International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 04 | Apr 2022 www.irjet.net p-ISSN: 2395-0072

SPEECH EMOTION RECOGNITION

[1]Vidya Pujari, [2]Nidhi Lulla, [3]Gaurav Sitlani, [4]Girish Chawla

[1] Professor, Dept. Of Information Technology, VESIT, MUMBAI
[234] Student, Dept. Of Information Technology, VESIT, MUMBAI

Abstract*—Emotion recognition of speech has consistently been a perplexing space of study in a system involving human-machine connection, since machines can never assess a speaker's feeling with their own a couple of frameworks were created which tried to recognize the speaker's feelings. SER's primary aim is to reinforce the human-machine interface. It might likewise be used in lie detectors to follow an individual's psycho physiological condition. Speech emotion recognition has as of late discovered applications in medication and criminology. Happy, Calm, Disgust, and Fearful are the four feelings recognized in this paper utilizing pitch and prosody highlights. The time space contains the heft of the speech highlights used in this study. The emotions were classified utilizing HistGradient Boosting Classifier, Multilayer Perceptrons Classifier, Extra Trees Classifier, LGBM Classifier, XGB Classifier, and CatBoost Classifier. The task is allocated to the Ravdess data set. The recognition rate was 83.33 percent, which is high. The paper that was used as a reference received an 81 percent recognition rate. When compared with the recently settled recognition system, the outcomes show that they are more compelling.*

***Keywords*—Emotion recognition, pitch, prosody, RAVDESS dataset, HistGradient Boosting Classifier, Multilayer Perceptrons Classifier, Extra Trees Classifier, XGB Classifier, and CatBoost Classifier.**

I. INTRODUCTION

Human Emotion in today's generation plays an important role in our life. Emotion can determine what someone has been feeling and if someone has felt something. In today's generation humans are working on recognition of emotions actively and are trying to find new solutions. Emotions can be displayed via facial expressions, hand gestures and can also reflected via speech. Determining an emotion with help of speech can be a tedious job because of different variables and can be tough to determine someone's emotion. The main purpose of this project was to accumulate all the researches we found and to create a better model which would help user easily understand or to get in touch with their emotion.

Speech emotion is an industry-based project which can be used in various places like in hospital by doctors or also therapist can use speech recognition model to determine a patient's emotion, in various call centers SER models are used to determine a customer's emotion while asking for their feedback or to understand them. Aside from them these models can be updated and also can be used along with artificial intelligence in creating virtual assistants that can recognize our emotions. Speech emotion recognition has many uses and in near future more advanced features would be created regarding this topic. But along with major advantages these topics can be pretty difficult to understand too and some of the concepts are challenging to understand like a speech recognition software to properly understand a human voice and accent also plays an important role as machine sometimes could not understand speech from every accent and in near future with more advances this problem could be solved or we can say that we could increase accuracy of these models.

So, our model consists of different classifiers which we have accommodated to increase accuracy and efficiency of our model. Accuracy and Efficiency are the two main factors that will decide the success of our Speech Emotion Recognition model. The suggested model would help the user to determine the speakers emotion easily and also it would also provide user with some recommendations which the user could choose from. Our model is divided into various stages. Initially we have collected a dataset and we have recognized our dataset. In our project we have used RAVDESS dataset. Then we have converted our dataset which consist of speech into text. After our speech is converted into text, we have cleaned our data. After the cleaning process we have plotted our data for in depth visualization. After cleaning and plotting is done, we have use extracting features to extract features from our clean data and it helps us to determine emotion. Once we have found our emotion it would be ran through different classifiers which would help us in determining the accuracy and efficiency of our model and also to help us know whether our model is working perfectly as expected.

II. LITERATURE SURVEY

[1] Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech Michael Neumann, Ngoc Thang Vu 2019 In this paper we got to know that if we represent our dataset with help of proper functions it can be used to increase the accuracy and also the efficiency of the model also because of graphs we could easily understand about our sound file. The paper represented SNE visualizations that revealed the discriminative strength of representation with highs and low arousals. How amplitude and frequency could be used to find pitch and with some functions even high-toned speech could be recognized. This paper leads to consistent improvements in accuracy of SER model.

