A PROJECT REPORT

on

Speech Emotion Recognition Using Librosa

Submitted in partial fulfillment of the

requirement for the award of the degree of Bachelor of Technology

Program Name: Computer Science and Engineering

Under The Supervision

Dr. Ashok Kumar
Associate Professor

Submitted By:
Utkarsh Saxena (19SCSE1010430)
Hargun Singh Dhall(19SCSE1010910)

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

I/We hereby certify that the work which is being presented in the thesis/project/dissertation, entitled **"Speech Emotion Recognition Using Librosa"** in partial fulfillment of the requirements for the award of the Bachelor of Technology submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out under the supervision of Dr. Ashok Kumar (Associate Professor), Department of Computer Science and Engineering/Computer Application and Information and Science, of School of Computing Science and Engineering , Galgotias University, Greater Noida

The matter presented in the thesis/project/dissertation has not been submitted by me/us for the award of any other degree of this or any other places.

Utkarsh Saxena (19SCSE1010430),

Hargun Singh Dhall (19SCSE1010910)

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Supervisor: Dr. Ashok Kumar

Designation: Associate Professor

CERTIFICATE

The Final Thesis/Project/ Dissertation Viva-Voce examination of Utkarsh Saxena(19SCSE1010430), Hargun Singh(19SCSE1010910) has been held on _____ and his/her work is recommended for the award of Bachelor of Technology in Computer Science and Engineering .

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Program Chair **Signature of Dean**

Date: May 2023 Place:

Greater Noida

Abstract

Emotion detection has become one of biggest marketing strategies in which mood of consumer plays an important role. So to detect current emotion of person and suggest the appropriate product or help him accordingly, the demand of the product will be increased or the company. The Emotion detection is natural for humans but it is very difficult task for machines. In today's world detecting emotions is one of the most important marketing strategy. For this purpose we decided to do a project in which we could detect a person's emotion just by their voice Using Deep and Convolutional Neural Networks for Accurate Emotion Classification on DEAP Dataset.

Speech emotion recognition, the best ever python mini project. The best example of it can be seen at call centers. If you ever noticed, call centers employees never talk in the same manner, their way of pitching/talking to the customers changes with customers. Now, this does happen with common people too, but how is this relevant to call centers? Here is your answer, the employees recognize customers' emotions from speech, so they can improve their service and convert more people. In this way, they are using speech emotion recognition.

Speech Emotion Recognition, abbreviated as SER, is the act of attempting to recognize human emotion and affective states from speech. This is capitalizing on the fact that voice often reflects underlying emotion through tone and pitch. This is also the phenomenon that animals like dogs and horses employ to be able to understand human emotion.
SER is tough because emotions are subjective and annotating audio is challenging**.**

In this paper, we propose a system that will analyze the speech signals and gather the emotion from the same efficient solution based on combinations. This system solely served to identify emotions present in the signal or speech using concepts of deep learning and algorithms of machine learning (ML). Using the above mentioned, the system will determine the eight emotions present in the speech signal; anger, sad, happy, neutral, calm, fearful, disgust and surprised. The system is built with the language python and librosa, sound file libraries, which are part of the more extensive scikit library used for specific applications of audio analysis. The system will receive the sound files from the dataset present on the internet called RAVDESS. It will then analyze the audio files' spectrograms in WAV format and return us the efficiency of the system, which is the intended Outcome. We have achieved an efficiency rate of 77%.

List of Tables

Table No.	Table Name	Page Number
1.	Table for Student Data	4
2.	Table for Faculty Data	4

S.No.	Name of the Student	Admission No.	Contact No.
1	Utkarsh Saxena	19SCSE1010430	8840013104
2	Hargun Singh Dhall	19SCSE1010910	9675979551

S.No.	Name of Faculty	Designation	Contact No.
1	Dr. Ashok Kumar	Associate Professor	9354866919
2			

Acronyms

B.Tech.	Bachelor of Technology
M.Tech.	Master of Technology
BCA	Bachelor of Computer Applications
MCA	Master of Computer Applications
B.Sc. (CS)	Bachelor of Science in Computer Science
M.Sc. (CS)	Master of Science in Computer Science
SCSE	School of Computing Science and Engineering

Table of Contents

Title			Page No.
Abstract			**I**
List of Table			**II**
List of Figures			**III**
Chapter 1	**Introduction**		**1**
	1.1	Introduction	**2**
	1.2	Formulation of Problem	**3**
		1.2.1 Tool and Technology Used	
Chapter 2	**Literature Survey/Project Design**		**5**
Chapter 3	**Methodology**		**11-22**
Chapter 4	**Functionality /Working of Project**		**23-24**
Chapter 5	**Result Analysis**		**25**
Chapter 6	**References**		**26-28**
	Publication/Copyright		

CHAPTER-1
Introduction

Speech emotion recognition(SER) is mostly beneficial for applications, which need human-computer interaction such as speech synthesis, customer service, education, forensics and medical analysis. Speech being a primary medium to pass information ,we humans can also understand the intensity and mood of the speaker by the speech data generated . Recognizing of emotional conditions in speech signals are so challengeable area for several reasons. First issue of all speech emotion methods is select the best features, which will be powerful enough to distinguish between different emotions. The presence of various language, accent, sentences, speaking style, speakers also add another difficulty because these characteristics directly change most of the extracted features includes pitch, energy, etc. Speech emotion recognition is tough because emotions are subjective and annotating audio is challenging. The idea of creating this project was to build a machine learning model that could detect emotions from speech we have with us all the time. In this we have used librosa and MLP classifier, here librosa is used for analyzing audio and music. It has flatter package layout, standardizes interfaces and names, backwards compatibility, modular functions, and readable code. The MLP-Classifier is used to classify the emotions from the given wave of learning rate to be adaptive. In this study we attempt to detect underlying emotions in recorded speech by analyzing the acoustic features of the audio data of recordings. There are three classes of features in speech namely, lexical , visual and acoustic features. The problem of speech emotion recognition can be solved by analyzing one or more of these features.

