TRJ VOL. 2 ISSUE 3 MAY-JUNE 2016 ISSN: 2454-7301 (PRINT) | ISSN: 2454-4930 (ONLINE)
DOI: 10.13140/RG.2.1.4268.1844

To Design a Speech Emotion Recognition System Using Feed Forward Neural Network

Ajay Kumar[1], Er.Deepika Chaudhary[2],Er. Viney Dhawan[3]
[1]*Student (M.Tech), KITM, Kunjpura, Karnal*
[2]*Assistant Professor, KITM, Kunjpura, Karnal*
[3]*Head of Department, KITM, Kunjpura, Karnal*

Abstract-In human machine boundary application, emotion acknowledgment from the speech signal has been research topic since many years. To recognise the emotions from the speech signal, many systems have been developed. Humans have the natural capability to use all their accessible senses for maximum awareness of the received message. Through all the available senses people essentially sense the emotional state of their announcement partner. The emotional detection is natural for humans but it is very problematic task for mechanism. Therefore the purpose of emotion recognition system is to use emotion associated knowledge in such a way that social machine announcement will be improved. Speech emotion recognition is nothing but the pattern acknowledgement system. This shows that the stages that are present in the pattern recognition system are also contemporary in the Speech emotion recognition system. The speech emotion recognition system contains five main components emotional speech input, feature extraction, feature selection, classification, and recognized emotional output. In this research paper speech emotion recognition based on the previous technologies which uses different classifiers for the emotion gratitude is studied. The classifiers are used to differentiate emotions such as anger, happiness, sadness etc. The database for the speech emotion recognition system is the emotional speech samples and the features removed from these speech samples are the liveliness, pitch, Mel frequency cepstrum coefficient. The optimize performance is based on extracted features (MFCC). The classification i.e.Feed Forward Neural Network performance is based on reduction (ACO). Inference about the performance and limitation of speech emotion recognition system based on the different classifiers are also discussed.The evaluate performance parameters like mean square error rate, signal to noise ratio and accuracy.

Keywords*:* Speech Emotion Recognition, Dissimilar types of emotions, Ant Colony Optimization, Mel Frequency Cepstrum coefficient and Feed Forward Neural Network.

I. INTRODUCTION

Emotional [1] speech recognition aims at involuntarily identifying the emotional or physical situation of a person being via his or her tone. A speaker has dissimilar stages throughout speech that are renowned as emotional facet of speech& are integrated inthe so named paralinguistic aspects. The linguistic content cannot adapt by emotional condition; in communication of entity this is a significant factor, since reaction information is provide in frequent appliance. Speech is probably the generally proficient way tocorrespond with each other[2].

Speech recognition process is basically done by the Speech Recognition System. In the speech recognition process, speech input signal is processed into recognition of speech as a text form. Speech Recognition System helps the technology to bring computers and humans more closely. There is basic terminology that one must know in order to implement or develop a Speech Recognition System [3].

- Utterances- User input speech is called utterances, in simple words when user speaks something it is called utterances.
- Pronunciations- Single word has multiple meanings and multiple recognitions. It all depends on pronunciation. A single word is uttered in different means in accordance to country, age etc.

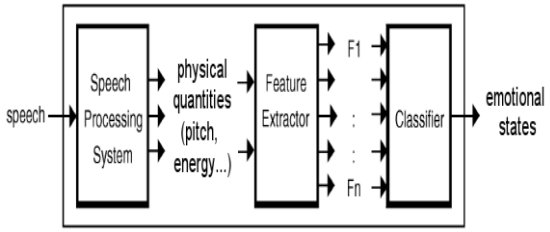

Fig.1: Speech Emotion Process

- Accuracy- It is the performance measurement tool. It is measured by number of means but in this case, if speaker utters "NO", then Speech Recognition System must recognise it as word "NO". If it is done precisely then accuracy of system is efficiently very good or else [4].

II. TYPES OF SPEECH

Separation of speech recognition scheme in different classes can be complete based on what type of utterance they have ability to recognize.

A. Isolated speech

Isolated phrase recognizer regularly set necessary condition that each sound having little or no noise on both sides of prototypical window. It requires particular utterance at a time. Often, these types of language have "Listen/Not-

DOI: 10.13140/RG.2.1.4268.1844

Listen states", where they require the speaker to have break between utterances. Remote word might be better name for this type [6].

