EMOTION DETECTION FROM SPEECH

SANJEEV N: 13EE144, DHANVINI GUDI: 13EE213, K SHASHANK HEGDE: 13EE224

GROUP 2

Abstract: During expressive speech, the voice is enriched to convey not only the intended semantic message but also the emotional state of the speaker. The intonation, tone, timing and energy of speech are all jointly influenced in a non trivial manner to express emotional message. The pitch contour (e.g., curvature, contour, shape and dynamics) is one of the important properties of speech that is affected by this emotional modulation. Analyzing this pitch, in turn the frequencies of different words and messages conveyed will help us in classifying it into a simple binary bifurcation indicating is the message conveyed has certain emotional attributes to it, or if it's simply a message conveyed. This project will be confined to this basic bifurcation as the boundaries between different emotions are blurred as they vary from one individual to another.

I. Introduction

Emotions play an important role in every-day life of human beings. Our social behaviour is quintessentially based on communication, both in terms of language and non-verbal interaction. Information from the environment determines individuals to mutually interpret other person's intentions, goals, thoughts, feelings and emotions and to change the behaviour accordingly. The term emotion stands for a concept that has been proved difficult to define. As claimed by the famous naturalist Charles Darwin, emotions emerged in the course of evolution as the means by which living creatures determine the significance of certain conditions to meet their urgent needs. The more developed and complex life organization is, the richer all sorts of emotion states developed by people are [2].

Different approaches for analysing emotions focused on essentially biological reactions through interpretive conventions with little or no biological contributions. The emotions most commonly acknowledged as basic are sadness, anger, fear, happiness, surprise and disgust. These are considered to be primary emotions and are the most studied categories in the literature. The more complex emotion categories could be represented by cultural conditioning or association combined with the prototypic forms of emotion [1].

In spoken dialogue research, it is beneficial to enable the systems not only to recognize the content encoded in user's response, but also to extract information about the emotional state of the user by analyzing how these responses have been spoken. Since we tackle the problem of speech emotion recognition as a pattern recognition task, we follow in broad lines the following approach:

- Consider an emotional model (e.g., discrete or continuous)
- Start analyzing one or more of the available speech emotion databases
- Extract a set of features
- Train a classifier in order to make statements on the test data

Each of these steps is actually a point where a decision needs to be made. The first two problems can be regarded as a whole, since there are not many available databases, so the emotional model in most cases will be the one used for the recording of the used database [4].

II. Emotional Corpus

The database used in this experiment is the EmoDB - German database of emotional speech. EmoDB is a recorded database of emotional utterances spoken by actors (i.e., simulated speech utterances). It is developed by the Technical University, Institute for Speech and Communication, Department of Communication Science, Berlin. This database contains recordings, sampled at 16 KHz, from 5 actors and 5 actresses, 10 different sentences of 7 kinds of emotions: anger, boredom, disgust, fear, happiness, sadness, and neutral [2][3][4].

Guide to Data Set:

Every utterance is named according to the same scheme

- Positions 1-2: number of speaker
- Positions 3-5: code for text
- Position 6: emotion (sorry, letter stands for German emotion word)
- Position 7: if there are more than two versions these are numbered a, b, c....

Example: 03a01Fa.wav is the audio file from speaker 03 speaking text a01 with the emotion "Freude" (Happiness).

Information about the speakers

- 03 - male, 31 years old
- 08 - female, 34 years
- 09 - female, 21 years
- 10 - male, 32 years
- 11 - male, 26 years
- 12 - male, 30 years
- 13 - female, 32 years
- 14 - female, 35 years
- 15 - male, 25 years
- 16 - female, 31 years

Code of Texts

Code	Text (German)	Translation
a01	Der lappen legit auf dem Eisschrank	The tablecloth is lying on the fridge
a02	Das will sie am Mittwoch abgeben	She will hand it on Wednesday
a04	Heute abend könnte ich es ihm sagen	Tonight I could tell him
a05	Das schwarze Stück Papier befindet sich da oben neben dem Holzstück	The black sheet of paper is located up there besides the piece of timber
a07	In sieben Stunden wird es soweit sein	In seven hours it will be
b01	Was sind den das für Tüten,die da unter dem Tisch stehen?	What about the bags standing there under the table?
b02	Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter	They just carried it upstairs and now they are going down again
b03	An den Wochenenden din ich jetzt immer nach Hause gefahren und habe Agnes besucht	Currently at the weekends I always went and saw Agnes
b09	Ich will das eben wegbringen und dann mit Karl was trinken gehen	I will just discard this and then go for a drink with Karl
b10	Die wird auf dem Platz sein, wo wir sie immer hinlegen	It will be in the place where we always store it

Code of emotions:

Letter	Emotion (German)	Letter	Emotion (English)
W	Ärger (wut)	A	Anger
L	Langeweil	B	Boredom
E	Ekel	D	Disgust
A	Angst	F	Fear
F	Freude	H	Happiness
T	Trauer	S	Sadness
N	Neutral	N	Neutral

III. Speech & Spectral Features

An important task in emotion classification from speech is to get a clear expression of emotions in the feature space. That is, several features should be selected according to their discriminatory capabilities. Since speech signal is time-varying, the analysis should be a time-frequency analysis.

- **Pitch Signal**: Pitch is an auditory perceptual property that allows the ordering of sounds on a frequency-related scale. Pitch may be quantified as a frequency, but pitch is not a purely objective physical property; it is a subjective psycho-acoustical attribute of sound. A pitch detector is an essential component in a variety of speech processing systems and provides necessary information about the nature of the excitation source for speech coding. The pitch contour of an utterance is useful for recognizing speakers, determination of their emotion state, for voice activity detection task, and many other applications.
- **Loudness:** Loudness is the quality of a sound that is primarily a psychological correlate of physical strength (amplitude). More formally, it is defined as ‘‘that attribute of auditory sensation in terms of which sounds can be ordered on a scale extending from quiet to loud’’. Sound energy is perceived as loudness and is related to emotional intensity.
- **Formants:** The formants are one of the quantitative characteristics of the vocal-tract. In the frequency domain, the location of vocal tract resonances depends upon the shape and the physical dimensions of the vocal tract. Since the resonances tend to ‘‘form’’ the overall spectrum, speech scientists refer to them as formants. A simple method to estimate formants relies on linear predictive analysis.
- **Mel-frequency cepstral coefficients:** Mel-frequency cepstral coefficients (MFCC) are coefficients that collectively make up the mel-frequency cepstrum. The mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. The mel scale is a perceptual scale of pitches judged by listeners to be equal in distance from one another [1].

