machine translation. Influence of emotional state of human speech in speaker recognition is very high [3]. The term "emotion" can refer to an extremely complex state associated with a wide variety of mental, physiological and physical events. Emotional speech database is valuable for this speaker recognition. In a generalized way, a speech emotion recognition system is an application of speech processing in which the patterns of derived speech features (MFCC, pitch) are mapped by the classifier (HMM) during the training and testing session using pattern recognition algorithms to detect the emotions from each of their corresponding patterns. The technique is synonymous to speaker recognition system but its different approach to detect emotions makes it intelligent and adds security to achieve better service in various applications [4]. Different techniques are used for the feature extraction which is the first step in emotion recognition. These features are the input to the classifier for the classification of the emotions. If the features extracted are chosen carefully it is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input.

II. STRUCTURE OF SPEECH RECOGNITION

The speech emotion recognition needs to extract short acoustic and prosody's feature parameters reflecting emotion, and distinguish through a variety of classifier means. The Speech Recognition System can be divided into the following categories as Signal Preprocessing, Feature Extraction, and Speech Classification. The block diagram for ER system is:



The speech files would be the input to the system. The input of the system will undergo some preprocessing. The next step is to extract the main features of the input speech that will differentiate between the different emotions After the extraction of the features, the feature selection, removal and optimization algorithms are applied to get the optimum feature vectors. The vector is then presented to the classifier in training and testing scheme. The final output is the classified emotion according to the input speech.

Signal Preprocessing

Recorded samples never produce identical waveforms as the length, amplitude, noise in the signal may vary. Therefore
It is necessary to perform signal pre-processing to extract only the speech related information. This means that giving right features to the classifier is crucial for successful classification.
Filtering stage – The samples of the speech can be recorded with a microphone. This signal contains a lot of distortion and noise due to the quality of the microphone or just because of picked up background noise. Therefore it becomes necessary to perform some filtering to eliminate low and high frequency of noise. This can be done by passing high and low frequency filters.

GFCC Feature Extraction

Feature Extraction gives the speakers emotion specific information from the given speech signal by performing complex transformations. Different levels of transformations are performed by using semantic and acoustic features. The acoustic features contain the characteristic information of the speech signal and are useful for recognizing the speaker. Widely used acoustic features are Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Cepstral Coefficients (LPCC) and Perceptual Linear Prediction Cepstral (PLPC). MFCC features [5] are derived from Fast Fourier Transform (FFT) power spectrum. The centre and the bandwidth of the filter bank are selected based on the Mel-frequency scale. Therefore, this feature gives more details on the low frequencies. Another acoustic features are GFCC features that exhibit exhibits superior noise robustness to commonly used Mel-frequency Cepstral coefficients (MFCC)[6]. GFCC employs frequency scale, ERB rate scale, and uses Gammatone filter.

Gammatone Filter Bank:

Gammatone filter-bank is a group of filters for the cochlea simulation. The impulse response of a Gammatone filter is similar to the magnitude characteristics of a human auditory filter. The membrane motion is 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 'f_c'.

$$g(t) = Kt^{n-1}e^{-2\pi Bt}cos(2\pi f_c t + \varphi),$$

Where K is the amplitude gain; *n* is the filter order is the filter's bandwidth;f_c is the center frequency in Hertz and φ is the phase shift. The fourth order Gammatone filter is similar to the function used to represent human auditory response [7].Hence it will be efficient to use fourth order Gammatone filters. the formula for fourth order Gammatone bandwidth B is:

$$B = 1.019 \times ERB(f_c),$$

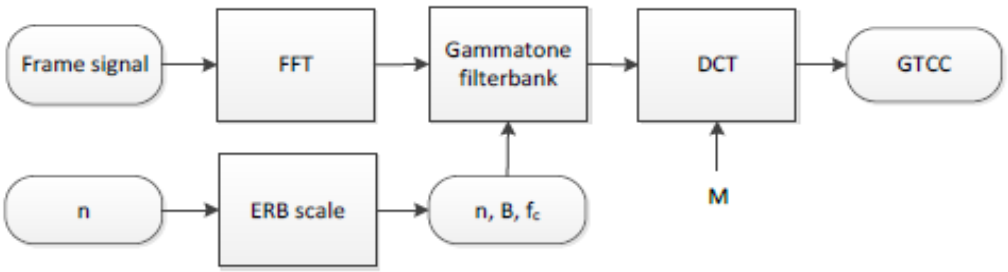
Where ERB is used for the Equivalent Rectangular bandwidth .The auditory filter's bandwidth is the value of ERB centered at frequency f_c.The relationship between ERB and can be modeled as:

$$ERB(f) = 24.7 + 0.108f,$$

With the Gammatone filter bank GFCC features can be calculated. The original audio signals are windowed into the frames. The default frame length is 25ms.and the default overlapping time is 10ms. After windowing the signal then FFT (fast Fourier transformation) is applied to each frame to analyze the spectrum. Then Gammatone filter bank is applied to the each Fast Fourier transformed Signal. The energy of each sub-band is calculated and denoted as X_n. the log function and the discrete cosine transform are applied to model the human loudness perception and decorrelate the logarithmic compressed filter outputs.

