HUMAN SPEECH EMOTION RECOGNITION

Maheshwari Selvaraj[#1] Dr.R.Bhuvana[#2] S.Padmaja[#3]
[#1,#2]Assistant Professor, Department of Computer Application, Department of Software Application,
A.M.Jain College,Chennai, India
[,#3]Assistant Professor, School of Computing Sciences,
Vels University, Chennai, India
[1] maheshwari.selvarajj@gmail.com
[3]Spadmaja.research@gmail.com
[2]bhuvanavr1981@yahoo.co.in

***Abstract* - Emotions play an extremely important role in human mental life. It is a medium of expression of one's perspective or one's mental state to others. Speech Emotion Recognition (SER) can be defined as extraction of the emotional state of the speaker from his or her speech signal. There are few universal emotions- including Neutral, Anger, Happiness, Sadness in which any intelligent system with finite computational resources can be trained to identify or synthesize as required. In this work spectral and prosodic features are used for speech emotion recognition because both of these features contain the emotional information. Mel-frequency cepstral coefficients (MFCC) is one of the spectral features. Fundamental frequency, loudness, pitch and speech intensity and glottal parameters are the prosodic features which are used to model different emotions. The potential features are extracted from each utterance for the computational mapping between emotions and speech patterns. Pitch can be detected from the selected features, using which gender can be classified. Support Vector Machine (SVM), is used to classify the gender in this work. Radial Basis Function and Back Propagation Network is used to recognize the emotions based on the selected features, and proved that radial basis function produce more accurate results for emotion recognition than the back propagation network.**

Index Terms – Speech Emotion Recognition, MFCC, Prosodic Features, Support Vector Machine, Radial Basis Function Network, Back Propagation Network.

I. INTRODUCTION

The importance of emotion recognition of human speech has increased in recent days to improve both the naturalness and efficiency of human - machine interactions. Recognizing human emotions is a very complex task in itself because of the ambiguity in classifying the acted and natural emotions. A number of studies have been conducted to extract the spectral and prosodic features which would result in correct determination of emotions. Nwe, T. L., et. al [13] explained about the emotion classification using human speech utterance based on calculated bytes.Chiu Ying Lay, et. al [6] explained about how to classify the gender using calculated pitch from human speech.Chang-Hyun Park, et. al [4] has discussed about extracting acoustic features from the speech and classify the emotions.Nobuo Sato et. al [11]gave their details regarding MFCC approach.The main intension of theirworkwas applyingMFCC to human speech and classifying the emotions more than 67% of accuracy.YixiongPan, et. al [15]have used Support Vector Machines (SVM), to the problem of emotion classification in an attempt to increase accuracy. Keshi Dai et. al [8] explained about recognizing the emotions using Support vector machines in neural network and gave more than 60% accuracy. Aastha Joshi [1] Speaks about the Hidden Markov Model and Support Vector Machine feature regarding speech emotion recognition. Sony CSL Paris [12], it presented the algorithms that allow a robot to express its emotions by modulating the intonation of its voice. BjörnSchuller, et. al [3] discussed about the approaches to recognize the emotional user state by analyzing spoken utterances on both, the semantic and the signal level. Mohammed E. Hoque, et. al [10] presented about robust recognition of selected emotions from salient spoken words. The prosodic and acoustic features were used to extract the intonation patterns and correlates of emotion from speech samples in order to develop and evaluate models of emotion. But even after so much of research, researchers have not gained much of success and the accuracy. Emotions can be classified as Natural and Artificial emotions and further can be divided into emotion set i.e. anger, sadness, neutral, happy, joy, fear. Different machine learning techniques have been applied to create recognition agents including k-nearest neighbour, radial basis function and back propagation of neural networks. Our simulation experiment results showed that radial basis function were effective in emotion recognition, and produce more accurate results. And regarding gender classification earlier all information in speech is in the range 200Hz to 8kHz. Humans discriminate voices between males and females according to the frequency. Females speak with higher fundamental frequencies than males. Therefore, by analysing the average pitch of the speech samples, an algorithm is being adopted for a gender classifier. To process a voice signal, there are techniques that can be broadly classified as either time-domain or frequency-

domain approaches. With a time-domain approach, information is extracted by performing measurements directly on the speech signal whereas with a frequency-domain approach, the frequency content of the signal is initially computed and information is extracted from the spectrum. Given such information, one can perform analysis on the differences in pitch and formant positions for vowels between male and female. This paper focused on identifying the emotion using the emotion set Anger, Happy, Sad and Neutral. However certain emotions have similar characteristics based on the set of features. An experimental study has been conducted to determine how well people recognize emotions in speech. Based on the results of the experimentthe most reliable utterances were selected for feature selection and for training recognizers. This paper is organized as five sections, section II explains about existing work, section III is about proposed work, section IV gives the experimental results and conclusion has been drawn in section V.

II. EXISTING METHODOLOGY

Speech sample [2, 4, 6, 9] is first passed through a gender reference database which is maintained for recognition of gender before it gets into the process. Statistical approach [5] is followed taking pitch as feature for gender recognition [9]. A lower and upper bound pitch for both male and female samples could be found using the reference database [14]. Input human voice sample was first broken down into frames of frame size 16 ms each. This was done for frame level classification in further steps.