*[2] Speech Emotion Recognition S. Lalitha, Abhishek Madhavan, Bharath Bhushan, Srinivas Saketh 2019.*In this paper we learned about a particular classifier SVM and how it can be used to recognize speech and determining emotion. As the proposed classifier uses compact feature set with a good recognition accuracy. This paper had a future scope of using hybrid classifier which would help us in increasing accuracy pretty much. The model was good enough to determine emotion but it wasn't easily able to incorporate more emotions and also it didn't provide a fairly high rate of recognition for few emotions. the main advantage of the given model was it was less complex compared to other different models but as we had accuracy and efficiency the two most important factors we couldn't use this model just to reduce complexity but this model provided us with some creative features which would help in cleaning of our data and also masking of our data.

[3] *Enhanced Human Speech Emotion Recognition Using Hybrid of PRNN and KNN J. Umamaheswari, A. Akila 2020.*This system was tested with different emotions and accuracy, precision rates were found using two different classifiers in a combined way so as to increase efficiency. The models used were Gausian mixture model and Hidden Markov Model. In this paper along with combined classifier's various function were also combined together to increase precision rates and it was carried out with different algorithms. PRNN and KNN algorithms were used feature extraction was made using cascaded structure comprising of MFCC and GLCM.the results of accuracy, precious rates and f-Measure were compared and we got to know that combining two different models help us in increasing the results for our model. With the help of this paper, we used 6 models in our research and combined our model and we increased our accuracy which was 61% to 83.3% by combining 4 models we get to know that accuracy and efficiency increases with more than one model as compared to standard algorithms

[4] *Analyzing Speech for Emotion Identification using Catboost Algorithm in Machine Learning Saranya CP Amal Mohan N , Lijo Tony A R , Naveen Chanth R and Vishnu K 2020* In this paper emotions are recognized by speech using the Catboost Algorithm The implemented algorithm took MFCC as the base feature and trained using the RAVDESS dataset. The algorithm's accuracy was tested for the emotions like neutral, happy, sad, angry, fear and produced an accuracy of 74.07%. The accuracy offered by this classifier on a reputed dataset shows that catboost is a great algorithm for emotion recognition. In future, this algorithm can be tested against the various datasets for testing its efficiency in the emotion recognition

[5] *Speech Emotion Recognition using MLP Classifier Sanjita. B. R , Nipunika. A , Rohita Desai 2019* in this paper we use machine learning techniques like Multilayer perceptron Classifier (MLP Classifier) which is used to categorize the given data into respective groups which are non-linearly separated. Mel-frequency cepstrum coefficients (MFCC), chroma and mel features are extracted from the speech signals and used for training MLP classifier. For achieving this objective, we use python libraries like Librosa, sklearn, pyaudio, numpy and sound file to analyze the speech modulations and recognize the emotion.

III. PROCESSED METHODOLOGY

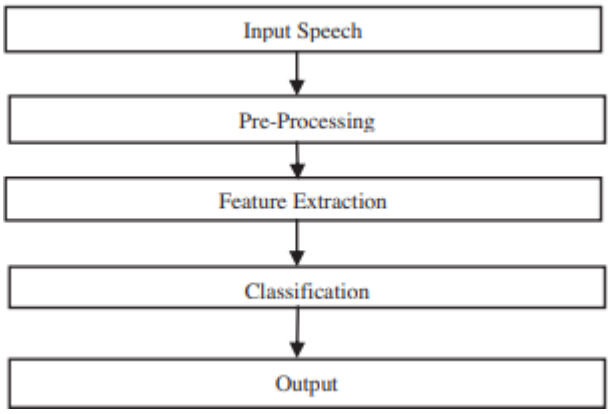
A. Input Speech

The input or source to this algorithm is a speech which is to be recognized. Our database is loaded which is the RAVDESS data set containing 1440 files: 60 trials per actor x 24 actors = 1440. The RAVDESS contains 24 actors 12 male and female respectively recording two literally matched statements in a neutral American accent.