IV. Feature Extraction

Mel-Frequency Cepstrum coefficients is the most important feature of speech it widely used spectral feature for speech recognition and speech emotion recognition which provides ease of calculation , reduction in noise , having better capability to distinguish. MFCC having high recognition rate and having low frequency region has a good frequency resolution. MFCC is based on the characteristics of the human ear's hearing & perception, which uses a nonlinear frequency unit to simulate the human auditory system [5]. The main steps in extracting the MFCC coefficients are as follows:

- Frame the signal into short frames. 25ms is standard. The next steps are applied to every single frame; one set of 12 MFCC coefficients is extracted for each frame. A short aside on notation: we call our time domain signal **s (n)** and framed signal as $\mathbf{s_i}$ **(n).**
- When we calculate the complex DFT, we get $\mathbf{S_i(k)}$ - where the **i** denotes the frame number corresponding to the time-domain frame. $\mathbf{P_i}$ **(k)** is then the power spectrum of frame **i**.

$$S_i = \sum_{n=1}^{N} \mathbf{s_i(n)h(n)}\, \mathbf{e}^{\frac{-\mathbf{j2\pi kn}}{\mathbf{N}}};\ \mathbf{1 \le k \le K}$$

where **h(n)** is an **N** sample long analysis window (e.g. hamming window), and **K** is the length of the DFT. The periodogram-based power spectral estimate for the speech frame $\mathbf{s_i(n)}$ is given by:

$$P_i(k) = \frac{1}{N} |S_i(k)|^2$$

- Apply the mel filterbank to the power spectra, sum the energy in each filter. This is a set of 20-40 (26 is standard) triangular filters that we apply to the periodogram power spectral estimate. To calculate filterbank energies we multiply each filterbank with the power spectrum, and then add up the coefficients. Once this is performed we are left with 26 numbers that give us an indication of how much energy was in each filterbank.
- Take the logarithm of all filterbank energies.
- Take the Discrete Cosine Transform (DCT) of the 26 log filterbank energies to give 26 cepstral coefficients.

V. Qualitative Analysis of Emotional Speech

Speech signals generally are composed of the content and the emotional aspects which tinge the content. Broadly, emotional speech is classified into a few basic classes. A qualitative comparison of the emotional features of speech signals is done on the basis of features such as spectrogram, pitch, intensity and formants to analyze visually, the variations of these features with different emotional states and different conditions.

The data set was worked with 10 different sentences, each spoken by 10 speakers, composed of both male and female speakers in different emotional states such as happiness, fear, anger, sadness and neutral. These speech samples were analyzed based on the above features discussed and were sought to be classified according to the emotional state perceived.

One particular sentence was considered and the spectrograms of 5 different emotions, namely, fear, sadness, happiness, anger and neutral tone were plotted. The pitch and intensity plots are plotted along with the spectrogram in blue and yellow lines respectively. To analyze the qualitative aspect of emotion, Praat was used. Praat is a free scientific computer software package used mainly in the analysis of speech in phonetics. The below spectrograms show different emotions for the recordings in “03a05” series.

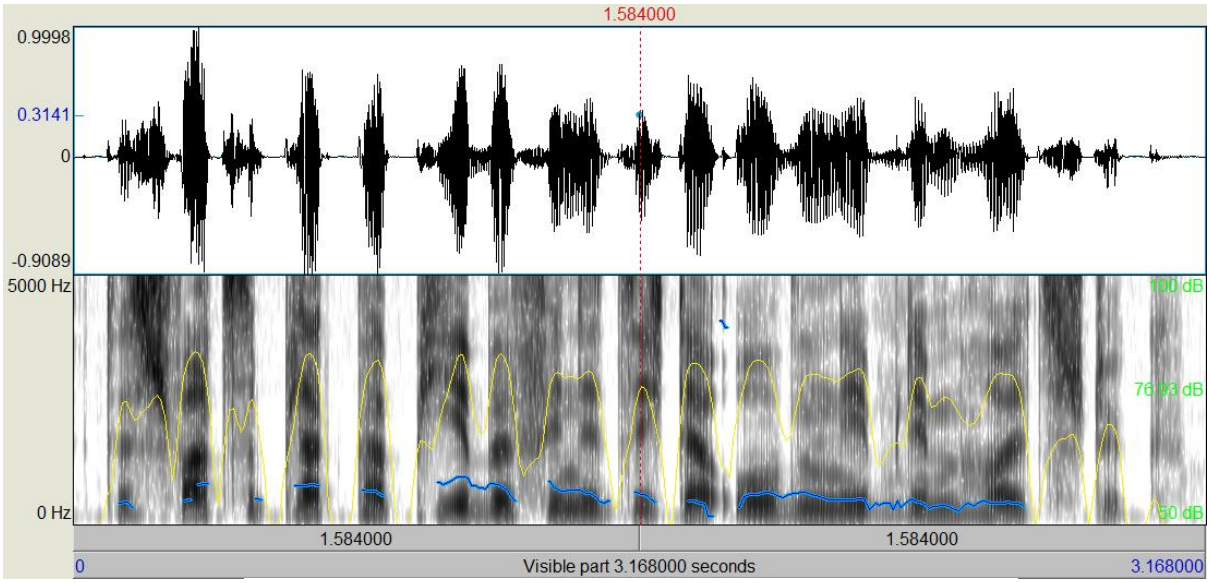

Fig 1a: Pitch and Intensity plot for Neutral or No emotion
Blue Line: Pitch; Yellow Line: Intensity

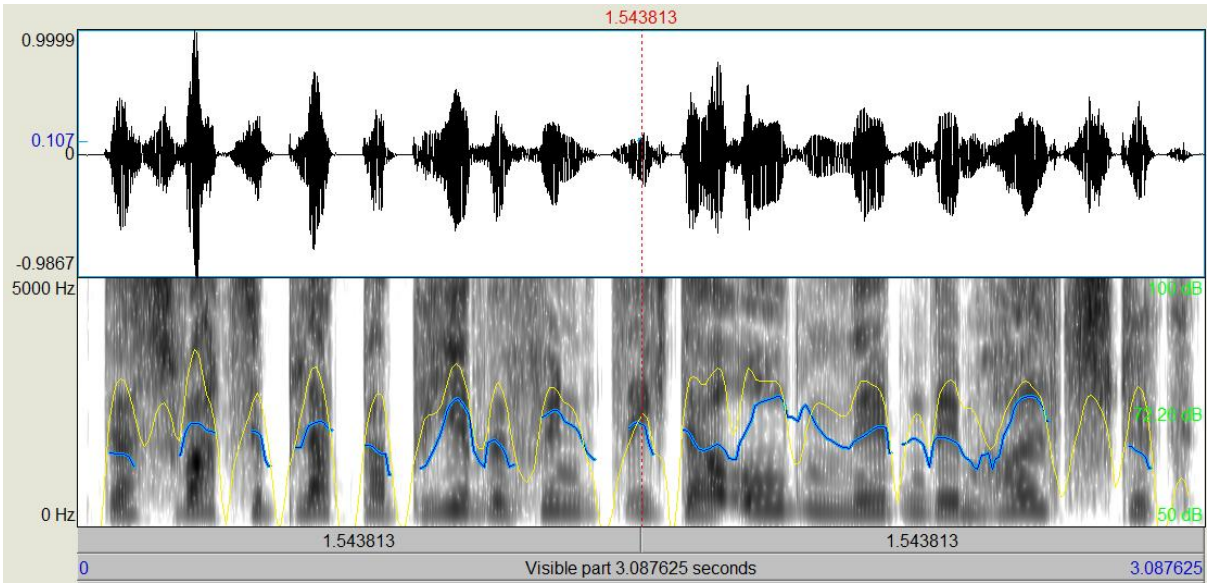

Fig 1b: Pitch and Intensity plot for the emotion Fear
Blue Line: Pitch; Yellow Line: Intensity