For each frame MFCC(Mel Frequency Cepstral Coefficient)was calculated as the main feature for emotion recognition. Reference database [14] is maintained which contains the MFCCs of emotions i.e. of Sad, Anger, Neutral and Happy.

MFCC of the frames were compared with the MFCCs stored in reference database and the distance was calculated between the comparable frames. Based on the distance of the analysis frame from the reference database, one can classify the frame as anger, happy or normal. The output is displayed in terms of emotional frame count.

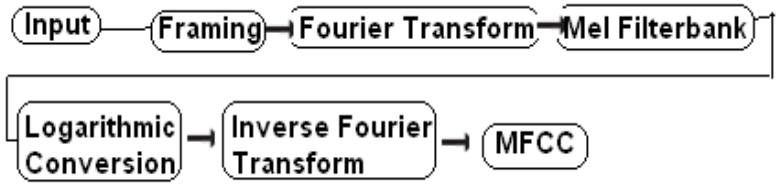

Figure 1. Standard MFCC Approach

III. PROPOSED METHODOLOGY

In the proposed work, human voice is given as the input. Then the input is converted into frames of frame size 60ms for every 50ms which means overlapping of data for 10ms. This is because for no missing of data. Fundamental frequencies are calculated based on pitch autocorrelation function [4,7]. Using Support Vector Machine's reference database, average pitch value is calculated based on which gender can be classified.

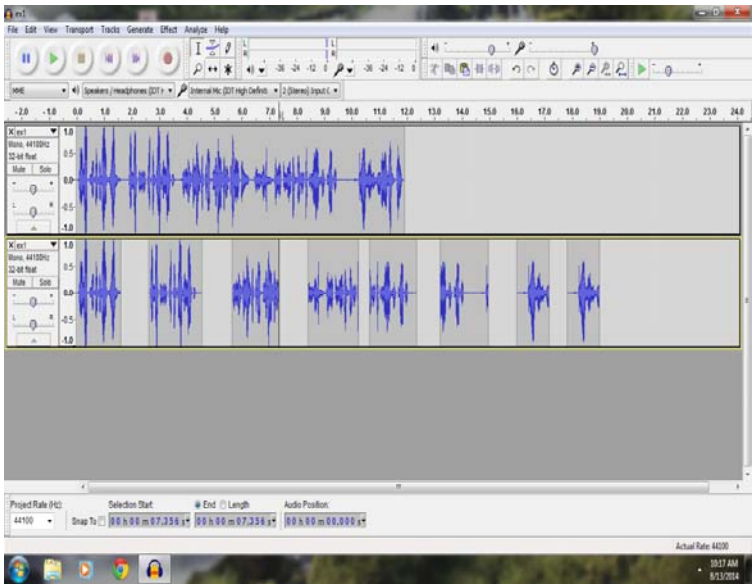

Figure 2. Frames of Data

For emotion recognition each frame can be entered into the proposed MFCC approach. Mel Frequency Cepstral Coefficient function contains group of four operations on human speech. Fast Fourier Transform will be applied to each for finding minimum and maximum frequencies [13]. Later Mel filter bank can be applied to map the powers of spectrum obtained above using overlapped triangular windows, after which logarithmic conversion will be done for finding amplitude values. Finally discrete cosine transform will be applied to get the missing data while compressing the audio clip, finally the MFCC values for each framewill be calculated[2, 11].

