JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY, Noida

Department of CSE

Bachelor of Technology, 6th Semester

MINOR PROJECT

Speech Emotion Recogntion

2020-2021

Submitted By:

1. Kunal sharma– 17103265(B7)
2. Mayur Bansal – 17103291(B7)
3. Aditya Maheshwari - 17103287(B7)

Submitted To:

Dr. Dharamveer

Ms. Mridula Sharma

Guidance Of:

Dr. Sangeeta Mittal

Certificate of Assurance

This is to certify that the work titled “Speech Emotion Recognition” submitted by Kunal Sharma, Mayur Bansal, Aditya Maheshwari in partial fulfilment for the award of degree of B.Tech, Jaypee Institute of Information Technology, Noida has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

Signature of Supervisor .....

Name of Supervisor .....

Designation .....

Date .....

Acknowledgement

We would like to express our deepest gratitude to our Supervisor. Dr. Sangeeta Mittal, Department of Computer Science and Engineering, Jaypee Institute of InformationTechnology, Noida for her constant guidance and encouragement in this project work. The credit for the successful completion of this project goes to her without whom our endeavour would have been futile.

We owe our sincere thanks to the panel in our project presentation who have contributed in improving our presentation skills with their comments and advice. Lastly, We would like to acknowledge the crucial role of the staff of the CSE Department who have always been willing to help us with their abilities whenever needed.

Signature of students			
Name of students			
Enrollment of students			

Date: ……

1. INTRODUCTION

Understanding human-speech has been an integral and fascinating part of AI as well as Digital Speech Processing for a long time. Human speech contains not only the linguistic content but also contains some emotions of the speaker. Human beings are always motivated to develop a machine which understands and communicates like them, for this necessary condition is a speech database.

Fig. 1, Speech Emotion Recognition

Speech technology plays a vital role in development of application for common public like Interactive Voice Recognition System for railways, agriculture, Counselling and determining client's medical state, during healthcare determining patients feeling and comfort level about the treatment, In the case of autism, struggling to interpret expressions, etc.

2. PROBLEM STATEMENT

Speech Emotion Recognition is a research area problem which tries to infer the emotion from the speech signals. This project gives a description of different types of studies conducted to analyze, perceive and recognize commonly occurring emotions in Hindi speech on the subset of IIT-KGP SEHSC: Simulated Emotion Hindi Speech Corpus.

3. LITERATURE SURVEY

3.1 Emotional Hindi Speech: Feature Extraction and Classification

By: Sweeta Bansal, Amita Dev

In this paper attempts are made to present the features of emotional Hindi language, for identification and classification.Emotional speech classification is performed on the basis of LPCC and MFCC features are also performed using K-mean cluster analysis.Efforts has been successful and found to be correct in 77%

in case of happy and 84% in case of sad, which is much higher than emotion recognition done on taking only any one cepstral coefficients on hindi speech database [1].

3.2 Identification of Hindi Dialects and Emotions using Spectral and Prosodic features of Speech

By: K Sreenivasa Rao, Shashidhar G Koolagudi

In this paper, researchers have explored speech features to identify Hindi dialects and emotions.Indian Institute of Technology Kharagpur Simulated Emotion Hindi Speech Corpus (IITKGP-SEHSC) is used for conducting the emotion recognition studies. Spectral features are represented by Mel frequency cepstral coefficients (MFCC) and prosodic features are represented by durations of syllables, pitch and energy contours. ANN and SVM are used for classification of the features, An accuracy of 78% and 81% was achieved [2].

3.3 Speech Emotion recognition System Using SVM and LIBSVM

By: Sujata B. Wankhede, Pritish Tijare, Yashpalsing Chavhan

This paper introduces an approach to emotion recognition from speech signals using SVM as a classifier. The speech features such as, Mel Frequency cepstrum coefficients (MFCC) and Mel Energy Spectrum Dynamic Coefficients (MEDC) are extracted from speech utterance. The recognition rates by implementing SVM are 62% and 71.66% for Berlin database and Hindi database respectively [3].

3.4 Speech emotion recognition of Hindi speech using statistical and machine learning techniques

By: Akshat Agrawal & Anurag Jain

This paper introduces an approach to emotion recognition from speech. Initial process is to extract the basic features of speech like Prosodic Features (PF) and Acoustic Features (AF) after considering all these facts, features extraction and classification using statistical technique Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) used and then emotional state analysis done by K-Nearest Neighbor (KNN) algorithm and Naïve Bayes Classifier (NBC). Due to no openly accessible database in focused language, the dataset was made with 4 speakers, 25 sentences for emotions(Anger, Happiness, Neutral, Sad, Surprise)[5].

3.5 Speech based Emotion Recognition using Machine Learning

By: Girija Deshmukh, Apurva Gaonkar, Gauri Golwalkar, Sukanya Kulkarni

In this paper, features like tone pitch and energy are used to train the model. The model is trained on The North American English language open source dataset was divided into training and testing manually. Further testing is done on manually recorded dataset of Indian English, marathi and hindi languages. Classification is done on features like MFCC, pitch and STE.The test dataset will undergo the extraction procedure following which the classifier (SVM is used here) would make a decision regarding the underlying emotion in the test audio.The accuracies obtained on all the features were 79.48% [6].

3.6 Emotion state detection via speech in spoken Hindi

By: Prakhar Kulshreshtha, Soumya Gayen

In this project, a simulated Hindi emotional speech database has been borrowed from a subset of IITKGP-SEHSC dataset(2 out of 10 speakers). The spectral features used are Mel Frequency Cepstral Coefficients(MFCCs) and Subband Spectral Coefficients(SSCs) The feature vector are used.This dataset is trained on multiple classifiers,such as SVM, ANN, MLP, KNN, Adaboost, Random forests.Support VectorMachines with Radial Basis Function kernel provides the best accuracy rates,with accuracy for male dataset being 89.08% and for female dataset being 83.16% [7].

3.7 Emotion Recognition fromHindi Speech Signal

By: Meenakshi Singh, Parul Khullar

This project is an effort to recognize emotion from Hindi speech signals. These signals were preprocessed and analyzed using various techniques like: cepstral , linear prediction coefficient etc.. In feature extraction of speech signals , the system uses various parameter features : fundamental frequency, pitch contour, formants, duration(pause length ratio) etc. to form a feature vector and then Knn Classifier were used to classify and recognize those emotions. The feature vector for each sample formed from large set features of different voices has an accuracy at an average of about 50% [8].

3.8 Visual Speech Recognition Using Optical Flow and Hidden Markov Model

By: Usha Sharma, Sushila Maheshkar, A. N. Mishra , Rahul Kaushik

The present work proposes audio-visual speech recognition with the use of Gammatone frequency cepstral coefficient (GFCC) and optical flow (OF) features with Hindi speech database.The OF refers to the distribution of apparent velocities of brightness pattern movements in an image. For classification Hidden Markov Model was used. The GFCC shows almost comparable results with MFCC in a clean environment; however, its performance goes down in a noisy environment. In the clean environment GFCC gives an accuracy of 93.76% (for audio only) while a combination of GFCC and OF gave an accuracy of 93.12% (for audio + video). While in a noisy environment the accuracy of GFCC ranges from 82.91% - 72.25% depending on the level of noise in audio only while combination of GFCC and OF gave an accuracy in the range of 88.78% - 80.05% on audio + video signals.[9]

3.9 Bagged support vector machines for emotion recognition from speech

By: Anjali Bhavan, Pankaj Chauhan, Hitkul, Rajiv Ratn Shah

The research is based on improving the current SVM technique using Ensemble learning. For that the features extracted are MFCC and the dataset used are EmoDB (Berlin Audio), RAVDESS (English Audio), IITKGP-SEHSC (Hindi Audio) and the classifiers used are SVM, Bagged Ensemble of SVM and AdaBoost ensemble of SVM. For the EmoDB Database, the accuracy are 92.45% and 87.32% respectively, for the RAVDESS database, the accuracy are 75.69% and 72.10% respectively and for the IITKGP-SEHSC Database, the accuracy are 84.11% and 77.19% respectively.

3.10 Deep features-based speech emotion recognition for smart affective services

By: Abdul Malik Badshah, Nasir Rahim, Noor Ullah, Jamil Ahmad, Khan Muhammad, Mi Young Lee, Soonil Kwon & Sung Wook Baik

In this paper,a study of speech emotion recognition based on the features extracted from spectrograms using a deep convolutional neural network (CNN) with rectangular kernels. Typically, CNNs have square shaped kernels and pooling operators at various layers, which are suited for 2D image data. However, in case of spectrograms, the information is encoded in a slightly different manner. Time is represented along the x-axis and y-axis shows frequency of the speech signal, whereas, the amplitude is indicated by the intensity value in the spectrogram at a particular position.The proposed scheme effectively learns discriminative features from speech spectrograms and performs better than many state-of the-art techniques when evaluated its performance on Emo-DB and Korean speech dataset. Features extracted were MFCC and classifiers used were SVM, Decision Tree, Random Forest, AlexNet and the above proposed model giving an accuracy of 61.48%, 63.24%, 71.15%, 79.14 and 86.54% respectively.[11]

In recent years, a great deal of research has been done to recognize human emotion using speech information. According to the research papers we studied, we came to know about the many available techniques to extract features, like for English and Berlin language, methods such as MFCC, MEDC, MS and many other from the available Librosa library have been used and while in case of Hindi language methods like MFCC, LPCC, SSC,etc. methods were used and for classification algorithms like k-NN, SVM, RNN, HMM, GMM etc. were used.

So, on the basis of our research and knowledge about the feature extraction and their classification, we have decided to do a comparative study on both English and Hindi dataset. We found that the most prominent feature extraction techniques were MFCC, SSC and LPC and similarly for classification were k-NN and SVM, and in case of neural networks we found CNN and MLP to be the most suitable for this research purpose. Our main aim is to predict better results for the Hindi dataset.

