**Automatic Speech Emotion Recognition**

Project Report submitted in partial fulfillment of

The requirements for the degree of

# BACHELOR OF TECHNOLOGY

In

# COMPUTER SCIENCE AND ENGINEERING

Of

# MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY

By

AYUSH, 84, 10900117099

IPSHA KUMARI, 72, 10900117087

PRIYANSHU RAJ, 49, 10900117064

PARIKSHIT, 54, 10900117069

Under the guidance of

Mr. Sourav Dutta

Assistant Professor

**DEPARTMENT OF COMPUTER SCIENCE**



## NETAJI SUBHASH ENGINEERING COLLEGE

**TECHNO CITY, GARIA, KOLKATA – 700152**

Academic year of pass out 2020-21

#### CERTIFICATE

This is to certify that this project report titled **Automatic Speech Emotion Recognition**  submitted in partial fulfillment of requirements for award of the degree Bachelor of Technology (B. Tech) in Computer Science and Engineering of Maulana Abul Kalam Azad University of Technology is a faithful record of the original work carried out by,

**Ayush**, **Roll no**. 10900117099, **Regd. No.**171090110029 OF 2017-2018

**Ipsha Kumari** , **Roll no**. 10900117087, **Regd. No.** 171090110041 OF 2017-2018

**Priyanshu Raj** , **Roll no**. 10900117064, **Regd. No.** 171090110064 OF 2017-2018

**Parikshit**, **Roll no**. 10900117069, **Regd. No.** 171090110059 OF 2017-2018

under my guidance and supervision.

It is further certified that it contains no material, which to a substantial extent has been submitted for the award of any degree/diploma in any institute or has been published in any form, except the assistance drawn from other sources, for which due acknowledgement has been made.

Date:………..

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dr. Piyali Chatterjee Mr. Sourav Dutta Head of the Department Project Guide

Computer Science and Engineering Computer Science and Engineering

## NETAJI SUBHASH ENGINEERING NETAJI SUBHASH ENGINEERING

## COLLEGE COLLEGE

TECHNO CITY, GARIA, TECHNO CITY, GARIA,

KOLKATA – 700 152 KOLKATA – 700 152

##### **DECLARATION**

##### We hereby declare that this project report titled **Automatic Speech Emotion Recognition** is our own original work carried out as an undergraduate student in Netaji Subhash Engineering College except to the extent that assistance from other sources are duly acknowledged.

##### All sources used for this project report have been fully and properly cited. It contains no material which to a substantial extent has been submitted for the award of any degree/diploma in any institute or has been published in any form, except where due acknowledgement is made.

##### Student’s names: Signatures: Dates:

##### ……………………….. ……………………….. ……………………

##### ……………………….. ……………………….. ……………………

##### ……………………….. ……………………….. ……………………

##### ……………………….. ……………………….. ……………………

##### **CERTIFICATE OF APPROVAL**

##### We hereby approve this dissertation titled **Automatic Speech Emotion Recognition** carried out by:

**Ayush**, **Roll no**. 10900117099, **Regd. No.** 171090110029 OF 2017-2018

**Ipsha Kumari** , **Roll no**. 10900117087, **Regd. No.** 171090110041 OF 2017-2018

**Priyanshu Raj** , **Roll no**. 10900117064, **Regd. No.** 171090110064 OF 2017-2018

**Parikshit** , **Roll no**. 10900117069, **Regd. No.** 171090110059 OF 2017-2018

under the guidance of Mr. Sourav Dutta of Netaji Subhash Engineering College, Kolkata in partial fulfillment of requirements for award of the degree Bachelor of Technology (B. Tech) in Computer Science and Engineering of Maulana Abul Kalam Azad University of Technology.

##### Date:………..

##### Examiners’ signatures:

1. ………………………………………….
2. ………………………………………….
3. ………………………………………….

**Acknowledgement**

The success and final outcome of this project required a lot of guidance and assistance from many people and we are extremely privileged to have got this all along the completion of the project. All that we have done is only due to such supervision and assistance and we would not forget to thank them. We owe our deep gratitude to our project guide Mr. Sourav Dutta who took keen interest in our project work and guided us all along, till the completion of the project work by providing all the necessary information for developing a good project.

We heartily thank our internal project guide, Dr. Piyali Chatterjee, HOD of CSE(NSEC) for her guidance and suggestions during this project work. We are thankful to and fortunate enough to get constant encouragement, support and guidance from all teaching staff of Netaji Subhash Engineering College which helped us in successfully completing the project work. Also, we would like to extend our sincere esteems to all staff in the laboratory for their timely support.

**Ayush**

**Ipsha Kumari**

**Priyanshu Raj**

**Parikshit**

Dated:…………………

**Abstract**

This project presents a comparative study of speech emotion recognition (SER) systems. Theoretical definition, categorization of affective state and the modalities of emotion expression are presented. To achieve this study, an SER system, based on different classifiers and different methods for features extraction, is developed. Mel-frequency cepstrum coefficients (MFCC) and modulation spectral (MS) features are extracted from the speech signals and used to train different classifiers. Feature selection (FS) was applied in order to seek for the most relevant feature subset. Several machine learning paradigms were used for the emotion classification task. A recurrent neural network (RNN) classifier is used first to classify seven emotions. Their performances are compared later to multivariate linear regression (MLR) and support vector machines (SVM) techniques, which are widely used in the field of emotion recognition for spoken audio signals. RAVDESS dataset is used as the experimental data set.

