Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019)
IEEE Xplore Part Number: CFP19K25-ART; ISBN: 978-1-5386-7808-4

Speech based Emotion Recognition using Machine Learning

Girija Deshmukh, Apurva Gaonkar, Gauri Golwalkar, Sukanya Kulkarni
Department of Electronics and Telecommunications Engineering
Bharatiya Vidya Bhavans Sardar Patel Institute of Technology
Mumbai, India.

***Abstract*—Emotion recognition from audio signal requires feature extraction and classifier training. The feature vector consists of elements of the audio signal which characterise speaker specific features such as tone, pitch, energy, which is crucial to train the classifier model to recognise a particular emotion accurately. The North American English language open source dataset was divided into training and testing manually. Speaker vocal tract information, represented by Mel-frequency cepstral coefficients (MFCC), was extracted from the audio samples in training dataset. Pitch, Short Term Energy(STE), and MFCC coefficients of audio samples in emotions anger, happiness, and sadness were obtained. These extracted feature vectors were sent to the classifier model. The test dataset will undergo the extraction procedure following which the classifier would make a decision regarding the underlying emotion in the test audio. The training and test databases used were North American English acted and natural speech corpus, real-time input English speech, regional language databases in Hindi and Marathi. The paper details the two methods applied on feature vectors and the effect of increasing the number of feature vectors fed to the classifier. It provides an analysis of the accuracy of classification for Indian English speech and speech in Hindi and Marathi. The achieved accuracy for Indian English speech was 80 percent.**

Keywords—Classification, Emotion recognition, Mel frequency cepstral coefficients, Pitch, Short term energy, Support Vector Machine

I. INTRODUCTION

Human speech conveys information and context through speech, tone, pitch and many such characteristics of the human vocal system. As human-machine interactions evolve, there is a need to buttress the outcomes of such interactions by equipping the computer and machine interfaces with the ability to recognize the emotion of the speaker. Today, a large amount of resources and efforts are being put into the development of artificial intelligence, and smart machines, all for the primary purpose of simplifying human life. If the machine is able to recognize the underlying emotion in human speech, it will result in both constructive response and communication.

T. Pao, C. Wang and Y. Li, in their paper, dated 2012 discussed 78 features extractable from a speech signal and classified a 13 feature set as being most suitable for a particular classifier[1]. M. S. Likitha, S. R. R. Gupta, K. Hasitha and A. U. Raju, outlined the process to determine the MFCC coefficients and checked deviation to determine amongst 3 emotions[2]. In the paper, by Chen and Luo, a text-dependent speaker verification system was implemented with the purpose of recognizing an imposter voice against an authentic user[3]. This was done through training the SVM with the help of speaker model and imposter model after extracting MFCC coefficients from the password spoken by the user.

There exists substantial research work on improving the accuracy of the classification results by implementing different classifier models. This paper focuses on the feature vectors in order to improve the accuracy of the classifier. The proposed methodology can be integrated in existing systems.

The remainder of the paper is organized as follows. Section II provides an overview of the system. In section III, the methodology implemented in feature extraction and classifier training is explained. Section IV constitutes the analysis of experimental results. We draw conclusions based on the research in Section V.

II. SYSTEM DESCRIPTION

Fig. 1 illustrates the overall system. The system is divided broadly into dataset formation, pre-processing, feature extraction and classification. The entire system is programmed using MATLAB R2014a. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)[4], which consisted of both male and female speech audio samples was used to test for 3 emotions- anger, happiness and sadness.

Firstly, the audio database is divided into training and testing sets. Each signal from both the sets is pre-processed to make it suitable for data gathering and analysis. In succession, the features are extracted from the pre-processed signal. The feature vectors are the input to the multiclass support vector machine (SVM) classifier which forms a model corresponding to every emotion. The test signal is tested with every model in order to classify and detect its emotion.