B. Connected word

Connected word need [5] minimum pause among utterances to make speech flow smoothly. They are almost alike to isolated words.

C. Constant speech

Constant speech is basically computer's dictation. It is normal human speech, with no silent pauses between words. This kind of speech makes machine understanding much more difficult.

D. Spontaneous speech

Spontaneous words can be attention of as speech that is natural sounding and no tried out before. An ASR method with spontaneous speech capability should be able to handle a diversity of natural speech structures such as words being run at the same time.

III. APPLICATION OF SPEECH RECOGNITION

- In addition to having fine speech recognition technology, efficient speech based applications heavily depend on several factors, excluding [7]:
- Good user interface which make the submission easy-to-use and robust to the good models of dialogue that keep the discussion moving forward, even in similar the task to the technology.
- Kinds of confusion that arise in person- machine communications by tone.
- Periods of large uncertainty on the part [8].
- Remind teach users what can be held at any point in the contact.
- Maintain reliability across features using a vocabulary that is 'almost always available'.
- Design for mistake.
- Provide the capacity to barge-in over prompts.
- Use implicit confirmation of voice input.

IV. RELATED WORK

PengPeng, Qian-Li Ma, Lei-Ming Hong, (2009) [9] presented a novel technique for solving method of Support Vector Machine method that is SMO that is an equivalent algorithm. According to this algorithm, primitive training setsare dispensed by master CPU to slave CPUs. Slave CPU run serial SMO on the relevant training sets. As buffer&shrink methods are as well selected, increase in speed of the parallel training algorithm is done, which is represented in the results of parallel SMO based onthe dataset of MNIST. The results of this work proved that by using SMO performance of solving large scale SVM is good.Rong-En Fan, Pai-Hsuen Chen, Chih-Jen Lin, (2005) [10] presented a new algorithm for selection of working set in SMO type decomposition method. It discussed that in training supportvector machines (SVMs), selection of working set in decomposition process is important. Fast convergence is achieved by using information of second order. Theoretical properties such as linear convergence are established. It is proved in results that proposed method provided better results in contrast to existing collection methods using first orderdata.Xigao Shao, KunWu, and Bifeng Liao, (2013) [11] proposed an algorithm for selection of working set in SMO-type decomposition. It showed that in training element, least square support vector machinery (LS-SVMs) the selection of working set in decomposition process is important. In the proposed method a single direction is selected to achieve the convergence of the optimality condition Experimental results represented that speed of training is faster than others but classification accuracy is not better than existing ones, it's almost same with others. S. K. Shevade, S. S. Keerthi, C. Bhattacharyya, and K. R. K. Murthy, (2000) [12]proposed Smola and Schölkopf's sequential minimal optimization(SMO) algorithm have some source of inefficiency which is pointed out for regression of support vector machine (SVM) that occurs by the use of a single threshold value. The KKT conditions for the dual problem is used, SMO modification is done on the basis of two threshold parameters that are employed for regression. This proposed algorithm with the modification in SMO performs faster thanthe original SMO.

Ibrahim Patel, Dr. Y. SrinivasRao, (2010) [13]described that recognition of speech signal can be done using frequency spectral informationwith Mel frequency.HMM based recognition is used to increase the results of selected approach for recognition. Speech signal is observed using Mel frequency move towardin given resolution which fallout in resolution feature overlapping thus resulting in recognition limit. Mapping approach for HMM is resolution disintegration with separating frequency based on speech recognitionsystem. Results from simulation represent the improvement in the excellence metrics of speech recognition with admiration to computational time, also progress the accuracy for a speech recognition system.