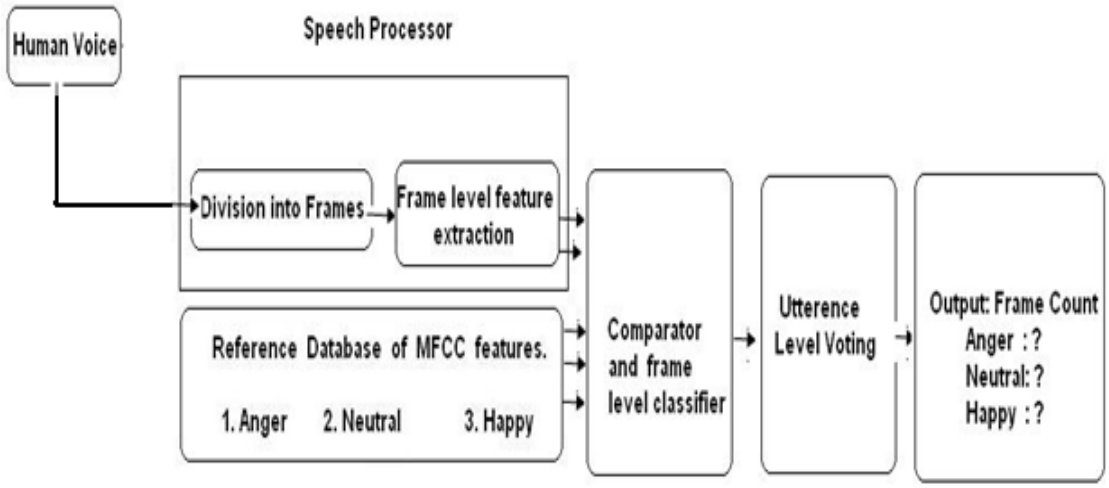

Figure 3. Standard Approach for Emotion Classification

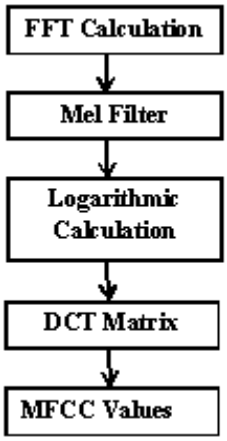

Figure 4. Proposed MFCC Approach

Then these values can be trained using Radial basis function network and back propagation network to obtain the average emotion values.

For RBF network the learning rate was taken as 0.001 with Gaussian activation function in the hidden layer and Identity activation function at the output layer for training the network.

For BPN network the learning rate was taken as 0.001 with Ramp activation function in the hidden layer and binary step function at the output layer for training the network.Finally these values can be matched with the speech emotion reference Berlin database [14], to classify the emotion set.