$$GFCC_m = \sqrt{\frac{2}{N}}\sum_{n=1}^{N} log_{10}(X_n)cos[\frac{\pi n}{N}(m-\frac{1}{2})], 1 \le m \le M$$

where is X_n the energy of the nth sub-band, N is the number of Gammatone filters, and M is the number of GFCC.



Block diagram of calculation of GFCC[8].

The brief description of the GFCC extraction is as follows:

1. Pass input signal through a 64-channel Gammatone filter bank
2. At each channel, fully rectify the filter response (i.e. take absolute value) and decimate it to 100 Hz as a way of time windowing. Then take absolute value afterwards. This creates a time frequency (T-F) representation that is a variant of cochlea gram .
3. Take cubic root on the T-F representation
4. Apply DCT to derive Cepstral features

Feature Selection or optimization

After feature extraction the optimal features are to be selected .This task can be done by various optimization algorithms as Sequential Minimization optimization. Bacterial Forging Optimization. Using the BFO with neural classifier for is the best combination .to increase the efficiency of the neural network Classifier it is necessary to give optimal inputs to the classifier. Thus using BFO as the feature selection algorithm is the good choice . Bacteria Foraging Optimization (BFO) algorithm is a new class of biologically encouraged stochastic global search technique based on mimicking the foraging behavior of E. coli bacteria. This method is used for locating, handling, and ingesting the food. During foraging, a bacterium can exhibit two different actions: tumbling or swimming. The tumble action modifies the orientation of the bacterium. During swimming means the chemo taxis step, the bacterium will move in its current direction. Chemo taxis movement is continued until a bacterium goes in the direction of positive-nutrient gradient. After a certain number of complete swims, the best half of the population undergoes the reproduction and eliminating the rest of the population. In order to escape local optima, an elimination-dispersion event is carried out where some bacteria are liquidated at random with a very small probability and the new replacements are initialized at random locations of the search space.

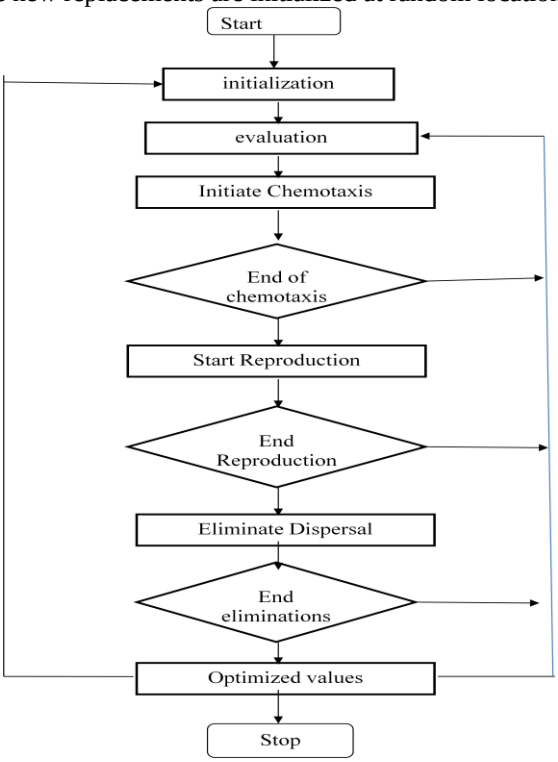
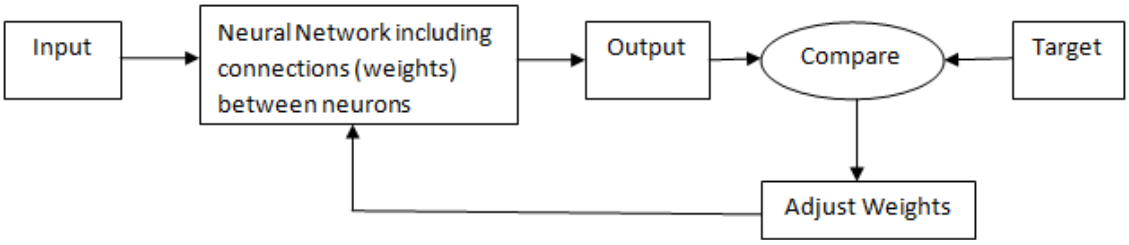