The emotional database contains six basic emotional classes such as

- Angry
- Happy
- Sad
- Neutral
- Surprise
- Fear

. Each expression is produced at emotional intensity , with an additional neutral expression.



FIGIII.1. METHODOLOGY

B. Pre-processing

Firstly, we have used many python libraries for creating our model we have called all the libraries that need to be used. After that we have uploaded our dataset using google

drive. Once database is loaded, we have used a Speech Emotion Recognition API which we use for finding any errors in our audio file. Speech Emotion Recognition API is not a powerful Api and it cannot recognize all audio file. So, with some function we have stated if the API cannot recognize a particular sound file it would display ERROR and if it would recognize audio file the audio file would be converted into text. After using API and converting speech to text we have plotted our graphs using basic time series data graphs and specto graphs just for understanding of our audio. In time series data graph on y axis, we had plotted sound amplitude and on x axis we had plotted time. Spectrogram is a visual way of representing strength of loudness of audio over frequency over period of time. After that we need to have in depth visualization of our audio file so that we could extract certain features for plotting. Then we have defined all our functions for the plotting graph then we have loaded our audio dataset to our plotting functions. On calling those functions graphs are plotted with the features we had defined for our audio file. After functions are applied, we had our dataset cleaned by using functions that could mask and it could downsample it. Once cleaning is done, we have called our dataset and a new cleaned dataset is created and then we decided to separate our new cleaned dataset into another folder which could be used to call later for feature extraction.

C. Feature extraction

The speech signal comprises a huge number of parameters that reflect the emotional characteristics and these parameters results in variations in emotion. Hence, feature extraction is one of the most vital aspects of every algorithm because choosing the relevant features is a crucial step in achieving high recognition performance. In this system we have used the librosa library and basic audio functions of the librosa library.

1. Mel Frequency Cepstral Coefficients (MFCCs):

MFCCs are the most commonly used spectral illustration of speech. Mel is a unit to measure the pitch or frequency of a tone. The Mel-scale is a mapping of the real frequency scale (Hz) to the perceived frequency scale (mels).

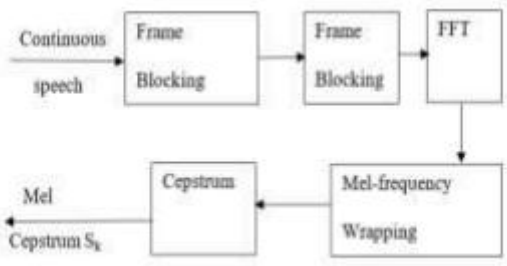


FIG III.2. MFCC

2.**Chroma**

Chroma-based features, which are a powerful tool for analyzing pitches The underlying observation is that humans perceive musical pitches as they are similar in color if they differ by octave. Tone height and Chroma are the two components in which pitch could be separated. Assuming we receive equal-tempered scale, one considers ten chroma values that are represented by the set.

{C, C♯, D, D♯, E , F, F♯, G, G♯, B}

After the feature extraction we have then split the dataset into training and testing dataset 75% for data for the training dataset and 25% for testing. Once using functions audio is predicted it is stored in csv files and then we use labels to define which emotion was predicted for which audio set. the process of mapping occurred which combined audiofiles were classified into labels predefined. After that we had used various functions for mapping our dataset with labels that need to be called later after prediction and we have also use functions to find shape of our dataset and using extract feature we also got to number of features extracted which were 180 in our case. After Feature Extraction we have used various classifiers for our model.

