**Detection of Emotional State By Speech Recognition**



**Project Team**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Reg. No.** | **Student Name** |
| **1.** | **17ETCS002002** | **AADESH SHARMA** |
| **2.** | **17ETCS002041** | **AYUSHI SINGHANIA** |
| **3.** | **17ETCS002029** | **APRATIM RAJ** |
| **4.** | **17ETCS002067** | **DIVYANSH MISHRA** |
| **5.** | **17ETCS002117** | **PARAKH CHAUDHARY** |

**Supervisors:** **Prof.** **DEEPAK VARADAM**

**July – 2020**

**B. Tech. in Computer Science and Engineering**

**Faculty of ENGINEERING AND TECHNOLOGY**

## M. S. Ramaiah University of applied sciences

### Bengaluru -560 054

## Faculty of engineering and technology



Certificate

**<One per student>**

*This is to certify that the Project titled “Detection of Emotional State By Speech Recognition” is a bonafide work carried out in the Department of Computer Science and Engineering by* ***Mr.******AADESH SHARMA*** *bearing Reg. No.* ***17ETCS002002****,* ***Ms.******AYUSHI SINGHANIA*** *bearing Reg. No.* ***17ETCS002041, Mr. APRATIM RAJ*** *bearing Reg. No.* ***17ETCS002029****,* ***Mr. DIVYANSH MISHRA*** *bearing Reg. No.* ***17ETCS002067****,* ***Mr. PARAKH CHAUDHARY*** *bearing Reg. No.* ***17ETCS002117*** *in partial fulfilment of requirements for the award of B. Tech. Degree in Computer Science and Engineering of Ramaiah University of Applied Sciences.*

**July – 2020**

**Supervisor**

**Mr. DEEPAK VARADAM**

**Professor, Dept. of CSE**

**Dr. Raghavendra V. Kulkarni Dr. M. Arulanantham**

**Professor and Head – Dept. of CSE Professor and Dean-FET**

# Declaration

*Detection of Emotional States by Speech Recognition*

The project work is submitted in partial fulfilment of academic requirements for the award of B. Tech. Degree in the Department of Computer Science and Engineering of the Faculty of Engineering and Technology of Ramaiah University of Applied Sciences. The project report submitted herewith is a result of our own work and in conformance to the guidelines on plagiarism as laid out in the University Student Handbook. All sections of the text and results which have been obtained from other sources are fully referenced. We understand that cheating and plagiarism constitute a breach of University regulations, hence this project report has been passed through plagiarism check and the report has been submitted to the supervisor.

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl. No.** | **Reg. No.** | **Student Name** | **Signature** |
| **1.** | **17ETCS002002** | **AADESH SHARMA** |  |
| **2.** | **17ETCS002041** | **AYUSHI SINGHANIA** |  |
| **3.** | **17ETCS002029** | **APRATIM RAJ** |  |
| **4.** | **17ETCS002067** | **DIVYANSH MISHRA** |  |
| **5.** | **17ETCS002117** | **PARAKH CHAUDHARY** |  |

**Date : 15 July 2020**

# Acknowledgements

First and foremost, we offer our sincere gratitude to our project supervisor **Mr. Deepak Varadam,** Professor, Department of Computer Science Engineering, MSRUASfor his valuable suggestions, guidance, and encouragement which he has provided throughout the duration of this project. The experience, which we have gained by working under him, is an invaluable possession.

We express our sincere thanks to **Dr. H.M. Rajashekara Swamy**, Dean, Faculty of Engineering and Technology, MSRUAS for providing necessary support for this project. We are also very thankful to **Dr.P.V.R. Murthy**, Head of Department, Computer Science Engineering, MSRUAS for providing all the help and facilities to carry out the research work.

We are grateful to our seniors and batch mates, who made suggestions, pointed out errors, responded to our questions, and helped in many ways directly and indirectly.

Finally, we deeply acknowledge our own teammates for their sheer hard work and drive that drove this project to what it is.

# Summary

In this project, we aimed at extracting emotion from human speech samples. This area has not been explored. Hence, we explore and then provide an in-depth analysis of various approaches we can employ in extracting emotion from speech.

This study can find its use in emotion recognition from acoustic signals picked up by Intelligent Agents. Current systems only predict the emotional state of the person based only on different features in a speech. We explore the acoustic parameters of sound samples and predict what the emotional state a person is.

The concept of Emotion Recognition through Speech is an area of great interest for human to computer interaction. To recognise the emotional states of a person when he or she is talking (by using extraction techniques) which can be very useful for modern industries, technologies and services hence this prototype is designed and developed.

Various steps are involved in emotion recognition through speech. The first step is preprocessing where the audio signal is converted into digital signal. Then comes feature extraction and selection where different features of the human voice is extracted and selected for processing using techniques like MFCC (Mel-Frequency Cepstrum coefficient), MS, (Modulation Spectral Features) LR-RFE (Linear Regression Recursive Feature Elimination). Lastly we classify the features of the voice sample that was taken as an input using classification techniques such as RNN (Recurrent Neural network), SVM (Support Vector Machines).