Fig 2: Flow chart of the BFO algorithm

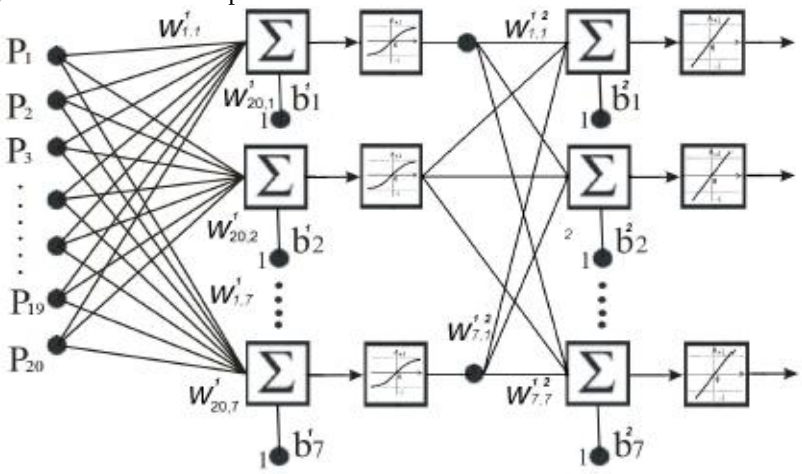
BPNN (Back Propagation Neural Network)

The back propagation algorithm in BPNN **can** be used as a classifier for the classification of the emotions in speech. Neural networks are composed of simple computational elements operating in parallel [1]. The network function is determined largely by the connections between elements. We can train a neural network so that a particular input leads to a specific target output Artificial Neural Network (ANN) is an efficient pattern recognition mechanism which simulates the neural information processing of human brain. The ANN processes information in parallel with a large number of processing elements called neurons and uses large interconnected networks of simple and nonlinear units [9].



Neural network[9]

The quantitative modeling and processing of data using neural networks is effectively performed using the Supervised Learning Neural Network Back-Propagation Algorithm. For a given set of training input -output pair, this algorithm provides a procedure for changing the weights in a back-propagation network (BPN) to classify the input patterns correctly. The aim of this neural network is to train the network to achieve a balance between the network's ability to respond (memorization) and its ability to give reasonable responses to the input that is similar but not identical to the one that is used in training (generalization) [7]. A BPNN is a multi-layer, feed-forward neural network consisting of an input layer, a hidden layer and an output layer. The hidden layers are used to classify successfully the patterns into different classes. The inputs are fully connected to the first hidden layer, each hidden layer is fully connected to the next, and the last hidden layer is fully connected to the outputs.



Example of the feed forward back propagation network[10]

The hidden layer consists of non-linear signoidal activation function neurons. The amount of neurons depends on some factors like the amount of input data and output layer neuron number, the needed generalization capacity of the network and the size of the training set.

III. CONCLUSION

In this paper we have focused emotion detection in speech. We also reviewed the entailed study of the structure of the emotion detection system ,the GFCC feature extracting GFCCs are more efficient that Any other techniques' he introduction to the BPNN classifier have also been reviewed Thus GFCC can be used with BPNN and the accuracy of the classifier can be enhanced by using different algorithms .

ACKNOWLEDGEMENT

I express my sincere gratitude to my guide Mr.Abhilash Sharma, and to Head of CSE department Dr. Anuj Gupta for his valuable guidance and advice. Also I would like to thank all the people who have given their heartwelling support in making this completion a magnificent experience.

REFERENCES

[1] J. Sirisha Devi , Y. Srinivas and Siva Prasad Nandyala *Automatic Speech Emotion and Speaker Recognition based on Hybrid GMM and FFBNN" in* International Journal on Computational Sciences & Applications (IJCSA) Vol.4, No.1, February 2014.

[2] M. Kockmann, L. Ferrer, L. Burget, E. Shriberg, and J. H. Cernocký, *"Recent progress in prosodic speaker verification,"* in Proc. IEEE ICASSP, (Prague), pp. 4556--4559, May 2011.