4. DATASET USED

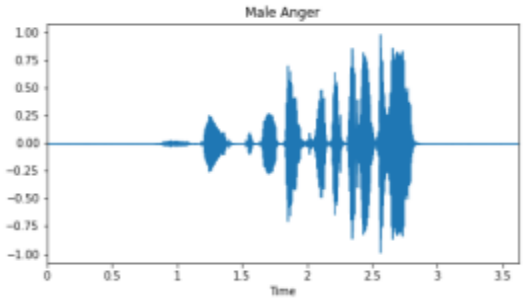
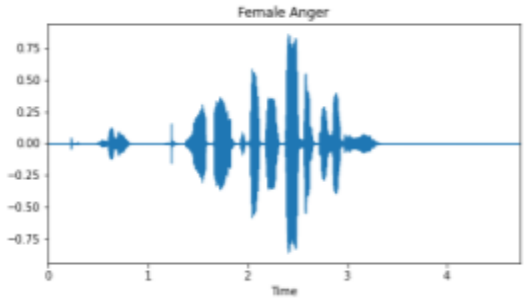
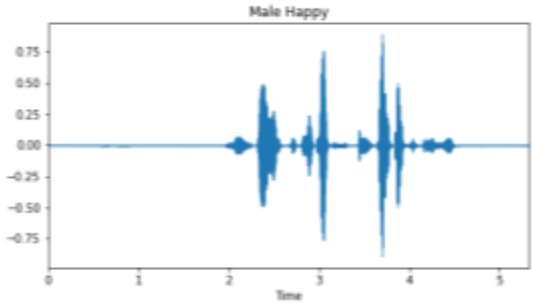
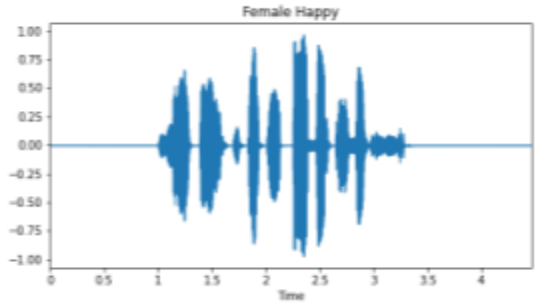
The dataset that we have used is a subset of IIT-KGP SEHSC: Simulated Emotion Hindi Speech Corpus (2 male and 2 female audio were provided)[4].

The proposed database is recorded using 10 (5 male and 5 female) professional artists from Gyan Vani FM radio station, Varanasi, India. All the sentences are emotionally neutral in meaning. Each of the artists has to speak 15 sentences in 8 basic emotions in one session. The number of sessions considered for preparing the database is 10. The total number of utterances in the database is 12000 (15 text prompts × 8 emotions × 10 speakers ×10 sessions). Each emotion has 1500 utterances. The eight emotions considered for collecting the proposed speech corpus are: anger, disgust, fear, happy, neutral, sad, sarcastic and surprise [4].

Below is a side-by-side comparison of 2 out of 8 emotions by Male and Female speakers by plotting waveform and spectrogram.

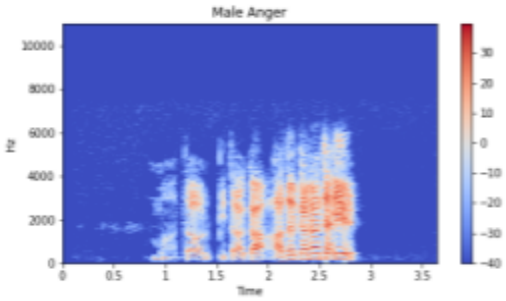
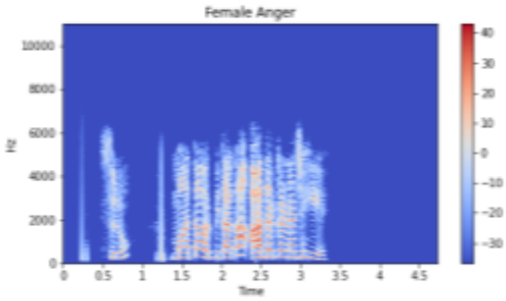
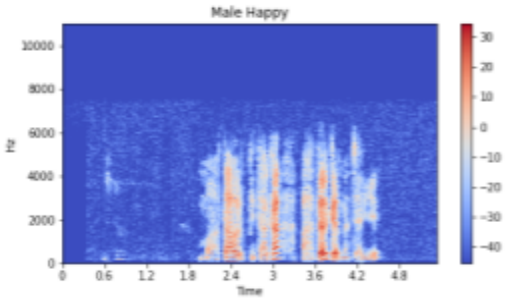
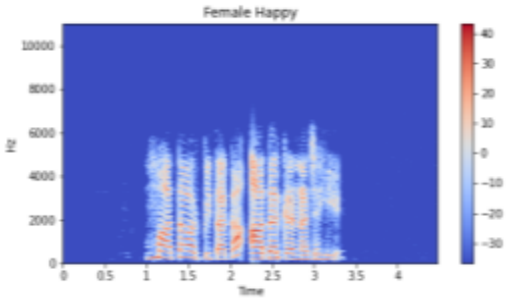
4.1 WAVEFORM

Waveform of a signal is the shape of its graph as a function of time. Louder the speaker produces a voice, the higher will be the peak of the plot at that instant of time.

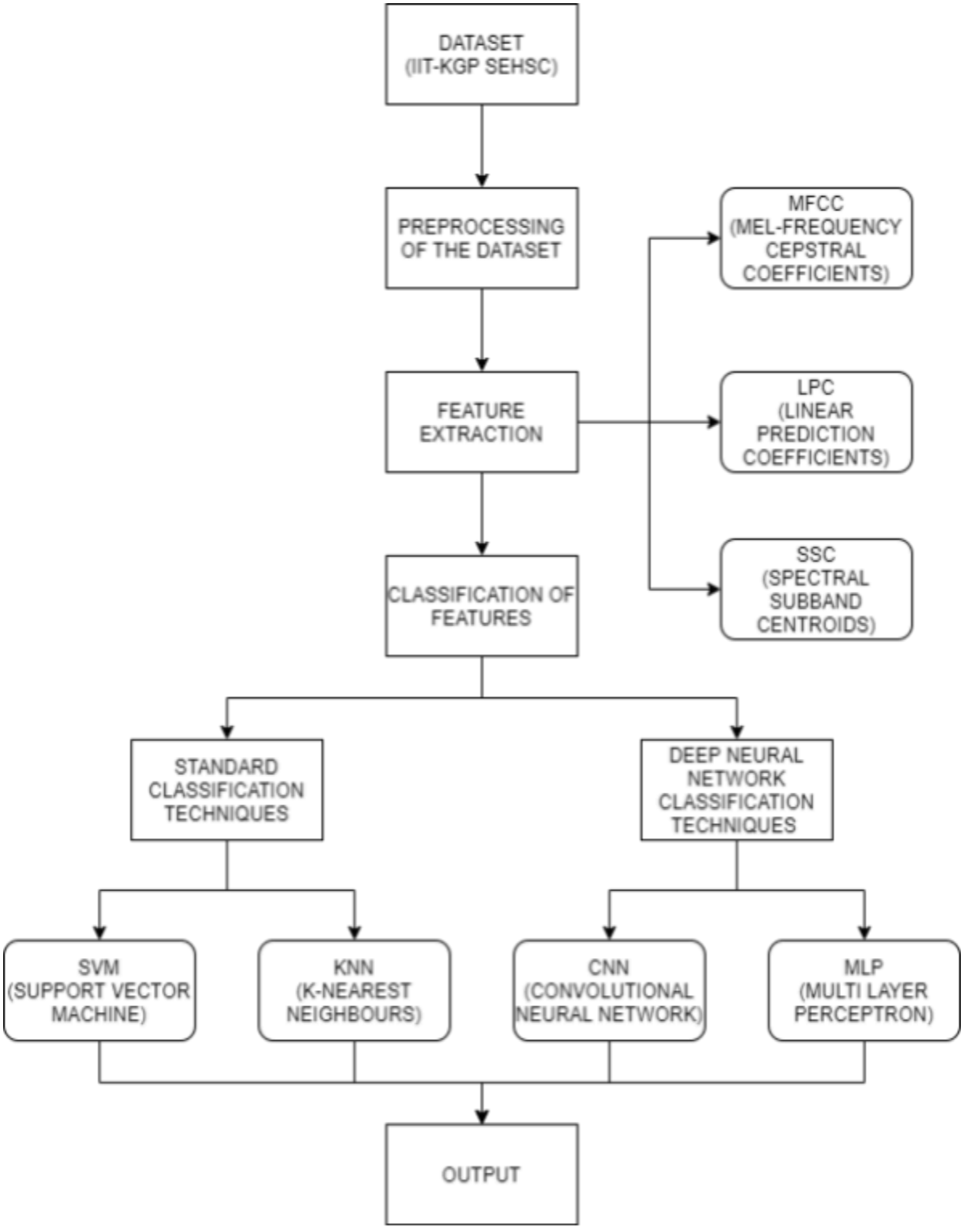
	MALE	FEMALE
ANGER		
HAPPY		

4.2 SPECTROGRAM

In the spectrogram view, the vertical axis displays frequency in Hertz, the horizontal axis represents time (just like the waveform display), and amplitude is represented by brightness. The blue background is silence, while the white curve is the sine wave moving up in pitch.

	MALE	FEMALE
ANGER		
HAPPY		

5.FLOW CHART


DATASET
(IIT-KGP SEHSC)
PREPROCESSING OF THE DATASET
FEATURE EXTRACTION
CLASSIFICATION OF FEATURES
MFCC
(MEL-FREQUENCY CEPSTRAL COEFFICIENTS)
LPC
(LINEAR PREDICTION COEFFICIENTS)
SSC
(SPECTRAL SUBBAND CENTROIDS)
STANDARD CLASSIFICATION TECHNIQUES
DEEP NEURAL NETWORK CLASSIFICATION TECHNIQUES
SVM
(SUPPORT VECTOR MACHINE)
KNN
(K-NEAREST NEIGHBOURS)
CNN
(CONVOLUTIONAL NEURAL NETWORK)
MLP
(MULTI LAYER PERCEPTRON)
OUTPUT

6. WORK PLAN

The problem of emotion speech analysis is basically divided into three steps, i.e. Preprocessing of the dataset, Feature Extraction and then finally Classifying the obtained features.

