Speech Emotion Recognition Using MFCC And Machine Learning For Human Computer Interaction

[1]Dr.Senthamizhselvi R, [2]S Supriya, [3]M Sivaranjani,[4]V Sangeetha

[1] Associate Professor, Easwari Engineering College, Chennai, India, senthamizhselvi.r@eec.srmrmp.edu.in, 9445088477
[2] UG Scholar, Easwari Engineering College, Chennai, India, supriyasundar2000@gmail.com, 8939109500
[3] UG Scholar, Easwari Engineering College, Chennai, India, ranjanisiva530@gmail.com, 7010722043
[4] UG Scholar, Easwari Engineering College, Chennai, India, sangeethaavijayalayan@gmail.com, 9944338359

Abstract: *Digital processing of speech signal and emotion recognition can significantly enhance performance of human-computer interaction systems. With rapid growth in smart systems, the need for emotion recognition by the automated machine has increased drastically. Existing methods involve extraction of signal features using different algorithms and it's classification using different machine learning models .While each model has its own merits and demerits, Hidden Markov Model(HMM) has achieved a maximum accuracy of 64.77% for speaker independent system so far. The proposed methodology involves a two stage approach of feature extraction using Mel Frequency Cepstral Coefficients(MFCC), Linear Prediction Coefficients(LPC) and Perceptual Linear Prediction Coefficients(PLPC) followed by classification model using K-Nearest Neighbour(K-NN) clustering. Parallelly, image form of the audio signal is directly analysed and classified using Convolutional Neural Networks(CNN).This methodology has successfully recognized seven different emotions, namely, Anger, Boredom, Disgust, Happiness, Neutral, Sadness and Fear and achieved better accuracy over other individual classifiers. The former model provided an accuracy of 74.64% while the latter image processing-CNN model proved better with an accuracy of 92.19% in speaker independent emotion recognition. The integration of the three feature extraction techniques proves to be better in terms of accuracy and performance parameters over other individual feature based classification systems.*

Keywords: Speech Emotion Recognition, Cepstral Coefficients, Prediction Coefficients, Classification, Extraction, k-nearest neighbour, convolutional neural networks

Introduction

Emotions can play a crucial role in how we think and behave. It has been proved that emotion has great impact on decision-making [1] and social communication. If a message doesn't make us feel something, we are unlikely to act thereon. Thus , humans tend to consider emotions as guidelines for each decision.

Human-computer interaction (HCI) is a multidisciplinary field of study focusing on the design of computer technology , especially, in the field of the interaction between the users and computers. Human-computer interaction can be defined as simply what happens when a computer and a human (user) interacts, and what exactly the computer understands from the input given by the user. Many parameters have to be understood by the computer to give a proper and exact solution to the user ; one such parameter is the emotion of the speaker. The computer has to understand in what context i.e., the type of emotion the user has displayed in his/her speech. This is where speech emotion recognition (SER) plays a major role. Speech emotion recognition(SER) is essentially the extraction of emotion from a person's speech. Speech Emotion Recognition(SER) system is a collection of methodologies that process and classify speech signals to detect emotions embedded in them.

It plays a vital role in social communication as well as decision making in instances of human-computer interaction. It is of dire need in the field of medicine in the case of autistic children and patients with Alzheimer who struggle to convey their emotions accurately and helps improve counselling and rehabilitation processes as well. This can also be applied in interactive smart classrooms to improve the quality of teaching ,it is also used to monitor the emotional quotient of a driver to ensure safety and it is used in designing interactive games to give the user a real time experience. Thus, Speech Emotion Recognition (SER) has applications in different facets of life.

There are lots of difficulties in detecting the exact emotion displayed by the user. Since the number of basic emotional labels are still arguable at present [2, 3] owing to fact that the same emotion can be defined in different ways depending on the situations. There exist many methods to detect emotions, this paper contains the combination of Mel Frequency Cepstral Coefficients (MFCC), Perceptual linear prediction (PLP) , and Linear prediction coefficients (LPC) for feature extraction and the corresponding emotion identified is classified using both convolutional neural networks (CNNs) & K-Nearest Neighbour (KNN) algorithm. In this paper optimal combination of algorithms has been tested for achieving better efficiency. The paper is hereafter segmented into two major sections, namely, signal processing and classification after analysis of exiting methodologies available in the domain.

Related Work

In the past few decades, there has been a lot of work in speech emotion recognition, which can be mainly described as the honest form of complex classifiers and features extracted by using MFCC to prove the improved accuracy. [4].Previous work in this area included use of varied classifiers like SVM,K-Nearest Neighbour (KNN), Gaussian Mixture Model (GMM),Radial Basis Function (RBF)and Binary SVM Classifier etc. are discussed.
Hany et al. [5] implemented six classifiers on ASSISTments dataset comprising of 15 features. They utilized IBK, VF1, Naïve Bayes Updateable, J48, ONER, and k means clustering classifiers to level the features. Results showed that k means clustering was the simplest in giving ranks to features and Naïve Bayes was better in giving prediction accuracy.