Tools and Technologies:

In this Python mini project, we will use the libraries librosa, soundfile, and sklearn (among others) to build a model using an MLP Classifier. This will be able to recognize emotion from sound files. We will load the data, extract features from it, then split the dataset into training and testing sets. Then, we'll initialize an MLP Classifier and train the model. Finally, we'll calculate the accuracy of our model.

- Librosa

  Librosa is a python library for analysing audio and music. It has a flatter package layout, standardizes interfaces and names, backwards compatibility, modular functions, and readable code.

- Jupyter Lab

  Jupyter Lab is an open-source, web-based UI for Project Jupyter and it has all basic functionalities of the Jupyter Notebook, like notebooks, terminals, text editors, file browsers, rich outputs, and more. However, it also provides improved support for third

party extensions.

- MLP Classifier

  Multi-Layer perceptron (MLP) is a network made up of perceptron. It has an input layer that receives the input signal, an output layer that makes predictions or decisions for a given input, and the layers present in connecting the input and output layer is called hidden layer. In the proposed methodology for Speech Emotion Recognition, the MLP network will have one input layer, 300,40,80,40 hidden layers and one output layer. The hidden layers will be large numbers and number of hidden layers can be changed as per requirements.

- MFCC

  The envelope of the temporal power spectrum of the speech signal is representative of the vocal tract and MFCC accurately represents this envelope.

- Chroma

  Chroma-based features, which are also referred to as "pitch class profiles", are a powerful tool for analyzing music whose pitches can be meaningfully categorized (often into twelve categories) and whose tuning approximates to the equal-tempered scale.

CHAPTER-2

Literature Survey

The primary aim of the present study is to contribute to the theoretical debate currently occupying music and emotion research by systematically comparing evaluations of perceived emotions using two different theoretical frameworks: the discrete emotion model, and dimensional model of affect. The importance of the comparison lies not only in the prevalence of these models in music and emotion studies, but also in the suggested neurological differences involved in emotion categorization and the evaluation of emotion dimensions, as well as in the categorically constrained affect space the excerpts have represented to date. Moreover, the various alternative formulations of the dimensional model have not been investigated in music and emotion studies before. A secondary aim is to introduce a new, improved set of stimuli – consisting of unfamiliar, thoroughly tested and validated non-synthetic music excerpts – for the study of music mediated emotions. Moreover, this set of stimuli should not only include the best examples of target emotions but also moderate examples that permit the study of more subtle variations in emotion.

In this work, we have presented a database for the analysis of spontaneous emotions. The database contains physiological signals of 32 participants, where each participant watched and rated their emotional response to 40 music videos along the scales of arousal, valence, and dominance, as well as their liking of and familiarity with the videos. We presented a novel semiautomatic stimuli selection method using affective tags, which was validated by an analysis of the ratings participants gave during the experiment. Significant correlates were found between the participant ratings and EEG frequencies. Single-trial classification was performed for the scales of arousal, valence and liking using features extracted from the EEG, peripheral and MCA modalities. The results were shown to be significantly better than random classification. Finally, decision fusion of these results yielded a modest increase in the performance, indicating at least some complementarity to the modalities. The database is made publicly available and it is our hope that other researcher will try their methods and algorithms on this highly challenging database.

Methodology

In this the emotions in the speech are predicted using neural networks. Multi-Layer Perceptron Classifier (MLP Classifier) and RAVDESS(Rayerson Audio-Visual Database of Emotional Speech and Song dataset) are used.

- Database Description RAVDESS dataset has recordings of 24 actors, 12 male actors, and 12 female actors, the actors are numbered from 01 to 24. The male actors are odd in number and female actors are even in number. The emotions contained in the dataset are as sad, happy, neural, angry, disgust, surprised, fearful and calm expressions. The dataset contains all expressions in three formats, those are: Only Audio, Audio Video and Only Video. Since our focus is on recognize emotions from speech, this model is trained on Audio-only data.

- A Multi-Layer perceptron (MLP) is a network made up of perceptron. It has an input layer

that receives the input signal, an output layer that makes predictions or decisions for a given input, and the layers present in between the input and output layer is called hidden layer. In the proposed methodology for Speech Emotion Recognition, the MLP network will have one input layer, 300,40,80,40 hidden layers and one output layer. The hidden layers will be large numbers and number of hidden layers can be changed as per requirements.