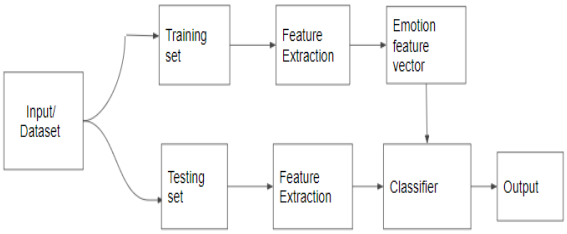

Fig. 1. System block Diagram

III. METHODOLOGY

This section describes the implementation of the feature extraction stage.

A. *Pre-processing*

1) Sampling: Sampling is the first and important step of signal processing. Signals which we used normally, are all analog signals i.e continuous time signals. Therefore, for processing purpose in computer, discrete signals are better. In order to convert these continuous time signals to discrete time signals, sampling is used[5].

$$f_s = \frac{1}{T} \tag{1}$$

Equation(1) denotes the relation between the sampling frequency(f_s) and the time period(T). Here, we are converting a continuous speech signal into a sequence of samples using MATLAB.

2) Pre-emphasis: The input signal often has certain low frequency components which will result in samples similar to their adjacent samples. These parts represent a slow variation with time and hence are not of any importance while extracting signal information. Therefore, we are performing pre-emphasizing by applying a high pass filter on the signal in order to emphasize the high frequency components which represent the rapidly changing signal. This will provide vital information. Equation(2) represents the pre-emphasis filter used where H(z) denotes pre-emphasized signal.

$$H(z) = 1 - 0.95z \tag{2}$$

3) De-silencing: Audio signals often contain regions of absolute silence occurring at the beginning or at the end and sometimes in between higher frequencies. It is required to remove this unwanted part from the signal and hence desilencing is performed. Silence removal is performed by applying a particular threshold to the signal. We get a signal in which the unvoiced parts which do not contain any relevant data are removed and the voiced parts are retained.

4) Framing: For the purpose of analysis, observable stationary signal is preferable. If we observe the speech signal on a short time basis, we are able to get a stationary signal. We divide the input signal into small constituent frames of a specific time interval. Generally, for speech processing, it was observed that frame duration of 20-30 ms is implemented. This ensures two things- firstly, that considerable stationary signal value is obtained for small time duration and secondly, the signal does not undergo too much changes in the interval. We have utilized a frame duration of 25 ms.

5) Windowing: Most of the digital signals are large and infinite that they cannot be analyzed entirely at the same time. For better statistical calculations and processing, values of signal at each point should be available. In order to convert large digital signal into a small set for processing and analyzing, and for smoothing ends of signal, windowing is performed. Different windowing techniques are available namely Rectangular, Hamming, Blackman etc. We have applied Hamming window on the framed signal. The Hamming window is represented by (3), where w(n) is windowed signal, M is the window length-1 and $0 \le n \le (M-1)$ [6].

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{M-1}\right) \tag{3}$$

B. *Feature Extraction*

1) Energy Feature Extraction for each Frame: For the speech signal, we need to work with stationary signals therefore we calculate the energies of each frame. The energy of the signal is related to the sample value s(m) as denoted through (4).

$$E_T = \sum_{n=-\infty}^{n=\infty} S^2(m) \tag{4}$$

Here, s(m) represents the samples within each frame and E_T represents energy of signal. A single energy value for each frame is obtained and then plotted simultaneously to obtain the short-term energy plot of the signal wherein the larger peaks represent high frequency frames.

2) MFCC feature vector extraction for each frame: MFCCs represent the short-term power spectrum envelope. This envelope in turn is representative of the shape of the human vocal tract which determines the sound characteristics. Therefore, MFCC is a vital feature for speech analysis. Fig. 2 gives the block diagram of the steps to extract MFCC feature vector.