Lei Yu [6] in their work initiated a feature selection algorithm which is particularly employed for peek dimensional data which is called as fast correlation base filter. This algorithm is for eliminating unrelated and unnecessary data. They implemented ReliefF, FCBF, ConSF, and corrF on four datasets and recorded the time period and number of features selected. Then they implemented C4.5 and NBC classification on the data.

Mohan Ghai et al. [7] established a system employing MFCC and examined on Berlin database. Proposed work is completed by employing an explicit classifier namely Support Vector Machine (SVM), Random Decision Forest (RDF) and Gradient Boosting. It identifies seven classes of emotion. SVM attained a mean accuracy of 55.89%, Gradient boosting achieved 65.23% and Random Decision Forest classifier attained 81.05%. Out of this three classifiers highest accuracy is attained during Random Decision Forest classifier (RDF) 81.05% accuracy is attained.

Milton et al. [8] established a system employing MFCC. During this method they used Berlin emotion database. It consists of 535 acted emotions in Germanic language with 7 different emotions and it's a multi-speaker database. during this proposed work they use Three Stage SVM classifier. And this classifier is additionally compared with another classifier which recognizes seven classes of emotions like: Happiness, Angry, Disgust, Fear, Boredom, Sadness, and Neutral. This author attained average accuracy of Three Stage SVM 68% and compared with SVM using Radial Basis Function (RBF), Linear Kernel i.e.55.4%, 65% and 68%. Out of those three classifiers highest accuracy of 68% is attained during SVM.

Frank Dellaert et al. [9] detected four emotions: happy, sad, anger and fear from their own datasets. They utilized 17 selected features from 5 groups and used three methods which were MLB classifier, KR and KNN where the utmost accuracy was found by KNN [9]. Tin Lay New et al. [10] used Hidden Markov model (HMM) to classify 6 categories of emotion and located average accuracy 70%. He utilized a database comprising of 60 emotional spoken words, each from twelve speakers [10]. He utilized log frequency power co-efficient (LFPC) to depict the speech signals and compared its performance with linear prediction Cepstral coefficients (LPCC) feature parameters and mel-frequency Cepstral coefficients (MFCC) feature parameters.

Methodology

The paucity in availability of emotional speech databases has played a significant role in the accuracy achieved in this research so far. However, with the right combination of optimal signal manipulation and classification techniques, an accuracy higher than prevalent ones has been achieved. The proposed system has been visualised in the block diagram that follows. The BERLIN database of emotional speech has been exploited for this study which comprises of seven contrasting emotions, namely, 7 emotions: neutral, anger, fear, happiness, sadness, disgust and boredom. However, this classification has been done under a controlled environment with known constraints; classification of real-time speech requires accurate signal processing.

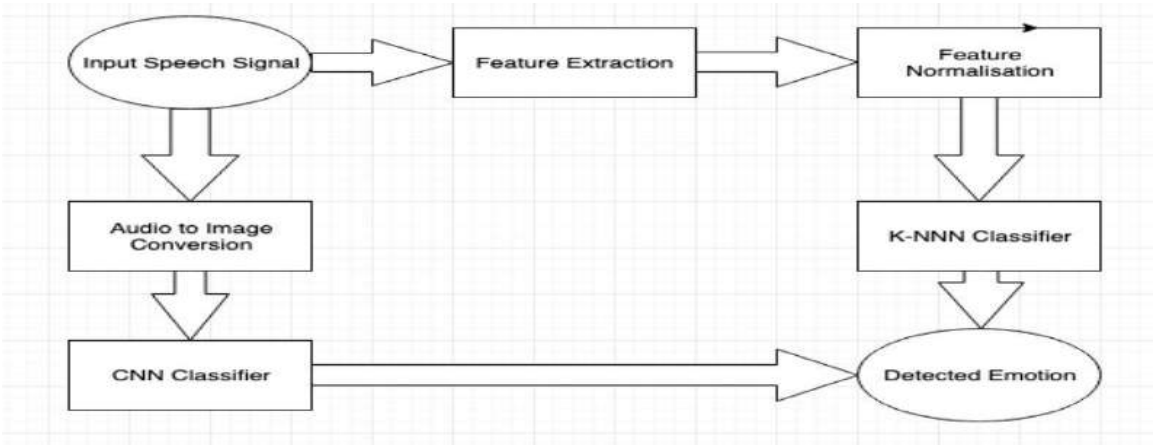

Figure 1: Block Diagram of overall process

Signal Processing

The raw input speech signal contains a varied range of features, each of which pertains to a specific emotional cue. Prosodic features, are commonly known to reflect the emotional state of the speaker with by analysing the pitch, duration and loudness of the speaker [11]. However, the proposed system utilises three distinct features as stated below in order to identify complex emotions and improve upon previous studies:

- Mel Frequency Cepstral Coefficients (MFCC)
- Linear Prediction Coefficients (LPC)
- Perceptual Linear Prediction Coefficients (PLPC)

Feature Extraction

The algorithm produced a total of 49 features composed of 39 Mel Frequency Cepstral Coefficients, 9 Linear Prediction Coefficients and a Perceptual Linear Prediction Coefficient for five hundred and twenty eight samples utilised in the system. Owing to the redundancy of such a large number of features and the impossibility of manual detection of intricate variances in the feature values, extraction of select features from the thousands of values obtained in a vital step in processing. Thus, optimal features are taken from the common variable where all the coefficients are stored in the subsequent stages to enable dimension reduction. The combination of the three feature extraction techniques helps improve accuracy and assures favourable classification.

Feature Normalisation

This process helps scale down and adjust the different cepstral and autocorrelation coefficients to a common scale so as to accomplish reliable classification. It also helps denoise the signal so as to acoustically match the test utterance fed into the classifier with the trained sample. Image form of the audio signal is normalised so as to change the range of pixel intensity values. This step provides consistency in the wide set of values used in signal manipulation.

Mel Frequency Cepstral Coefficients

This step gives the significant portion of cepstral coefficients needed for classification subsequently. MFCC constitutes 39 features, each for every training and testing sample utilised in the system. These cepstral coefficients give the rate of change in spectral bands in the quefrency domain. The unique nature of every individual's vocal system from the tongue and teeth to the vocal cords distinguishes the nature of sound that he or she produces. MFCC denotes the short time power spectrum that is essentially a manifestation of the vocal tract. This thus gives an accurate representation of the phoneme being produced [12]. Frequency being measured in Hertz(Hz) is converted into the Mel Scale so that even minute variations can be discerned by the human ear. This is done with the help of the below formula:

$$Mel(f) = 2595\, log\left(1 + \frac{f}{700}\right) \qquad (2.1.1)$$

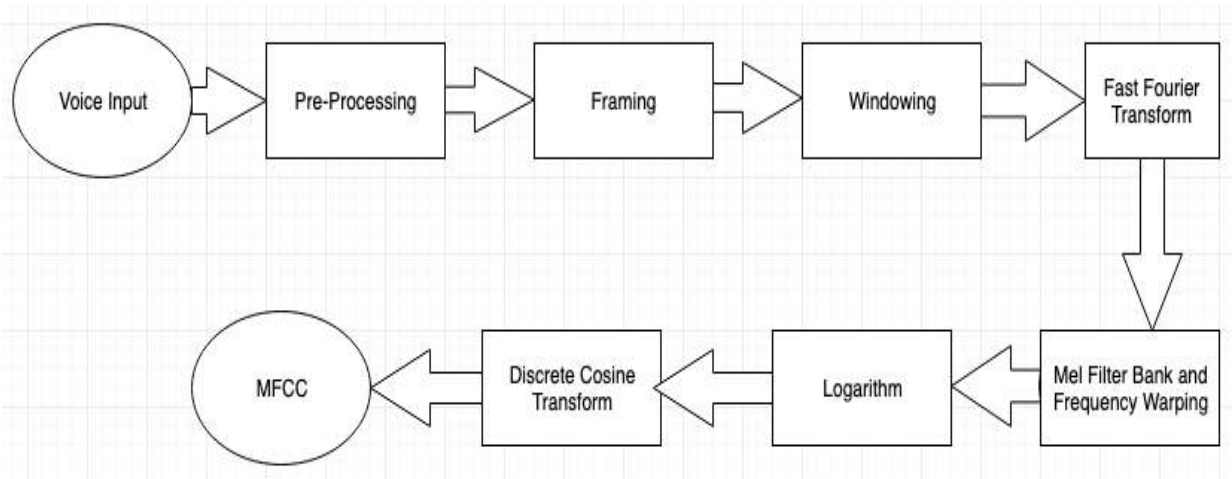

Figure 2 : Block Diagram of MFCC

Step 1: *Pre-Processing*
This aids in removal of harmonics i.e., the distortion of a sinusoidal waveform by waveforms of different frequencies. This assures removal of any residue noise that may have persisted from the previous stages and helps forward only the required content.
Step 2 : *Framing*
The signal is framed into N samples width since short-time, stable signal is needed in place of constantly changing audio for evaluation and analysis. The digital signal is segmented into frames of 20-40ms intervals each.
Step 3: *Windowing*
Hamming window is employed to reduce spectral leakage whilst considering the subsequent processing steps. This window helps combine frequency lines closest to each other. The Hamming window equation is :