- Features: The data was acquired directly from the group of Audio files and they were transformed in 264 vectors of features. A wide range of possibilities exist for parametrically representing a speech signal and its content in a vector, with intention to extract a relevant information from it. The learning process covers two steps as shown in Figure 1, the first step is a forward processing of input data by the neurons that produces a forecasted output, the second step is the adjustment of weights within neuron layers, in order to minimize the errors of forecasted solution compared with the correct output.

CHAPTER -3
Implementation

System Architecture*:*

In this Python mini project, we will use the libraries librosa, soundfile, and sklearn (among others) to build a model using an MLP Classifier. This will be able to recognize emotion from sound files. We will load the data, extract trait from it, then split the dataset into training and testing sets. Then, we will initialize a MLP Classifier and train the model. Finally, we will calculate the accuracy of our model.

- Librosa is a python library for analyzing audio and music.
- Jupyter Lab is an open-source, web-based UI for Project Jupyter and it has all basic functionalities of the Jupyter Notebook, like notebooks, terminals, text editors, file browsers, rich outputs, and more.

neural networks.

1) Multi-Layer Perceptron Classifier (MLP Classifier)
2) RAVDESS (Rayerson Audio-Visual Database of Emotional Speech and Song dataset) used.
3) TESS (Toronto emotional speech set) is also used for testing variable accuracy.

Database Description:

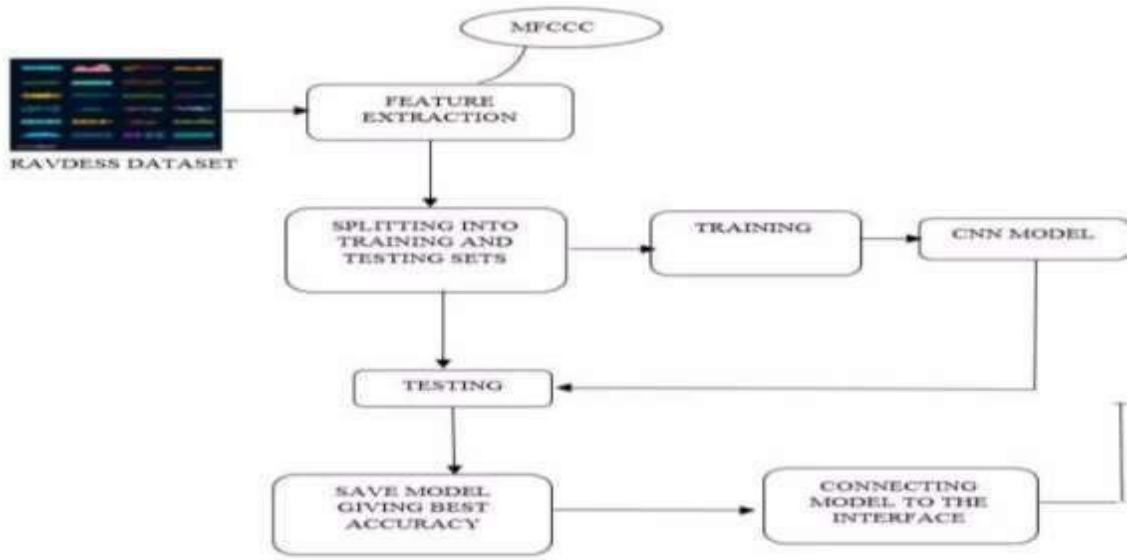
- RAVDESS dataset has recordings of 24 actors, 12 male actors, and 12 female actors, the actors are numbered from 01 to 24. The male actors are odd in number and female actors are even in number. The emotions contained in the dataset are as sad, happy, neural, angry, disgust, surprised, fearful and calm expressions. The dataset contains all expressions in three formats, those are: Only Audio, Audio Video and Only Video. Since our focus is on recognizing emotions from speech, this model is trained on Audio-only data.

- In TESS dataset there is a set of 200 target words were spoken in the carrier phrase "Say the word _' by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). There are 2800 data points (audio files) in total. The dataset is organized such that each of the two female actor and their emotions are contained within its own folder. And within that, all 200 target words audio file can be found. The format of the audio file is a WAV format.