D. Classification

1.Catboost Algorithm-

Catboost stands for Categorical Boosting. This algorithm can handle the various categories of data in the form as they appear using the process of encoding. Since our dataset is not in text format and falls under another category as an audio type/wav file, Catboost can handle such category without any hassle. Catboost algorithm also depends on the ordered boosting technique. Ordered boosting technique is a form of gradient boosting algorithm. Gradient boosting algorithms have the problem of prediction shifting. Prediction shifting happens due to a special kind of target leakage. Ordered boosting was implemented to diminish the problem of shifting. Catboost was implemented on the basis of – processing categorical features and the ordered boosting technique. Catboost makes use of the gradient boosted decision trees. Numerous decisions are built during training. Decision trees formed in the upcoming iteration will be having low loss compared to the trees in previous iterations. Iteration will be continued till there is no significant reduction in the loss function.

2.XGB

eXtreme Gradient Boosting (XGBoost) is an improved version of the gradient boosting algorithm which was created for increasing both efficiency, computational speed and also model performance. XGB is an open-source library

and also one of the import parts of Machine Learning Community. The software and hardware capabilities were designed to enhance the existing boosting techniques along with accuracy in the shortest duration.

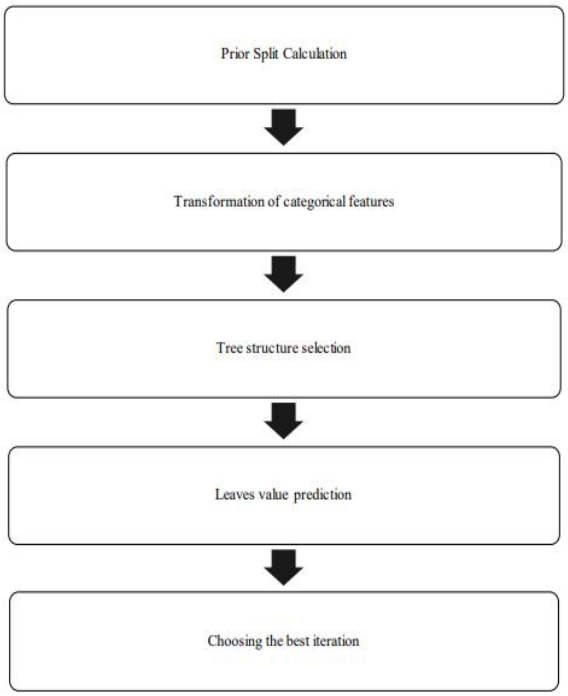


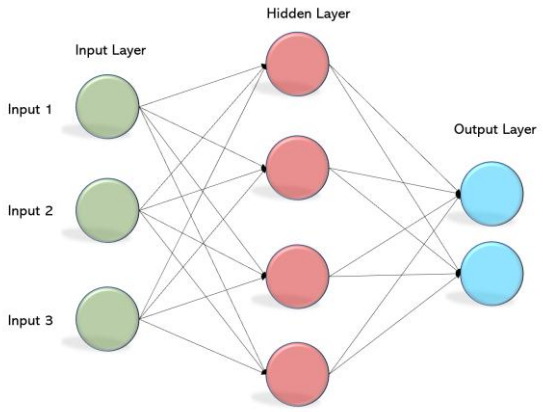
Fig III.3: Catboost Algorithm Workflow

3.Hist Gradient

Gradient boosting produces a prediction model which is in the form of an ensemble that consist of weak prediction models. It involves three elements like loss function, weak learner and additive model. Gradient boosting framework is a general framework in which any differentiable loss function can be utilized. In gradient boosting the weak learner are the decision trees. To minimize the cost, trees are assembled in a greedy manner by splitting in best purity scores points. It is an over fitting algorithm. Weak learners can be constrained in some specific ways like nodes, splits, leaf nodes or a maximum number of layers. Additive model means trees are added for once at a time. The aim of the Gradient boosting procedure is to minimize the loss while adding trees. In this procedure, loss is calculated in each step and after calculating the loss, a tree which reduces the loss must be added and weights are updated. We also use scikit learn ensemble which is applied in this emotion recognition system used for gradient boosting.