Table of Contents

[Declaration i](#_Toc45720852)

[Acknowledgements ii](#_Toc45720853)

[Summary iii](#_Toc45720854)

[List of Tables vi](#_Toc45720855)

[List of Figures vii](#_Toc45720856)

[Nomenclature ix](#_Toc45720857)

[Abbreviation and Acronyms x](#_Toc45720858)

[1. Introduction 1](#_Toc45720859)

[1.1. Introduction: 1](#_Toc45720860)

[1.2. Motivation: 2](#_Toc45720861)

[1.3. Scope: 2](#_Toc45720862)

[1.4. Organization of the report 3](#_Toc45720863)

[2. Background Theory 5](#_Toc45720864)

[2.1. Logistic Regression using Python 5](#_Toc45720865)

[2.2. Speech Emotion Recognition with Convolutional Neural Network 5](#_Toc45720866)

[2.3. Analysis of CNN-Based Speech recognition system using raw speech as input 5](#_Toc45720867)

[2.4. Emotion Detection through Speech 5](#_Toc45720868)

[2.5. Convolutional Neural Network: 6](#_Toc45720869)

[2.6. Natural Language Processing: 7](#_Toc45720870)

[3. Aim and Objectives 9](#_Toc45720871)

[3.1. Title 9](#_Toc45720872)

[3.2. Aim 9](#_Toc45720873)

[3.3. Problem statement 9](#_Toc45720874)

[3.4. Objectives 9](#_Toc45720875)

[3.5. The Main Objectives are: - 10](#_Toc45720876)

[3.6. Methods and Methodology/Approach to attain each objective 11](#_Toc45720877)

[4. Problem Solving 13](#_Toc45720878)

[4.1. Design of the Speech Emotion Recognition Model: 13](#_Toc45720879)

[4.2. Major obstacles faces in developing the Speech Emotion Recognition Model: 15](#_Toc45720880)

[4.3. Data Collection for designing and implementing the Speech Emotion Recognition Model: 15](#_Toc45720881)

[5. Problem Solving & Software Implementation 17](#_Toc45720882)

[5.1. Functional and Non- Functional Requirements: 17](#_Toc45720883)

[5.2. Functional Diagram of the Software Model: 18](#_Toc45720884)

[5.3. Implementation of the Software Model: 20](#_Toc45720885)

[5.4. Testing and Analysis of the Software Model: 29](#_Toc45720886)

[6. Results 35](#_Toc45720887)

[7. Project Costing 40](#_Toc45720888)

[8. Conclusions and Suggestions for Future Work 41](#_Toc45720889)

[7.1. Conclusion 41](#_Toc45720890)

[7.2 Suggestions for future work: 42](#_Toc45720891)

[References 43](#_Toc45720892)

[Appendix 46](#_Toc45720893)

[Appendix-A 46](#_Toc45720894)

[Appendix-B 46](#_Toc45720895)

[Appendix-C 47](#_Toc45720896)

[Appendix-D 47](#_Toc45720897)

# List of Tables

[Table 1: Method and Methodologies Table. 12](#_Toc45721094)

[Table 2: Total Number of Data Taken to Train the Model. 16](#_Toc45721095)

[Table 3: List of Features Implemented.[7] 25](#_Toc45721096)

[Table 4: Test Case of Different Input and Output. 30](#_Toc45721097)

[Table 5: Result of the Output File with Correct Emotion Detected. 35](#_Toc45721098)

[Table 6: Overall Cost Table. 40](#_Toc45721099)

# List of Figures

[Figure i: Accuracy of CNN in Training vs Testing according to work by Derek Hung and Mitesh puthran. 6](#_Toc45720524)

[Figure ii: Layers of CNN. 7](#_Toc45720525)

[Figure iii: Raw Figure of design Approach 13](#_Toc45720526)

[Figure iv: Architecture of Speech Emotion Recognition. 14](#_Toc45720527)

[Figure v: State Diagram of the Speech Emotion Recognition Model. 19](#_Toc45720528)

[Figure vi: DFD Level 0 diagram for Speech Emotion Recognition Model. 19](#_Toc45720529)

[Figure vii: DFD Level-1 Diagram for Speech Emotion Recognition Model. 20](#_Toc45720530)

[Figure viii: Import the necessary python library for training the model. 20](#_Toc45720531)

[Figure ix: Code implementation for knowing the waveform of one of audio files. 21](#_Toc45720532)

[Figure x: Implementation of the waveform of the audio file(.wav). 21](#_Toc45720533)

[Figure xi: Code implementaion of showing the spectogram of the audio file. 21](#_Toc45720534)

[Figure xii: Getting the Spectrogram of the audio file. 22](#_Toc45720535)

[Figure xiii: Code Implementation to organizing the audio files and setting labels. 22](file:///F:\Cse%206th%20Sem\project\UG%20Group%20Project%20Report%20Writing%20Template.docx#_Toc45720536)

[Figure xiv: Coding implementation of Feature Extraction using librosa library. 24](#_Toc45720537)

[Figure xv: Training the datasets after extracting its features. 24](#_Toc45720538)