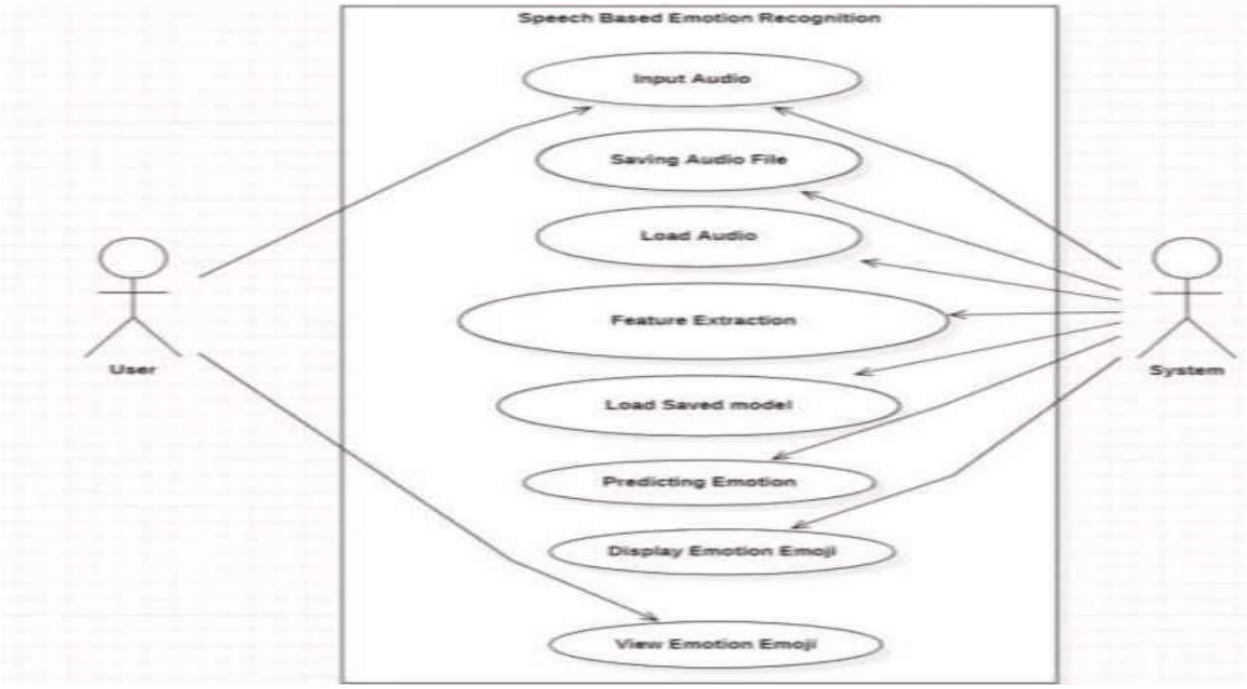
Multi-Layer perceptron (MLP) is a network made up of perceptron. It has an input layer that receives the input signal, an output layer that makes predictions or decisions for a given input, and the layers present in connecting the input and output layer is called hidden layer. In the proposed methodology for Speech Emotion Recognition, the MLP network will have one input layer, 300,40,80,40 hidden layers and one output layer. The hidden layers will be large numbers and number of hidden layers can be changed as per requirements.

Traits : The data was acquired directly from the group of Audio files and they were transformed in 264

vectors of trait. A wide range of possibilities exist for parametrically representing a speech signal and its content in a vector, with intention to extract admissible information from it.



Use Case Diagram:



Features used in this study

From the Audio data we have extracted three key features which have been used in this study, namely, MFCC (Mel Frequency Cepstral Coefficients), Mel Spectrogram and Chroma. The Python implementation of Librosa package was used in their extraction.

Choice of features

- MFCC was by far the most researched about and utilized features in research papers and open source projects.
- Mel spectrogram plots amplitude on frequency vs time graph on a "Mel" scale. As the project is on emotion recognition, a purely subjective item, we found it better to plot the amplitude on Mel scale as Mel scale changes the recorded frequency to "perceived frequency".
- Researchers have also used Chroma in their projects as per literatures, thus we also tried basic modeling with only MFCC and Mel and with all MFCC, Mel, Chroma. The model with all of the features gave slightly better results, hence we chose to keep all three features.

MFCC (Mel Frequency Cepstral Coefficients)

In the conventional analysis of time signals, any periodic component (for example, echoes) shows up as sharp peaks in the corresponding frequency spectrum (i.e. Fourier spectrum. This is obtained by applying a Fourier transform on the time signal). Any spectrum feature is obtained by applying Fourier Transform on a spectrogram. The special characteristic of MFCC is that it is taken on a Mel scale which is a scale that relates the perceived frequency of a tone to the actual measured frequency. It scales the frequency in order to match more closely what the human ear can hear. The envelope of the temporal power spectrum of the speech signal is representative of the vocal tract and MFCC accurately represents this envelope.

Mel Spectrogram

A Fast Fourier Transform is computed on overlapping windowed segments of the signal, and we get what is called the spectrogram. This is just a spectrogram that depicts amplitude which is mapped on a Mel scale.

Chroma

A Chroma vector is typically a 12-element feature vector indicating how much energy of each pitch class is present in the signal in a standard chromatic scale.

Pre Processing

As the typical output of the feature extracted were 2D in form, we decided to take a bi-directional approach using both a 1D form of input and a 2D form of input as discussed below.