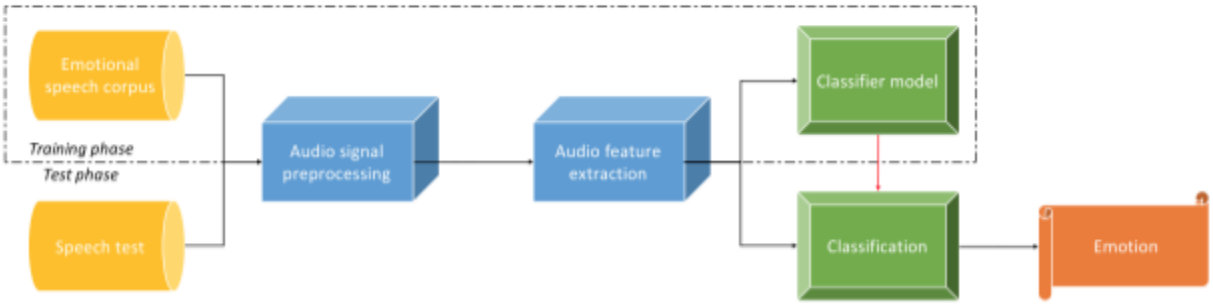


Fig. 11 Flow of recognising emotion from speech

6.1 PREPROCESSING OF DATASET

Preprocessing is the very first step after collecting data that will be used to train the classifier in a SER system. Some of these preprocessing techniques are used for feature extraction, while others are used to normalize the features so that variations of speakers and recordings would not affect the recognition process.

6.2 FEATURE EXTRACTION

6.2.1 Comparison with Languages other than Hindi

There are two types of features considered in an audio signal,

- **Cepstral Features** - Cepstral based Features, which typically represent the magnitude properties of speech spectrum, are widely used in speech processing.
- **Prosodic Features** - Generally prosody means, “the structure that organizes sound”. Tone, loudness and the rhythm (tempo) structures are the main components of prosody.

In this research, we have used the most widely used Cepstral Features MFCC, SSC and LPC. Since the features are based on the magnitude properties, the features are not language specific, as the features are extracted using the magnitude properties. So the same techniques can be used for Audio Datasets other than Hindi. We have taken into account this as future work and would love to compare the results for other languages with hindi results in the future.

But while considering the Prosodic Features, we have to keep in mind the language and manipulate the features based on that only, as Prosodic Features takes Tone, loudness and rhythm into account, which can vary significantly depending upon the language chosen.

We have used the following feature extraction techniques :

6.2.2 MFCC - One popular audio feature extraction method is the Mel-frequency cepstral coefficients (MFCC) which have 39 features. The feature count is small enough to force us to learn the information of the audio. 12 parameters are related to the amplitude of frequencies. It provides us with enough frequency channels to analyze the audio.
Fig 12 is the flow of extracting the MFCC features.

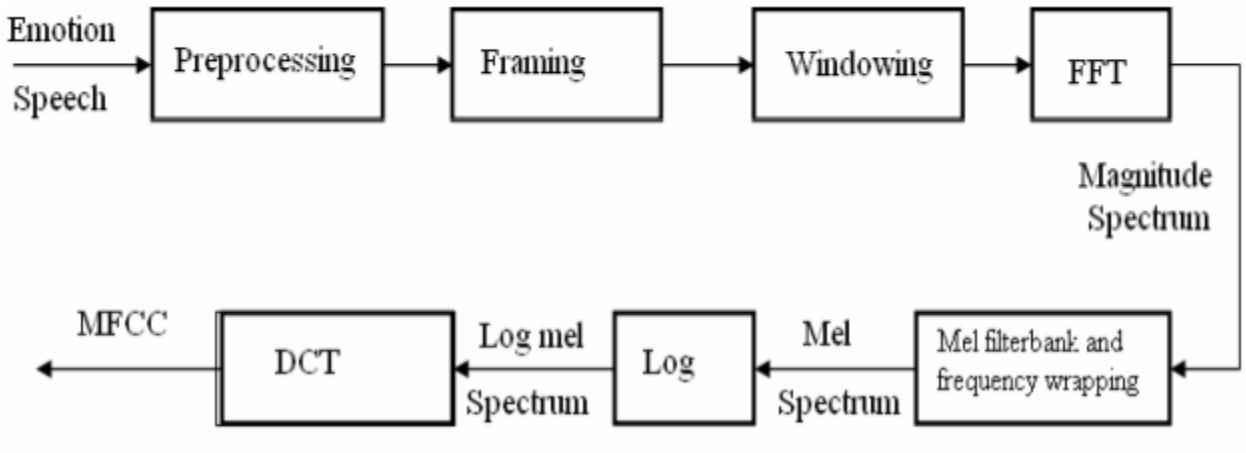


Fig. 12, Flow of extracting MFCC features

6.2.3 SSC - Spectral Subband Centroids (SSC) are computed as the centroid frequencies of subband spectra and they give the locations of the local maxima of the power spectrum. Recognition accuracy of SSCs is lower in noise-free conditions compared with MFCCs. However, SSCs can outperform MFCCs in noisy conditions.

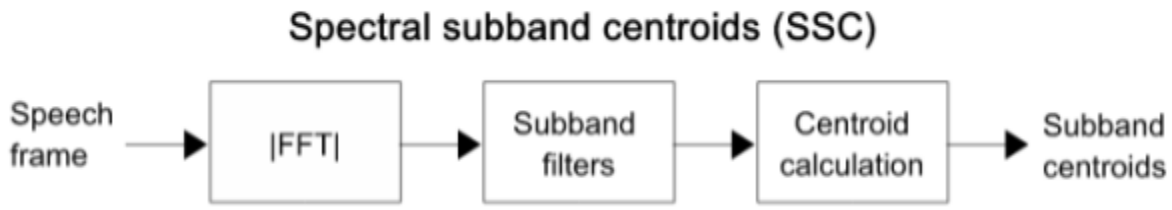


Fig. 13, Flow of extracting SSC features

6.2.4 LPC - Linear prediction coefficients (LPC) imitates the human vocal tract and gives robust speech features. It evaluates the speech signal by approximating the formants, getting rid of its effects from the speech signal and estimates the concentration and frequency of the left behind residue. The result states each sample of the signal as a direct incorporation of previous samples. The coefficients of the difference equation characterize the formants, thus, LPC needs to approximate these coefficients. LPC is a powerful speech analysis method and it has gained fame as a formant estimation method. Below is a flow of extracting LPC features.

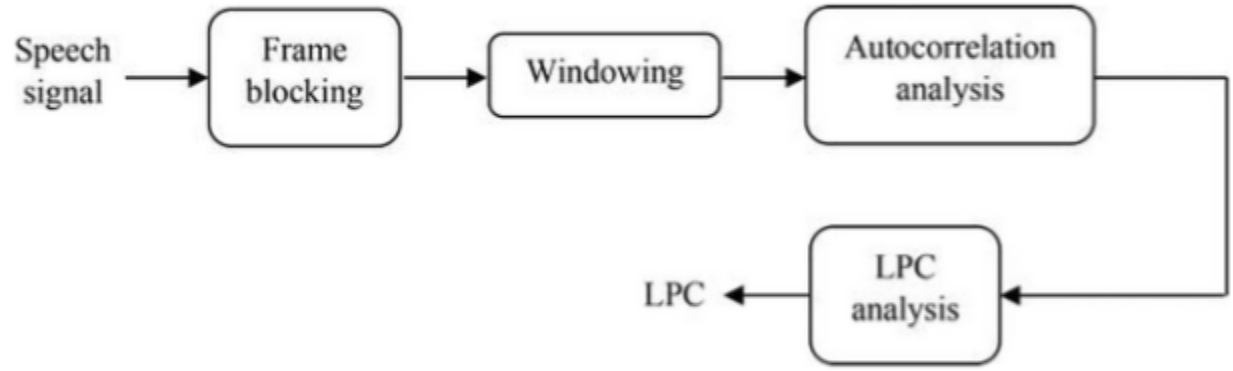


Fig. 14, Flow of extracting LPC features

6.3 CLASSIFICATION OF EXTRACTED FEATURES

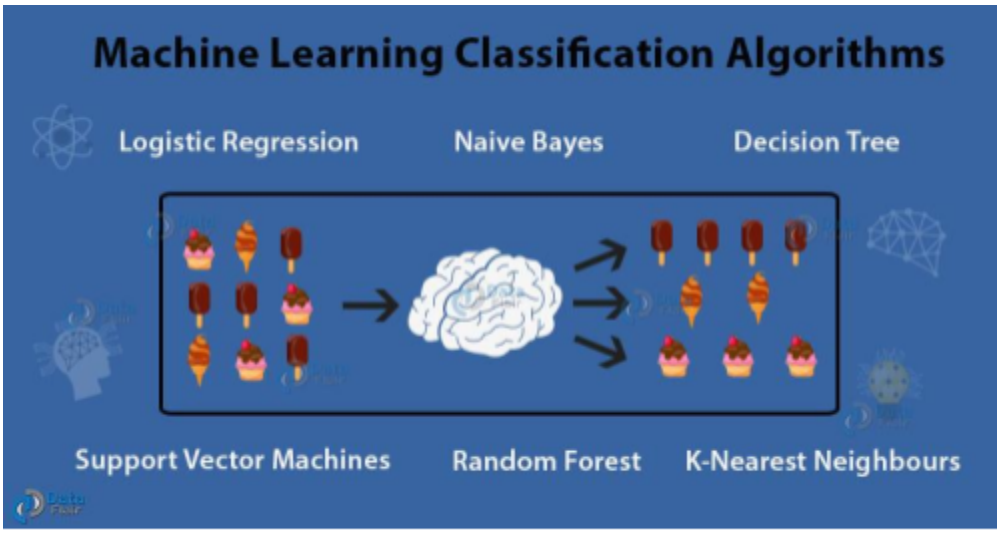


Fig. 15, ML Classification Algo

We have divided the classification into two types and that is by using:

6.3.1 Standard Classification Techniques

6.3.2 Deep Neural Network Classification Techniques

6.3.1 STANDARD CLASSIFICATION TECHNIQUES

6.3.1.1 SVM - The Support Vector Machine (SVM) is trained according to labeled features. The SVM kernel functions are used in the training process of SVM. Binary classification can be viewed as the task of separating classes in feature space. SVM is a binary classifier, but it can also be used as a multiclass classifier.