**Table of Contents**

1. **Introduction ……………………………………………………….… 7**

1.1 Features used in this study …………………………………………...... 8

1.2 MFCC ……………………………………………………………….… 8

1.2.1 MEL Spectrogram ………………………………………………..… 8

1.2.2 Chroma ……………………………………………………………... 9

1.2.3 Methodologies Used ………………………………………………... 9

1.3 DATASET ………………………………………………………....…… 9

1. **Literature Review ……………………………………………..……. 10**

2.1 Berlin Database Of Emotional Speech …………………………...……. 10

2.2 RAVDESS ………………………………………………………...…… 10

2.2.1 File Naming Convention ……………………………………... 10

2.2.1.1 File Name Identifier …………………………………...…. 10

2.2.1.2 File Name Example ………………………………………. 11

1. **Proposed Work ……………………………………………………... 12**

3.1 Emotional Speech Databases ……………………………………….…. 12

3.2 Hardwares and Softwares Used …………………………………..…… 13

3.2.1 Librosa ……………………………………………………………... 13

3.2.2 Jupyter Notebook …………………………………………………... 13

3.2.3 Kernel ……………………………………………………………..... 13

3.2.4 Notebook Dashboard …………………………...…………………... 14

3.3 System Requirements ………………………………………………..… 14

3.4 Algorithm …………………………………………………………....… 14

1. **Results ……………………………………………………………….. 16**
2. **Conclusion ………………………………………………………...… 20**
3. **Future Work ……………………………………………………...…. 21**
4. **References ………………………………………………………...… 22**

**Chapter 1:** **Introduction**

Emotion plays a significant role in daily interpersonal human interactions. This is essential to our rational as well as intelligent decisions. It helps us to match and understand the feelings of others by conveying our feelings and giving feedback to others. Research has revealed the powerful role that emotion plays in shaping human social interaction. Emotional displays convey considerable information about the mental state of an individual. This has opened up a new research field called automatic emotion recognition, having basic goals to understand and retrieve desired emotions. In prior studies, several modalities have been explored to recognize the emotional states such as facial expressions, speech, physiological signals, etc. Several inherent advantages make speech signals a good source for effective computing. For example, compared to many other biological signals (e.g., electrocardiogram), speech signals usually can be acquired more readily and economically. This is why the majority of researchers are interested in speech emotion recognition (SER). SER aims to recognize the underlying emotional state of a speaker from his/her voice. The area has received increasing research interest all through current years. There are many applications of detecting the emotion of the persons like in the interface with robots, audio surveillance, web-based E-learning, commercial applications, clinical studies, entertainment, banking, call centers, cardboard systems, computer games, etc. For classroom orchestration or E-learning, information about the emotional state of students can provide focus on the enhancement of teaching quality. For example, a teacher can use SER to decide what subjects can be taught and must be able to develop strategies for managing emotions within the learning environment. That is why the learner's emotional state should be considered in the classroom.

Three key issues need to be addressed for a successful SER system, namely,

(1) choice of a good emotional speech database,

(2) extracting effective features, and

(3) designing reliable classifiers using machine learning algorithms.

In fact, the emotional feature extraction is a main issue in the SER system. Many researchers have proposed important speech features which contain emotion information, such as energy, pitch, formant frequency, Linear Prediction Cepstrum Coefficients (LPCC), Mel-frequency cepstrum coefficients (MFCC), and modulation spectral features (MSFs). Thus, most researchers prefer to use a combining feature set that is composed of many kinds of features containing more emotional information. However, using a combining feature set may give rise to high dimension and redundancy of speech features; thereby, it makes the learning process complicated for most machine learning algorithms and increases the likelihood of overfitting. Therefore, feature selection is indispensable to reduce the dimensions of redundancy of features. A review for feature selection models and techniques is presented in. Both feature extraction and feature selection are capable of improving learning performance, lowering computational complexity, building better generalizable models, and decreasing required storage. The last step of speech emotion recognition is classification. It involves classifying the raw data in the form of utterance or frame of the utterance into a particular class of emotion on the basis of features extracted from the data. In recent years in speech emotion recognition, researchers proposed many classification algorithms, such as Gaussian mixture model (GMM), hidden Markov model (HMM), support vector machine (SVM), neural networks (NN), and recurrent neural networks (RNN). Some other types of classifiers are also proposed by some researchers such as a modified brain emotional learning model (BEL) in which the adaptive neuro-fuzzy inference system (ANFIS) and multilayer perceptron (MLP) are merged for speech emotion recognition. Another proposed strategy is a multiple kernel Gaussian process (GP) classification, in which two similar notions in the learning algorithm are presented by combining the linear kernel and radial basis function (RBF) kernel. The Voiced Segment Selection (VSS) algorithm also deals with the voiced signal segment as the texture image processing feature which is different from the traditional method. It uses the Log-Gabor filters to extract the voiced and unvoiced features from the spectrogram to make the classification.

### 1.1. 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 the Librosa package was used in their extraction.

**Choice of features**

* MFCC was by far the most researched 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 literature, 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.

**1.2. 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 cepstrum 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.

##### **1.2.1. 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.

##### **1.2.2. 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.

## 1.2.3. Methodologies used:

* **Preprocessing:**

In this process, the removal of unwanted voice signals from the speech takes place by the following processes :

1. Silent Removal

2. Background Noise Removal

3. Windowing

4. Normalization

* **Feature Extraction:**

In this process we extract the feature from the audio file. It is used to identify the following features :

-Pitch -Loudness -Rhythm etc.

For feature extraction we make use of the [LibROSA](https://librosa.github.io/librosa/) library in python which is one of the libraries used for audio analysis.

* **Classification:**

Once the features of the voice have been extracted in the feature selection process, the project matches the feature with corresponding emotions.

### 1.3. Datasets:

Made use of one dataset:

1. [RAVDESS](https://zenodo.org/record/1188976): This dataset includes around 1500 audio file input from 24 different actors. 12 male and 12 female where these actors record short audios in 8 different emotions i.e 1 = neutral, 2 = calm, 3 = happy, 4 = sad, 5 = angry, 6 = fearful, 7 = disgust, 8 = surprised.  
    Each audio file is named in such a way that the 7th character is consistent with the different emotions that they represent.

**Chapter 2 :** **Literature Review**

In previous related works, a system for the recognition of seven acted emotional states (anger, disgust, fear, joy, sadness, and surprise) is presented. To do that, extraction of the MFCC and MS features were done and they were used to train three different machine learning paradigms (MLR, SVM, and RNN). It was demonstrated that the combination of both features has a high accuracy of 72% on the RAVDESS database. All previously published works generally use the Berlin database. To our knowledge, the RAVDESS emotional database has not been popular much. For this reason, it has been chosen to compare them as it is in the english language.