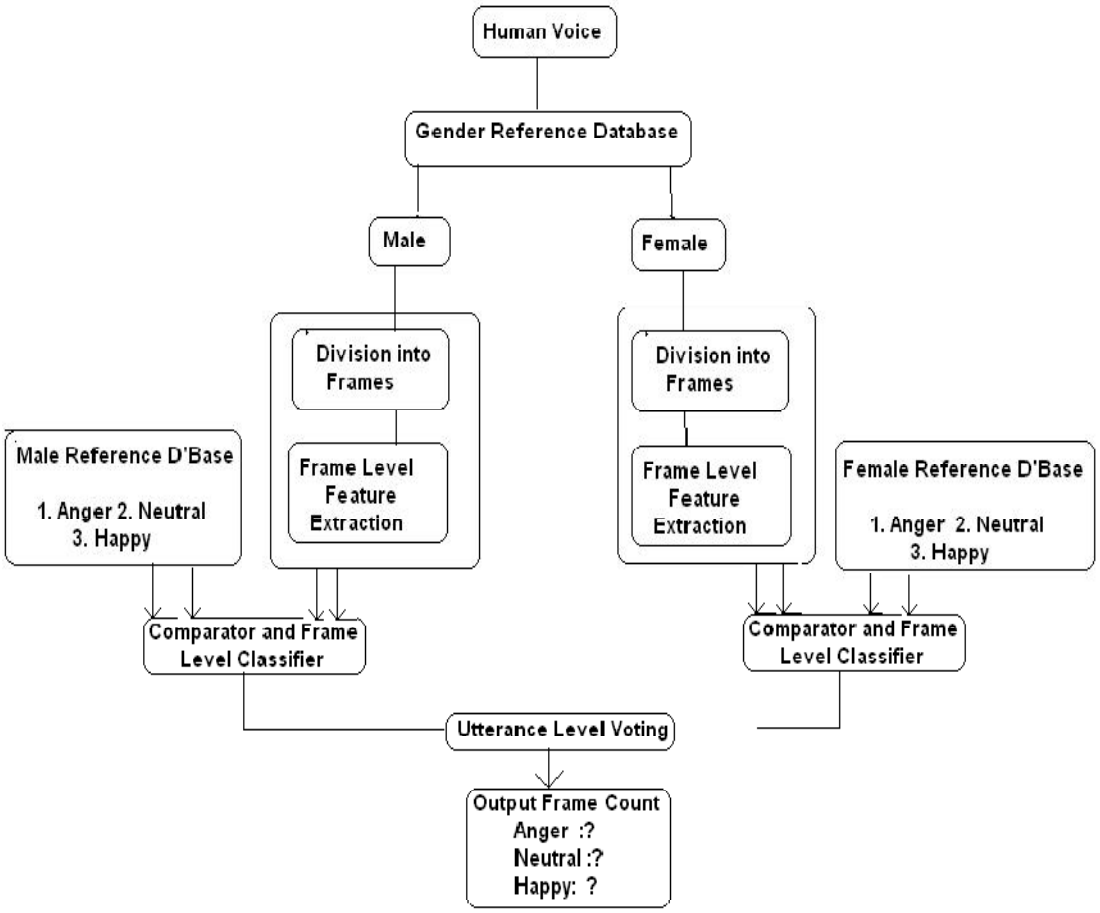

Figure 5. Standard Approach for Emotion Recognition

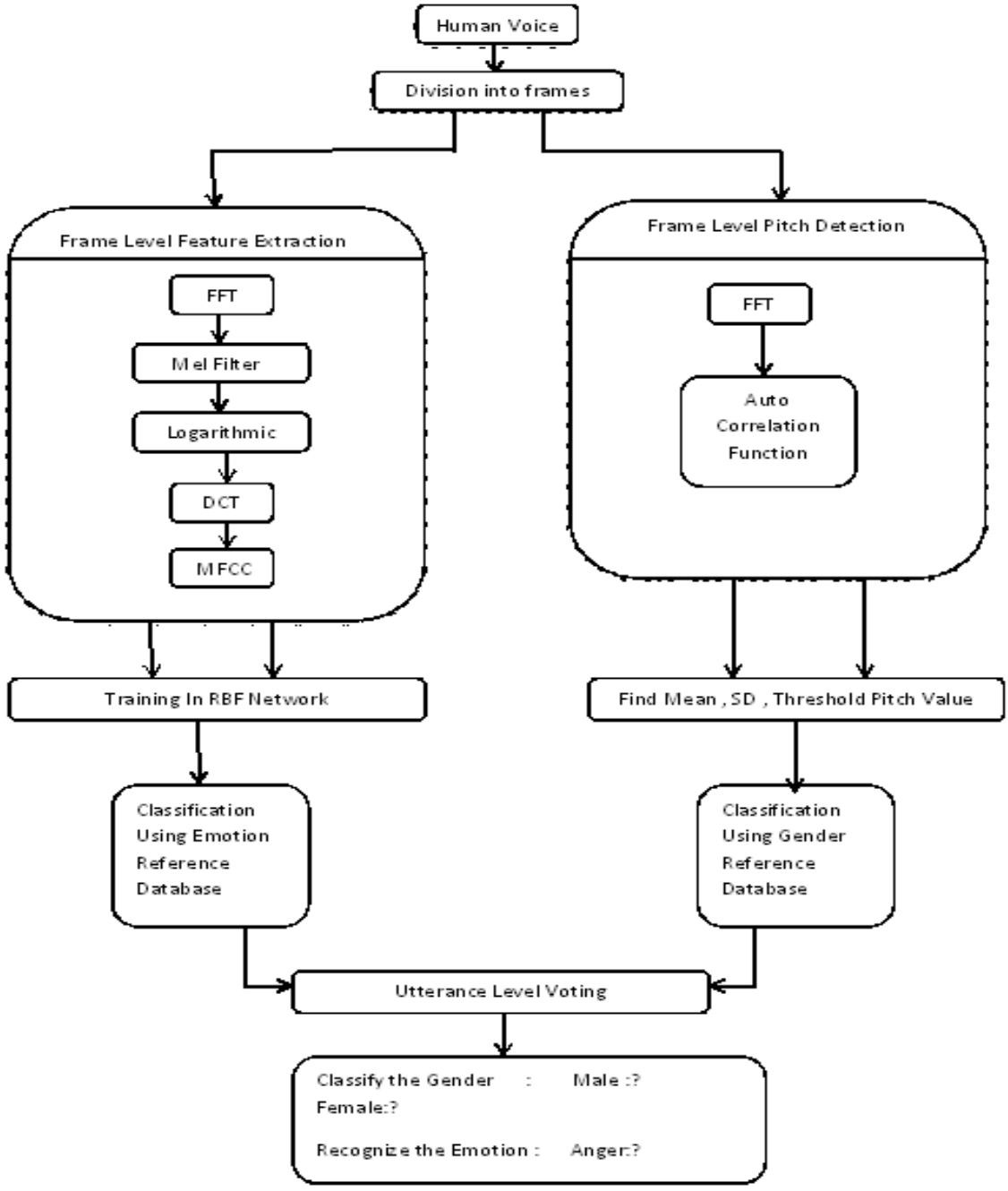

Figure 6. Proposed Approach for Emotion Recognition

IV. EXPERIMENTAL RESULTS

Input for the experiment is taken from the Berlin speech database [14]. Forty sets of input male and female alternatively was applied for the experiment. The model that have been chosen for gender identification is pitch extraction via autocorrelation [4] since human ears mainly differentiate by pitch.

The algorithm for pitch extraction is as follows:

Step :1 Divide the speech into 60ms frame segments. Each segment is extracted at every 50ms interval. This implies that the overlap between segments is 10ms.

Step:2 Use Pitch Autocorrelation at each segment to estimate the fundamental frequency for that segment.

Step:3 For each segment calculate pitch autocorrelation and apply the centre clipping algorithm.

Typical autocorrelation function is given by

$$R(K) = \sum_{n=-\infty}^{\infty} x[n].x[n+k] \qquad \text{----------(1)}$$

R(k) => Correlation of the signal "k".

n => Index of the signal (n=0,1,2…N-1).

x => Source of Signals.

After clipping, the short-time energy function is computed.

Energy can be defined as

$$W[n] = \sum |x[n]| \cdot w[n-m] \qquad \text{----------(2)}$$

W[n] => Window of the signal "n" (n =0,1,2,…N-1).

m => no of Unvoiced speech.

Step:4 Apply median filtering for every 3 segments so that it is less affected by noise.

Step:5 Calculate the average of all fundamental frequencies

Forty samples have been selected from the calculated pitch list then average fundamental frequencies (pitch) are computed for both male class and female class. A threshold is obtained by getting the mean of the male and female average fundamental frequencies. The standard deviation (SD) for each class is computed. The values used as parameters of the classifier are tabulated in table 1.