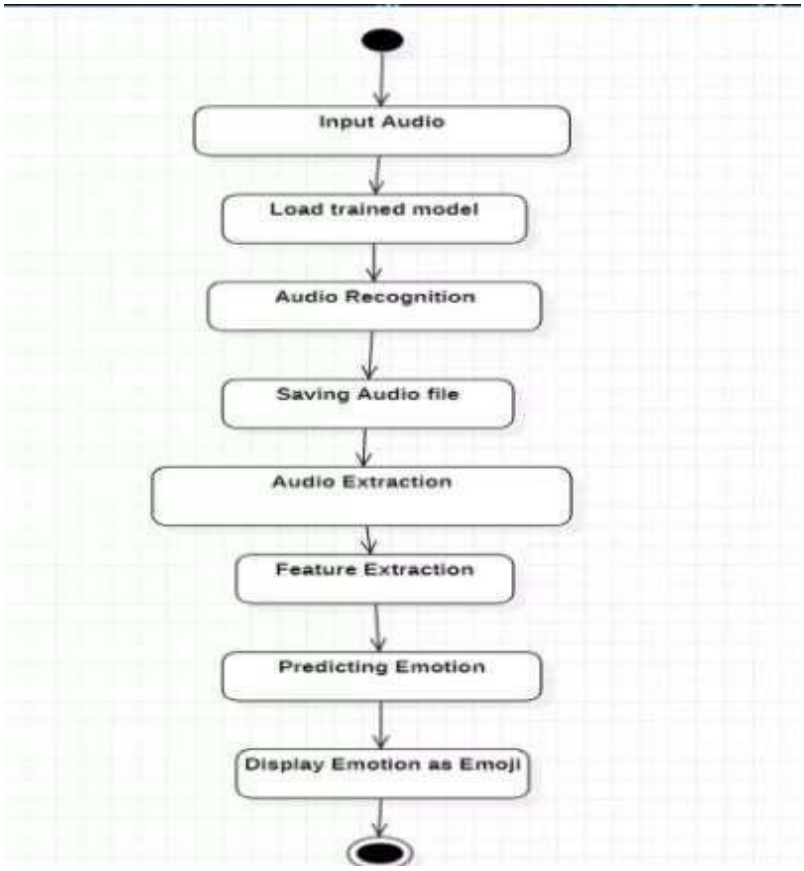
1D Data Format

These features obtained from extraction from audio clips are in a matrix format. To model them on traditional ML algorithms like SVM and XGBoost or on 1-D CNN, we considered converting the matrices into the 1-D format by taking row means and column means. Upon preliminary modelling the results obtained from the array of row means turned out to be better than the array of column means, so we proceeded with the 1 -D array obtained from row means of the feature matrices.

2D Data Format

The 2D features were used in the deep learning model (CNN). The y-axis of the feature matrices obtained depends on the n_mfcc or n_mels parameter we choose while extracting data. The x-axis depends upon the audio duration and the sampling rate we choose while feature extraction. Since the audio clips in our datasets were of varying lengths ranging from just under 2 seconds to over 6 seconds, steps like choosing one median length where we'll clip all audio files and pad all shorter files with zeroes to maintain dimensions wouldn't be feasible. This is because this would have resulted in the loss of information for longer clips and the shorter clips would be just silence for the latter half of their audio length. To check this problem, we decided to use different sampling rates in extraction in accordance with their audio lengths. In our approach any, audio file greater or equal to 5 seconds was clipped at 5 seconds ,sampled at 16000 Hz and the shorter clips were sampled such that the audio duration * sampling rate multiple remains 80000. In this way, we were able to maintain the dimensions of the matrix for all audio clips without losing much of the information.

Activity Diagram



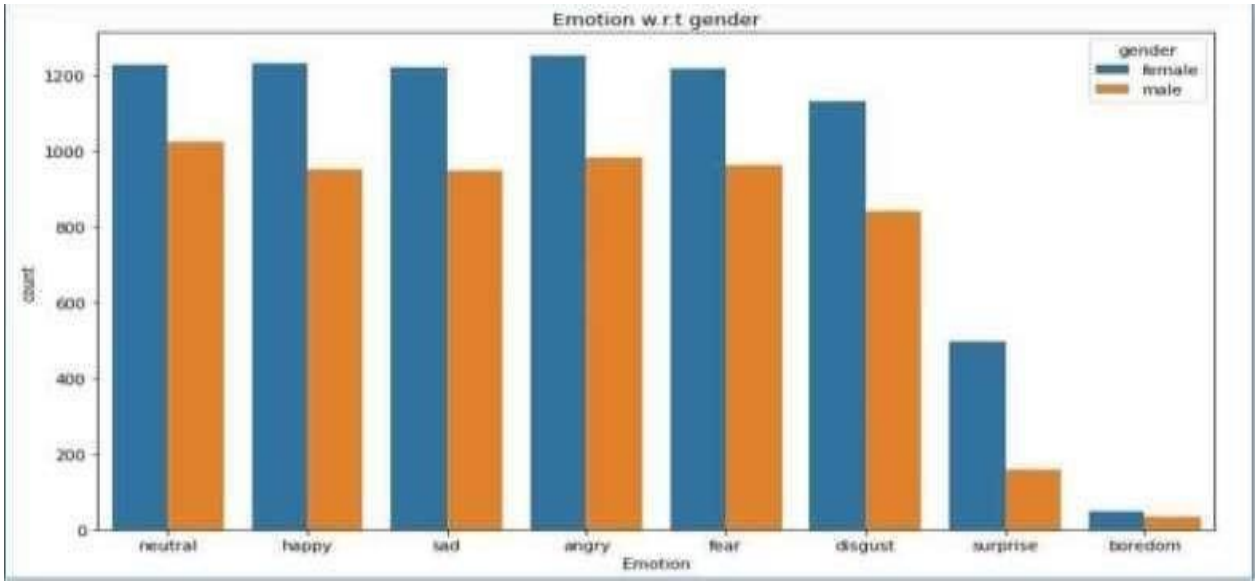
Exploratory Data Analysis

The combined data set from the original 5 sources is thoroughly analysed with respect to the following aspects

- Emotion distribution by gender
- Variation in energy across emotions
- Variation of relative pace and power across emotions
- We checked the distribution of labels with respect to emotions and gender and found that while the data is balanced for six emotions viz. neutral, happy, sad, angry, fear and disgust, the number of labels was slightly less for surprise and negligible for boredom. While the slightly fewer instances of surprise can be overlooked on account of it being a rarer emotion, the imbalance against boredom was rectified later by clubbing sadness and boredom together due to them being similar acoustically. It's also worth noting that boredom could have been combined with neutral emotion but since both sadness and boredom are negative emotions, it made more sense to combine them.