DOI: 10.13140/RG.2.1.4268.1844

V. PROPOSED MODEL

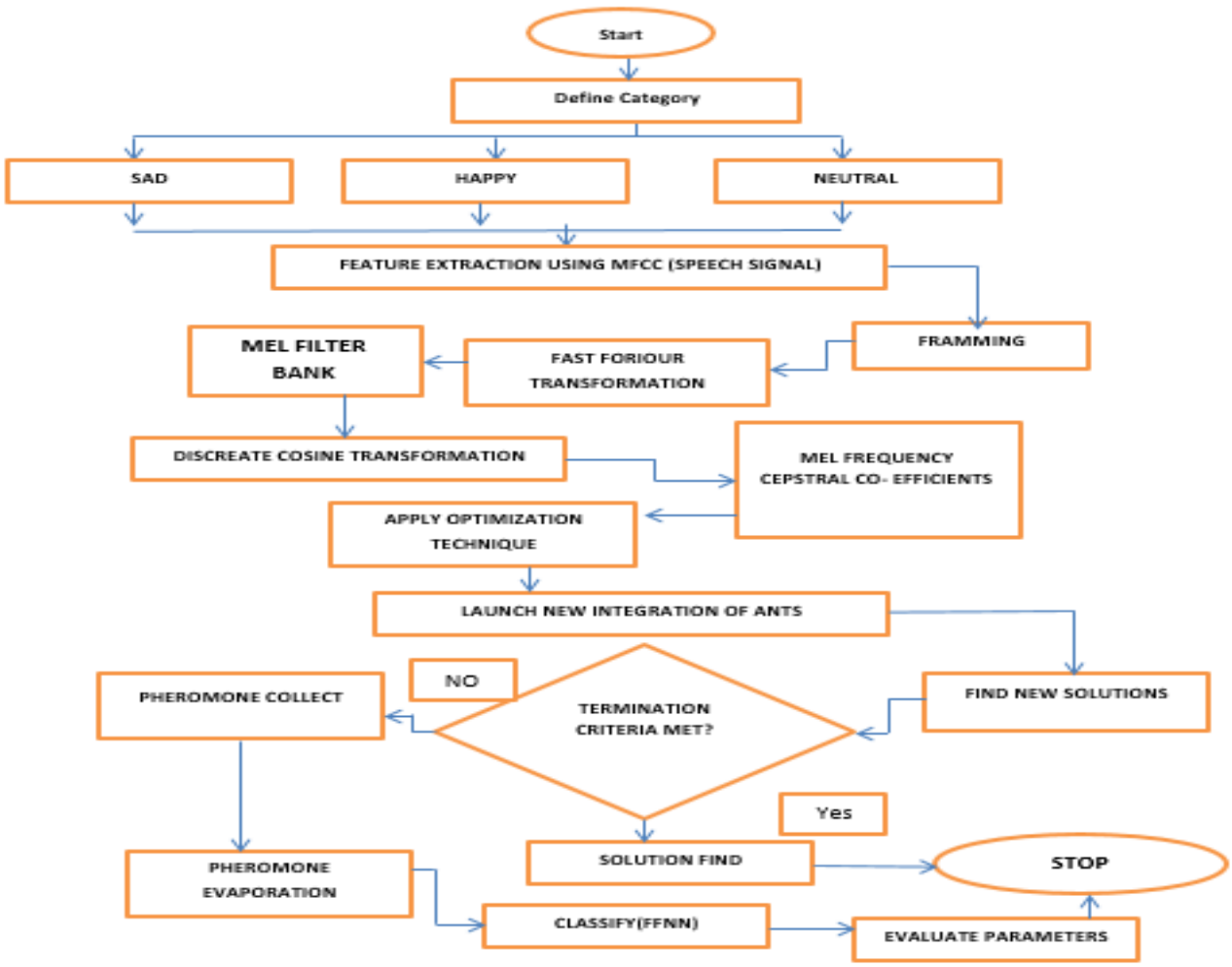

Fig.2: Proposed Flow Work

Steps in Proposed Model described below:

A. Upload the speech Categories

i) Sad
ii) Joy
iii) Aggressive

B. Apply for feature extraction using MFCC algorithm:

Steps:

i) Pre-Emphasis

The speech signal x(n) is sent to a high-pass filter :

$$y(n) = x(n) - a * x(n - 1) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where y(n) is the output signal and the value of a is usually between 0.9 and 1.0. The Z transform of this equation is given by:

$$H(z) = 1 - a * z -1 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

The goal of pre-emphasis is to compensate the high-frequency part that was suppressed during the sound production mechanism of humans. Moreover, it can also amplify the importance of high-frequency formants.

ii) Framing and blocking

In this step the permanent 1D signal are blocked into small frames of N samples, with next frames divided by M samples (M<N) with this the adjacent frames are overlap by N-M samples. As per many research the standard value taken for N= 256 and M= 100 with a reason of separating the given 1D signal into small frames having enough samples[9] to get enough information.