Table 1. Threshold Value for Gender Classification

Mean pitch for male	308.2934 Hz
SD for male	17.5583 Hz
Mean pitch for female	322.7154 Hz
SD for female	17.9643 Hz
Threshold	315.5049 Hz

The threshold is the determinant for the gender class. If the pitch of a voice sample falls below the threshold, the classifier will assign it as male. Otherwise, it will assign as female. The model that have chosen for EMOTION classification is Mel Frequency Cepstral Coefficient approach via spectral features of the voice signal.

The algorithm for MFCC is as follows:

1.Frame Level Break Down : Input human voice sample is first break down into frames of frame size 60ms each. This is done for frame level classification in further steps.

2. Frame Level Feature Extraction :For each frame got in '1', will calculate MFCC as the main feature for emotion recognition.

3. Comparator and Frame Level Classifier :Forty set of samples including male and female voices have been selected from MFCC then that are trained using RBF network then average emotion values can be calculated

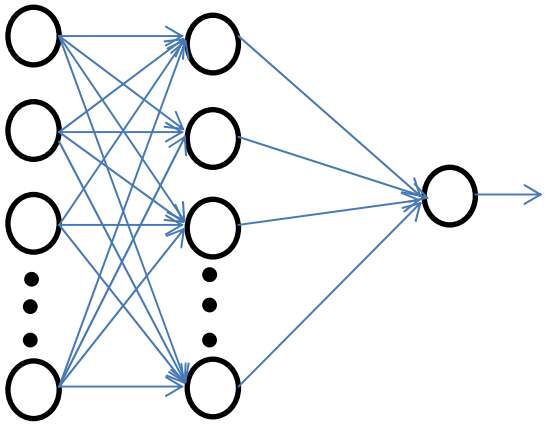

Figure 7. RBF Training Network

Number of input patterns for theRBF network was no of frames in a input data, attributes in each input patterns has taken as bytes in each frames, number of hidden neurons for the network was calculated as no of frames plus twenty, number of output neurons were two. For training based on RBF network, the parameter used for weight [-1,+1], with learning rate 0.001.Activation Function used for output layer was Identity function, and for hidden layer was Gaussian function.

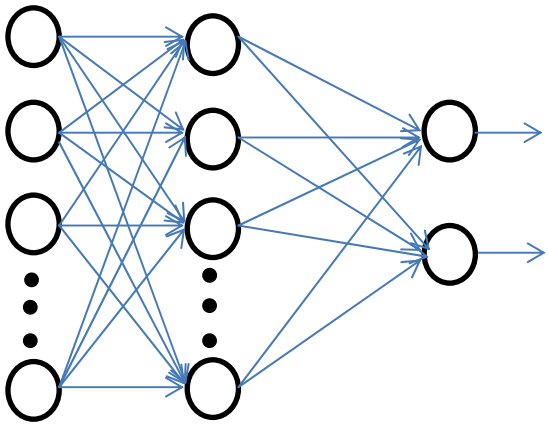

Figure 8. BPN Training Network

Number of input patterns for the BPN network was no of frames in a input data, attributes in each input patterns has taken as bytes in each frames, number of hidden neurons for the network was calculated as no of frames plus twenty, number of output neurons were one. For training based on BPN network, the parameter used for weight [-1,+1], with learning rate 0.001.Activation Function used for output layer was Identity function, and for hidden layer was Ramp function.

4. Utterance-Level Voting:Comparing the average value with the reference database, classify the frame as anger, happy, sad or normal. The output is displayed in terms of emotional frame count.

Table 2. Training Results for Twenty Male Samples using RBF Network

Sample name (Input – Male)	Pitch Observed	Gender Classified Correctly	Emotion Frame Count	Emotion Classified Correctly
Angry	108.0	Yes	255.6	Yes
Angry	188.0	Yes	267.8	Yes
Angry	67.0	Yes	273.9	Yes
Angry	171.0	Yes	250.3	Yes
Angry	72.0	Yes	265.1	Yes
Sad	35.0	Yes	35.9	Yes
Sad	347.0	No	46.8	Yes
Sad	89.0	Yes	78.9	Yes
Sad	171.0	Yes	198.4	No
Sad	188.0	Yes	76.9	Yes
Neutral	209.0	Yes	102.4	Yes
Neutral	134.0	Yes	143.3	Yes
Neutral	190.0	Yes	127.8	Yes
Neutral	455.0	No	109.2	Yes
Neutral	98.0	Yes	145.3	Yes
Happy	189.0	Yes	167.8	Yes
Happy	112.0	Yes	153.4	Yes
Happy	69.0	Yes	155.9	Yes
Happy	41.0	Yes	89.0	Yes
Happy	45.0	Yes	176.3	Yes

RBF error rate have been calculated for the forty samples and the plot diagrams has been shown in Figure 9 and Figure 10 :