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \quad 0 \le n \le N-1 \qquad (2.1.2)$$

Step 4 : *Fast Fourier Transform*
The Fast Fourier Transform is computed on each frame to calculate the periodogram, i.e., estimate of spectral density of signal. The frame of N samples is converted from the time domain to the frequency domain. This identifies which frequencies are present in the frame. This is demonstrated by the below equation :

$$Y(w) = FFT\,[h(t) * X(t)] = H(w) * X(w) \qquad (2.1.3)$$

Here, H(w) and X(w) are the Fourier Transform of h(t) and X(t) respectively. It represents the convolution of the vocal tract impulse response, h(n) and the global pulse.
Step 5: *Mel Filter Bank and Frequency Warping*
Frequency warping transforms the values from one frequency scale to another while the set of 20-40 triangular filter, i.e., Mel filter bank summates the periodogram to compute the amount of energies in the frequency regions

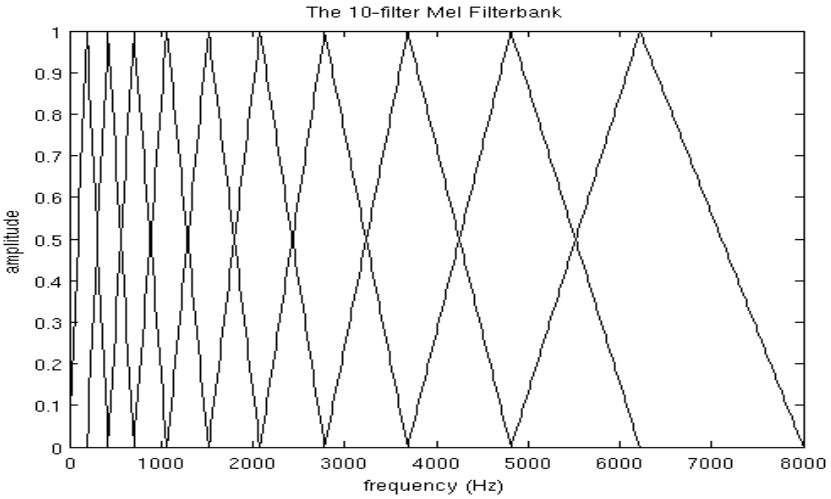

Figure 3 : Mel Filter bank with 10 filters

Step 6: *Logarithm*
This step allows use of cepstral mean subtraction. It is a form of compression that helps model features close to human hearing range by taking logarithm of the filter bank energies since variations become negligible in loud speech signals.
Step 7: *Discrete Cosine Transform*
It decorrelates the overlapped filter bank energies by converting the log Mel spectrum into time domain. The result obtained is the MFCC or acoustic vectors. Discrete Cosine Transform of the frame is achieved by the following equation:

$$S_i(k) = \sum s_i(n)h(n)e^{-\frac{2j\pi kn}{N}} \quad 1 \le k \le K \qquad (2.1.4)$$

where h (n) is an N sample long analysis window (hamming window), and K is the length of the Discrete Fourier Transform. The

Linear Prediction Coefficients

This computation aids in inverse filtering by removal of formants and facilitates provision of concentration of the residue or frequency estimate. These coefficients give the spectral envelope of the signal in compressed form and each value induced by the computation points to an emotional frequency [13]. LPC can be derived by the formula given below where a_m is the linear prediction coefficient, k_m is the reflection coefficient:

$$a_m = log\left(\frac{1 - k_m}{1 + k_m}\right) \qquad (2.1.5)$$

By utilising a sliding window and determining the corresponding peak of the spectrum of the linear prediction filter, the coefficients and correspondingly, the position of format frequencies are computed. This parameter helps predict the future features based on previous features as well.

Perceptual Linear Prediction Coefficient

This computation combines the critical bands, intensity-to-loudness compression and equal loudness pre-emphasis for extraction of relevant data from the audio signal. This step is necessary to compensate for inequal perception of loudness of input signal in dynamic frequency intervals. It also eliminates speaker dependent features and gives short term spectral values similar to MFCC [13]. Thus, it sheds the unwanted information from the speech signal and significantly aids in optimal processing. The perceptual linear prediction coefficients are obtained from the linear prediction coefficients by performing perceptual processing; post which, cepstral conversion is done. The following recursion is done to compute PLPC:

$$c_q = ln(Q) \ : q = 0 \qquad (2.1.6)$$

$$c_q = \ -b_q + \frac{1}{m}\sum_{q=1}^{Q} -(m - q)\, b_q c_{m-q} \ : q > 0 \qquad (2.1.7)$$

Classification

The next step is the analysis and classification of the mean value of the 49 features loaded into an empty matrix from the previous stage. The classification model is trained with a data set of 400 samples taken from the BERLIN Database . The remaining 135 samples in the said database were set aside for testing in accordance with the 3:1 training:testing ratio to achieve higher accuracy.