DOI: 10.13140/RG.2.1.4268.1844

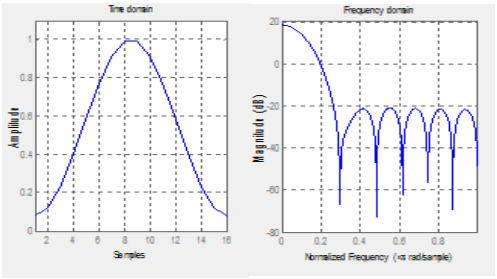

Fig.3: Time and Frequency Domain in line pattern

Because, if the frame size lesser than this size is taken then the number of samples in the frames will not be sufficient to get the reliable information and with large size frame it can cause frequent change in order inside the frame.

iii)Windowing

Windowing is done for minimize the disruption at the starting and at the end of the frame, the frame and window function is being multiply. If the window being defined is

W_n (m), 0 _ m _ N_m-1

where Nm stands for the amount of samples within every surround, the production after windowing the signal will be presented as

Y (m) = X (m) W_n (m), 0 _ m _ N_m -1

where Y (m) denote the output signal after increase the input signal represent as (X_m) and Hamming window signify by W_n (m). Essentially, many window functions exist such as rectangular window, even top window and hamming window other than, mostly hamming window is applied for carrying out windowing which typically represented as:

W_n (m) = 0.54 – 0.46

Cos (2_m/ (N_m-1)), 0 _ m_ N_m -1……(1)

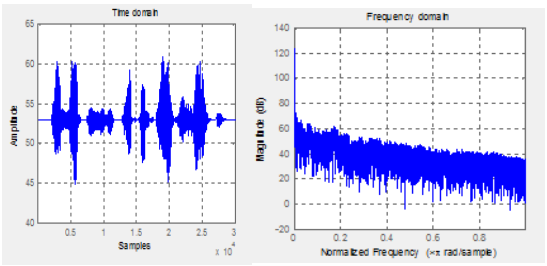

Fig.4: Time and Frequency Domain in Spectrum of the Speech

iv) Fast Fourier Transform

Spectral analysis shows that unlike timbres in speech signals parallels to different energy circulation over frequencies. Therefore FFT is achieved to obtain the magnitude frequency answer of each frame. When FFT is achieved on a frame, it is supposed that the signal within a frame is periodic, and nonstop when wrapping around. If this is not the case, FFT can still be achieved but the discontinuity at the frame's first and last points is likely to familiarise undesirable effects in the incidence response. To deal with this problem, we multiply each frame by a hamming window to increase its continuousness at the first and last points.

v) Triangular Bandpass filters

The magnitude incidence response is multiplied by a set of 40 three-sided band pass filters to get the log energy of each triangular band pass filter. The positions of these filters are correspondingly spaced along the Mel frequency. From centre frequencies from 133.33 Hz to 1 kHz, there are 13 touching (50%) linear filters, while for centre frequencies from 1 kHz to 8 kHz there are 27 overlapping filters spread out logarithmically.

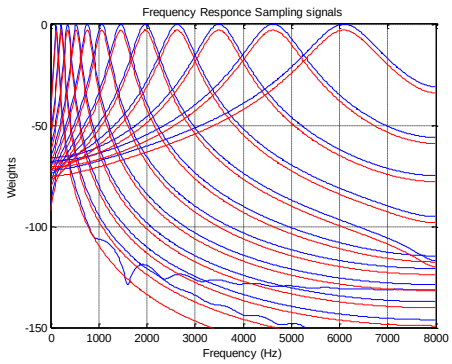

Fig.5: Frequency Response Sampling Signals

vi) Discrete Fourier Transform

In this step, DCT is applied to the output of the N triangular bandpass filters to obtain L mel-scale cepstral coefficients. The formula for DCT is,

$$C(n) = \sum Ek * \cos(n * (k - 0.5) * \pi/40)) \quad \ldots\ldots(5)$$

Where n = 0,1,..to N.

Where N is the number of triangular bandpass filters, L is the number of mel-scale cepstral coefficients. In this project, there are

N = 40 and L = 13.

Since we have performed FFT, DCT transforms the frequency domain into a time-like domain called quefrency domain. The found features are similar to cepstrum, thus it is referred to as the mel-scale cepstral numbers, or MFCC. MFCC alone can be used as the feature for speech recognition.