4.MLP

We have used the RAVDESS dataset of audio files which has speeches of 24 people with variations in parameters. For the training, we store the numerical values of emotions and their respective features in different arrays. MLP Classifier had been initialized and input was given in array format. The Classifier identifies different categories in the datasets and classifies them into different emotions. The model will now be able to understand the ranges of values of the speech parameters that fall into specific emotions. For testing of model, we enter the unknown test dataset as an input, and then it will retrieve the parameters which can help us to predict the emotion as per training dataset values. The accuracy of the system is displayed in the form of percentage which is the final result of our project.



FigIII.4-Neural Network - MLP

Once we had used our classifiers our prediction of emotion using trained data, we tuned the model and predicted emotions for our test data and model is loaded again to run test data and its result was stored in .csv file with labels for mapping individual predicted result with their file name. After that we tried a live demo prediction for which we could record voice of anyone via microphone and we can store it in .wav file format. So, that later we can load our model to predict result of a particular .wav file.

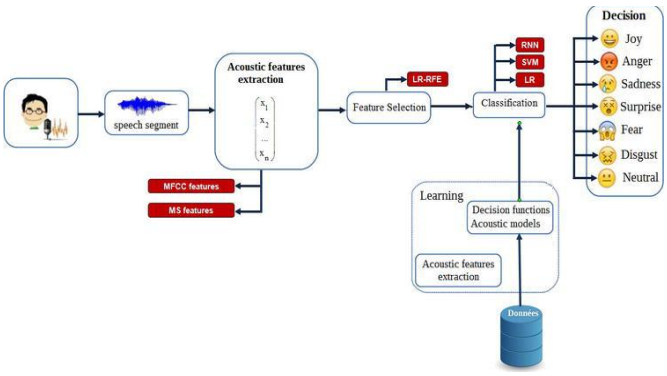
E. Output:

Once we have trained and tested our models our model could be ran for the recorded model in a series where graph would be plotted for custom audio and using extracting features of librosa and mapping those features with labels and passing through our classifier an emotion would be predicted and then we could print those emotion that is found and can be displayed it to user. After displaying emotion user has multiple option to do more on our interface.

Block Diagram

Firstly when the user wants to use our model they need to record their voice, the voice recorded would be passed through our trained model which would first recognize the voice and using matplotlib functions a graph would be plotted and displayed to us for reference and the voice

would be down sampled and cleaned using functions cleaned audio sample would be created and using function like mel,chroma,mfccs emotion would be recognized and it would passed through our model which would provide us with an emotion depicted from the given test model.



FigIII.5 -block Diagram

Modular Diagram

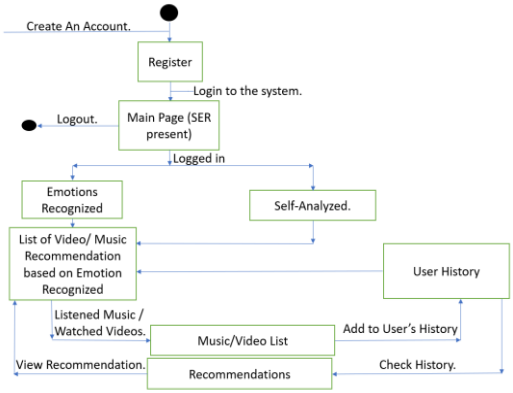


Fig- modular Diagram

Our Modular Diagram represents the interfaces firstly that the use would visit would be our homepage where the user can record their voice and their emotion would be predicted there. Along with Prediction User could also use recommender once their Emotion is Predicted. Along with SER model and Recommender if user wants to read about the project and if user wanted to contact developers with any issues and any suggestion could contact us.

IV. Research Analysis

The Following Photos tell us about various outputs that we have captured in our research. The photos represent both our model on front end and back end. We have obtained an accuracy of 84%..