K-Nearest Neighbour

It is a supervised machine learning algorithm that is based on the principle that there exists a direct correlation between similarity of features and their proximity. [14] It does not make any assumptions and classifies purely based on the existing data. All the features procured from the previous stages are analysed and grouped according to their shortest Euclidean distance(ED). This is computed by the formula :

$$ED = \sqrt{(x_2 - x_1)(x_2 - x_1) + (y_2 - y_1)(y_2 - y_1)} \qquad (2.2.1)$$

The matching patterns are inferred with respect to the training samples fed into the model and predicted estimate is displayed as the desired output. This approached produced a modest accuracy of 74.64 % .

Convolutional Neural Network

This classifier is fed with the image of the Fast Fourier Transform of the audio signal as input without any additional processing techniques [14]. The said methodology architecture is composed of six convolutional layers, each accompanied by its own normalisation and ReLU layer, followed by Max Pooling and a fully connected layer. Image processing is completed with grey scaling and resizing for displaying the final output. Convolution is done with Kernels and appropriate padding ; this involves sliding the filter over the input.

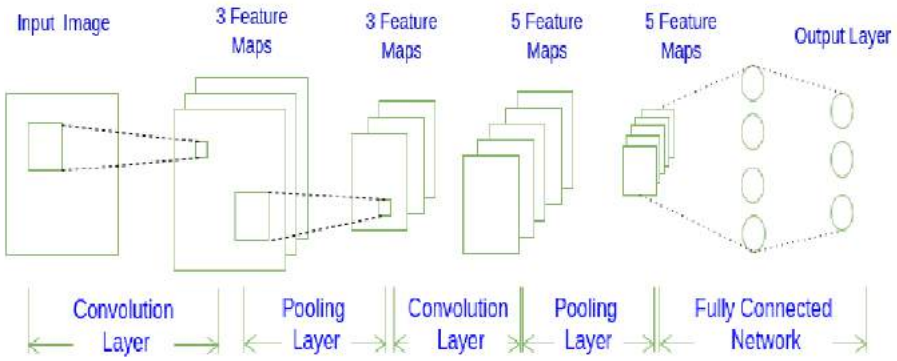

Figure 4 : Architecture of Convolutional Neural Network used

At every location, a matrix multiplication is performed and the result is added onto the feature map. Numerous convolutions are computed on the input, where each operation uses a different filter, producing different feature maps. Finally, all of these feature maps are combined as the final output of the convolution layer. The training part of this classifier can be visually observed using MATLAB library functions.

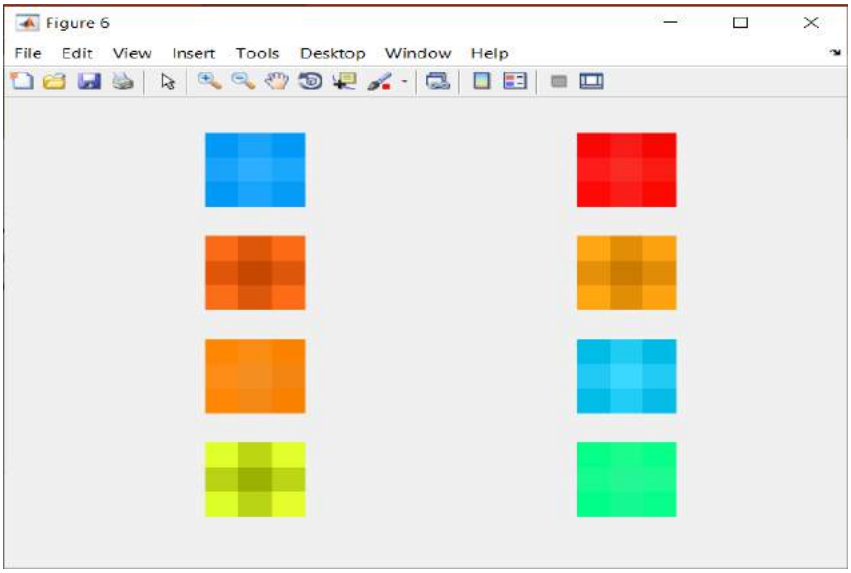

Figure 5 : Layers used in the proposed model

The training of the samples over a period of 15 iterations has successfully showcased an accuracy of 92.19%.