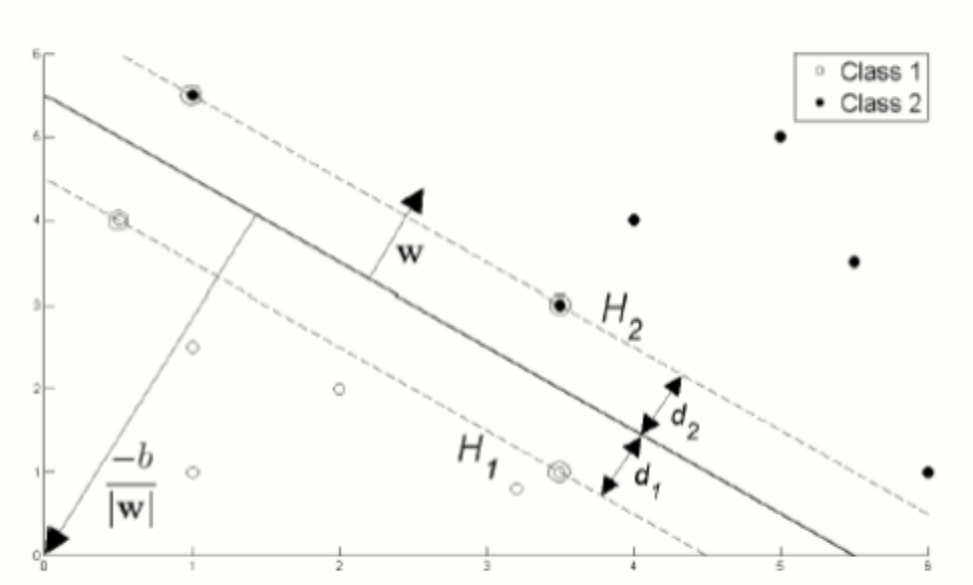


Fig. 16, Support Vector Machine (SVM) Classifier

6.3.1.2 KNN - The K-Nearest Neighbors is a non-parametric method used for classification and regression. In both cases, the input consists of the *k* closest training examples in the feature space.

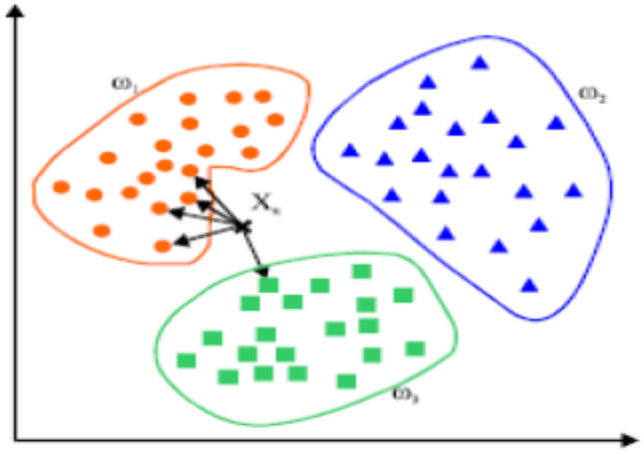


Fig. 17, KNN Classifier Algorithm

6.3.2 DEEP NEURAL NETWORK CLASSIFICATION TECHNIQUE

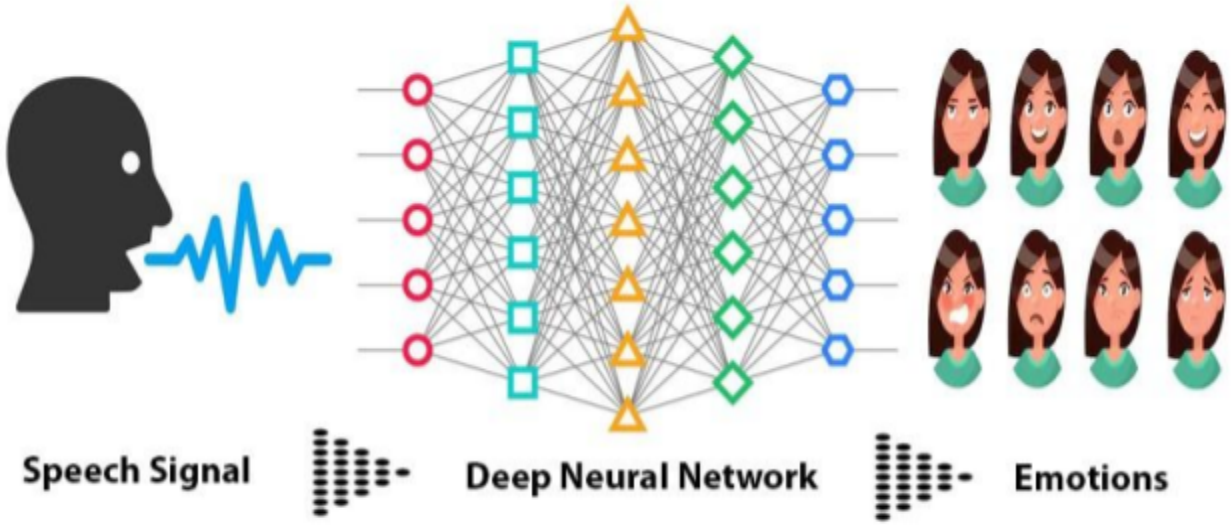


Fig. 18, SER using deep neural network

6.3.2.1 CNN - CNNs(Convolutional Neural Network) are the popular variants of deep learning that are widely adopted in ASR(Automatic Speech Recognition) systems. CNNs have many attractive advancements, i.e., weight sharing, convolutional filters, and pooling. Therefore, CNNs have achieved an impressive performance in ASR. CNNs are composed of multiple convolutional layers. Fig shows the block diagram of CNN.

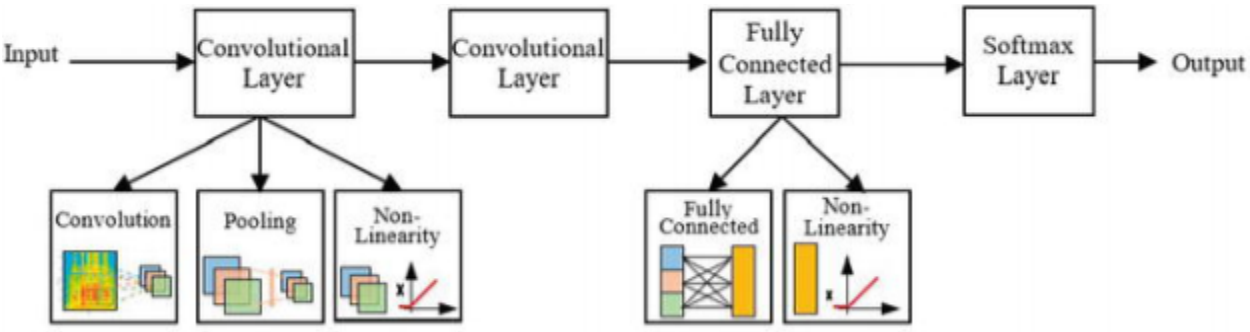


Fig. 19, flow of CNN

6.3.2.2 MLP - Multi-layer perceptron, fig 20, is a supervised learning algorithm that learns a function f(X)=Rn:Rn→R0 by training on a speech dataset, where n is the number of dimensions for input and 0 is

the number of dimensions for output. Given a set of features X=x1,x2,…,xn , it can learn a nonlinear function approximator for either classification.

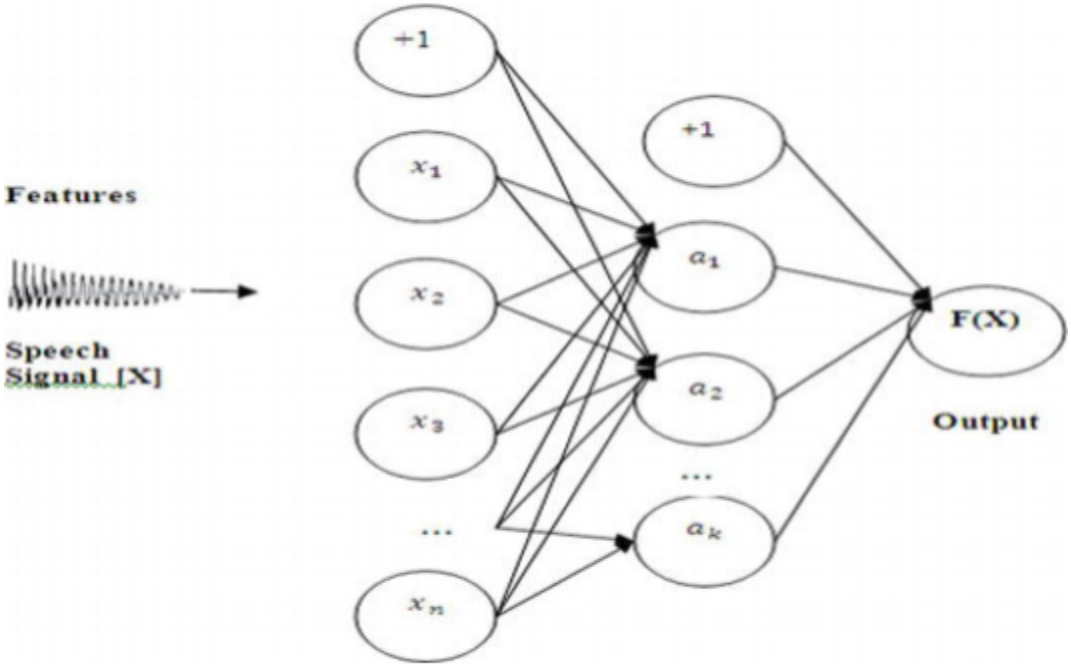


Fig. 20, MLP Classifier Structure

7. RESULTS:

We used the above described techniques to get the best results and these are the results we got by using the following:

7.1 SVM : We used the SVM classifier with three different kernels,

7.1.1 Linear

7.1.2 Poly

7.1.3 Rbf

The result for all the kernels with all features combined gave the accuracy bar graph presented in fig 21.