**2.1. Berlin Database of Emotional Speech (EMO-DB)** is a simulated dataset composed of 10 German sentences, five short sentences, and five long sentences. Ten speakers, five females, and five males were employed to create the dataset. Each one of the speakers had expressed ten sentences, five long and five short, with different emotions. Along with the recording of the audio signal, to be able to extract prosodic and voice quality more precisely, the electroglottograms were also recorded along with the voices. The whole dataset contains 700 samples, of 10 sentences acted with seven emotions. The emotions chosen for this dataset were neutral, anger, fear, joy, sadness, disgust, and boredom. This study shows that for the Berlin database all classifiers achieve an accuracy of 83% when a speaker normalization (SN) and a feature selection are applied to the features.

**2.2. Ryerson Audio-Visual Database of Emotional Speech and Song(RAVDESS) Database**

It is a large dataset with an audio and video database. The original size of this full dataset of speech and song, audio and video is 24.8 GB. But we will use a smaller portion of it and not the whole dataset. This will help us to stay focused, train our model faster and to keep things simple.

This portion of the RAVDESS contains 1440 files: 60 trials per actor x 24 actors = 1440. The RAVDESS contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech emotions include neutral,calm, happy, sad, angry, fearful, surprise, and disgust expressions.

**2.2.1. File naming convention**

Each of the 1440 files has a unique filename. The filename consists of a 7-part numerical identifier (e.g., 03-01-06-01-02-01-12.wav). These identifiers define the stimulus characteristics:

**2.2.1.1. Filename identifiers**

* Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
* Vocal channel (01 = speech, 02 = song).
* Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
* Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the 'neutral' emotion.
* Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").
* Repetition (01 = 1st repetition, 02 = 2nd repetition).
* Actor (01 to 24. Odd numbered actors are male, even numbered actors are female).

**2.2.1.2. Filename example**: 03-01-06-01-02-01-12.wav

* Audio-only (03)
* Speech (01)
* Fearful (06)
* Normal intensity (01)
* Statement "dogs" (02)
* 1st Repetition (01)
* 12th Actor (12)
* Female, as the actor ID number is even.

For the RAVDESS database, the best accuracy (72 %) is achieved by RNN classifier without SN and with FS.

In this project we have used the Multi Layer Perceptron(MLP) classifier and increased the accuracy to 75-80%. MLP uses a neural network model to optimize the log-loss function using Limited memory BFGS or stochastic gradient descent.

MLP Classifier trains iteratively since at each time step the partial derivatives of the loss function with respect to the model parameters are computed to update the parameters.

It can also have a regularization term added to the loss function that shrinks model parameters to prevent overfitting.

This implementation works with data represented as dense numpy arrays or sparse scipy arrays of floating point values.

Our accuracy score varies between 70-80%(it came 82.76% also), and that is pretty impressive. We usually get a similar score after fitting the model multiple times. We do think that this is a satisfying score for an emotion recognition model, which was trained by audio recordings. Thanks to machine learning and artificial intelligence model developers.

**Chapter 3 : Proposed Work**

**3.1. Emotional Speech Databases :**

For every machine learning task, we need to have a training set of samples; SER is not different from the rest. The process of creating a training dataset for SER needs human agents to label the samples by hand, and different people perceive emotions differently. For example, one might tag an emotional voice as angry whilst the other perceives it as excited. This ambiguity means to label the samples we must have more than one agent reviewing each and then having a system to choose the labels for available samples confidently. There are three types of databases specifically designed for speech emotion recognition, simulated, semi-natural, and natural speech collections. The simulated datasets are created by trained speakers reading the same text with different emotions. Semi-natural collections are made by asking people or actors to read a scenario containing different emotions. Moreover, natural datasets are extracted from TV shows, YouTube videos, call centers, and such, and then labeled the emotions by human listeners. Simulated data sets such as EMO-DB (German) , DES (Danish) , RAVDESS, TESS, and CREMA-D are standardized collections of emotions, which makes comparing results very easy. Although their numbers of distinct emotions are significant, as they have synthesized emotions, they tend to have overfitted models around emotions slightly different than what is happening in day-to-day conversations. Semi-natural collections of emotions include IEMOCAP, Belfast, and NIMITEK. This group has the advantage of being very similar to the natural utterances of speech. However, even though they are based on scenarios and the speech is happening in a contextual setting, they are artificially created emotions, especially when speakers know that they are being recorded for analysis reasons. Additionally, due to the limitations of the situations in scenarios, they have a limited number of emotions in comparison to the previous group. The last group is the natural corpora of emotional speech databases such as VAM , AIBO , and call center data. These are entirely natural, and they can be safely used to model emotion recognition systems without hesitation about being artificially made. However, modeling and detection of the emotions with this type of datasets can be complicated due to the consciousness of emotions and their dynamic variation during the course of the speech, and the existence of concurrent emotions together, and the presence of background noise. Additionally, because the sources of the data were limited, the number of different emotions found in these corpora is limited. Moreover, there can be potential copyright and privacy issues that arise using this type of corpora. The major challenge in using this type of dataset is the need for noise reduction. Earlier examples of databases for emotional speech used to contain a limited number of samples with a limited number of actors, but newer databases tend to create a larger number of samples and a wider range of speakers.

We have taken ideas through many approaches. One of them is

**Berlin Database(EMO-DB)**:

The Berlin Database of Emotional Speech (EMO-DB) is one of the most widely used datasets for speech emotion recognition. It is a simulated dataset composed of 10 German sentences, five short sentences, and five long sentences. Ten speakers, five females, and five males were employed to create the dataset. Each one of the speakers had expressed ten sentences, five long and five short, with different emotions. Along with the recording of the audio signal, to be able to extract prosodic and voice quality more precisely, the electroglottography was also recorded along with the voices. The whole dataset contains 700 samples, of 10 sentences acted with seven emotions. The emotions chosen for this dataset were neutral, anger, fear, joy, sadness, disgust, and boredom.