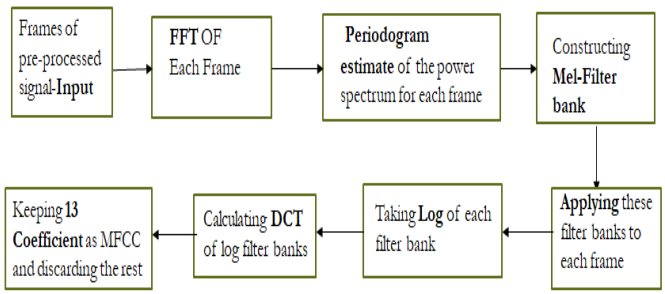

Fig. 2. MFCC extraction steps

a) Fourier transform of windowed signal (FFT): MFCC is a spectral feature which is not extractable in the time domain. Hence conversion to frequency domain is required. In order to obtain the periodogram estimate, it is necessary to convert the signal into frequency domain using FFT. Hence, we obtain the Fourier transform of each frame and get the FFT points. The output is obtained using (5).

$$X[K] = \sum_{n=0}^{N-1} x[n] W_N^{nk} \tag{5}$$

Here, x[k] is FFT of windowed signal w(n).

b) Determination of Power Spectral Density: After calculating FFT, the power spectrum of each frame was calculated. The periodogram estimation helps in finding frequencies present in each frame and identifying how much energy is present in different regions of frequencies. We have used welch method to identify power of signal at different frequencies.

$$S_x^W(\boldsymbol{\omega_k}) \triangleq \frac{1}{K} \sum_{m=0}^{K-1} P_{x_m, M(\omega k)} \tag{6}$$

PSD of signal is calculated using (6) where $S(W_k)$ is PSD of signal.

c) Mel filter bank: The Mel scale relates the perceived sound frequency to its actual frequency. An upper and lower limiting frequency is determined. It is then converted into Mel scale using (7) [7].

$$M(f) = 1125 \ln\left(1 + \frac{f}{700}\right) \tag{7}$$

Considering 26 filter banks, 24 values are obtained between the upper and lower Mel converted frequencies. Each of these are converted back to frequency domain using (8).

$$M^{-1}(m) = 700\left(\exp\left(\frac{m}{1125}\right) - 1\right) \tag{8}$$

Each filter bank points are calculated using (9).

$$\begin{aligned} H_m(k) &= 0 && k < f(m-1) \\ &= k - \frac{f(m-1)}{f(m) - f(m-1)} && f(m-1) \le k \le f(m) \\ &= \frac{f(m+1) - k}{f(m+1) - f(m)} && f(m) \le k \le f(m+1) \\ &= 0 && k > f(m+1) \end{aligned} \tag{9}$$

Each filter is a triangular function and are plotted overlaid on each other.

d) Application of the Mel filters on the frame: Each Mel filter is applied to the frames to get 26 output values for each frame. The filter bank is multiplied with the power spectrum and coefficients are added up. The resultant is 26 coefficients representing filter bank energies.

e) Logarithm of obtained frame energies: Logarithm of each of the filter bank energies is performed. This is because loudness is not perceived linearly. For loud sound, variations in energy may sound the same. Hence such compression or normalization is performed to bring perceived signal nearer to actual signal.

f) Discrete Cosine Transform to obtain real MFCC coefficients: After taking logarithm of obtained filter banks, DCT is applied on each of them. Due to the overlapping filter banks, these energies are correlated with each other. In order to decorrelate these energies, DCT is performed.

g) MFCC Feature vector extraction: DCT coefficients are related to energies of filter bank. High DCT values represents high rate of change of filter bank energies. For better performance and extraction, only 13 values of DCT coefficients are kept and rest are discarded. These values are the relevant MFCC coefficients.

3) *Pitch of Audio signal:* Pitch is an important feature of voiced signal. As perceived by the human ear, emotions have a certain pitch which is described as 'high' or 'low'. The pitch for each frame of every training and testing sample is extracted. On each signal frame is applied the Hamming window function. FFT of the log magnitude spectrum of each windowed signal is calculated. The pitch is calculated by seeking the peak of the signal during the frame duration[8]. The cepstrum provides lag wherein the highest energy indicates the dominant frequency. This gives the pitch of the signal in a particular time duration[9].