[3] Marius Vasile Ghiurcau , Corneliu Rusu,Jaakko Astola,(2011) *"A Study Of The Effect Of Emotional State Upon Text-Independent Speaker Identification",* Published in ICASSP

[4] Rahul.B.Lanjewar, D.S.Chaudhari,(2013) *"Speech Emotion Recognition:A"Review* International Journal of Innovative Technology and Exploring Engineering, ISSN:2278-3075, Vol.2,Issue-4

[5] K.A. Senthildevi and E. Chandra(2007) *"Speech Data Mining & Document Retrieval,*" publication of the IEEE signal processing

[6] *Xiaojia Zhao and DeLiang Wang "Analyzing noise robustness of mfcc and gfcc features in speaker identification*".in ICASSP 2013.

[7] M. Slaney "*An efficient implementation of the Patterson-Holdsworth auditory filter bank,"* Apple Computer, Perception Group, Tech. Rep, 1993.

[8] Jia-Ming Liu1, Mingyu You1*, Guo-Zheng Li1, Zheng Wang1, Xianghuai Xu2, Zhongmin Qiu2, Wenjia Xie1, Chao An1, Sili Chen1 "*cough signal recognition with gammatone cepstral coefficients"*in Signal and information Processing 2013 IEEE China Summit and international Confernece on 6^{-7} july 2013

[9] Firoz Shah. A, Raji Sukumar. A, and Babu Anto. P, "Discreet Wavelet Transforms and Artificial Neural Networks for Speech Emotion Recognition", International Journal of Computer Theory and Engineering, Vol. 2, No. 3, 1793-8201, June 2010, pp.319-322

[10] Neural Networks used for Speech Recognition Wouter Gevaert, Georgi Tsenov, Valeri Mladenov, *Senior Member, IEEEJOURNAL OF AUTOMATIC CONTROL, UNIVERSITY OF BELGRADE, VOL. 20:1-7, 2010©*

[11] Priyanka Abhang, Shashibala Rao, Bharti W. Gawali, Pramod Rokade, *"Emotion Recognition using Speech and EEG Signal – A Review"*, International Journal Of Computer Applications (0975 – 8887) Volume 15– No.3, February 2011.

[12] N. Murali Krishna, P.V. Lakshmi, Y. Srinivas J.Sirisha Devi, *"Emotion Recognition using Dynamic Time Warping Technique for Isolated Words",* IJCSI International Journal Of Computer Science Issues, Vol. 8, Issue 5, No 1, September 2011.

[13] Krishna Mohan Kudiri, Gyanendra K Verma and Bakul Gohel*," Relative amplitude based feature for emotion detection from speech",* IEEE Transactions On Audio, Speech, And Language Processing , 2010.

[14] Emily Mower, Maja J Mataric, Shrikanth Narayanan," *A framework for automatic human emotion classification using emotion profiles",* IEEE Transactions On Audio, Speech, And Language Processing, Vol. 19, No. 5, July 2011.

[15] Jagvir Kaur, Abhilash Sharma" *a review of automatic speechemotion recognition"*published in International Journal of Advanced and Innovative Research (2278-7844) / # 308 / Volume 3 Issue 4

[16] Joder, Cyril ;Schuller *"Exploring Nonnegative Matrix Factorization forAudio Classification: Application to Speaker Recognition"*published in SpeechCommunication; 10. ITG Symposium, 26-28 Sept. 2012

[17] Garg, Vipul, Kumar, Harsh; Sinha, Rohit*,"Speech based Emotion Recognition based on hierarchical decision tree with SVM, BLG and SVR classifiers"*in Communications (NCC), 2013 National Conference.

[18] Tobias May, Steven van de Par, and Armin Kohlrausch*,"Noise-Robust Speaker Recognition Combining Missing Data Techniques and Universal Background Modeling",* IEEE Transactions On Audio, Speech, And Language Processing, Vol. 20, No. 1, January 2012.

[19] Nitisha and Ashu Bansal, *"Speaker Recognition Using MFCC Front End Analysis and VQ Modelling Technique for Hindi Words using MATLAB",* Hindu College of Engineering, Haryana, India.

[20] Alejandro Bidondo, Shin-ichi Sato, Ezequiel Kinigsberg, Adrián Saavedra, Andrés Sabater, Agustín Arias, Mariano Arouxet, and Ariel Groisman (2013) *"Speaker recognition analysis using running autocorrelation function parameters",* POMA - ICA Montreal Volume 19, pp. 060036

[21] David A. van Leeuwen and Rahim Saeidi*," Knowing The Non-Target Speakers: The Effect Of The I-Vector Population For Plda Training In Speaker Recognition",* ICASSP 2013.

[22] Akshay S. Utane, Dr. S. L. Nalbalwar, *"Emotion Recognition through Speech Using Gaussian Mixture Model and Support Vector Machine"*, International Journal of Scientific & Engineering Research, Volume 4, Issue 5, May-2013.