But here in this one we are taking a different approach:

**Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS):**

The RAVDESS is a dataset consisting of happy, sad, angry, fearful, surprised, disgusted, calm, and neutral emotions performed by 24 actors, 104 sentences each actor. Totaling 2496 clips, RAVDESS is very rich in variations of the samples; also, every emotion is performed in two different intensities and both with a normal voice and singing voice. This is one of the crucial features of RAVDESS, and only a few data sets can Sensors 2021, 21, 1249 6 of 27 claim to have such a feature. Moreover, RAVDESS is among the few datasets that contain, North American English accent, and this could be important in cases where the American English accent makes the evaluation test set.

**3.2. Hardwares and Softwares used:**

**3.2.1. Librosa:**

librosa is a python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems.

Librosa’s load function will read in the path to an audio file, and return a tuple with two items. Librosa’s load function will read in the path to an audio file, and return a tuple with two items. The first item is an ‘audio time series’(type: array) corresponding to the audio track. The second item in the tuple is the sampling rate that was used to process the audio.

**3.2.2. Jupyter notebook:**

The Jupyter Notebook App is a server-client application that allows editing and running notebook documents via a web browser. The Jupyter Notebook App can be executed on a local desktop requiring no internet access (as described in this document) or can be installed on a remote server and accessed through the internet.

In addition to displaying/editing/running notebook documents, the Jupyter Notebook App has a “Dashboard” (Notebook Dashboard), a “control panel” showing local files and allowing them to open notebook documents or shutting down their kernels.

**3.2.3. Kernel**:

A notebook kernel is a “computational engine” that executes the code contained in a Notebook document. The ipython kernel, referenced in this guide, executes python code. Kernels for many other languages exist (official kernels).

When we open a Notebook document, the associated kernel is automatically launched. When the notebook is executed (either cell-by-cell or with menu Cell -> Run All), the kernel performs the computation and produces the results. Depending on the type of computations, the kernel may consume significant CPU and RAM. Note that the RAM is not released until the kernel is shut-down.

**3.2.4. Notebook dashboard:**

The Notebook Dashboard is the component which is shown first when you launch the Jupyter Notebook App. The Notebook Dashboard is mainly used to open notebook documents, and to manage the running kernels (visualize and shutdown).

The Notebook Dashboard has other features similar to a file manager, namely navigating folders and renaming/deleting files.

**3.3. System Requirements:**

**Jupyter notebook requirements:**

With PySpark (Team Studio version 6.2 and later):

* Memory and disk space required per user: 1GB RAM + 1GB of disk + .5 CPU core.
* Server overhead: 2-4GB or 10% system overhead (whatever is larger), .5 CPU cores, 10GB disk space.
* Port requirements: Port 8000 plus 5 unique, random ports per notebook.

Without PySpark (Team Studio version 6.0 or 6.1):

* Memory and disk space required per user: 512MB RAM + 1GB of disk + .5 CPU core.
* Server overhead: 2-4GB or 10% system overhead (whatever is larger), .5 CPU cores, 10GB disk space.
* Port requirements: Port 8000.

**3.4. Algorithm:**

Step 1 : First thing’s first, we will install the libraries we need by using :

*pip install soundfile librosa numpy sklearn*

Step 2 : After the installation of the project, we can directly open a new text editor(preferably Jupyter Notebook).

Step 3 : Then we will install all the necessary libraries like Librosa, SoundFile, Numpy, Scikit-learn etc. :

*import librosa as lb*

*import soundfile as sf*

*import numpy as np*

*import os, glob, pickle*

*from sklearn.model\_selection import train\_test\_split*

*from sklearn.neural\_network import MLPClassifier*

*from sklearn.metrics import accuracy\_score*

Step 4 : Then we will create a dictionary of emotion categories to use when training the machine learning model.

*emotion\_labels = {*

*'01':'neutral',*

*'02':'calm',*

*'03':'happy',*

*'04':'sad',*

*'05':'angry',*

*'06':'fearful',*

*'07':'disgust',*

*'08':'surprised'*

*}*

*focused\_emotion\_labels = ['happy', 'sad', 'angry']*

Note : Here we have mainly focused on 3 emotions (i.e., happy, sad and angry) as it is hard to predict all the emotions because one speech can sound in more than one emotion simultaneously.

Step 5 : The next step that we will do is extracting the features from the audio recordings by defining a function which will extract the features namely mfcc,mel and chroma, and store them horizontally using a numpy hstack method.

Step 6 : Next we are going to define a function to load our dataset. First, we are loading the data and then extracting the features using the function defined in the previous step. While features are extracted, we are assigning the features with the label's emotions.

Features as our input (x) and the labeled emotion as an output (y).

This is a well-known machine learning model, also known as Supervised Learning.

Step 7 : Then we are going to split our dataset using the *‘train\_test\_split()’* function. This way we can define how much dataset we need for the training purpose and how much dataset we need to require for the testing purpose.

Step 8 : Finally we will start calling the functions we defined earlier and recognize the emotions from speech audio recordings.

Step 9 : Then we will save the model by using the ‘joblib’ library.