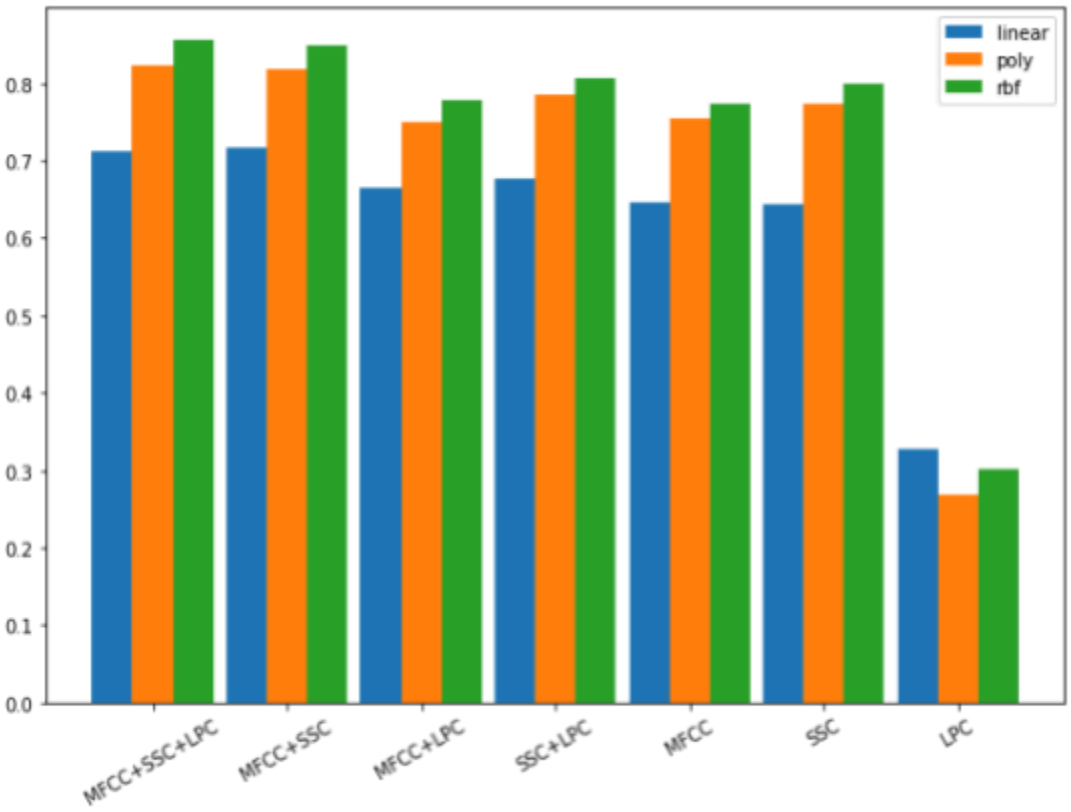


Fig. 21, SVM Classifiers Accuracy Comparison

7.1.1 Linear Kernel: The maximum accuracy achieved for this kernel is by combining MFCC and SSC and is 71.68 %. Confusion matrix for the same is shown in fig 22.

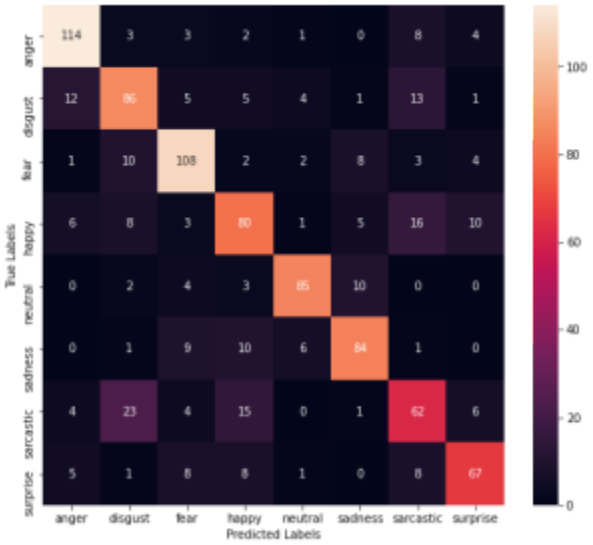


Fig. 22, Confusion Matrix for SVM (Linear)

7.1.2 Poly Kernel: The maximum accuracy achieved for this kernel is by combining all three features MFCC, SSC and LPC and is 82.23 %. Confusion matrix for the same is shown in fig 23.

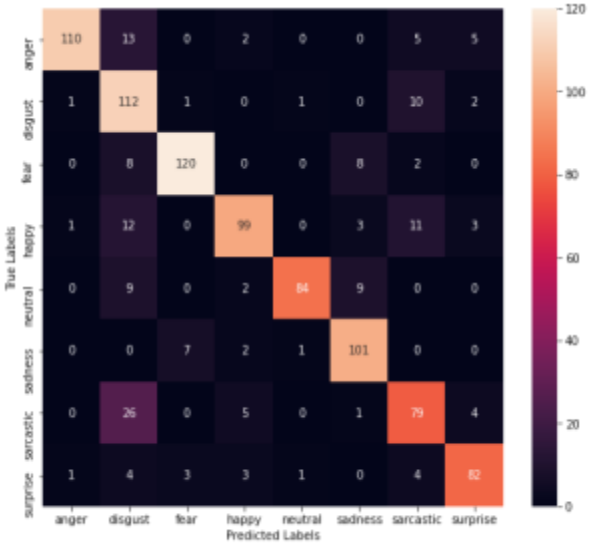


Fig. 23, Confusion Matrix for SVM (Poly)

7.1.3 RBF Kernel: The maximum accuracy achieved for this kernel is by combining all three features MFCC, SSC and LPC and is 85.47 %. Confusion matrix for the same is shown in fig 24.

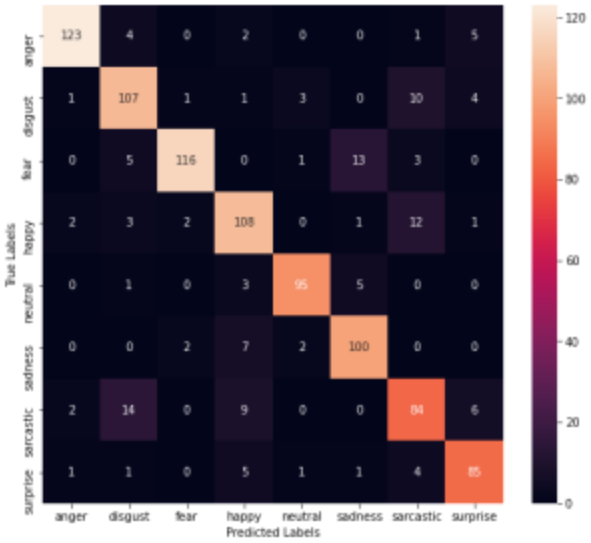


Fig. 24, Confusion Matrix for SVM (RBF)

7.2 KNN: Classifying all the 3 features(MFCC, SSC, LPC) with KNN by considering the neighbors from 1 to 30, the testing results are shown in fig 25.

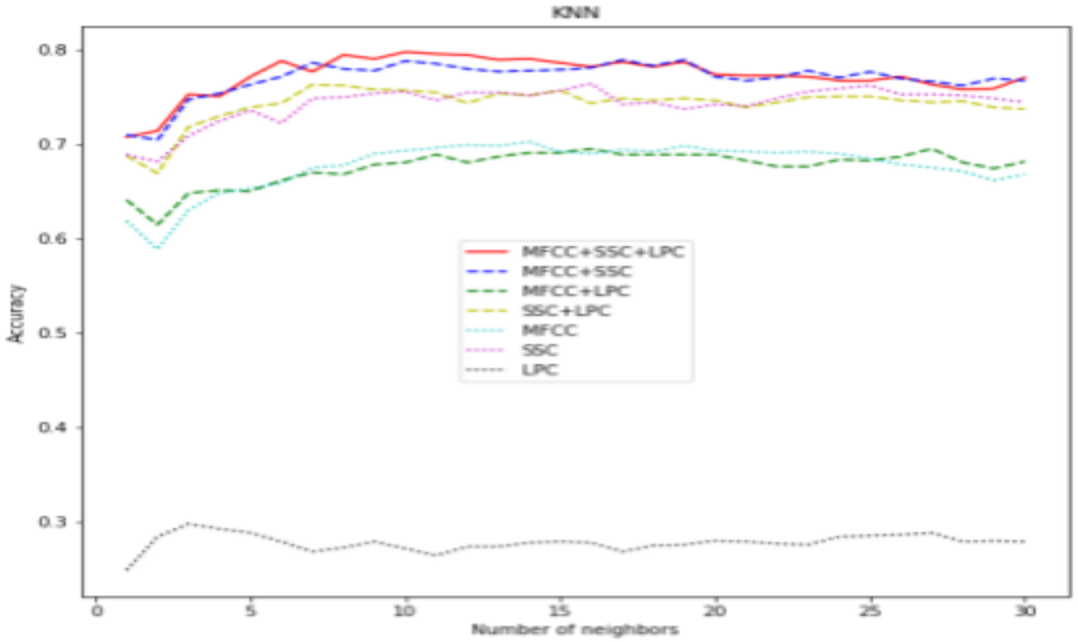


Fig. 25, KNN Classifier Accuracy Comparison with no. of neighbors

As seen from the fig 25, the maximum accuracy by using the KNN Classifier is achieved by combining all the three features MFCC+SSC+LPC and by taking neighbors=10 and is 79.72 %. Confusion matrix for the same is shown in fig 26.