C. *SVM Classifier Implementation:*

Classification is performed using the Support Vector Machine linear classification algorithm. SVM serves as a non-probabilistic linear classifier. It is an algorithm which classifies between two classes. Hence we build models of every emotion versus the rest. The feature vectors are plotted in space and the two classes are separated by a hyperplane which is widened as much as possible. The hyperplane can be represented as the set of points x which satisfy (10).

$$\vec{w} \cdot \vec{x} - b = 0, \tag{10}$$

Here, w is vector normal to hyperplane. Another part of the training dataset is the set y_j which consists of the categories of the training signals. The margin or width of the hyperplane is determined from the support vectors. They form the hyperplane boundary points x_i which satisfy (11) [10].

$$y_i(< w_i x_i > + b) = 1 \tag{11}$$

There are 3 classes (emotions) - Anger, Happiness and Sadness. Labels are assigned to the training signals. The two feature sets i.e. training and testing, and the label set is sent to

the classifier. The classifier forms a model for each class and the signal is tested against each model. The classifier output is the classified emotion of the tested signal.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

Accuracy was calculated for the classification of emotions by mode and mean method by using American english speech corpus. Here, 80% of the database was given to training set and 20% to testing. Accuracy for real time input was also calculated. Regional language dataset in Hindi and Marathi languages was created by recording the audio input of speakers in the age range 18-25. The emotion classes were anger, happiness, and sadness.

Fig. 3 represents signal on applying pre-emphasis high pass filter.

Fig. 4 represents de-silenced signal. When we compare pre-emphasized signal with de-silenced one, we realise that the silence part of signal at the beginning as well as end is removed and hence the whole signal is shifted towards the Y-axis.

Fig. 5 represents plot of a single frame. It is observed that there are 1200 samples in the frames which matches with the calculated value.

Fig. 6 represents the frame after applying hamming window. It can be observed that the shape is similar to that of the hamming window with maximum attenuation towards the ends and minimum at the centre.

Fig. 7 represents the Mel filter bank of 26 filters overlapped on each other.

Fig. 8 represents the plot of the extracted energy signal (short-term energy) of an input audio signal.