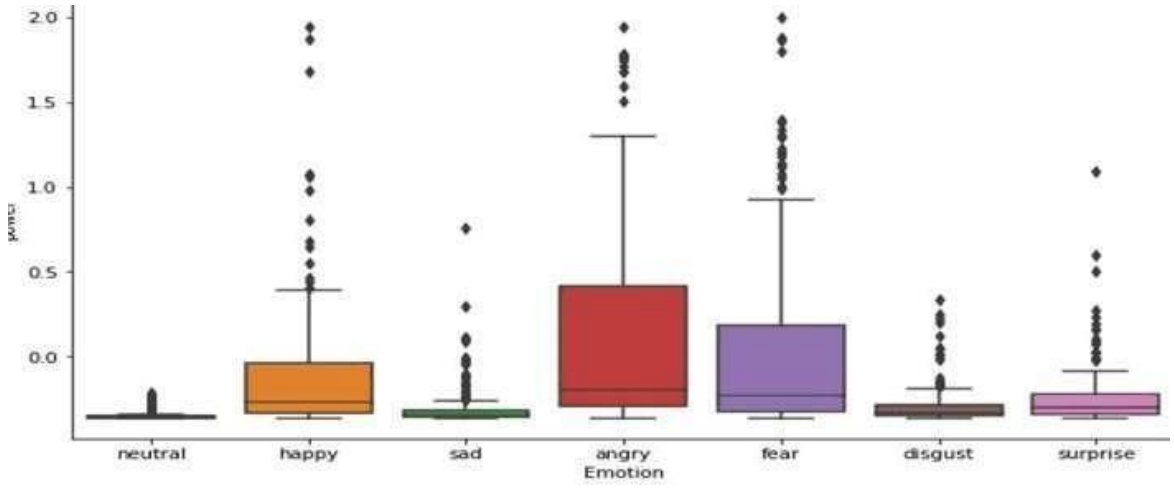
Emotion Distribution of Gender

Regarding the distribution of gender, the number of female speakers was found to be slightly more than the male speakers, but the imbalance was not large enough to warrant any special attention. Refer Figure.



Variation in Energy Across Emotions

To ensure uniformity in our study of energy variation as the audio clips in our dataset were of different lengths, a power which is energy per unit time was found to be a more accurate measure. This metric was plotted with respect to different emotions. From the graph See Fig. 2) it is quite evident that the primary method of expression of anger or fear in people is a higher energy delivery. We also observe that disgust and sadness are closer to neutral with regards to energy although exceptions do