Results and Discussion

The BERLIN database contains 127 angry, 71 happy, 81 bored, 79 neutral, 46 disgusted, 69 afraid, and 62 sad emotional audio samples. The 400 samples used for training the classifier consists of 20 samples each of every emotion in the database except fear, where 15 samples have been utilised for training purposes.
The results obtained by the two different classifiers are analysed in detail in this section.

Performance Evaluation

Empirical parameter of accuracy is primarily used to assess the grade of performance of the classifier. Precision and Recall has also been generated in case of K-NN classifier to further asses the system. The elaborate analysis of the same is given in the following sections.

Results of K Nearest Neighbour

The highest accuracy is for neutral voice with an accuracy of 74% while the lowest accuracy is for the emotion fear with an accuracy of 70%. The emotions of fear and anger have been significantly misinterpreted as anger and joy respectively for around 20% of the samples. It achieves an overall accuracy of 74.64% and has recall and

precision values of 0.751 and 0.718 respectively.The empirical tools exploited in the assessment of the K-Nearest Neighbour (K-NN) classifier are accuracy, precision, and recall which are computed as follows:

$$Accuracy = \left(\frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions\ made}\right)100 \quad (3.1.1)$$

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)} \quad (3.1.2)$$

$$Recall = \frac{True\ Positive}{(True\ Positive + False\ Negative)} \quad (3.1.3)$$

```
accuracy = 74.64%

ans =

precision = 0.718559


ans =

Recall = 0.751067
```

Figure 6 : Results of K-NN classifier efficiency

Results of Convolutional Neural Network

A training cycle with 15 epochs ensures 45 iterations. This is accomplished over a brief period of 1 minute and 25 seconds and produces an overall accuracy of 92.19%. The primary accuracy at the first iteration is 10.94% which has increased with every iteration of convolution as the image layers were intrinsically analysed and parameters for classification were adjusted.

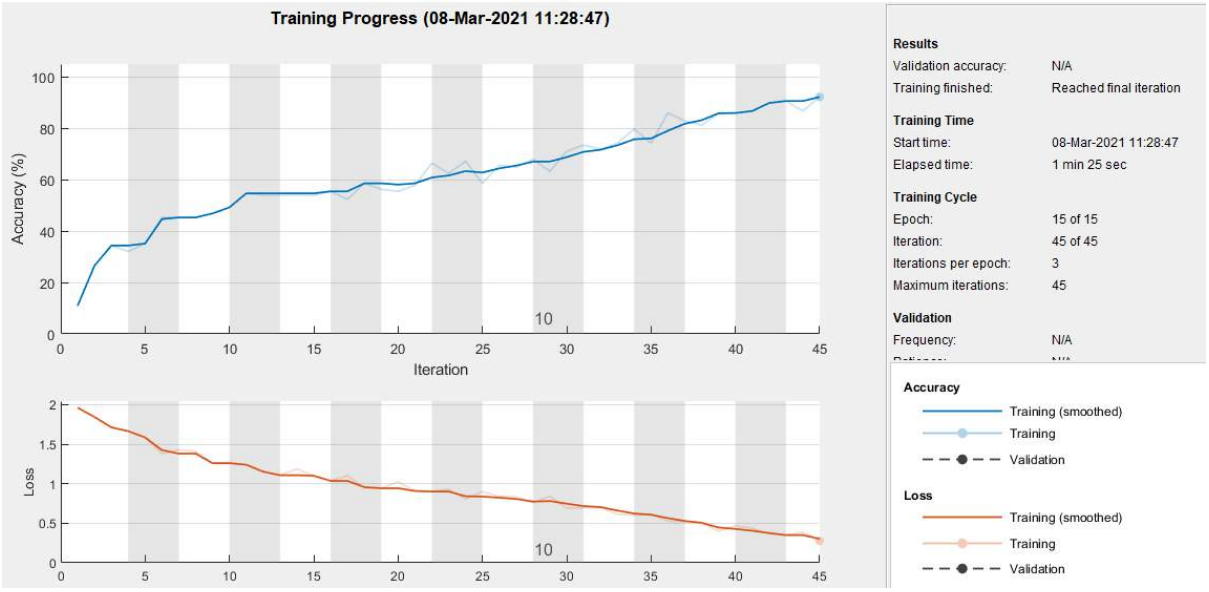

Figure 7 : Line Graph of Training Progress of sample using CNN classifier

```
Training on single CPU.
Initializing image normalization.
|=========================================================================================|
|  Epoch  |  Iteration  |  Time Elapsed  |  Mini-batch  |  Mini-batch  |  Base Learning  |
|         |             |   (hh:mm:ss)   |   Accuracy   |     Loss     |      Rate       |
|=========================================================================================|
|       1 |           1 |       00:00:02 |       10.94% |       1.9610 |          0.0100 |
|      15 |          45 |       00:01:25 |       92.19% |       0.2777 |          0.0100 |
|=========================================================================================|
```

Figure 8 : Results of CNN classifier efficiency

Objective Evaluation

Significant part of the test utterances fed to the system has been successfully identified. The emotion detected is displayed post computation in an individual window over the input signal drafted. Thus, the primary objective of the methodology has been met with.