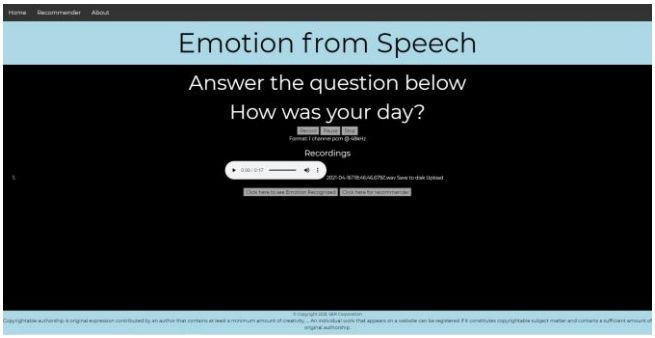


Fig 2-Front end Recording

Fig 2 talks about how recording takes place once we click on the recording button and the user has to provide us with access to the mic through which we can record the user's voice that would be used for prediction.

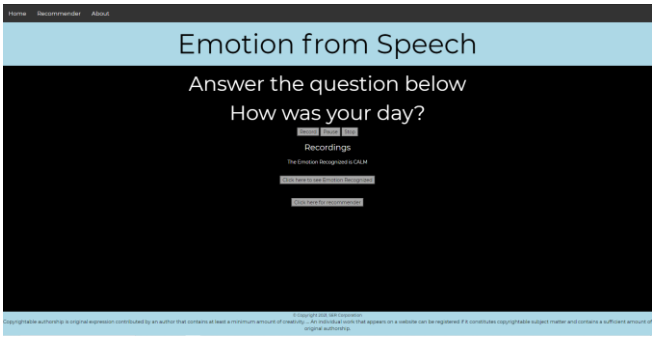


Fig 3-UI interface

Figure 3 provides us with a display of UI interface where users can go to the About section to know more about the project and also can use our recommender to navigate through.

Fig 4: BackEnd Output

Fig 4 represents how after testing the recorded voice is passed through the model and the emotion is printed once the model has predicted it.

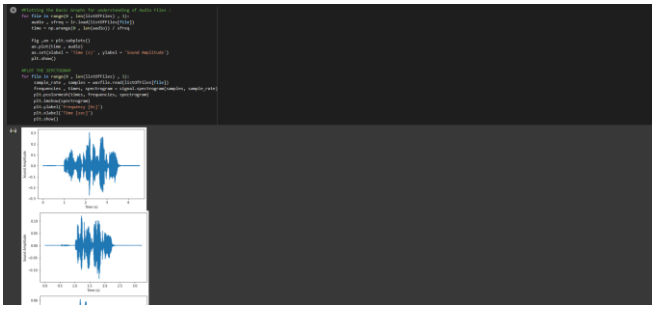


Fig 5: Plotting Output

Fig 5 displays plotting outputs which can be seen for in depth visualization to understand tone, pitch, loudness and other important factors of a sound.

```
[ ]  accuracy_score(y_test_le , yhat)

     0.84375
```

Fig 6: Accuracy Score

Fig 6 represents the accuracy score of our model after combining different models to increase efficiency of our model.

Fig 7: Using Mat Plot for Plotting Test Case

Additional front end :

We have also created a recommender for the user. Recommender consist of recommendations regarding every emotion. Calm, Happy, Sad, Fear, Neutral, surprise every emotion consist of various movies, music and article. User can select any option he likes and can enjoy.

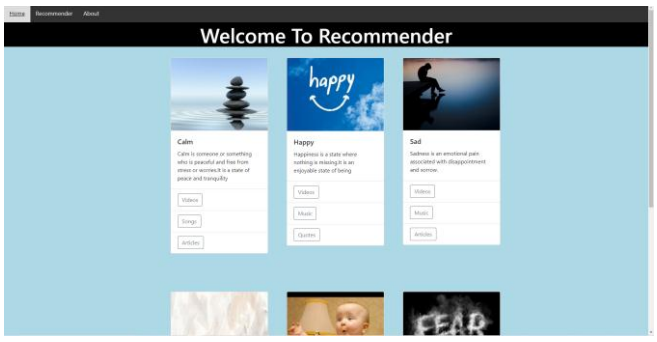


Fig 8: Recommender

User could browse through the recommender and can find a playlist of music movies and articles for himself baked on their emotion predicted.