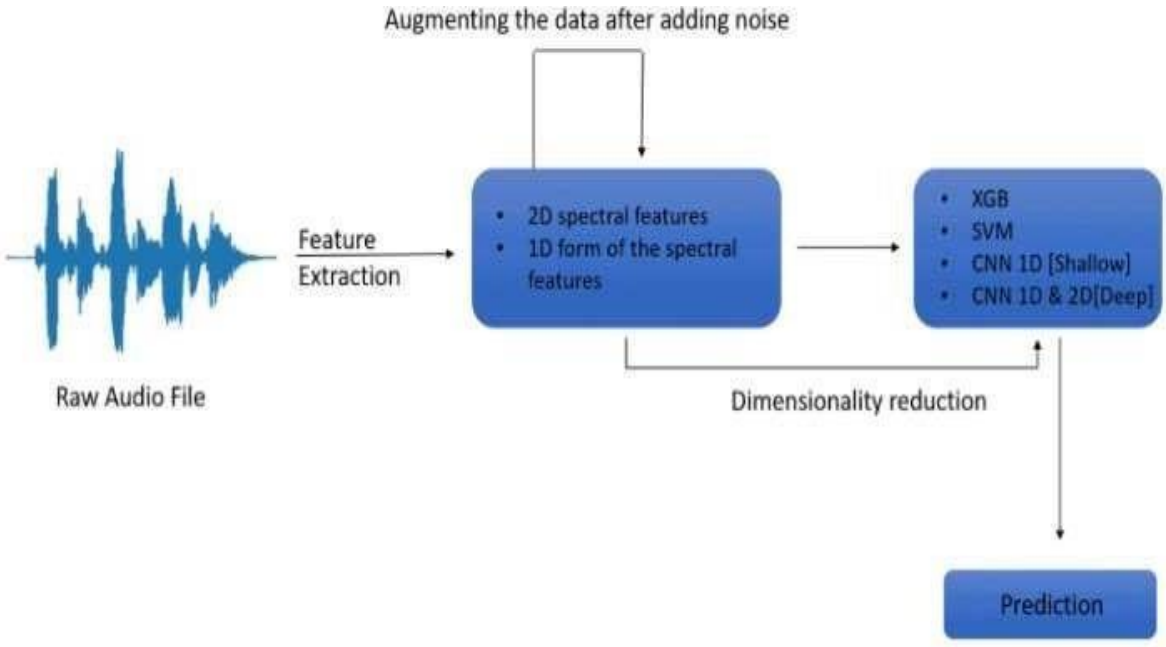


Variation of Relative Pace and Power with respect to Emotions

A scatter-plot of power vs relative pace of the audio clips was analyzed and it was observed that the 'disgust' emotion was skewed towards the low pace side while the 'surprise' emotion was skewed more towards the higher pace side. As mentioned before, anger and fear occupy the high power space and sadness and neutral occupy the low power space while being scattered pace-wise. Only, the RAVDESS dataset was used for plotting here because it contains only two sentences of equal length spoken in different emotions, so the lexical features don't vary and the relative pace can be reliably calculated.

Solution Pipeline

The solution pipeline for this study is depicted in the schematic shown in Fig. 4. The raw signal is the input which is processed as shown. At first the 2D features were extracted from the datasets and converted into 1 -D form by taking the row means. A measure of noise was added to the raw audio for 4 of our datasets (except CREMA-D as the others were studio recording and thus cleaner). The features were then extracted from those noisy files and our dataset was augmented with them. Post feature extraction we applied various ML algorithms such as SVM, XGB, CNN- 1D(Shallow) and CNN-1D on our 1D data frame and CNN-2D on our 2D- tensor. As some of the models were overfitting the data, and taking into consideration a large number of features (181 in 1D) we tried dimensionality reduction to check overfitting and trained the models again.



Dimensionality Reduction Approach

In order to rectify the overfitting of the models we used a dimensionality reduction approach. PCA technique was employed for dimensionality reduction in 1D features and dimensions were reduced from 180 to 120 with an explained variance of 98.3%. Dimensionality reduction made the model slightly less accurate but reduced the training time, however it didn't do much to reduce overfitting in the deep learning model. From this we deduced that our dataset is simply not big enough for a complex model to perform well and realised the solution was limited by lack of a larger data volume.

Insights from Testing

- We tested the developed models on user recordings, from the test results we have the following observations
- An ensemble of CNN-2D and CNN-1D (shallow and deep) based on a soft voting gave best results on user recordings.
- The model often got confused between anger and disgust.
- The model also got confused among low energy emotions which are sadness, boredom and neutral.
- If one or two words are spoken in higher volume than other words, especially at start or end of a sentence, it almost always classifies as fear or surprise.
- The model seldom classifies an emotion as happy.
- The model isn't too noise sensitive, meaning it doesn't falter as long as background noise is not too high.

Grouping Similar Emotions

Since the model was confusing between similar emotions like anger- disgust and sad-bored, we tried combining those labels and training the model on 6 classes which were neutral, sadness/boredom, happy, anger/disgust, surprise and fear. The accuracies certainly improved on reducing the number of classes, but this introduced another problem with regards to class imbalance. After, combining anger-disgust and sad- boredom, the model developed a high bias towards the anger-disgust. This may have happened because the number of instances of anger-disgust became disproportionately more than the other labels. So, it was decided to stick with the older model.

CHAPTER-4

Functionality/Working of Project

Code:

```
import sys
!{sys.executable} -m pip install librosa import librosa
import soundfile import os, glob,
pickle import numpy as np
from sklearn.model_selection import train_test_split from
sklearn.neural_network import MLPClassifier from sklearn.metrics import
accuracy_score
def extract_feature(file_name, mfcc, chroma, mel): with
    soundfile.SoundFile(file_name) as sound_file:
        X = sound_file.read(dtype="float32")
        sample_rate=sound_file.samplerate if chroma:
            stft=np.abs(librosa.stft(X)) result=np.array([])
        if mfcc:
            mfccs=np.mean(librosa.feature.mfcc(y=X, sr=sample_rate, n_mfcc=40).T, axis=0)
            result=np.hstack((result, mfccs)) if chroma:
            chroma=np.mean(librosa.feature.chroma_stft(S=stft, sr=sample_rate).T,axis=0)
            result=np.hstack((result, chroma)) if mel:
            mel=np.mean(librosa.feature.melspectrogram(X, sr=sample_rate).T,axis=0)


result=np.hstack((result, mel)) return result
emotions={ '01':'neutral',
  '02':'calm',
  '03':'happy',
  '04':'sad',
  '05':'angry',
  '06':'fearful',
  '07':'disgust',
  '08':'surprised'
}
observed_emotions=['calm', 'happy', 'fearful', 'disgust'] def
load_data(test_size=0.2):
    x,y=[],[]
    for file in glob.glob("D:\\DataFlair\\ravdess data\\Actor_*\\*.wav"):
        file_name=os.path.basename(file) emotion=emotions[file_name.split("-")[2]]
        if emotion not in observed_emotions: continue
        feature=extract_feature(file, mfcc=True, chroma=True, mel=True)
        x.append(feature) y.append(emotion)
```

```
    return train_test_split(np.array(x), y, test_size=test_size, random_state=9)
x_train,x_test,y_train,y_test=load_data(test_size=0.25) print((x_train.shape[0],
x_test.shape[0]))
print(f'Features extracted: {x_train.shape[1]}') model=MLPClassifier(alpha=0.01, batch_size=256,
epsilon=1e-08, hidden_layer_sizes=(300,), learning_rate='adaptive', max_iter=500)
model.fit(x_train,y_train)
y_pred=model.predict(x_test) accuracy=accuracy_score(y_true=y_test, y_pred=y_pred)
print("Accuracy: {:.2f}%".format(accuracy*100))
```