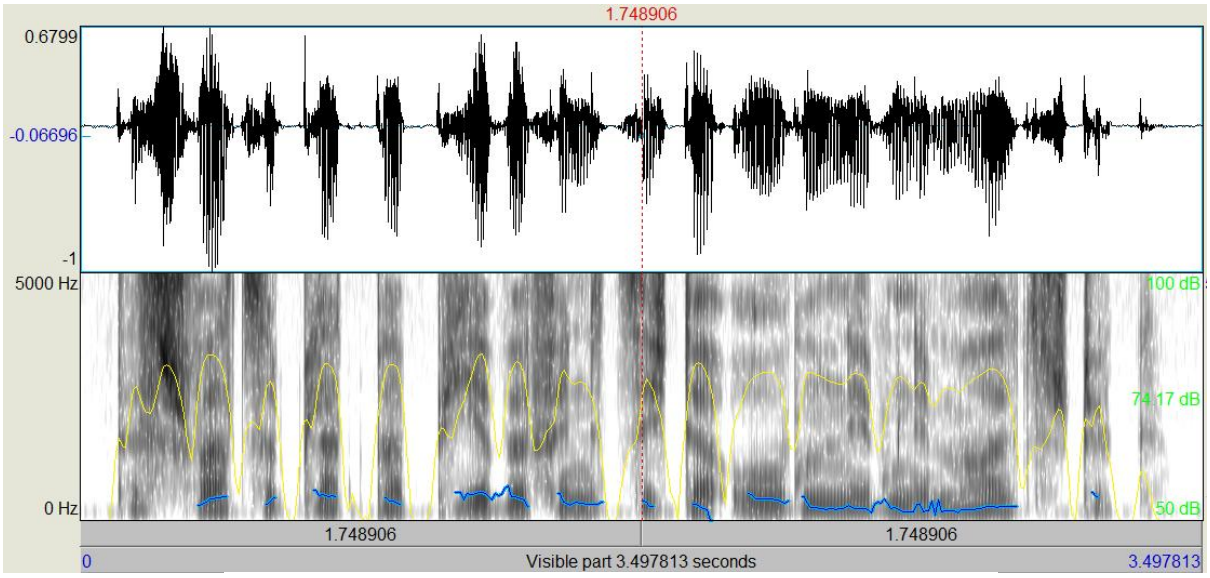

Fig 1c: Pitch and Intensity plot for the emotion Sadness
Blue Line: Pitch; Yellow Line: Intensity

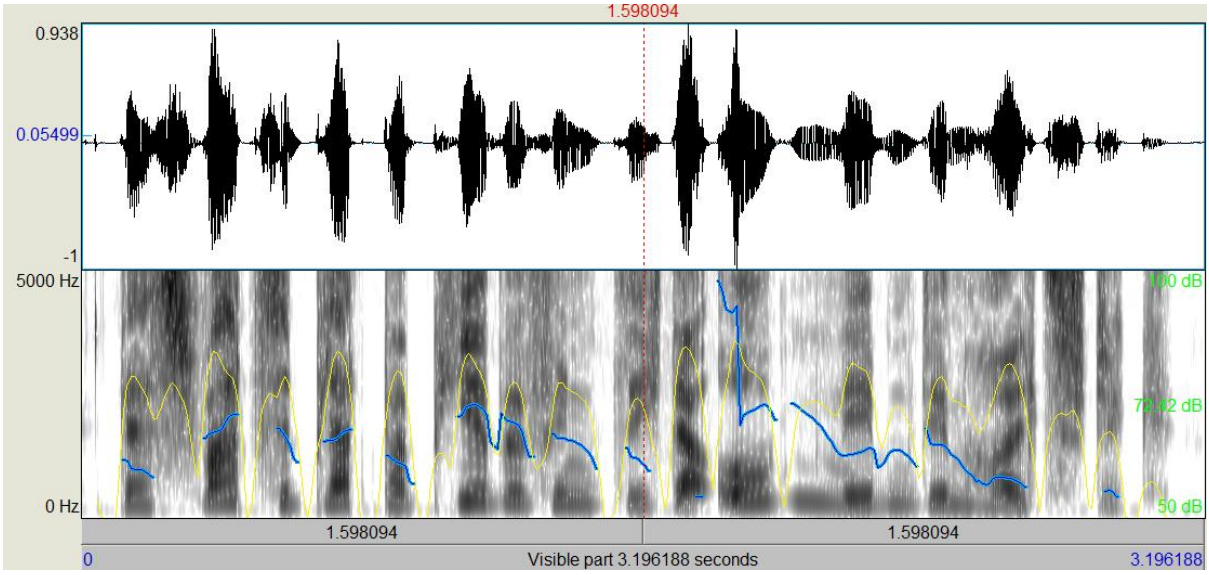

Fig 1d: Pitch and Intensity plot for the emotion Happiness
Blue Line: Pitch; Yellow Line: Intensity