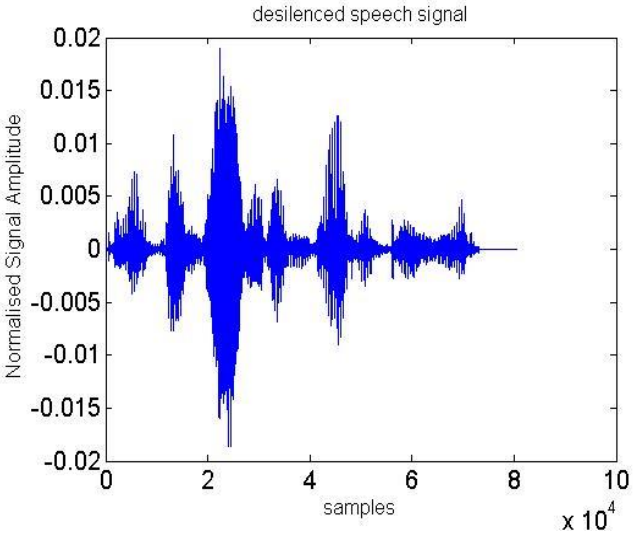

Fig. 4. Desilenced Signal

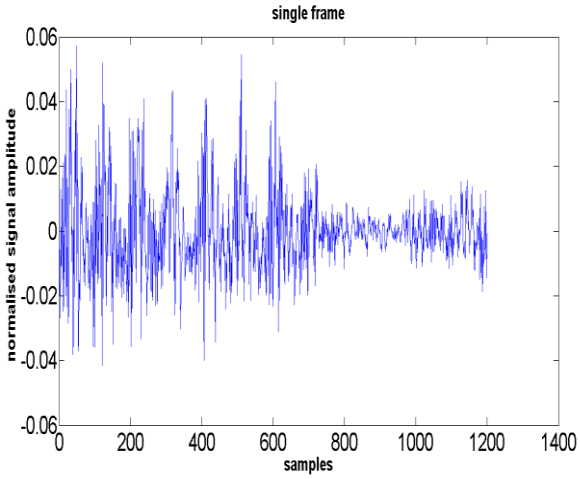

Fig. 5. Single Frame

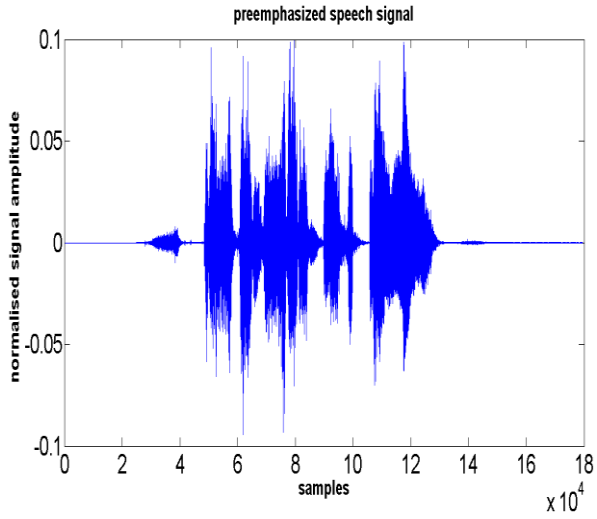

Fig. 3. Pre-emphasized speech signal

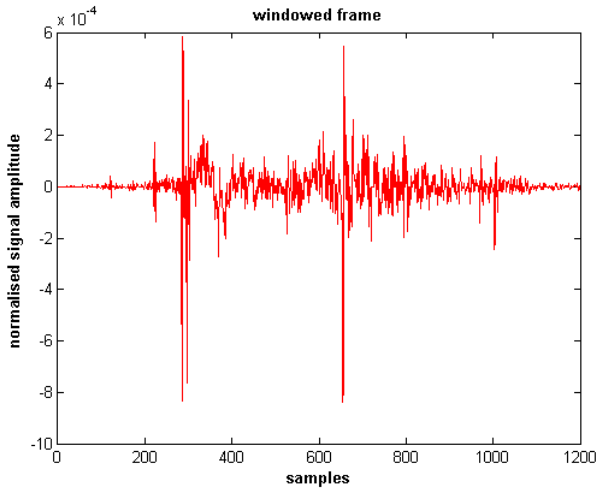

Fig. 6. Windowed Frame

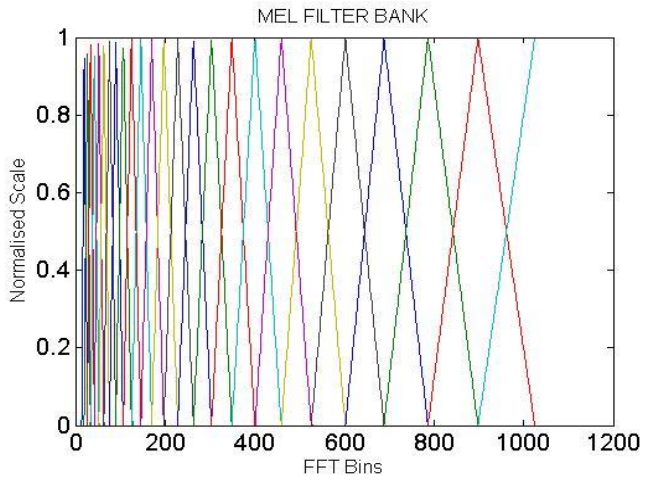

Fig. 7. Mei-Filter bank of 1024 FFT bins

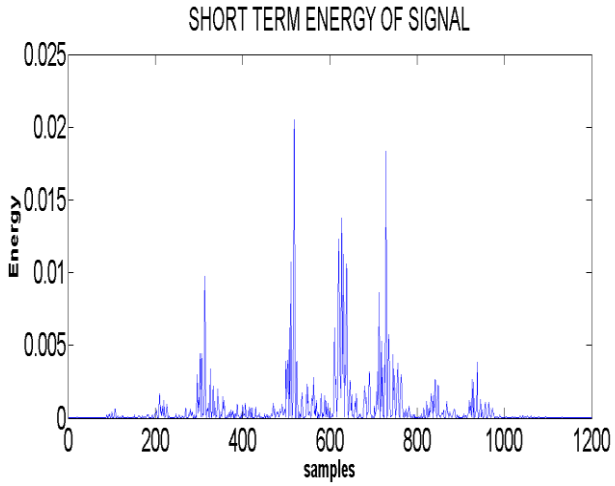

Fig. 8. Short Term energy of Signal

A. Formation of Feature Vector set

1) Mode Method: Three features STE, MFCC, Pitch were extracted for each audio signal input and given as input to classifier. In mode method, mode of result for each signal was found out and the accuracy calculated accordingly. Accuracy of mode method was not found to be satisfactory. A dataset can have more than one modal value, meaning there can be more than one value occuring at highest frequency. Therefore, bimodal, trimodal values can be obtained. Mode as a central tendency of data works accurately for categorical variables such as level of education[11].

2) Mean Method: In mean method, the three features STE, MFCC, Pitch were extracted for each input audio signal. An average feature dataset per sample was calculated and this data was sent to the classifier. Accuracy was calculated from the result. The overall accuracy of mean method was found to be satisfactory in comparison to the mode method. Mean of a dataset is calculated by taking in consideration every value available in the dataset, how much ever it might be skewed[12].