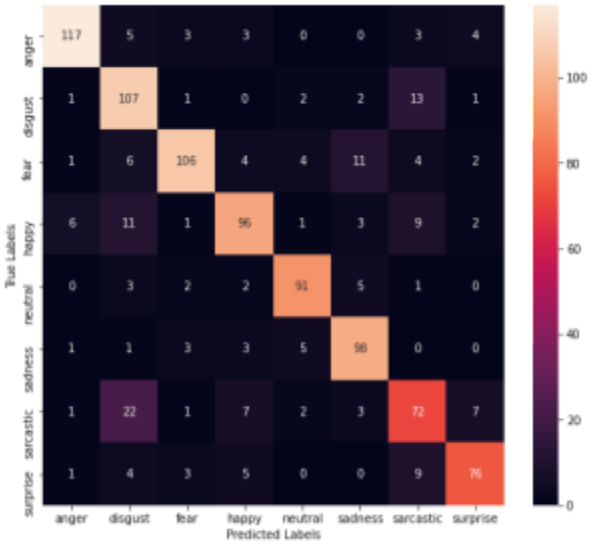


Fig. 26, Confusion Matrix for KNN (n=10)

7.3 CNN

CNN model applied on *150* epochs with *20* batch size, *sparse_categorical_crossentropy* as loss function and *accuracy* as evaluation metric. when the model converges the following accuracies were obtained on the combination of various features. The testing results are shown in fig 27.

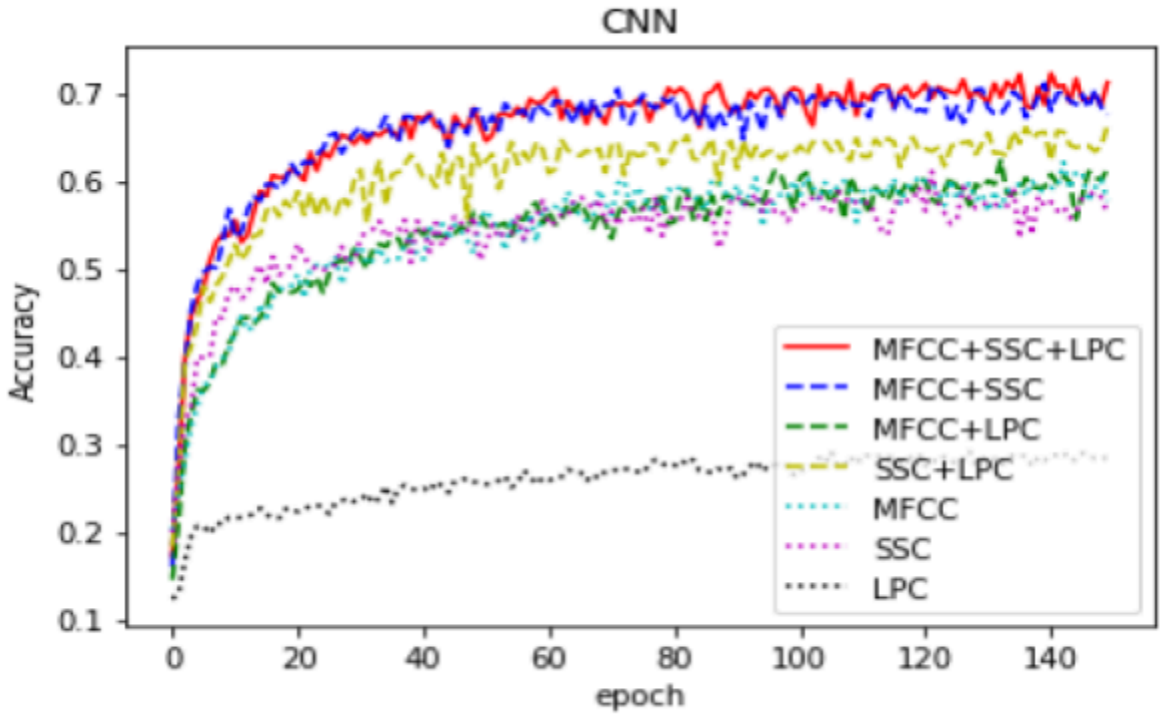


Fig. 27, CNN Accuracy Results Comparison with no. of epoch

Varying accuracies were obtained when different combinations of features were considered. Maximum Accuracy achieved is 71.16% by combining all three features MFCC, SSC and LPC. Accuracy bar graph obtained over features using CNN model is shown in fig 28.

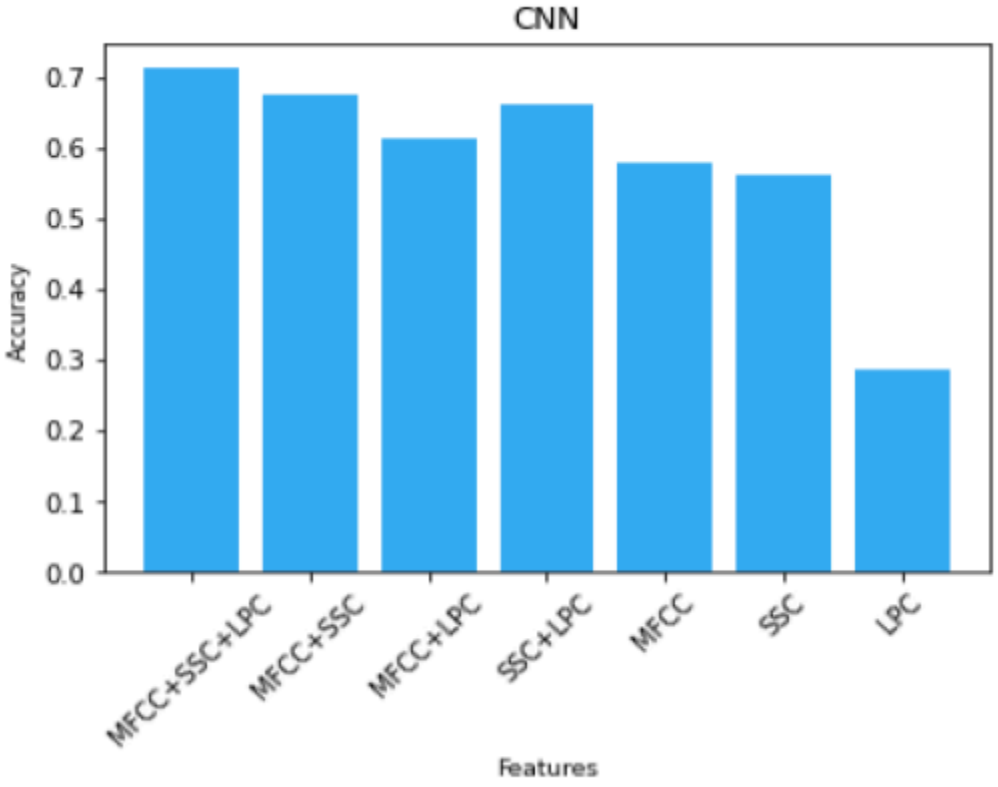


Fig. 28, CNN Results

Fig 29, is the confusion matrix for combination of features(MFCC, SSC, LPC) giving a maximum accuracy of 71.16%.

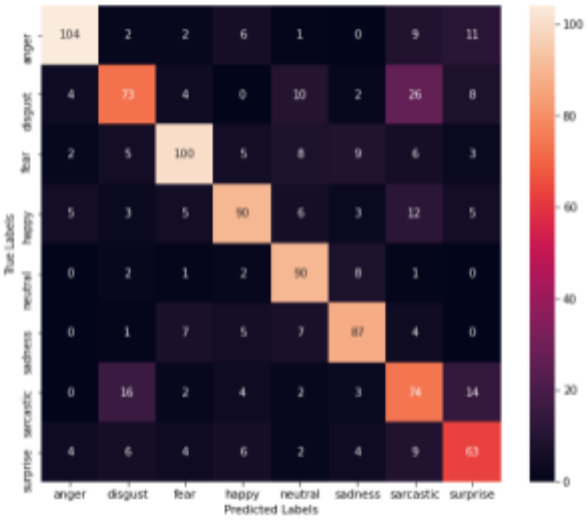


Fig. 29, Confusion Matrix for CNN

7.4 MLP

Multi Layer Perceptron was applied for 500 iterations with a learning rate of 0.01 and a batch size of 256 and when the model converges the following accuracies were obtained on the combination of various features. Accuracy bar graph obtained over features using MLP model is shown in fig 30.

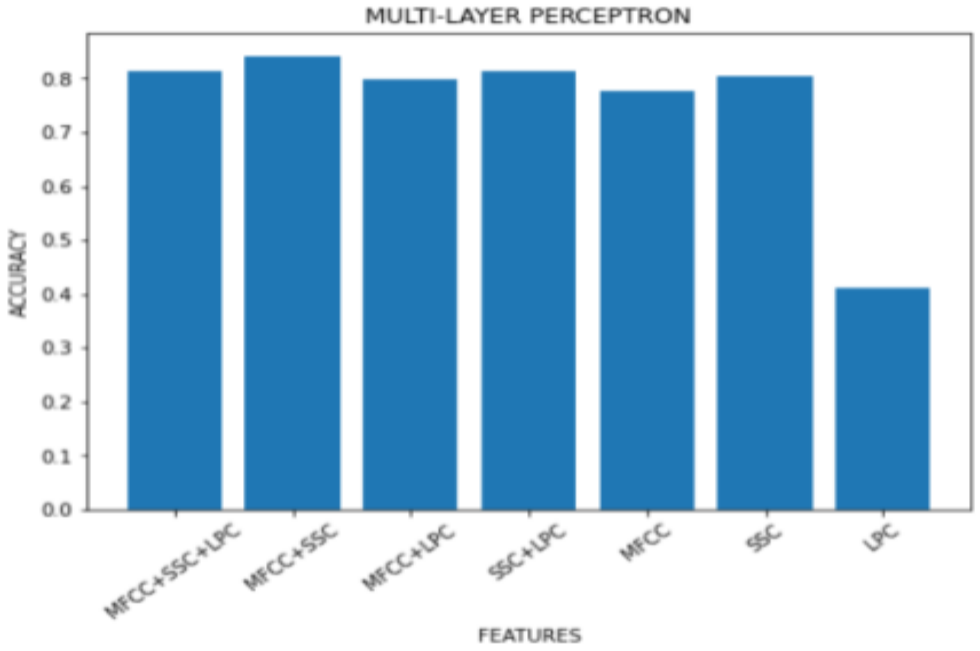


Fig. 30, MLP Results

As seen from the fig 30, The max accuracy was obtained for a combination of 2 features, i.e., MFCC + SSC and it came out to be 84.22%.