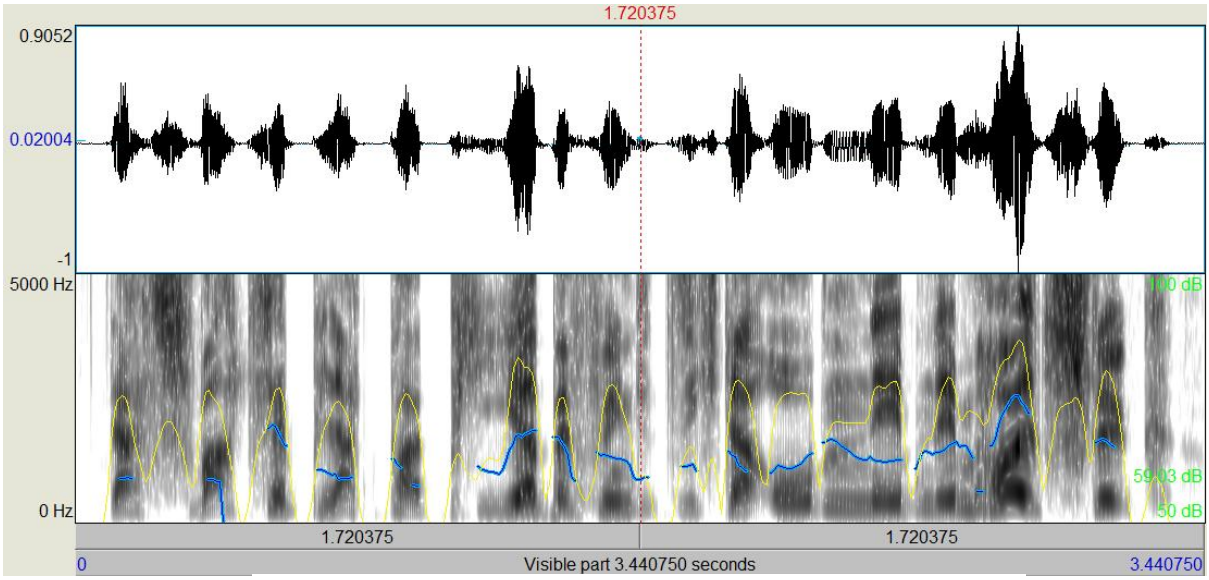

Fig 1e: Pitch and Intensity plot for the emotion Anger
Blue Line: Pitch; Yellow Line: Intensity

Intensity of the above signals can be analyzed in all the above cases. It is observed that the intensity graph is very similar for both neutral and sad speech samples. However, for fear, happiness and anger, the intensity plots vary a little more than neutral intensities, with happiness and anger having quick changing and slightly higher intensities than the other speech signals.

Furthermore, it is observed that sad speech has the lowest pitch, with neutral speech signal having similar pitch, but of a higher frequency. Both these emotions have similar looking roughly single frequency plots. The pitch in the above two cases is much lower than that of the other emotions. Fear has a higher pitch pattern than neutral speech. Angry speech has a lot of higher and lower pitches as compared to the neutral, and there are a lot more variations. Happy speech has the most discontinuous looking pitch patterns, with a huge rang of pitch frequencies.

Formant Analysis:

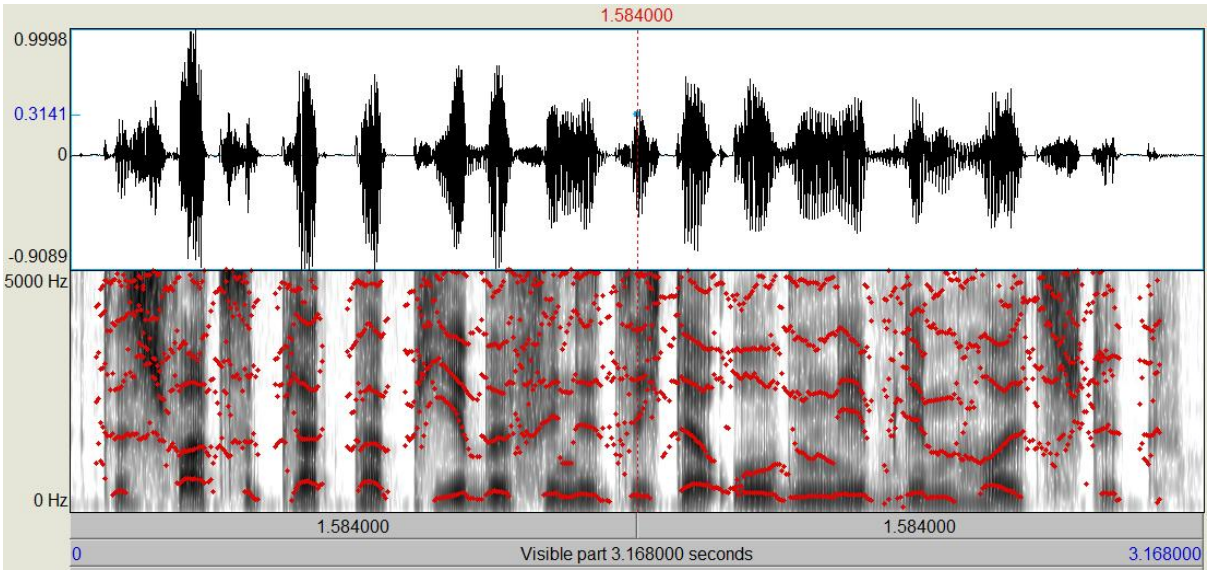

Fig 2a: Spectrogram for Neutral or No emotion
Scatter points in red denotes the energy level of the formants

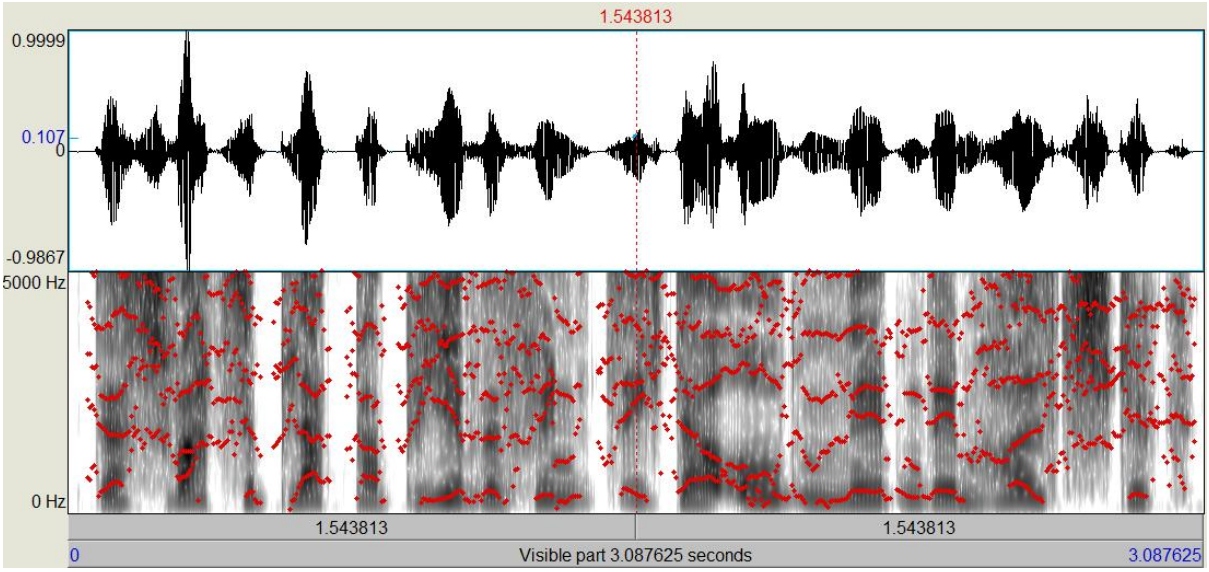

Fig 2b: Pitch and Intensity plot for the emotion Fear
Scatter points in red denotes the energy level of the formants