B. Accuracy results

Fig. 9 is the bar graph for the accuracy obtained using three features- Pitch, MFCC, STE. Overall accuracy using two features was in range 60%-65%. After including third feature i.e. Short-term Energy not only for individual feature but overall accuracy was also found to be improved. In this second case the overall accuracy was obtained as 80%.

C. Real Time Input

For considering real time audio input, our voices were recorded and then tested. The system was able to classify in all three emotions. Fig. 10 displays the graph of accuracy for real time testing.

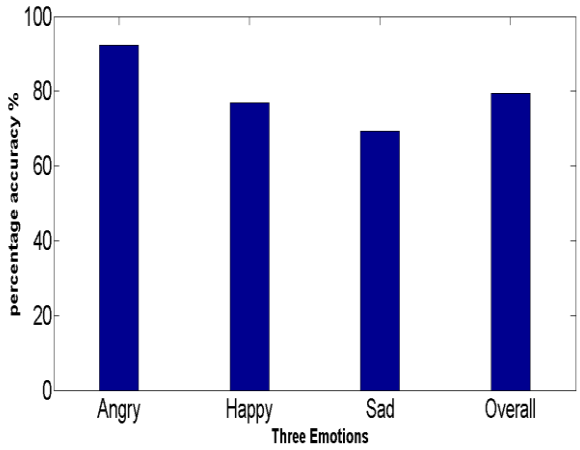

Fig. 9. Classification Accuracy with three features

Fig. 10. Classification Accuracy with real time audio input

Table I gives the tabular representation of the accuracy of two feature vector method and three feature vector method.

Table II provides the classification accuracy and misclassification accuracy in terms of percentage for three emotions.

Fig. 12 is the graph of accuracy results for emotion classification from Hindi speech input.

TABLE I. COMPARISON OF ACCURACY

Emotion	Two feature accuracy (%)	Three feature accuracy (%)
Anger	69.8462	92.3077
Happiness	61.5385	76.9231
Sadness	69.2308	69.2308
Overall	61.5385	79.4872

TABLE II. PERCENT CLASSIFICATION INTO DIFFERENT EMOTION CLASSES

Emotion	Anger	Happiness	Sadness
Anger	92.31%	7.69%	-
Happiness	7.69%	76.92%	15.38%
Sadness	-	30.77%	69.23%