Fig 31, is the confusion matrix for combination of features(MFCC, SSC) giving a maximum accuracy of 84.22%.

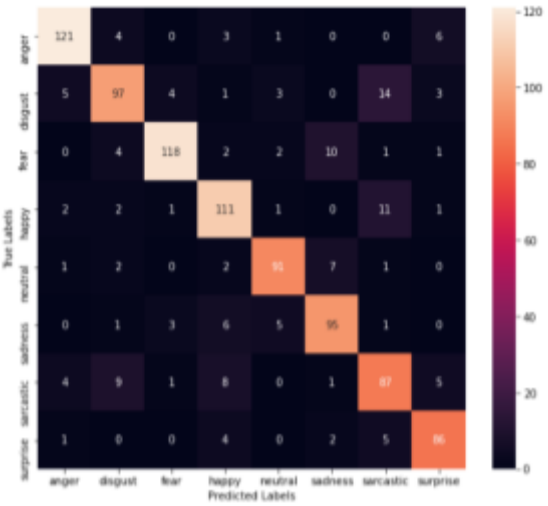


Fig. 31, Confusion Matrix for MLP

ACCURACY FOR ALL CLASSIFIERS(%)

FEATURES / CLASSIFIERS	MFCC, SSC, LPC	MFCC, SSC	MFCC, LPC	SSC, LPC	MFCC	SSC	LPC
SVM (Linear)	71.26	**71.68**	66.56	67.60	64.57	64.36	32.81
SVM (Poly)	**82.23**	81.81	75.02	78.47	75.54	77.42	26.75
SVM (RBF)	**85.47**	84.84	77.84	80.66	77.42	79.83	30.09
KNN	**79.72 (k=10)**	78.89 (k=19)	69.48 (k=16)	76.28 (k=7)	70.21 (k=14)	76.38 (k=16)	29.78 (k=3)
CNN	**71.16**	67.50	61.23	66.25	57.68	56.22	28.53
MLP	81.40	**84.22**	79.73	81.40	77.85	80.36	41.07

8. COMPARISON WITH WORK FROM RESEARCH PAPERS

Author	Year of Publication	Dataset source & emotions considered	Language of Speech	ML Technique	Features Extracted	Accuracy (%)	Best Advantage	Worst limitation
Comparative analysis of different speech recognition system								
Sweeta Bansal , Amita Dev [1]	2015	IEC gzb. Dataset is created by 4 students using their phone. Emotions are anger, fear, joy,sadness.	Hindi	K-mean cluster	LPCC, MFCC	84	global features are efficient only in distinguishing between high-arousal emotions,e.g. anger, fear, and joy	The speech data is contributed by the four students of engineering college, not accurate. temporal information from speech is lost due to global features.

Author	Year of Publication	Dataset src. & emotions considered	Language of speech	ML technique	Features Extracted	Accuracy (%)	Best Advantage	Worst limitation
K Sreenivasa Rao, Shashidhar G Koolagudi [2]	2011	Dataset (IITKGP-SEHSC) is used. Emotions are anger, disgust, fear, happy, neutral and sad.	Chhattisgarhi,Bengali, Marathi, General Hindi, Telugu	AANN, SVM	prosodic features, MFCC	AANN-78, SVM-81	prosodic features contains more dialect specific informa-tion compared to spectral features.	component of speech (i.e., excitation source) not been considered.
Sujata B. Wankhede, Pritish Tijare, Yashpalsing Chavhan [3]	2011	Dataset created from movies audio clips, Emotions are anger, happiness, sadness, neutral, fear	Hindi, Berlin	SVM, LIBSVM(RBF kernel)	MFCC, MEDC	1.SVM Hindi-71.6, Berlin-62 2.LIBSVM Hindi-78.33, Berlin-99.39	Majority dataset contains emotion for anger,hence they got max accuracy for anger.	Hindi dataset is created from bollywood movies dialogues,contains noise.
Prakhar Kulshreshtha, Soumya Gayen [7]	2016	Dataset (IITKGP-SEHSC) is used. Emotions are anger, fear, disgust, happiness, surprise, neutral, sadness and sarcastic.	Hindi	SVM, ANN, MLP, KNN, Adaboost, Random forests	MFCC, SSC	SVM-80.46, ANN-77.83, MLP-80.31, KNN-74.45, Adaboost-59.18, Random forests-74.14	The best performance was given by SVM with RBF Kernel.	The ER system is speaker dependent on which dataset it was trained.
Girija Deshmukh, Apurva Gaonkar, Gauri Golwalkar, Sukanya Kulkarni [6]	2019	The North American English language open source dataset.	North American English, Indian English, marathi, hindi	SVM	MFCC, pitch,STE	79.48	Mean data values provided better results than mode values. Highest accuracy on 3 features set.	The paper basically classifies only 3 emotions, i.e. happiness, anger and sadness.
Meenakshi Singh, Parul Khullar [8]	2018	Dataset (IITKGP-SEHSC) is used. Emotions are happy, angry, sad, depress, bored, anxious, fear and nervous.	Hindi	ANN	standard deviation of fundamental frequency, energy and pitch contour, formant frequencies, duration, zero crossing	50	Previous ER in hindi speech didn't include all acoustical features such as formants , zcr etc. which are also involved here.	Majority of local features are combined together to get results but getting less accuracy. Global features could have been used.

Author	Year of Publication	Dataset src. & emotions considered	Language of speech	ML technique	Features Extracted	Accuracy (%)	Best Advantage	Worst limitation
Akshat Agrawal & Anurag Jain [5]	2019	Dataset was made with 4 speakers, emotions(Anger, Happiness, Neutral, Sad, Surprise).	Hindi	KNN,NBC	Prosodic Features (PF) and Acoustic Features (AF)	62.1	Statistical techniques like PCA, LDA are used for reducing dimensionality of data. making the model evaluate faster.	Combinations of local features and global features not considered. Only local features considered.
Usha Sharma, Sushila Maheshkar, A. N. Mishra , Rahul Kaushik [9]	2019	Dataset was made with 24 speakers with total of 240 samples(each speaker being recorded with utterances of hindi digits from 1-10)	Hindi	HMM (Hidden Markov Model)	GFCC , OF	For Clean GFCC- 92.775 GFCC+OF - 93.12 For Noisy data GFCC- 82.91 GFCC + OF - 88.78	Give better results for recognition of hindi digits(1-10) only.	Limited size dataset was used(of only 240 samples). Only trained for recognition of hindi digits from 1-10 (not even any word in hindi). Limited features are being considered.
Anjali Bhavan, Pankaj Chauhan, Hitkul, Rajiv Ratn Shah [10]	2019	EmoDB, RAVDESS, IITKGP - SEHSC	Berlin, English, Berlin	Bagged ensemble of SVM, AdaBoost ensemble of SVM	MFCC	EmoDB - 92.45%, 87.32%, RAV DESS - 75.69%, 72.10%, IITKGP - SEHSC - 84.11%, 77.19%	The results for the Berlin Database are very good.	Worked on only MFCC features, did not use a combination of features.
Abdul Malik Badshah, Nasir Rahim, Noor Ullah, Jamil Ahmad, Khan Muhammad, Mi Young Lee, Soonil Kwon, Sung Wook Baik[11]	2017	Emo-DB and Korean speech dataset.	Korean, Berlin	SVM, Decision Tree, Random Forest, AlexNet, CNN with rectangular kernels.	MFCC	SVM- 61.48, Decision Tree- 63.24, Random Forest- 71.15, AlexNet - 79.14 CNN model- 86.54	The proposed model seems to predict results quite well. Rectangular kernels proved to be better than square kernels on spectrograms.	Worked on only MFCC features, did not use a combination of features.

Research having dataset column colored with ◯ have same dataset as we have considered i.e, (IITKGP-SEHSC)

Comparative analysis of our speech recognition system							
Author	**Dataset src. & emotions considered**	**Language of speech**	**ML techniq ue**	**Feature s extracte d**	**Accuracy (%)**	**Best advantage**	**Worst limitation**
Kunal Sharma, Mayur Bansal, Aditya Maheshwari	Dataset (IITKGP-SEH SC) is used. Emotions are anger, fear, disgust, happiness, surprise, neutral, sadness and sarcastic.	Hindi	KNN, CNN, MLP, SVM(Linear kernel, Rbf kernel, Poly kernel)	LPC, MFCC, SSC	KNN-79.72 , CNN-71.16, MLP-84.22, SVM(Linear kernel-71.6 8, Rbf kernel-85.4 7, Poly kernel-82.2 3)	Model is not trained gender specifically still giving better results than research under consideration with accuracy of 85.47% .	Local features such as pitch, duration,energy etc not considered.

9. TESTING WITH CUSTOM INPUT

8.1 Input

Audio is recorded and tested on the best performing model, SVM (RBF Kernel). Fig 32 describes the input waveform of input audio signal considered for testing.