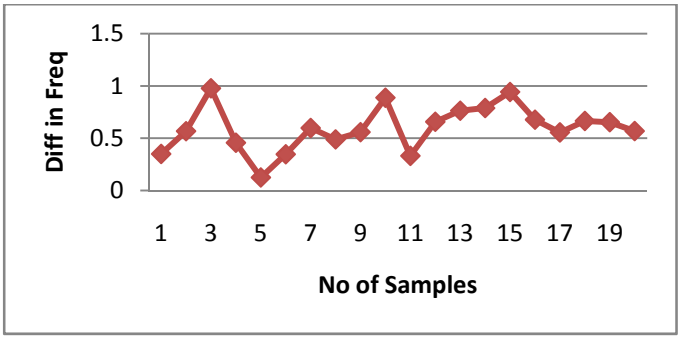

Figure 9. RBF Error rate for Male Samples

Table 3. Training Results for Twenty Female Samples using RBF Network

Sample name (Input - Female)	Pitch Observed	Gender Classified Correctly	Emotion Frame Count	Emotion Classified Correctly
Angry	388.0	Yes	298.3	Yes
Angry	331.0	Yes	312.5	Yes
Angry	121.0	No	303.2	Yes
Angry	345.0	Yes	278.9	Yes
Angry	72.0	No	289.4	Yes
Sad	335.0	Yes	35.9	Yes
Sad	347.0	Yes	79.0	Yes
Sad	389.0	Yes	59.6	Yes
Sad	341.0	Yes	298.1	No
Sad	338.0	Yes	37.8	Yes
Neutral	393.0	Yes	105.6	Yes
Neutral	334.0	Yes	139.8	Yes
Neutral	397.0	Yes	125.6	Yes
Neutral	355.0	Yes	133.5	Yes
Neutral	348.0	Yes	126.7	Yes
Happy	459.0	Yes	187.2	Yes
Happy	112.0	No	176.3	Yes
Happy	469.0	Yes	199.0	Yes
Happy	451.0	Yes	155.8	Yes
Happy	435.0	Yes	167.8	Yes

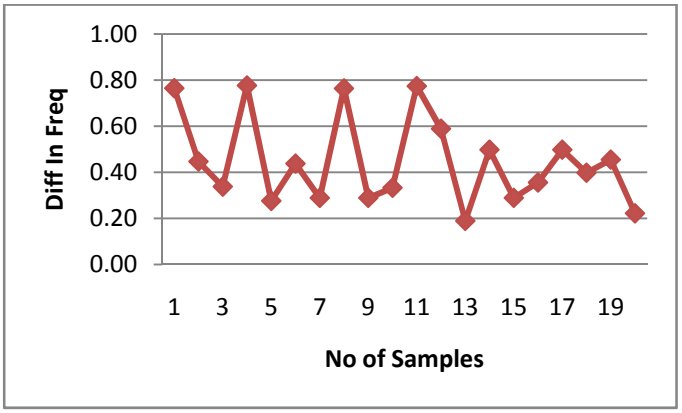

Figure 10. RBF Error rate for Female Samples

Table 4. Training Results for Twenty Male Samples using BPN Network

Sample name (Input – Male)	Pitch Observed	Gender Classified Correctly	Emotion Frame Count	Emotion Classified Correctly
Angry	108.0	Yes	251.2	Yes
Angry	188.0	Yes	267.1	Yes
Angry	67.0	Yes	273.9	Yes
Angry	171.0	Yes	151.3	No
Angry	72.0	Yes	263.1	Yes
Sad	35.0	Yes	35.2	Yes
Sad	347.0	No	41.4	Yes
Sad	89.0	Yes	78.2	Yes
Sad	171.0	Yes	98.5	Yes
Sad	188.0	Yes	176.1	No
Neutral	209.0	Yes	101.4	Yes
Neutral	134.0	Yes	143.9	Yes
Neutral	190.0	Yes	127.5	Yes
Neutral	455.0	No	109.1	Yes
Neutral	98.0	Yes	145.2	Yes
Happy	189.0	Yes	167.4	Yes
Happy	112.0	Yes	173.5	Yes
Happy	69.0	Yes	154.3	Yes
Happy	41.0	Yes	89.6	No
Happy	45.0	Yes	176.3	Yes

BPN error rate have been calculated for the forty samples and the plot diagrams has been shown in Figure 11 and Figure 12 :