C. Ant Colony Optimization Technique

It is motivated by social insects, such as ants and termites, or previous animal societies, such as fish train& bird flocks. Although each separate has only limited capabilities, the whole swarm exhibitions complex on the wholeactivities. Therefore, the intelligent behaviour can be seen as an emergent unique of the swarm. When spotlight on ant colonies, it can be perceived that ants communicate only in anindirect method through their surroundings by dropping a material called pheromone.

DOI: 10.13140/RG.2.1.4268.1844

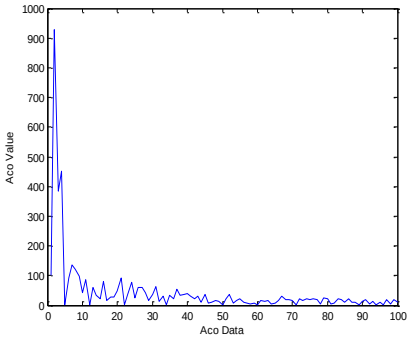

Fig.6: Ant Colony Optimization in Speech Emotion

Paths with higher pheromone levels will more possible be preferred & thus reinforced, while the pheromone intensity of pathways that are not chosen is decreased by desertion. This form of not direct statement is identified as stigmergy, and offers the ant colony shortest-path finding capacity.ACO employs imitation ants that work together to find good solutionsfor discrete optimization complications. These software agents mimic the foraging performance of their biological complementsin finding the shortest-path to the food source.

D. Feed Forward Neural Network

This technique applies for classification using Feed Forward Neural Network. This is creating the two module like training and testing module. Feed Forward Neural Network is an organically stimulated organization algorithm. It consists of amount of simple neuron like processing units, arranged in layers. Each unit in a layer is linked with all the units in the preceding layer [10]. These connections are not all equal: each joining may have a dissimilar strength or weight. The weights on these contacts encode the information of a network. Frequently the units in a neural network are also called nodes.

VI. RESULT ANALYSIS

The subsequent Development Tools has been used in the expansion of this work. There may also be other tools which can be used in this development as it depends individual to person and his interest. Therefore the used tools are : i) least amount of 3 GB of RAM ii) Intel Pentium III Processor or over and iii) MATLAB R2013a.

Table no: 1 Proposed Work:

Speech Category	MSE	FAR	FRR	Accuracy	SNR
Sad	1.345	0.000244	0.0003234	99.23	70.334
Aggressive	1.2423	0.000456	0.000123	98.89	70.123
Joy	1.1403	0.00035068	0.00011039	99.09	71.1192

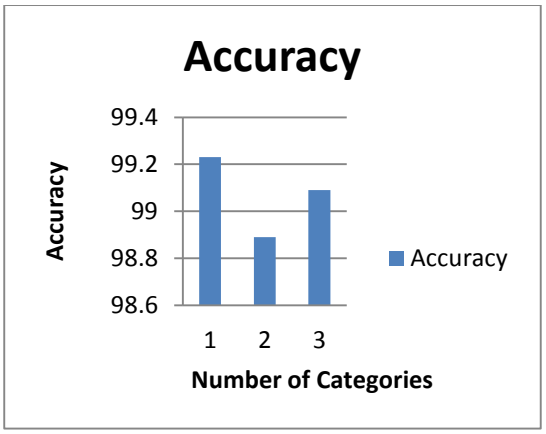

Fig.7: Accuracy

Figure defines the accuracy description of the random errors, measures of statistical variability. Accuracy is how nearby a measured value is to the actual (true) value. Above figure shows the accuracy value for proposed method and it has been clearly seen that accuracy for proposed method is good.

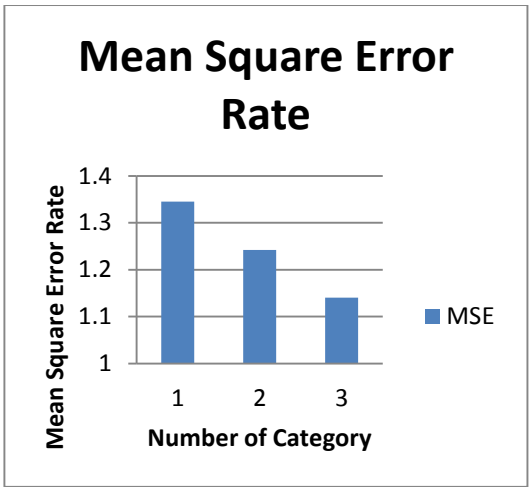

Fig.8: Mean Square Error in different category

Figure shows that, mean square error the computed average of percentage errors by which estimates of a model differ from real values of the quantity being forecast.