V. RESULT

After Using Different Classifiers, we got to know that Catboost Classifier obtained us with accuracy of 73.4% because catboost did not support some of the plotting features.

Using Lgbm Model we acquired an efficiency of 76.5% it sensitive to overfitting and easily overfit data. There is no threshold on number of rows

Using MLP classifiers we occupied a frequency of 77% because of too many parameters connection is a bit dense.

Using Extra Trees Classifier, we acquired an accuracy of 73.4% it creates many extra trees and samples it without replacement and it splits on random subsets.

Using the XGB classifier we acquired an accuracy of 78.1%. It is scalable and provides us with accurate performance and it has fast computational speed.

Using Hist Gradient Boosting Classifier we acquired an accuracy of 82.29%. Using Histogram boosting as it the reduced feature substantially helps us to train faster. Also it is more efficient in memory consumption and training speed is very fast compared to other classifiers. Using Classifiers can cause few problems and some function also needs to be called during using those classifiers but jump in accuracy and less time frame of model would be the main reason to use multiple classifiers. We used 5 Classifiers and combined them in our model each with their own specific functions and we achieved an accuracy of 84%.

Classifier	Accuracy
Catboost Classifier	73.4%
Lgbm Model	76.5%
Using MLP classifier	77%
Extra Trees Classifier	73.4%
XGB classifier	78.1%
Hist Gradient Boosting Classifier	82.29%

VI. ACKNOWLEDGMENT

Our work has been supported by Vivekanand Education Society's Institute of Technology. WE would like to thank Professor Vidya Pujari, Information Technology Engineering, Vivekanand Education Society's Institute of

Technology for her support and guidance. We would also like to thank our teachers for their constant guidance.

VII. Conclusion and future scope

From the given research we can conclude that speech representation is a tedious job but with the help of various plotting functions and various python libraries we could map our speech into graphical representation and with the help of extracting features we can differentiate between different voice models. Efficiency of the model could be increased if we accommodate various models. This gave us a better result which was 83.3% accurate. MLP and Gradient Boost were two classifiers which gave us the most accuracy.

Although the reference paper provides a better accuracy, the work in this paper could be improved if we get more databases. Since emotions are also dependent on the accent of the language, accent-based classifiers might also be developed for the proficient analysis of emotions. To be more precise and accurate, voice assistants can go through a particular person's speeches and create a machine learning model for that individual which may result in an optimized assistance.

VIII. References

[1]Speech Emotion Recognition S.Lalitha, Abhishek Madhavan, Bharath Bhushan, Srinivas Saketh

[2]IMPROVING SPEECH EMOTION RECOGNITION WITH UNSUPERVISED REPRESENTATION LEARNING ON UNLABELED SPEECH Michael Neumann, Ngoc Thang Vu

[3]An Enhanced Human Speech Emotion Recognition Using Hybrid of PRNN and KNN J.Umamaheswari, A.Akila

[4]Analysing Speech for Emotion Identification using Catboost Algorithm in Machine Learning Saranya CP Amal Mohan N , Lijo Tony A R , Naveen Chanth R and Vishnu K

[5]Speech Emotion Recognition using MLP Classifier Sanjita. B. R , Nipunika. A , Rohita Desai

[6]A Comparative Analysis of XGBoost Candice Bent´ejac, Anna Cs¨org˝o, Gonzalo Mart´ınez-Mu˜noz.

[7]A Real-time Emotion Recognition from Speech using Gradient Boosting Aseef Iqbal, Kakon Barua.