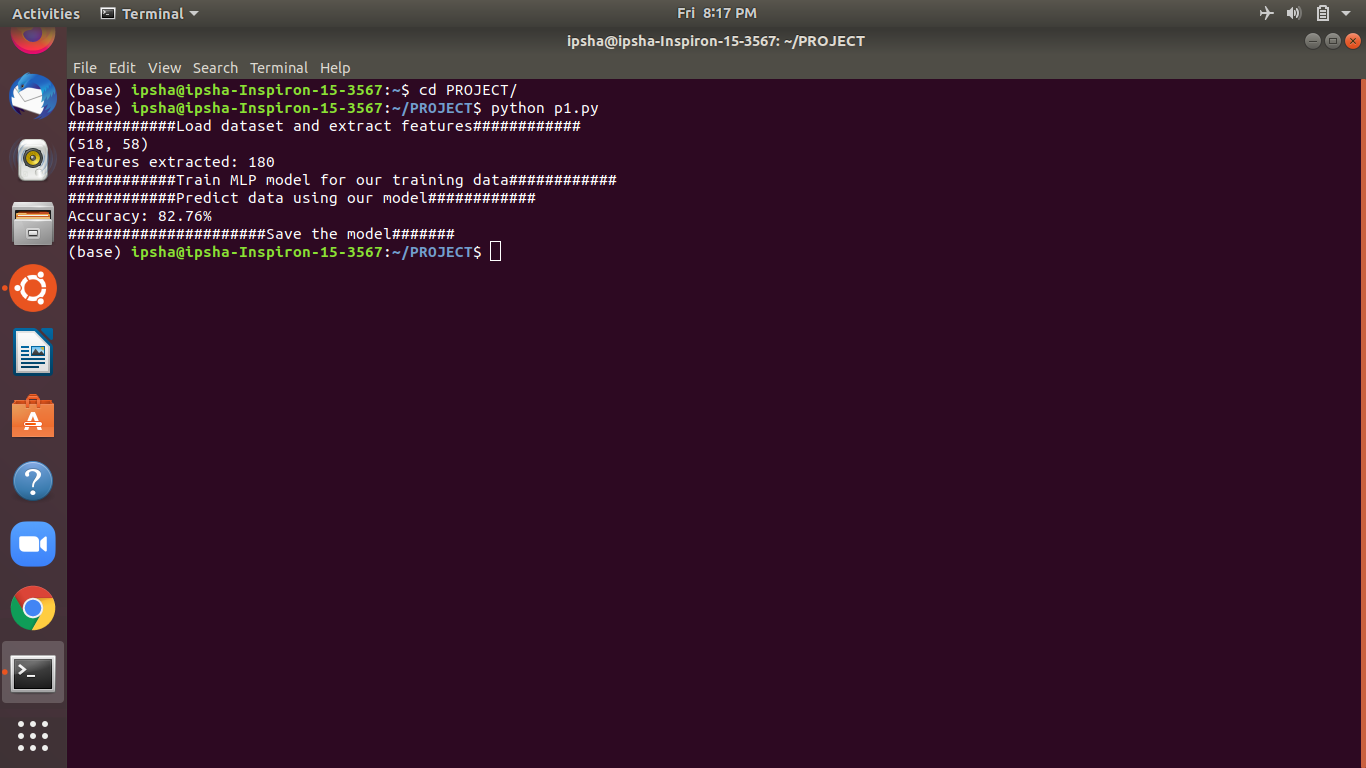
Step 10 : Now for the testing purpose, we will load the model that has been already saved.

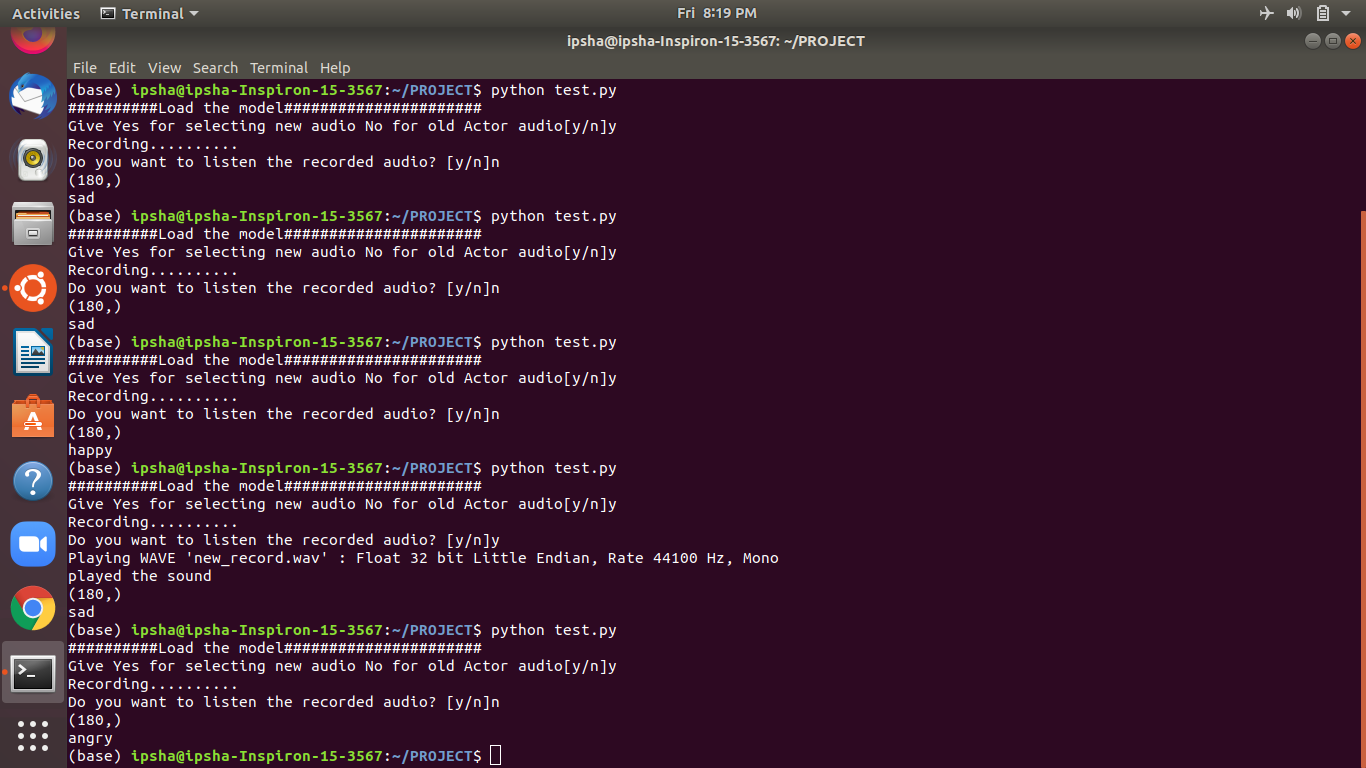
Step 11 : After all this, we will require the user to tell us if it wants to recognise the emotion of the recorded clips of the actors or if it wants to give its live audio for the recognition.