CHAPTER 5
Result Analysis and Conclusions

Through this project, we showed how we can leverage Machine learning to obtain the underlying emotion from speech audio data and some insights on the human expression of emotion through voice. This system can be employed in a variety of setups like Call Centre for complaints or marketing, in voice-based virtual assistants or chatbots, in linguistic research, etc. A few possible steps that can be implemented to make the models more robust and accurate are the following

- An accurate implementation of the pace of the speaking can be explored to check if it can resolve some of the deficiencies of the model.
- Figuring out a way to clear random silence from the audio clip.
- Exploring other acoustic features of sound data to check their applicability in the domain of speech emotion recognition. These features could simply be some proposed extensions of MFCC like RAS-MFCC or they could be other features entirely like LPCC, PLP or Harmonic spectrum.
- Following lexical features based approach towards SER and using an ensemble of the lexical and acoustic models. This will improve the accuracy of the system because in some cases the expression of emotion is contextual rather than vocal.
- Adding more data volume either by other augmentation techniques like time-shifting or speeding up/slowing down the audio or simply finding more annotated audio clips.

The result is based on the accuracy metrics in which there is a comparison between predicted values and the actual values. A confusion matrix is created which consists of true positive (TP), true negative (TN), false positive (FP), and false negative (FN). From confusion metrics, we have calculated accuracy as follows:

$$\text{accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

In this work we learned to recognize emotions from speech. We used an MLP Classifier(Multi Layer Perceptron) for this and made use of the sound file library to read the 30 audio files, and the librosa library to extract traits from it.

```
accuracy=accuracy_score(y_true=y_test, y_pred=y_pred)
print("Accuracy: {:.2f}%".format(accuracy*100))
```

```
Accuracy: 77.08%
```

```
import pandas as pd
df=pd.DataFrame({'Actual':y_test,'Predicted':y_pred})
df.head(10)
```

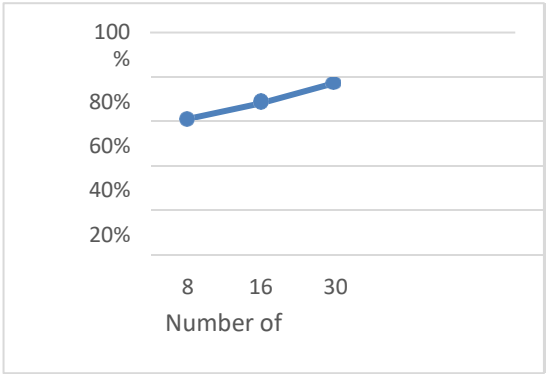
	Actual	Predicted
0	happy	happy
1	calm	calm
2	happy	happy
3	happy	happy
4	disgust	disgust
5	calm	calm
6	happy	disgust
7	happy	disgust
8	disgust	disgust
9	happy	happy

Fig 1. Shows a correlation between actual and predicted values of the model

The training dataset is kept at 75% and the testing dataset is kept at 25% while splitting.

Table 1.

Samples	Accuracy
8	61%
16	68%
30	77%



Chapter 6

References

Blogs and Documentations:

Librosa – https://librosa.org/librosa/tutorial.html

MFCC

- https://medium.com/prathena/the-dummys-guide-to-mfcc- aceab2450fd
- http://practicalcryptography.com/miscellaneous/machine- learning/guide-mel-frequency-cepstral-coefficients-mfccs/
- Mel Spectrogram- https://towardsdatascience.com/getting-to- know-the-mel-spectrogram-31bca3e2d9d0

Dataset References:

- SAVEE: http://kahlan.eps.surrey.ac.uk/savee/Download.html
- TESS: https://tspace.library.utoronto.ca/handle/1807/24487
- RAVDESS: https://zenodo.org/record/1188976#.XvbvZudS_IU
- BERLIN: http://www.emodb.bilderbar.info/download/
- CREMA-D: https://github.com/CheyneyComputerScience/CREMA-D

Literature References

- Ittichaichareon, C. (2012). Speech recognition using MFCC. … Conference on Computer …, 135–138.

  https://doi.org/10.13140/RG.2.1.2598.3208

- Al-Talabani, A., Sellahewa, H., & Jassim, S. A. (2015). Emotion recognition from speech: tools and challenges. Mobile Multimedia/Image Processing, Security, and Applications

2015, 9497(May 2020), 94970N.
https://doi.org/10.1117/12.2191623

- Sezgin, M. C., Gunsel, B., & Kurt, G. K. (2012). Perceptual audio features for emotion detection. EURASIP Journal on Audio, Speech, and Music Processing, 2012(1),

  https://doi.org/10.1186/1687-4722-2012-16

- Kernel References
  https://github.com/marcogdepinto/emotion-classification-from- audio-files?fbclid=IwAR2T4hhtWWfKdU4FwLS8LOAnF5sBwnmfc6PQHTGidzLaLl1uUVOvicx7TVw