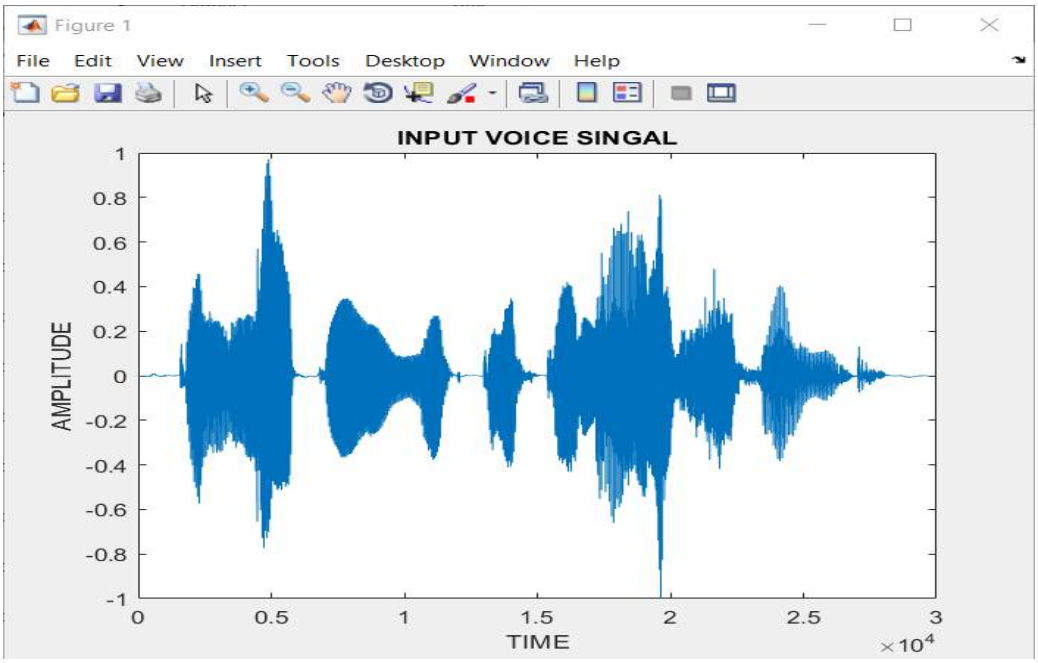

Figure 9 : Input signal – Fear

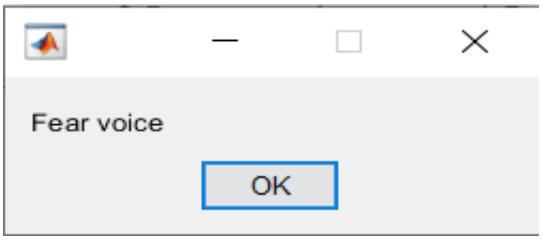

Figure 10 : Output window - Fear

3.3 INFERENCE

K-NN classifier proves to be the poorer one owing to its dependency on Euclidean distance metric. Each feature has to be weighted as per correlation with the accurate class. CNN on the other hand does not require redundant modifications since each training cycle involves numerous convolutions and filtering with the vast range of layers. Hence, under the given controlled environment, devoid of noise and other variable factors, CNN proves to be better emotional voice signal classifier.

Table 1 : Comparison of accuracy achieved by the two classifiers

Emotion	Accuracy with K-NN classifier	Accuracy with CNN classifier
Anger	70.4	91.6
Joy	71	90.33
Neutral	74	92.1
Sadness	73	91.2
Fear	70.6	90.7
Disgust	73.6	92.99
Boredom	73.99	91.45