Step 12 : Accordingly our program will tell the emotion of the selected sound clip and then redirect to the movie recommendation page according to the emotion fetched.

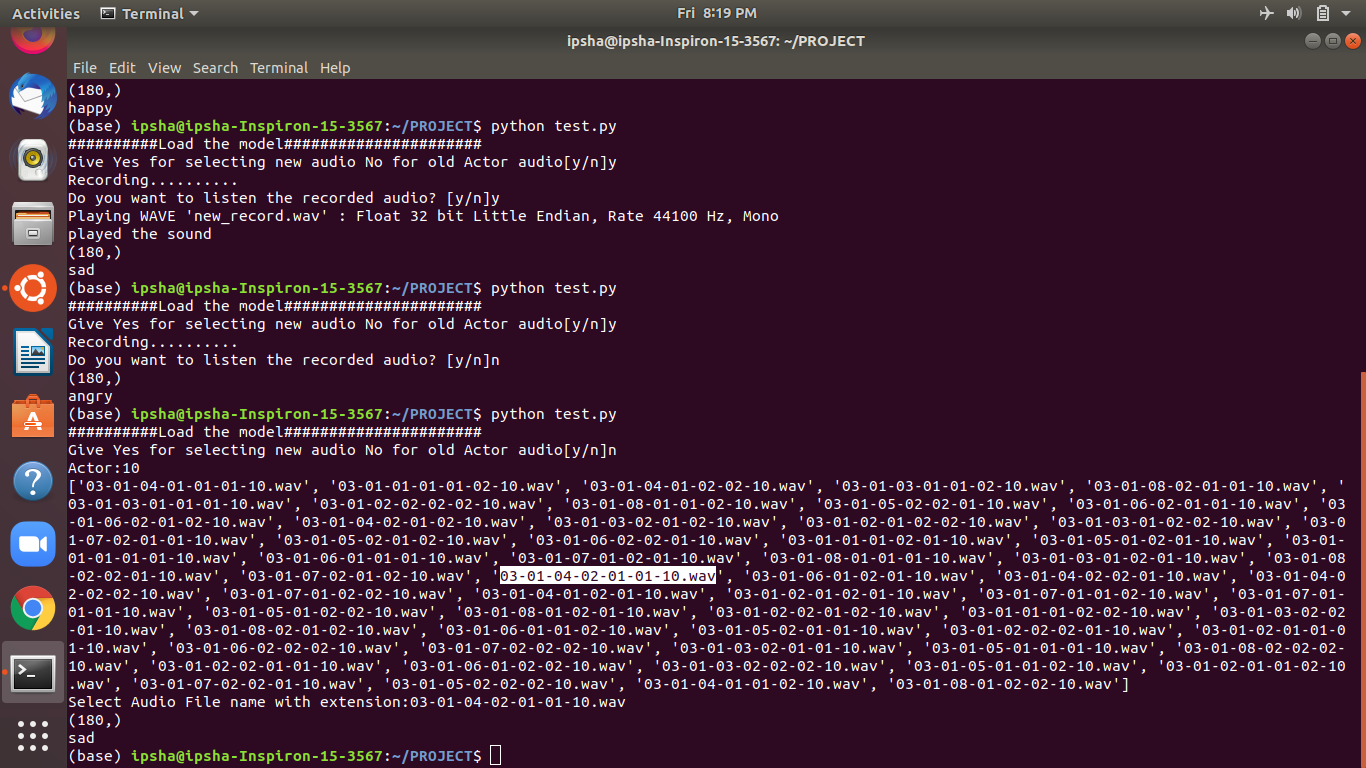
**Chapter 4 : Results**

Our project got an accuracy of 82.76%.