DOI: 10.13140/RG.2.1.4268.1844

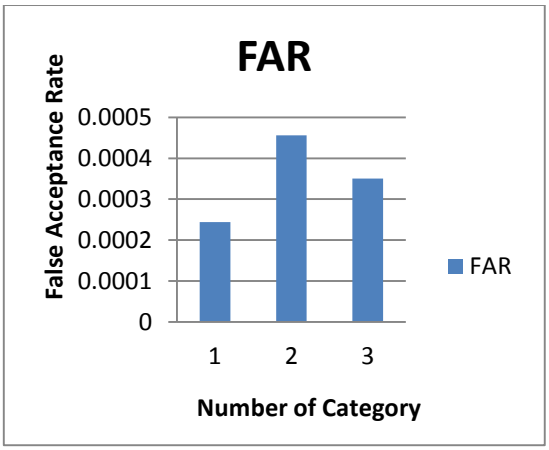

Fig.9: False Acceptance Rate

A system's FAR characteristically is stated as the ratio of the number of untruthful acceptances divided by the number of identification attempts. Same here, we plot a graph which uses the FAR parameter for the proposed approach.

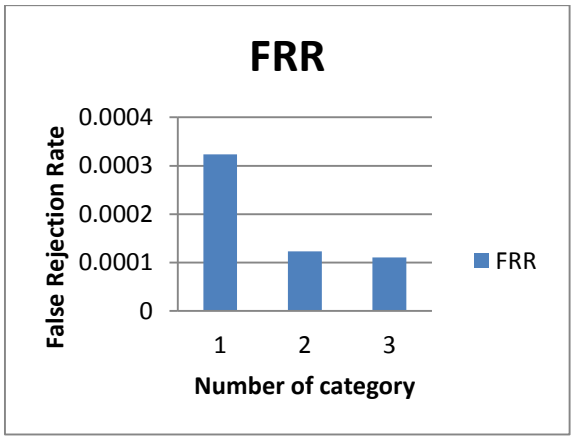

Fig.10: False Rejection Rate

A system's FRR typically is stated as the ratio of the number of false rejections divided by the number of identification attempts. Above figure shows the rate of FRR for proposed approach.

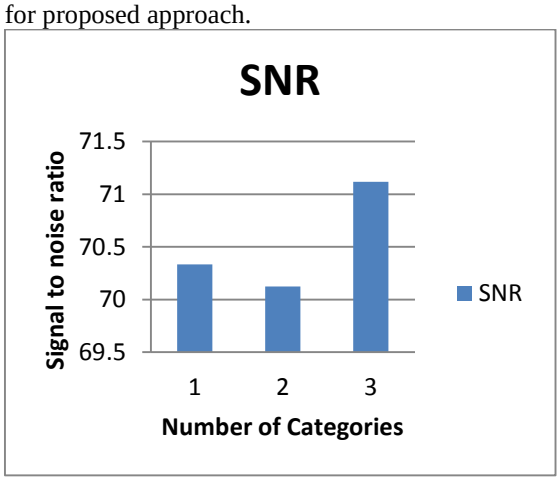

Fig.11: Signal To Noise Ratio

Figure defines that the, Generally, it is the ratio of signal magnitude to thermal noise for the signal bandwidth which you are examining. Bit error rate is a measure of the errors one gets over time for a given digital signal.