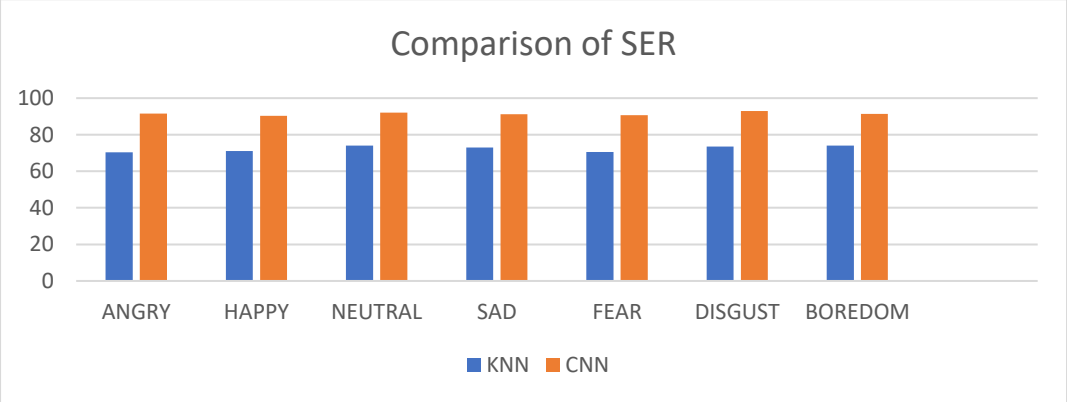

Figure 11 : Bar Graph illustration of accuracy comparison of two classifiers

Conclusion

The methodology proposed in this paper has satisfactorily identified the seven different emotions of anger, boredom, disgust, fear, sadness, happiness, and neutrality from the BERLIN database. The accuracy achieved with each approach is contingent on the features selected from the input. LPC aids prediction of future features as well, imitating the function of training process of classifier, thus enhancing accuracy. The employment of the combined feature extraction technique has ensured optimal selection of autocorrelation and cepstral coefficients; ensuring improved results over single feature based systems. While the K-NN classifier has achieved a decent improvement over the existing systems, CNN classifier has produced a high accuracy owing to the size and ratio of the dataset and the number of iterations. Thus, the primary objective of the research has been met with and this work may be further enhanced upon by experimenting with different classifiers in correlation with these coefficients.

References

[1].A.Damasio,Descartes,“Error: Emotion, Reason, and the Human Brain,” London, U.K.: Putman, 1994.

[2].A. Ortony, G. Clore, and A. Collins, “The Cognitive Structure of Emotions.” Cambridge, U.K.: Cambridge Univ. Press, 1988.

[3].R. Plutchik, “The Psychology and Biology of Emotion,” New York: HarperCollins College, 1994.

[4]. Chandra prakash, prof. V. B. Gaikwad, “analysis of emotion recognition system through speech signal using KNN, GMM AND SVM Classifier”, IJECS VOLUME.4

[5]. Hany M. Harb1, Malaka A. Moustafa, “Selecting optimal subset of features for student performance model”, IJCSI Vol. 9, Issue 5, No 1, September 2012, 1694-0814

[6].Lei Yu leiyu,Huan Liu, “Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution”, (ICML-2003), Washington DC, 2003.

[7].Mohan Ghai, Shamit Lal, Shivam Dugga l and Shrey Manik, “Emotion Recognition On Speech Signals Using Machine Learning”, 978-1-5090-6399-4/17/$31.00_c 2017 IEEE,2017 International Conference On Big Data Analytics and computational Intelligence (ICBDAC)

[8].A. Milton, S. Sharmy Roy,S. Tamil Selvi, “SVM Scheme for Speech Emotion Recognition using MFCC Feature”, International Journal of Computer Applications.

[9].F. Dellaert, T. Polzin, and A. Waibel, “ Recognizing emotion in speech,” in Proc. ICSLP , Philadelphia, PA, 1996, pp. 1970 – 1973.

[10].T. L. Nwe, S. Wei Foo, and Liyanage C. De Silva., “Speech emotion recognition using hidden Markov models,” Elsevier Speech Communications Journal Vol. 41, Issue 4, pp. 603-623,November 2003.

[11].X. Arputha Rathina, K. M. Mehata, M. Ponnavaikko, "A Study of Prosodic Features of Emotional Speech", Second International Conference on Computer Science, Engineering and Applications., AISC 166

[12]. Md. Afzal Hossan, Sheeraz Memon, Mark A Gregory, "A novel approach for MFCC feature extraction", 4th International Conference on Signal Processing and Communication Systems, 2010

[13]. A. N. Mishra, M. C. Shrotriya, S. N. Sharan, "Comparative Wavelet, PLP and LPC Speech Recognition Techniques on the Hindi speech digits Database", The International Society for Optical Engineering ,7546, February 2010

[14]. B. Srinivas, G. Sasibhushana Rao, "A Hybrid CNN-KNN Model for MRI brain Tumor Classification", International Journal of Recent Technology and Engineering (IJRTE)ISSN: 2277-3878, Volume-8 Issue-2, July 2019