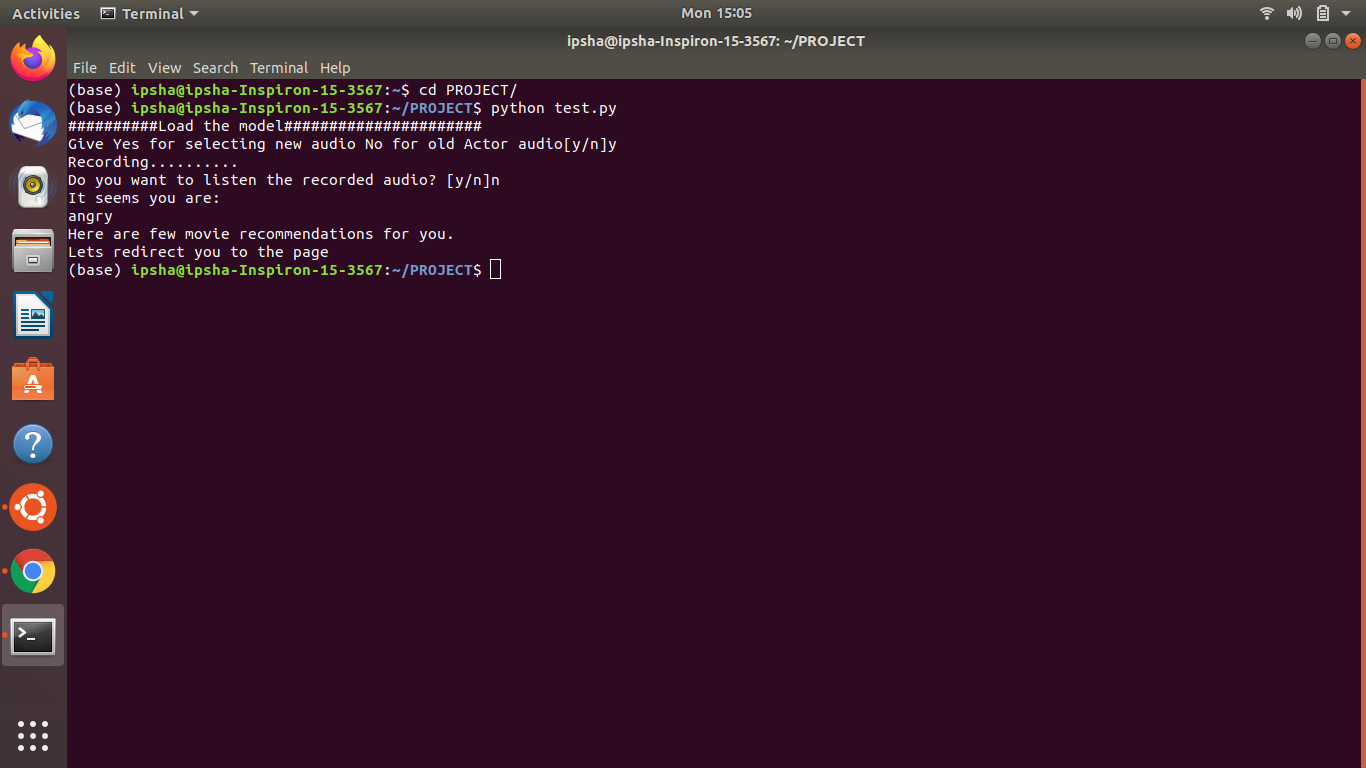


It shows the emotion of the live recorded audio.

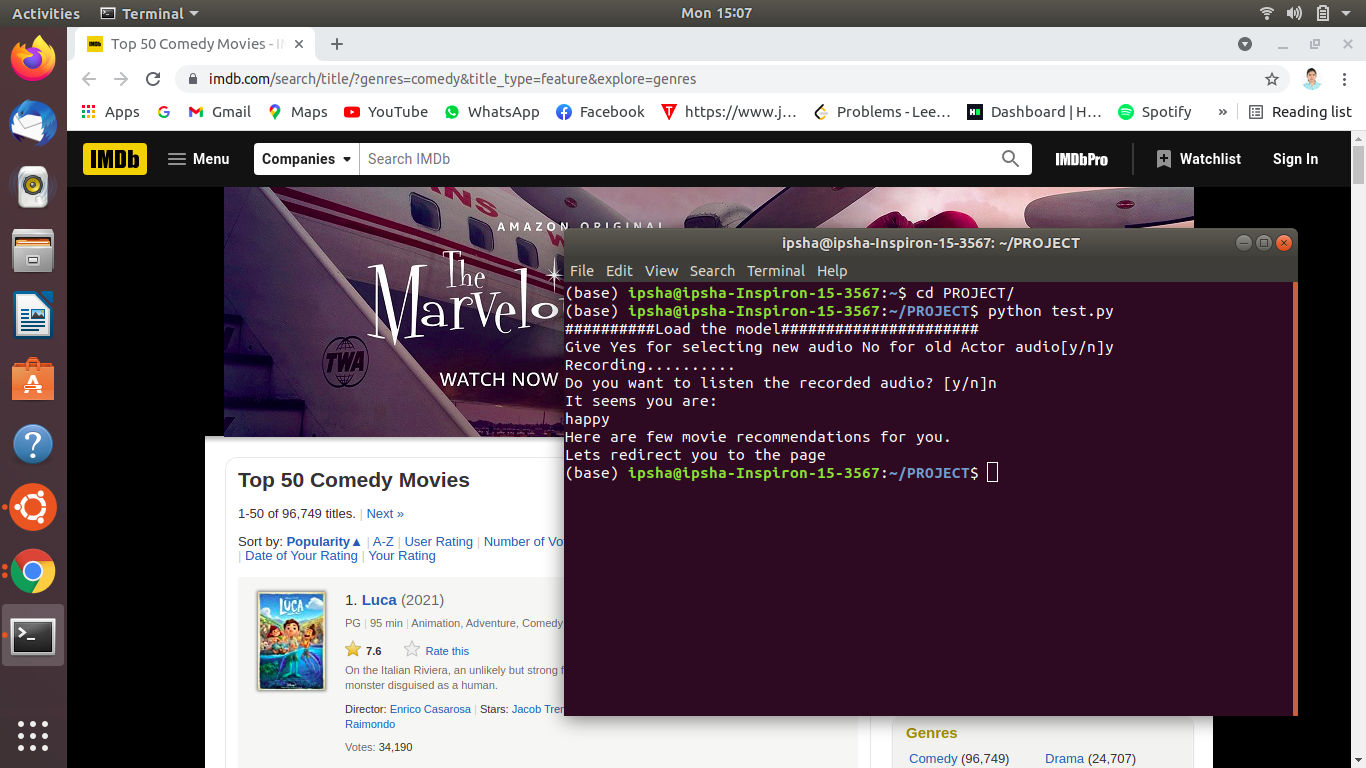
It is also showing the emotion of the recorded audios of the actors of the RAVDESS dataset which we can choose on our will.