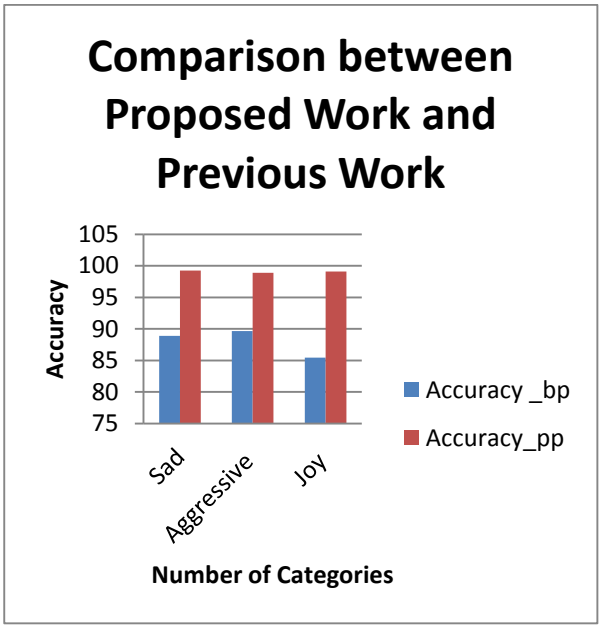

Fig.12: Comparison Between Proposed Work and Existing Work

Above figure define that the comparison between proposed work between previous work. We achieved the accuracy in proposed work is 99.89 and previous work is 88.88 value.

Table no: 2 Comparisons between Proposed Work and Base Work

Speech Category	Accuracy _bp	Accuracy_pp
Sad	88.88	99.23
Aggressive	89.67	98.89
Joy	85.45	99.09

VII. CONCLUSION AND FUTURE SCOPE

The goal of this paper was to give an inclusive survey on the geographies and the classification techniques complex in speech emotion detection. From the training it was seen that features like pitch, extent, MFCC and ACO optimized features are used to detect the emotions transported through speech. It was also seen that classifiers like BPNN and comparison between ANN have been expansively studied for their efficiency in classifying emotions.We are at an evolution point where voice and natural-language considerate are at the forefront. Customer care enters already have cultured systems that help pick up Irate customers and transfer their calls to customer service congresses. From the examination done we can understand that in most studies showed, the speech recognition model was built using either produced data base, where specialized and non-professional speakers were asked to

DOI: 10.13140/RG.2.1.4268.1844

speak as how they would talk under stress or recordings from real-time emergency circumstances where people are naturally strained. Above all, these models are not being realized effectively as it should be or in a way that can be of provision to common people.A future work should integrate an improved segmentation unit by employing fuzzy logic and neural networks so as to have a better classification of voiced and unvoiced segments and hence help fetch better features and lead to improvement parameters. Using a hybrid mixed classification collective can help in enhancement of performance parameters.

VIII. REFERENCES

[1]http://www.informatik.uniaugsburg.de/lehrstuehle/hcm/projects/tools/emovoice/

[2]http://en.wikipedia.org/wiki/Speech_recognition

[3]Schuller, Björn, Gerhard Rigoll, and Manfred Lang. "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture." Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on.Vol. 1.IEEE, 2004.

[4]Ververidis, Dimitrios, and Constantine Kotropoulos. "Emotional speech recognition: Resources, features, and methods." Speech communication48.9 (2006): 1162-1181.

[5]de Krom, Guus. "Consistency and reliability of voice quality ratings for different types of speech fragments." Journal of Speech, Language, and Hearing Research 37.5 (1994): 985-1000.

[6]Schuller, Björn, Gerhard Rigoll, and Manfred Lang. "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture." Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on.Vol. 1.IEEE, 2004.

[7]Hansen, John HL, and Mark A. Clements. "Constrained iterative speech enhancement with application to speech recognition." IEEE Transactions on Signal Processing 39.4 (1991): 795-805.

[8]S. K. Shevade, S. S. Keerthi, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to the SMO Algorithm for SVM Regression", IEEE Transactions on Neural Networks, Vol. 11, No. 5, September 2000.

[9] PengPeng, Qian-Li Ma, Lei-Ming Hong, "The Research Of The Parallel Smo Algorithm For Solving Svm", Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, 12-15 July 2009.

[10] Rong-En Fan, Pai-HsuenChen,Chih-Jen Lin, "Working Set Selection Using Second Order Informationfor Training Support Vector Machines", Journal of Machine Learning Research 6 , pp.1889–1918, 2005.

[11] Xigao Shao, KunWu, and Bifeng Liao, "Single Directional SMO Algorithm for Least Squares SupportVector Machines", Computational Intelligence and Neuroscience, Article ID 968438. 2013.

[12] PengPeng, Qian-Li Ma, Lei-Ming Hong, "The Research Of The Parallel Smo Algorithm For Solving Svm", Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, 12-15 July 2009.

[13] Rong-En Fan, Pai-HsuenChen,Chih-Jen Lin, "Working Set Selection Using Second Order Informationfor Training Support Vector Machines", Journal of Machine Learning Research 6 , pp.1889–1918, 2005.