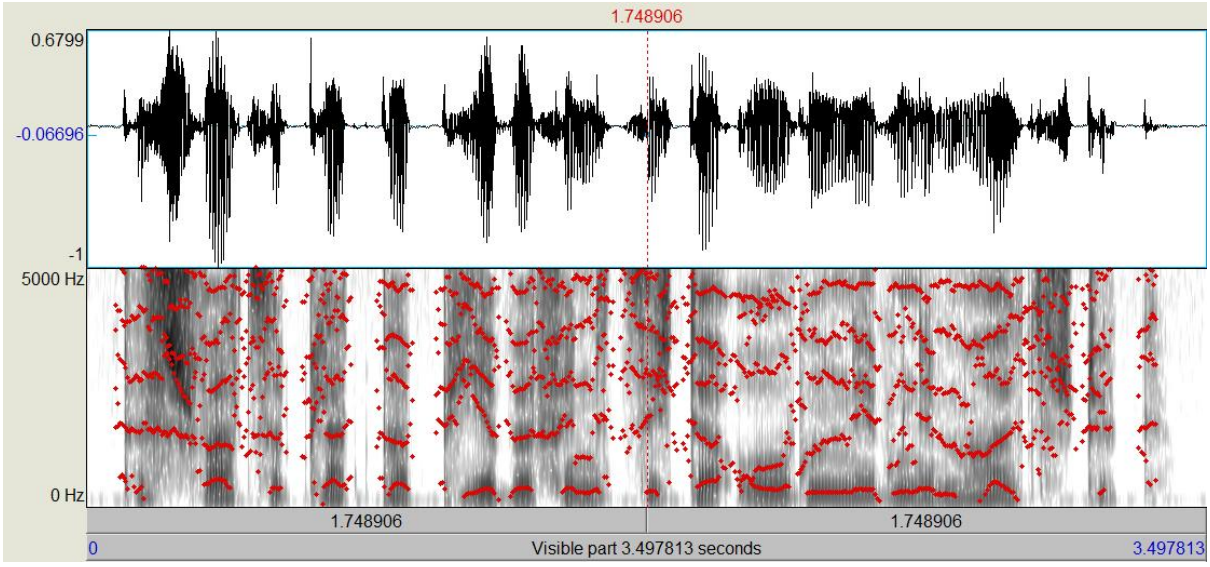

Fig 2c: Pitch and Intensity plot for the emotion Sadness
Scatter points in red denotes the energy level of the formants

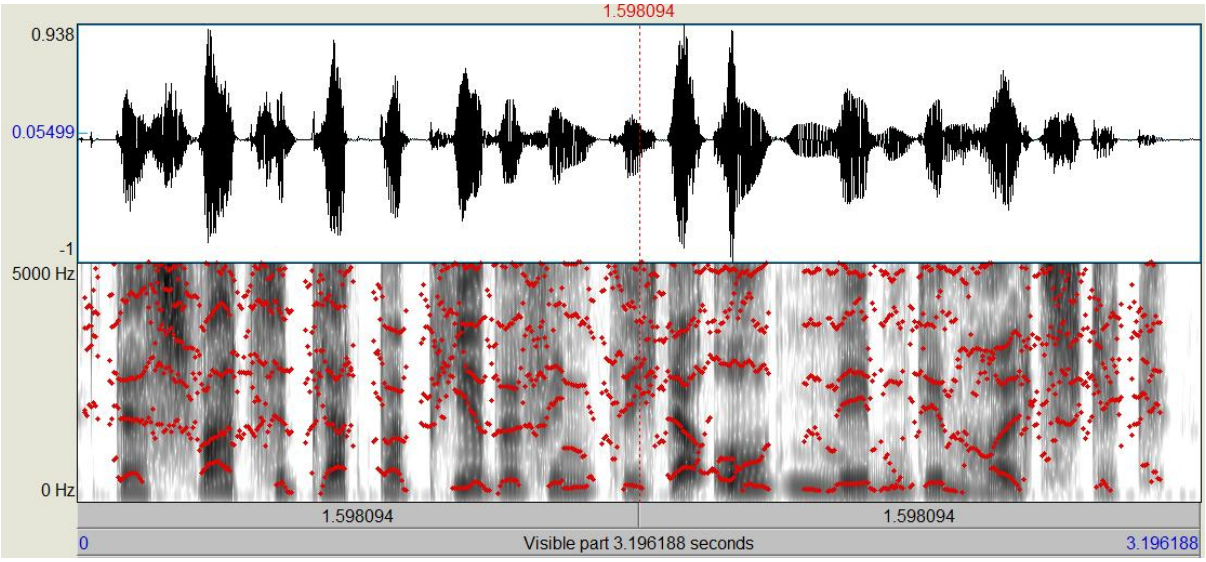

Fig 2d: Pitch and Intensity plot for the emotion Happiness
Scatter points in red denotes the energy level of the formants

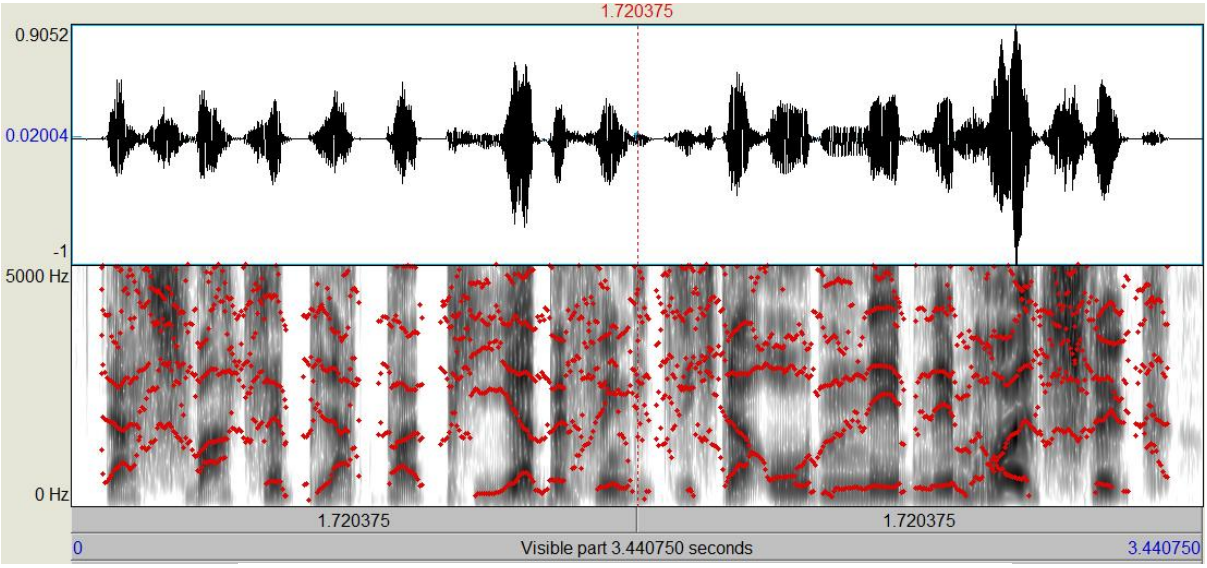

Fig 2e: Pitch and Intensity plot for the emotion Anger
Scatter points in red denotes the energy level of the formants