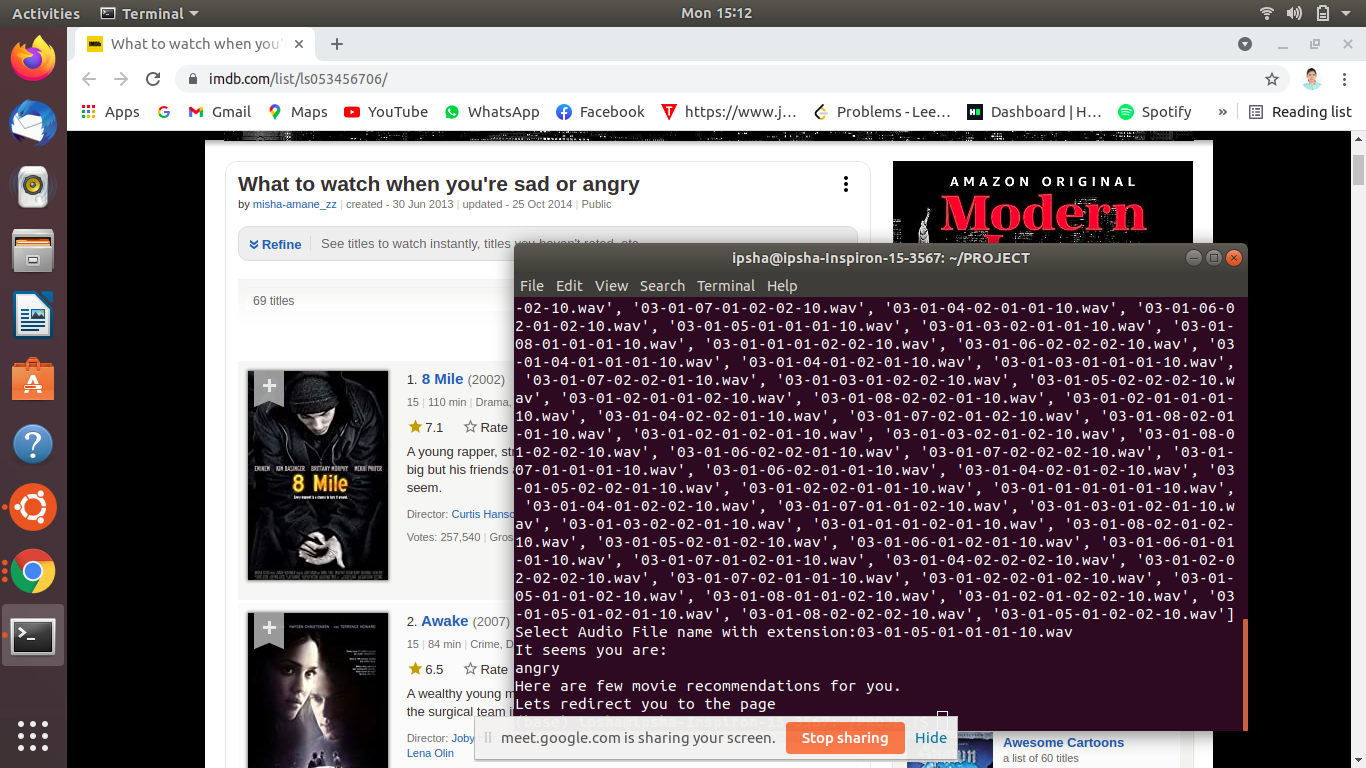
According to the emotion fetched from the recorded audio or the live audio the program is showing the movie recommendations by redirecting to the page of IMDB.



As we can see, when we got the emotion as ‘happy’, the program got redirected and started showing the movie recommendations of the comedy movies.



Also we can see, for the emotion, when fetched ‘angry’, the program is showing the movie recommendations accordingly.



**Chapter 5 : Conclusion**

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.

In this current study, we presented an automatic speech emotion recognition (SER) system using three machine learning algorithms (MLR, SVM, and RNN) to classify seven emotions. Here features are extracted from the RAVDESS database and presented. In fact, we study how classifiers and features impact recognition accuracy of emotions in speech. A subset of highly discriminant features is selected. Feature selection techniques show that more information is not always good in machine learning applications. The machine learning models were trained and evaluated to recognize emotional states from these features. RNN often performs better with more data and it suffers from the problem of very long training times. Therefore, we concluded that the SVM and MLR models have a good potential for practical usage for limited data in comparison with RNN .

Enhancement of the robustness of emotion recognition systems is still possible by combining databases and by fusion of classifiers. The effect of training multiple emotion detectors can be investigated by fusing these into a single detection system. We aim also to use other feature selection methods because the quality of the feature selection affects the emotion recognition rate: a good emotion feature selection method can select features reflecting emotion state quickly. The overall aim of our work is to develop a system that will be used in a pedagogical interaction in classrooms, in order to help the teacher to orchestrate his class. For achieving this goal, we aim to test the system proposed in this work.

**Chapter 6: Future Work**

To solve the problem of SER, we need to address the challenges mentioned earlier. Additionally, one of the significant hurdles to SER is the limited size of the datasets. To solve this problem, one option is to create a deep learning friendly database, meaning a vast number of samples. This is a viable but costly method. We also have the option of combining some of the datasets to create a superset. At the same time, this is possible; there could be problems because of different methods and techniques in creating different databases. As a suitable solution, we suggest exploring the creation of an entirely synthetic dataset using generative techniques trained by available datasets. GANs structures would be an excellent candidate for such a system, as they have been used already and proven successful for other applications. Another challenge that can be addressed in SER is the difference in emotion expressions in different languages. We believe using transformers, we can build a language-aware model that adapts to the language to classify emotions, and the same concept can be used for different accents in a language. Furthermore, as we discussed laboratory-generated data and noise in real-life situations, we can use generator models as has been explored in the methods reviewed to create noisy samples and try to design a noise-robust model for SER. Another point that we can improve the robustness of SER models is to create models that classify continuous speech emotions. For this reason, we can design architectures that are keeping a sliding window and measure the emotional content of the slide and decide based on that. Additionally, to improve the SER model’s robustness, a similar concept can be employed to learn and classify not only fully emoted emotions but also the transition states of the feelings, and based on emotion transition models, we can gain more confidence in recognized emotions.

**Chapter 7 : References**

**7.1. Dataset References :**

* RAVDESS: [https://zenodo.org/record/1188976#.XvbvZudS\_IU](https://zenodo.org/record/1188976" \l ".XvbvZudS_IU)

##### **7.2. Blogs and Documentations :**

* Librosa –[https://medium.com/@patrickbfuller/librosa-a-python-audio-libary-60014eeaccf](https://medium.com/@patrickbfuller/librosa-a-python-audio-libary-60014eeaccfb)b ; Patrick Fuller; 29th May 2019

##### **7.3. Kernel References:**

* <https://data-flair.training/blogs/python-mini-project-speech-emotion-recognition/> ;Data Flair