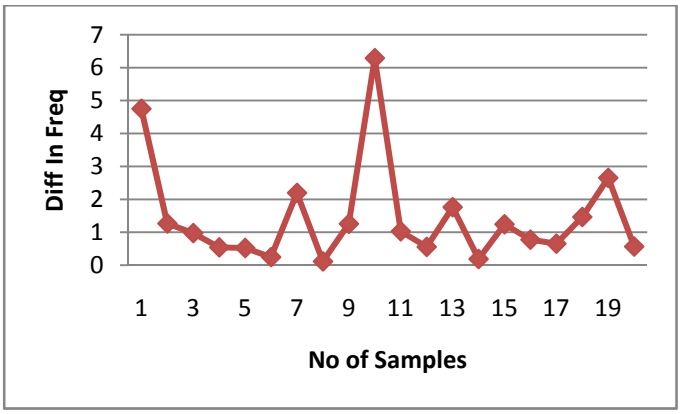

Figure 11. BPN Error rate for Male Samples

Table 5. Training Results for Twenty Female Samples using BPN Network

Sample name (Input - Female)	Pitch Observed	Gender Classified Correctly	Emotion Frame Count	Emotion Classified Correctly
Angry	288.0	Yes	198.5	No
Angry	231.0	Yes	311.9	Yes
Angry	121.0	No	301.9	Yes
Angry	245.0	Yes	278.5	Yes
Angry	72.0	No	289.5	Yes
Sad	235.0	Yes	32.5	Yes
Sad	347.0	Yes	75.1	Yes
Sad	289.0	Yes	159.9	No
Sad	341.0	Yes	98.1	Yes
Sad	238.0	Yes	39.1	Yes
Neutral	293.0	Yes	105.9	Yes
Neutral	334.0	Yes	139.3	Yes
Neutral	397.0	Yes	124.5	Yes
Neutral	255.0	Yes	182.4	No
Neutral	348.0	Yes	126.7	Yes
Happy	259.0	Yes	186.1	Yes
Happy	112.0	No	110.1	No
Happy	469.0	Yes	199.5	Yes
Happy	451.0	Yes	155.2	Yes
Happy	235.0	Yes	167.1	Yes

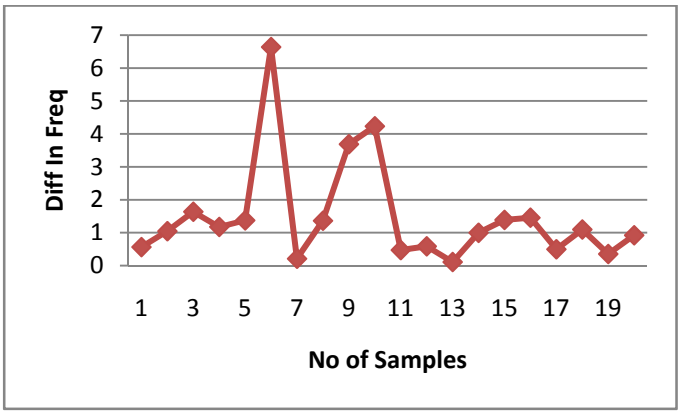

Figure 12. BPN Error rate for Female Samples

The difference between actual and calculated output in RBF network is less than one(<1). But in the case of BPN network the difference is more than one and less than ten (<10). The experiment shows that RBF network recognizing the emotions more accurately than the BPN network.

Table 6 shows the difference between actual and expected value for both the RBF and BPN network for the fortySamples :

Table 6. Comparison between RBF and BPN Network in Error Rates

Expected Value	Actual Value for RBF Network	Difference	Actual Value for BPN Network	Difference
255.95	255.6	0.35	251.2	4.75
268.368	267.8	0.568	267.1	1.268
274.878	273.9	0.978	273.9	0.978
250.757	250.3	0.457	251.3	0.543
167.925	167.8	0.125	167.4	0.525
123.747	123.4	0.347	123.5	0.247
156.498	155.9	0.598	154.3	2.198
89.489	89.0	0.489	89.6	0.116
36.459	35.9	0.559	35.2	1.259
47.688	46.8	0.888	41.4	6.288
79.232	78.9	0.332	78.2	1.032
99.058	98.4	0.658	98.5	0.558
103.164	102.4	0.764	101.4	1.764
144.089	143.3	0.789	143.9	0.189
128.743	127.8	0.943	127.5	1.243
109.878	109.2	0.678	109.1	0.778
145.856	145.3	0.556	145.2	0.656
77.566	76.9	0.666	76.1	1.466
265.754	265.1	0.654	263.1	2.654
176.869	176.3	0.569	176.3	0.569
299.065	298.3	0.77	298.5	0.565
312.947	312.5	0.447	311.9	1.047
303.538	303.2	0.338	301.9	1.638
279.677	278.9	0.777	278.5	1.177
187.476	187.2	0.276	186.1	1.376
176.738	176.3	0.438	170.1	6.638
199.289	199	0.289	199.5	0.211
156.564	155.8	0.764	155.2	1.364
36.189	35.9	0.289	32.5	3.689
79.333	79	0.333	75.1	4.233
60.374	59.6	0.774	59.9	0.474
98.689	98.1	0.589	98.1	0.589
105.789	105.6	0.189	105.9	0.111
140.298	139.8	0.498	139.3	0.998
125.889	125.6	0.289	124.5	1.389
133.856	133.5	0.356	132.4	1.456
127.198	126.7	0.498	126.7	0.498
38.198	37.8	0.398	39.1	1.098
289.855	289.4	0.455	289.5	0.355
168.022	167.8	0.222	167.1	0.922