From the above plots, it is easily observed that the energy of speech along with formant features. It is seen that for happiness and anger, the formant features are more scattered and found more along the higher frequency spectrums when compared to sad and fearful speech signals, which have a more continuous formant band evenly spread out. Neutral speech has a continuous band like structure in lower frequencies and discontinuous points at higher frequencies. On analyzing the density of the spectrogram, it is observed that happiness and anger have the highest density, ie, the darkest spectrogram plots. This implies that the energy contents in angry and happy speeches are much higher than that of neutral speech. The least energy is observed to be seen in the spectrogram of sad speech, which has a very light spectrogram plot.

VI. Artificial Neural Network Model

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. A trained neural network can be thought of as an "expert" in the category of information it has been given to analyses. An artificial neuron is a device with many inputs and one output. The neuron has two modes of operation; the training mode and the using mode. In the training mode, the neuron can be trained to fire (or not), for particular input patterns. In the using mode, when a taught input pattern is detected at the input, its associated output becomes the current output.

In the speech data shown previously, anger, disgust, boredom, happiness, sadness, fear and emotionless speech are the data set used. To draw a line between classes like boredom and sadness is rather tricky and is out of the scope of this experiment. But there is a very well defined boundary between emotions like anger and emotionless speech. The difference is noticeable not only to the human ear but is also shown in the following ML system. Given that the number of sample in each of these classes is nearly comparable, and that there are only two classes, using a Neural Network classifier is a prudent choice.

The methodology of this work is quite simple. Data base has to be trained for different categories. As explained above, features will be extracted for the wave files. The extracted features will be saved to the database for every category processed. The neural network works on two scenarios namely the training and the testing part. The integrated neural network of MATLAB needs to get a target set, the target set will be helpful in indentifying what exactly the files are. Then randomly a file will be uploaded to test the training scenario of the neural network. The neural network will intake the features of the uploaded file and would set the targets of the value provided earlier. It would classify the wave files according to the target set and would provide the exact result. The false predictions would be called as far. The number of neurons in the hidden layer was set to 10. The 13 MFCC coefficients obtained from each window from each sample were concatenated to form one 13*N matrix, where N was the sum of number of windows of all samples. This matrix is the input data to the neural network. Corresponding to this matrix, another 2*N matrix was formed. This matrix was used as the target data (to specify which sample belongs to which class). The contents were as follows:

[0 1] --------- ANGER

[1 0]---------- NO-EMOTION

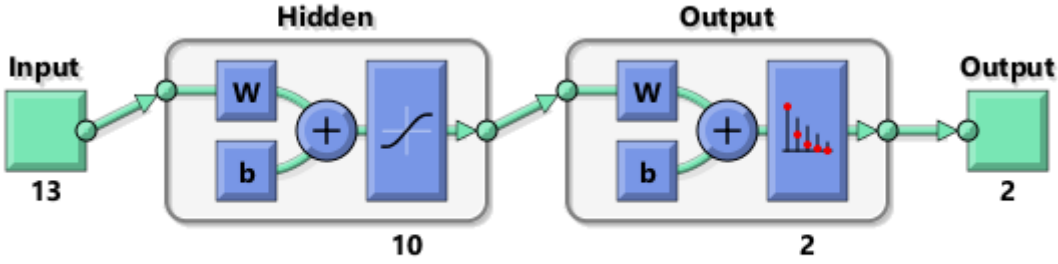

Fig 3: Neural Network Model

Hence, the neural network had 13 inputs (MFCC coefficients) and two outputs. After specifying the data input and the target matrices, the network was trained with 70% of the data till suitable convergence is reached. This network was tested by 50% of the remaining data and was validated by the remaining 50%. After complete training the following figures were obtained:

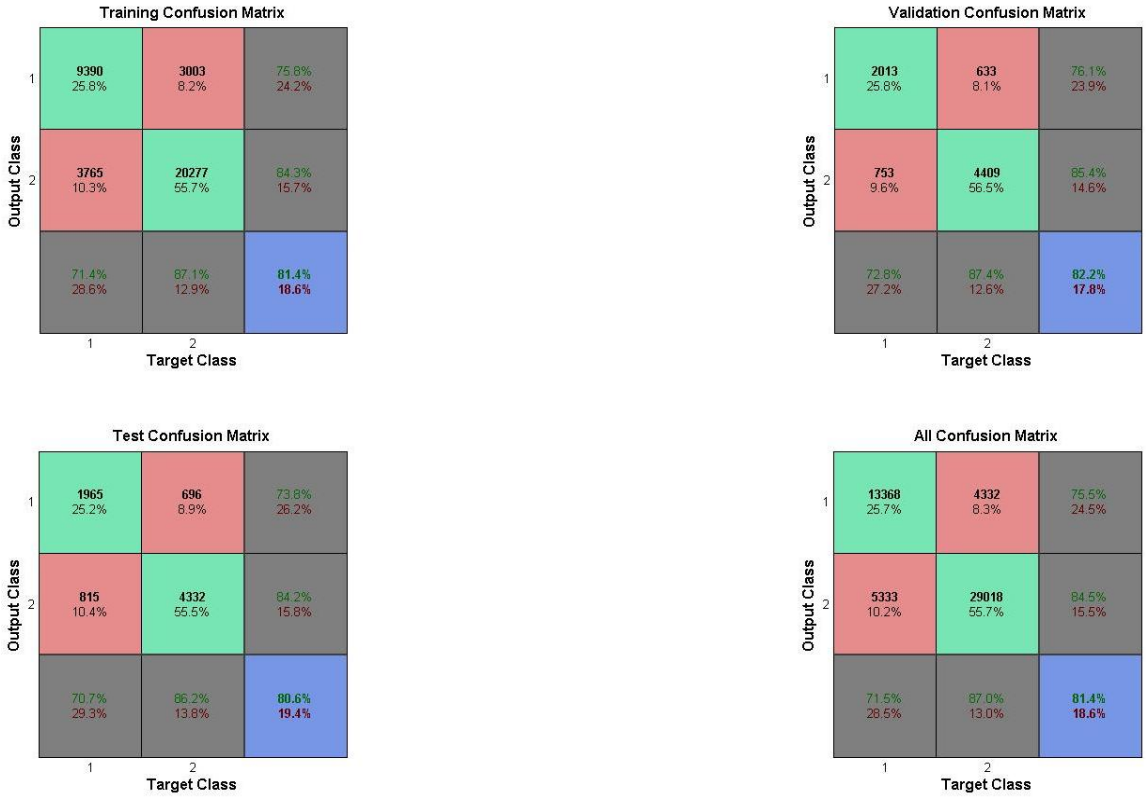

Fig 4: Confusion Matrices- On the confusion matrix plot, the rows correspond to the predicted class (Output Class), and the columns show the true class (Target Class). The diagonal cells show for how many (and what percentage) of the examples the trained network correctly estimates the classes of observations. The off diagonal cells show where the classifier has made mistakes. The column on the far right of the plot shows the accuracy for each predicted class, while the row at the bottom of the plot shows the accuracy for each true class. The cell in the bottom right of the plot shows the overall accuracy.

From these tables we see the efficiency of classification. Class 1 is no Emotion and class 2 is anger. The target class is the actual class the sample belongs to and the output class is what it has been classified as. Hence green squares represent correct classification.

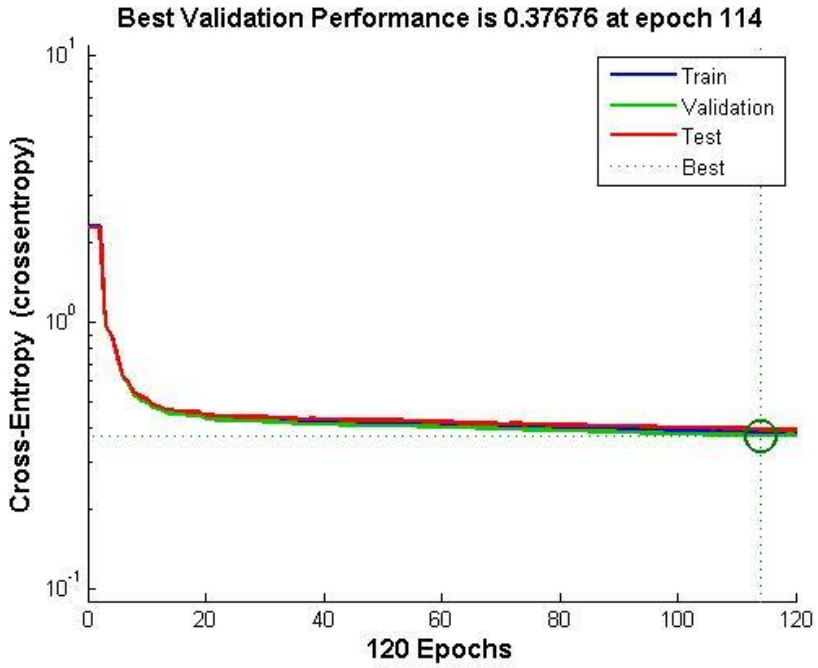

Fig 5: Cross entropy graph- shows a network performance given targets and outputs, with optional performance weights and other parameters.

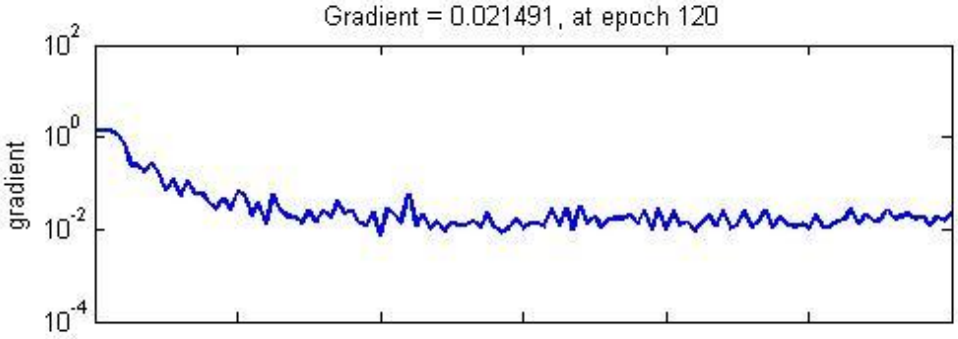

Fig 6: Gradient of the error function

A good classifier has lower values of cross-entropy. This quantity can be casually considered equivalent to error. We can see that as number if iterations in training increase, cross-entropy decreases. Also it is seen that the gradient of the error function keeps reducing gradually, this shows convergence.

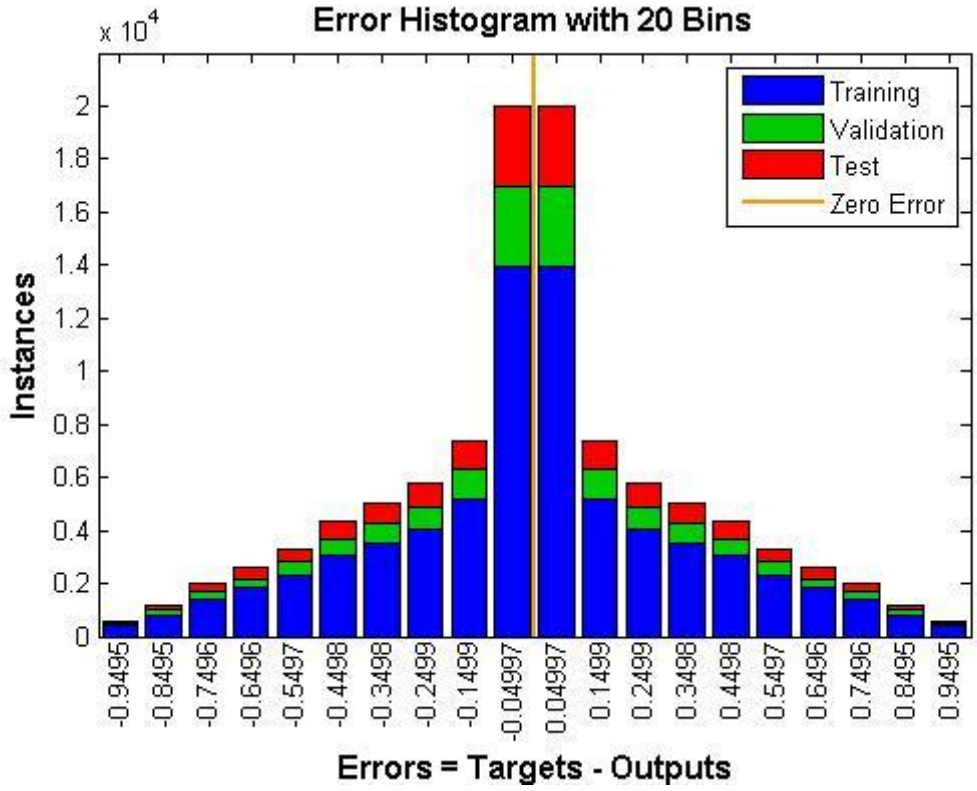

Fig 7: Error Histogram- The histogram can give you an indication of outliers, which are data points where the fit is significantly worse than the majority of data

This plot shows that error was kept closer to zero and classification was more than satisfactory.