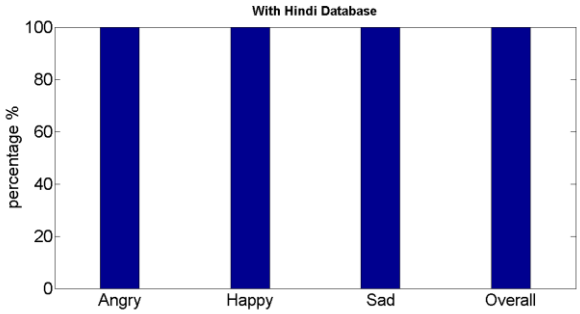

Fig. 12. Emotion classification accuracy for Regional language Hindi

V. CONCLUSION

In this paper, three emotions- anger, happiness, and sadness, were classified using three feature vectors. Pitch, Mel frequency cepstral coefficients, Short Term Energy were the three feature vectors extracted from audio signals. Open source North American English acted speech corpus and recorded natural speech corpus were used as input. The dataset used for training and testing consisted of audio samples in male and female voice and divided in ratio 4:1.

The mean method provided greater accuracy over the mode method. Mean of a set of data values takes into consideration every value present while there can be two or more values at the highest frequency.

The classification accuracy for all three emotions was found to have increased by 20% by using three features as against using two features. Anger and happiness emotions classification accuracy increased by 15%-20% with the help of STE feature vector. The emotion sadness did not improve its accuracy of classification despite using STE feature vector. Sadness is more susceptible in being misclassified as happiness emotion. Happiness is misclassified as angry due to close values for STE, and sadness due to its lower pitch.

Classification of emotions for regional Indian languages namely Hindi and Marathi was implemented. The accuracy of classification of real time input audio for regional language Hindi was obtained 100%. Further, accuracy could be affected according to the natural tone of speaker. Analysis on the basis of age group of speakers is another area of future work.

REFERENCES

[1] T. Pao, C. Wang and Y. Li, "A Study on the Search of the Most Discriminative Speech Features in the Speaker Dependent Speech Emotion Recognition." *2012 Fifth International Symposium on Parallel Architectures, Algorithms and Programming*, Taipei, 2012, pp. 157-162.

[2] M. S. Likitha, S. R. R. Gupta, K. Hasitha and A. U. Raju, "Speech based human emotion recognition using MFCC." *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, Chennai, 2017, pp. 2257-2260.

[3] Chen, S.-H. and Y.-R. Luo (2009), "Speaker Verification Using MFCC and Support Vector Machine," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, March 18 - 20, 2009, Vol I IMECS 2009.

[4] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391.

[5]Thomas Zawistowski & Paras Shah, "An Introduction to Sampling Theory." Internet: http://www2.egr.uh.edu/~glover/applets/Sampling/Sampling.html, [Feb.24, 2019].

[6] Akhilesh Chandra Bhatnagar, R. L. Sharma, Rajesh Kumar, "Analysis of Hamming Window Using Advance Peak Windowing Method." *International Journal of Scientific Research Engineering &Technology,* vol.1 issue 4, pp 015-020, July 2012.

[7] Practical Cryptography, "Mel Frequency Cepstral Coefficients (MFCC) tutorial." Internet:http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/, [Feb.27, 2019].

[8] Naotoshi Seo, "Project: Pitch Detection." Internet: http://note.sonots.com/SciSoftware/Pitch.html, [Feb.25, 2019].

[9] Vocal Technologies, "Pitch Detection using Cepstral Method." Internet: https://www.vocal.com/perceptual-filtering/pitch detection/, [Feb.25, 2019].

[10] Sunil Ray, Analytics Vidhya, "Understanding Support Vector Machine algorithms from examples." Internet: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/, Sept.13, 2017 [Mar. 10, 2019].

[11] Institute for Digital Research and Education, UCLA, "What is the difference between categorical, ordinal and interval variables?" Internet:https://stats.idre.ucla.edu/other/multpkg/whatstat/what-is-the-difference-between-categorical-ordinal-and-interval-variables/, [Feb.27, 2019].

[12] Laerd Statistics, "Measures of Central Tendency." Internet:https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median.php, [Feb.27, 2019].