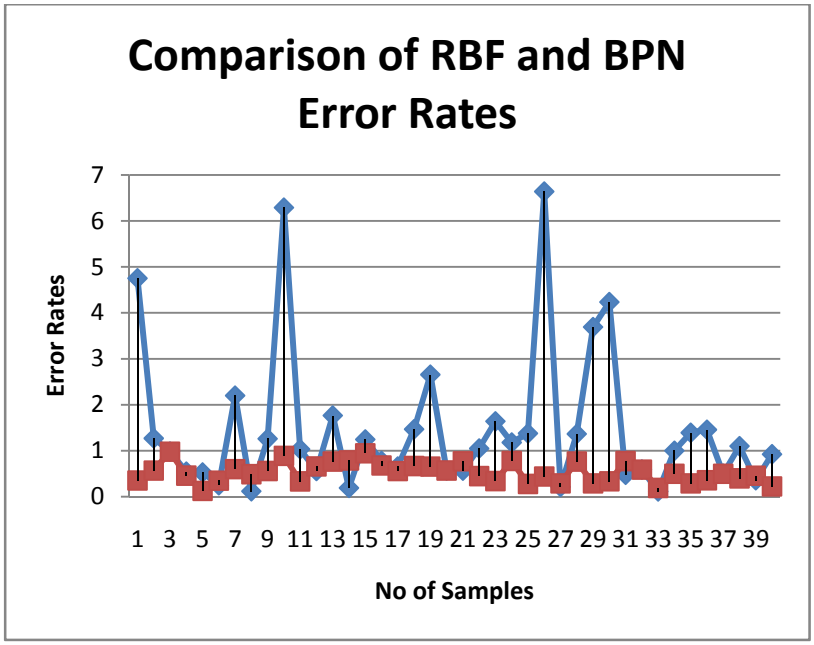

Figure 13

The above mentioned figure describes the difference in accuracy of emotion recognition using both the BPN and RBF Network. RBF classifies the Emotion more accurately than the BPN network.

IV. CONCLUSION

In this paper, the concept implemented was emotion recognition using MFCC approach using Radial basis function network. Support vector machine is used to classifying the gender in this work. Gender speech classifier is based on pitch analysis. MFCC approach for emotion recognition from speech is a stand-alone approach which does not require calculation of any other acoustic features and produce more accurate results. Hence proved that the Radial basis function network recognize emotions more accurately than the Back Propagation Network.

Table 7. Success Rate for Gender Classification

Category	Gender		
	Correctly Classified	Mis Classified	Success Rate
Male	88	12	88%
Female	87	13	87%

Table 8. Success Rate for Emotion Classification

Category	Emotion(RBF Network)			Emotion(BPN Network)		
	Correctly Classified	Mis Classified	Success Rate	Correctly Classified	Mis Classified	Success Rate
Male	89	11	89%	83	17	83%
Female	90	10	90%	80	20	80%

REFERENCES

[1] AasthaJoshi "Speech Emotion Recognition Using Combined Features of HMM & SVM Algorithm", National Conference on August 2013.

[2] AnkurSapra, Nikhil Panwar, SohanPanwar "Emotion Recognition from Speech", International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 2, pp. 341-345, February 2013.

[3] BjörnSchuller, Manfred Lang, Gerhard Rigoll "Automatic Emotion Recognition by the Speech Signal", National Journal on 2013, Volume 3, Issue 2, pp. 342-347.

[4] Chang-Hyun Park and Kwee-Bo Sim. "Emotion Recognition and Acoustic Analysis from Speech Signal" 0-7803-7898-9/03 Q2003 IEEE, International Journal on 2003, volume 3.

[5] Chao Wang and Stephanie Seneff "Robust Pitch Tracking For Prosodic modeling In Telephone Speech" National Conference on "Big data Analysis and Robotics" in 2003.

[6] Chiu Ying Lay, Ng Hian James. "Gender Classification from speech", (2005)

[7] Jason Weston "Support Vector Machine and Statistical Learning Theory", International Journal on August 2011, pp. 891-894.

[8] Keshi Dai1, Harriet J. Fell1, and Joel MacAuslan2"Recognizing Emotion In Speech Using Neural Networks", IEEE Conference on "Neural Networks and Emotion Recognition" in 2013.
[9] Margarita Kotti and Constantine Kotropoulos "Gender Classification In Two Emotional Speech Databases" IEEE Conference on 2004.
[10] Mohammed E. Hoque1, Mohammed Yeasin1, Max M. Louwerse2 "Robust Recognition of Emotion from Speech" , Internation Journal on October 2011, Volume 2, pp. 221-225.
[11] Nobuo Sato and YasunariObuchi. "Emotion Recognition using MFCC"s" Information and Media Technologies 2(3):835-848 (2007) reprinted from: Journal of Natural Language Processing 14(4): 83-96 (2007)
[12] Sony CSL Paris "The production and recognition of emotions in speech: features and algorithms" in September 2001.
[13] T L Nwe'; S W Foo L C De Silva, "Detection of Stress and Emotion in speech Using Traditional And FFT Based Log Energy Features" 0-7803-8185-8/03 2003 IEEE (2003)
[14] Webreference:http://emodb.bilderbar.info/start.htmlhttp://database.syntheticspeech.de http://sg.geocities.com/nghianja/CS5240.doc.
[15] Yixiong Pan, Peipei Shen and Liping Shen, "Speech Emotion Recognition Using Support Vector Machine" International Journal on 2012, Issue 3, pp. 654-659.