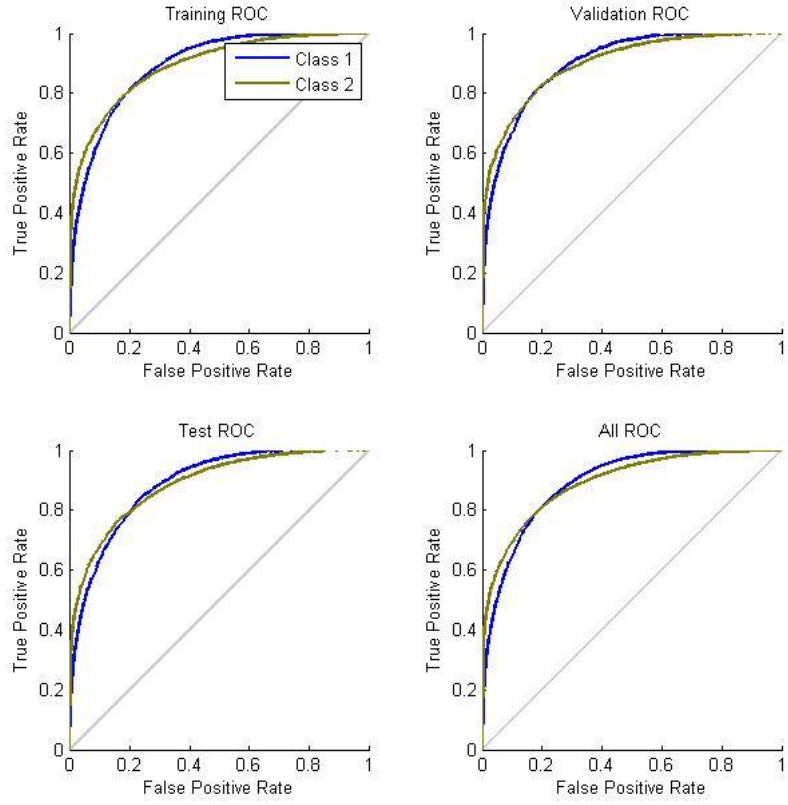

Fig 8: The ROC or receiver operating characteristic is a metric used to check the quality of classifiers. For each class of a classifier, roc applies threshold values across the interval [0, 1] to outputs. For each threshold, two values are calculated, the True Positive Ratio (the number of outputs greater or equal to the threshold, divided by the number of one targets), and the False Positive Ratio (the number of outputs less than the threshold, divided by the number of zero targets).

The above chart shows the relation between true positive rate and false positive rate. A true positive is when a sample is classified as a target correctly. A true negative is when a sample is classified as a non target correctly. A false positive is a false alarm or a wrong classification as a target. A false negative is a 'miss' or a misclassification as non target. Ideally this graph must be like a step function, but anything above the 45 degree line is considered as satisfactory classification. Given these facts this classifier is above standards.

$$\boldsymbol{Precision = Positive\ predictive\ value = \frac{TP}{TP + FP} = p}$$

$$\boldsymbol{Recall = True\ positive\ rate = Sensitivity = \frac{TP}{TP + FN} = r}$$

$$\boldsymbol{F\ Measure = \frac{2pr}{p + r}}$$

From the above, we obtain that

- precision p = 84.5%
- recall r = 87%
- f-measure = 85.73%

Conclusions

This report deals mainly with two aspects of Emotion detection. The former is qualitatively analyzing the aspects of speech and mining information from the same. The latter is building an algorithm in the textbook method of basic machine learning algorithm that involves feature extraction, training, validating and testing. The features extracted from speech signals, are key parameters for the design of a speech emotion recognition system. They influence not only the performance of the system but also its overall structure. Cepstral features have been used extensively in the literature, as well as prosodic features, but their combinations should be further examined. Finding a set of features that is optimal and data-independent is still an unsolved problem. In this study, it was observed that using different types of features, different classifiers and in the end fusing everything in an optimal manner can lead to strong improvements of the results. Though the human brain can perceive the emotional aspects of the speech clearly and distinctly to near perfect accuracy, it is imperative at the same time to feed such qualities to a machine as well. The race for making computer programs to be able to read emotions has been motivated by the potentially wide range of applications which involve human-machine interfaces. The emotion recognition mechanisms which now represent true engineering milestones for research area will definitely represent standards to empower useful devices in every-day life of people in the near future.

References

1. Dragoş Datcu, "Multimodel Recognition of Emotions", Technische Universiteit Delft, 2009.
2. Alexandros Georgogiannis, "Automatic Speech Emotion Recognition", Technical University of Crete, Greece, 2011.
3. S Lalitha, D Geyasruti, R Narayanan, Shravani M, "Emotion detection using MFCC and cepstrum features, 4th International Conference on Eco-friendly Computing and Communication Systems.
4. S. Demircan and H. Kahramanl, "Feature Extraction from Speech Data for Emotion Recognition", Journal of Advances in Computer Networks, Vol. 2, No. 1, March 2014.
5. Chandra Prakash, Prof. V B Gaikwad, Dr. Ravish R Singh, Dr. Om Prakash, " Analysis of Emotion Recognition System through Speech Signal using K-NN and G-MM Classifier", IOSR Journal of Electronics and Communication Engineering (IOSR-JECE) e-ISSN: 2278-2834,p- ISSN: 2278-8735.Volume 10, Issue 2, Ver.1 (Mar - Apr.2015), PP 55-61
6. Raunaq Shah and Michelle Hewlett, "Emotion Detection from Speech" , Stanford University, 2007.
7. Tomas Pfister, "Emotion Detection from Speech" , Gonville & Caius College, 2010.
8. Jagvir Kaur, Abhilash Sharma, "Emotion Detection Independent of user using MFCC Feature extraction", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 6, June 2014.
9. Florian Eyben, Felix Weninger, Florian Gross, Björn Schuller: "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor", In Proc. ACM Multimedia (MM), Barcelona, Spain, ACM, ISBN 978-1-4503-2404-5, pp. 835-838, October 2013. doi:10.1145/2502081.2502224.