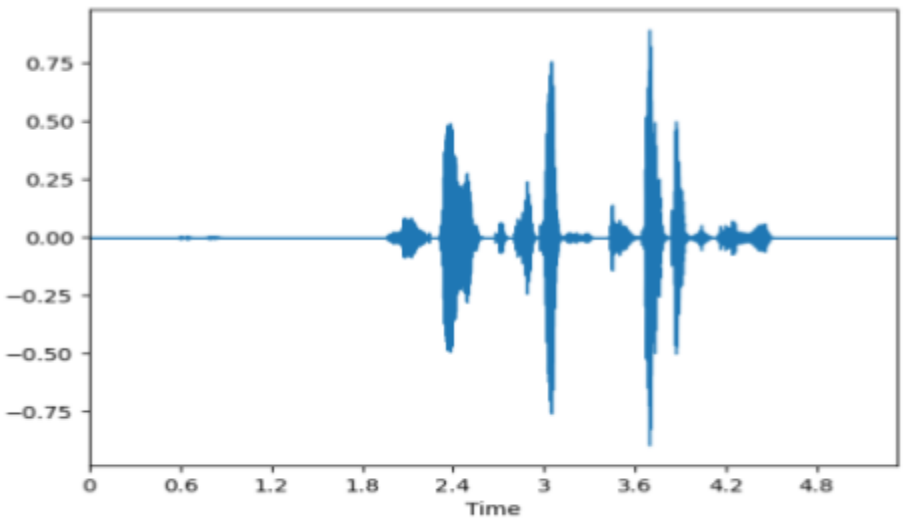


Fig. 32, Input audio signal waveform

8.2 Feature Extraction

8.2.1 Process of Extraction Of MFCC Features

Step 1: The audio file is loaded and the mfcc features are extracted from the audio using the library python_speech_features.

```
def extract_mfcc(audio_file):
  data, sr = wav.read(audio_file)
  mfcc_ = mfcc(sr, numcep=13)
```

Fig. 33, Reading the Audio File

The top 13 cepstral features (standard) are extracted from the audio and the output looks like this,

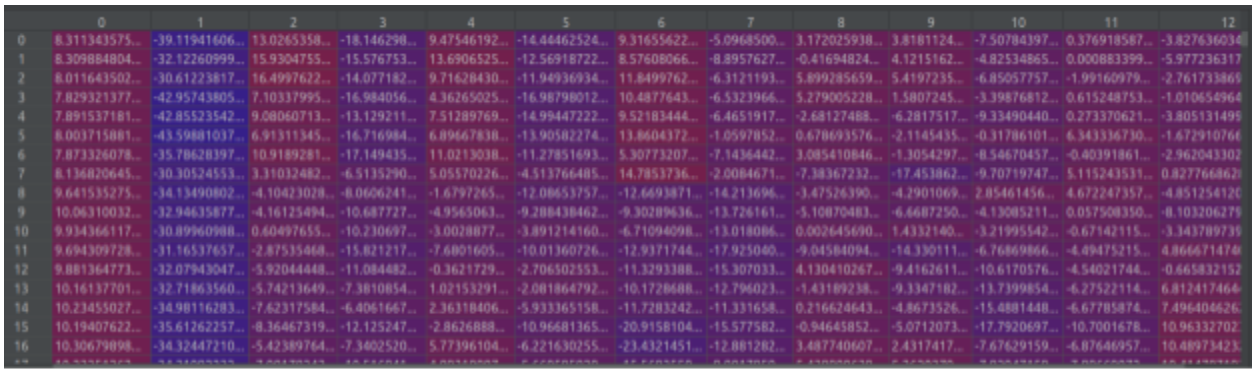


Fig. 34, Extracted MFCC Features Table

There are a total of 13 columns, each column representing each cepstral feature. The number of rows depends upon the length of the audio. For e.g. An audio of length 4 sec would have around 362 columns.

Step 2: Now we convert the 2D-matrix for a single audio as seen in fig. 34, into a single matrix for further classification and to do this, we took 7 different kinds of values from each column for all the 13 columns and those are:

- Mean of the column
- Variance of the column
- Maximum value of the column
- Minimum value of the column
- Variance of Gradient of column
- Mean of upper half of the column
- Mean of lower half of the column

```
mfcc_features = []
for i in range(mfcc_.shape[1]):
  column = mfcc_[:,i]
  features  = [np.mean(column), np.var(column), np.amax(column), np.amin(column),
  np.var(np.gradient(column)), np.mean(column[0:column.shape[0]//2]), np.mean(column[column.shape[0]//2:column.shape[0]])]
  mfcc_features.append(np.array(features))
```

Fig. 35, Extracting the Significant Values from the MFCC Features table

The 7 features extracted for the first column of the table shown in fig 34 are shown in fig 36

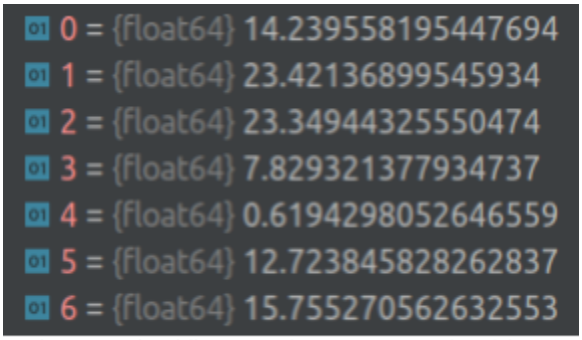


Fig. 36, Significant Values Extracted Table

The data after extracting all the above features for all the 13 columns are as shown in fig 37,

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	14.239558...	23.421368...	23.3494432...	7.8293213...	0.619429...	12.72384...	15.755270...	-19.162014...	198.32815...	7.465828...	-43.598810...	10.673585...	-20.756327...	-17.56770...	-16.212847...

Fig. 37, Final Features for an Audio File

The data would have 13x7 = 91 values. So for each audio length of extracted MFCC features is 91.
We do the Step1 and Step2 for all the audio files, to create our database for further classification.
Similarly SSC and LPC features are extracted.

8.3 Classification

The emotion has been classified using the SVM (RBF Kernel) in the website with the waveform plotted, Fig. 38, describes the emotion obtained on the tested audio sample.

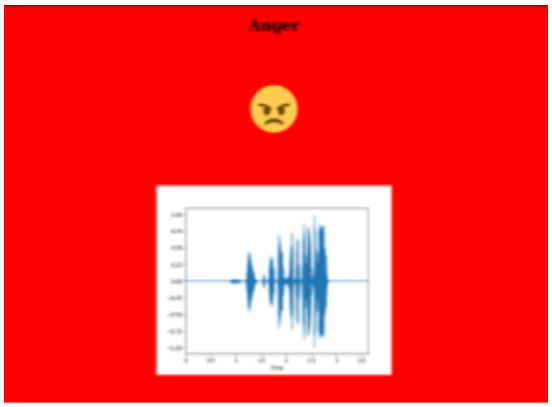


Fig. 38, Output Emotion for Test Audio File

9. FUTURE WORK:

There is a lot of work we wish to add in this research in future,

- The complete IITKGP-SEHSC dataset consists of 10 artists (5 male, 5 female) but we have got only 4 artists (2 male and 2 female) for this research. We wish to get the full dataset and get the best possible results.
- We wish to use different audio datasets like English, Berlin and do a comparative study.
- Tuning the parameters of the models further can also lead to better accuracy that can be taken into future work.
- Supplementing the audio dataset with visual cues like facial features may help in predicting the broader spectrum of emotion. This can be taken further as a future work in the research to improve the accuracy of the predictions

10. CONCLUSION:

Speech recognition technology which is an increasingly popular concept in recent years that attracts attention from organizations to individuals; the technology being widely used for the various advantages it provides. It brings the ability for a machine to listen and understand what people are talking or what users are commanding. The research and implementation of an entire speech recognition model for the Hindi language provided great insight into the technology and future scope for research.

11. REFERENCES:

[1] S. Bansal and A. Dev, "Emotional Hindi speech: Feature extraction and classification," 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2015, pp. 1865-1868.

[2] K Sreenivasa Rao and Shashidhar G Koolagudi, “Identification of Hindi Dialects and Emotions using Spectral and Prosodic features of Speech”, SYSTEMICS, CYBERNETICS AND INFORMATICS, Vol. 9 No. 4, 2011.

[3] Sujata B. Wankhade, Pritish Tijare, Yashpalsing Chavhan, “Speech Emotion Recognition System Using SVM AND LIBSVM”, International Journal Of Computer Science And Applications, Vol. 4, No. 2, June July 2011.

[4] S. G. Koolagudi, R. Reddy, J. Yadav and K. S. Rao, "IITKGP-SEHSC : Hindi Speech Corpus for Emotion Analysis," 2011 International Conference on Devices and Communications (ICDeCom), Mesra, 2011, pp. 1-5.

[5] Akshat Agrawal & Anurag Jain (2020) Speech emotion recognition of Hindi speech using statistical and machine learning techniques, Journal of Interdisciplinary Mathematics, 23:1, 311-319, DOI: 10.1080/09720502.2020.1721926
[6] G. Deshmukh, A. Gaonkar, G. Golwalkar and S. Kulkarni, "Speech based Emotion Recognition using Machine Learning," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 812-817, doi: 10.1109/ICCMC.2019.8819858.
[7] Prakhar Kulshreshtha, Soumya Gayen, “Emotion state detection via speech in spoken Hindi”, IIT Kanpur, April 20, 2016.
[8] Meenakshi Singh, Parul Khullar, “Emotion Recognition fromHindi Speech Signal”, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 6, Issue 5, May 2018
[9] Sharma, U., Maheshkar, S., Mishra, A.N. *et al.* Visual Speech Recognition Using Optical Flow and Hidden Markov Model. *Wireless Pers Commun* 106, 2129–2147 (2019).
[10] Anjali Bhavan, Pankaj Chauhan, Hitkul, Rajiv Ratn ShahBagged support vector machines for emotion recognition from speech (2019)
[11] Badshah, A. M., Rahim, N., Ullah, N., Ahmad, J., Muhammad, K., Lee, M. Y., … Baik, S. W. (2017). *Deep features-based speech emotion recognition for smart affective services. Multimedia Tools and Applications.* doi:10.1007/s11042-017-5292-7