[Figure xvi: Coding Implementation of CNN Model. 26](#_Toc45720539)

[Figure xvii: Model Summary of the build CNN model. 26](#_Toc45720540)

[Figure xviii: Five Layer CNN Network. 27](#_Toc45720541)

[Figure xix: here we compile and fit the model to get the accuracy after training. 27](#_Toc45720542)

[Figure xx: Graph which shown the model accuracy. 28](#_Toc45720543)

[Figure xxi: Implementation after saving and loading the model in which the model achieved 72.73% accuracy. 28](#_Toc45720544)

[Figure xxii: User Registered Successfully. 30](#_Toc45720545)

[Figure xxiii: User Login Successfully. 31](#_Toc45720546)

[Figure xxiv: User Unable to Login 31](#_Toc45720547)

[Figure xxv: Recording and waveform of the output file. 32](#_Toc45720548)

[Figure xxvi: Neutral Emotion Detected Successfully. 32](#_Toc45720549)

[Figure xxvii: Recording and the waveform of the output file. 32](#_Toc45720550)

[Figure xxviii: Surprise Emotion Detected Successfully. 33](#_Toc45720551)

[Figure xxix: Recording and the Waveform of the output file. 33](#_Toc45720552)

[Figure xxx: Sad Emotion Detected Successfully. 33](#_Toc45720553)

[Figure xxxi: Recording and The waveform of the output file. 34](#_Toc45720554)

[Figure xxxii: Happy Emotion Detected Successfully. 34](#_Toc45720555)

[Figure xxxiii: Recording and the Waveform. 34](#_Toc45720556)

[Figure xxxiv: Failed to detect the anger emotion. 35](#_Toc45720557)

[Figure xxxv: Recording and waveform of the output.wav file. 36](#_Toc45720558)

[Figure xxxvi: Neutral Emotion Detected Successfully. 36](#_Toc45720559)

[Figure xxxvii: The output.wav file 36](#_Toc45720560)

[Figure xxxviii: Recording and The waveform of the output1.wav file. 37](#_Toc45720561)

[Figure xxxix: Happy Emotion Detected Successfully. 37](#_Toc45720562)

[Figure xl: The output1.wav file 37](#_Toc45720563)

[Figure xli: Recording and the waveform of output2.wav file 37](#_Toc45720564)

[Figure xlii: Sad or Worry emotion detected. 38](#_Toc45720565)

[Figure xliii: The Output2.wav file 38](#_Toc45720566)

[Figure xliv: Recording and the waveform of the output3.wav file. 38](#_Toc45720567)

[Figure xlv: Surprise Emotion Detected Successfully. 38](#_Toc45720568)

[Figure xlvi: The Output3.wav 38](#_Toc45720569)

[Figure xlvii: Recording and the Waveform of the output4.wav file. 39](#_Toc45720570)

[Figure xlviii: Sad Emotion Detected Successfully. 39](#_Toc45720571)

[Figure xlix: The Output4.wav file. 39](#_Toc45720572)

# Nomenclature

*FE* Feature Extraction

*FS* Feature Selection

*NN* Neural Network

*sr Speech Recognition*

*tf Tensorflow*

*pd Pandas*

# Abbreviation and Acronyms

**RAVDESS** The Ryerson Audio-Visual Database of Emotional Speech and Song

**SAVEE** Surrey Audio-Visual Expressed Emotion

**MFCC** Mel Frequency Cepstral Coefficient

**CNN** Convolutional Neural Network

**SVM** Support Vector Machines

**ASR** Automatic Speech Recognizer

**NLP** Natural Language Processing

# 1. Introduction

This Chapter attempts to understand the different aspects of the human voice and learn to gain deeper knowledge about the problem statement. This Chapter will help to explore the challenges involved in creating an app for emotion recognition through speech patterns and provide an appropriate solution to the problem.

## 1.1. Introduction:

The characteristics of human voice such as the pitch, loudness, and tone make human voice a versatile to communicate. It can be observed that humans can also express their emotions by varying the stated characteristics. This allows for identifying human emotion by analysing speech.

With different emotions and moods, not only does the tonal quality vary, but the associated speech patterns change too. For instance, people may tend to talk in loud voices when angry and use shrill or high-pitched voices when in a scared or panicked emotional state. Some people tend to ramble when they get excited or nervous. On the contrary, when in a pensive emotional state, people tend to speak slowly and make longer pauses, thereby indicating an increase in time spacing between consecutive words of their speech.

The characteristics of human voice such as the pitch, timbre, loudness, and tone make human voice a versatile to communicate. It can be observed that humans can also express their emotions by varying the stated characteristics. This allows for identifying human emotion by analysing speech. With different emotions and moods, not only does the tonal quality vary, but the associated speech patterns change too. For instance, people may tend to talk in loud voices when angry and use shrill or high-pitched voices when in a scared or panicked emotional state. Some people tend to ramble when they get excited or nervous. On the contrary, when in a pensive emotional state, people tend to speak slowly and make longer pauses, thereby indicating an increase in time spacing between consecutive words of their speech.

**Data Set Used**: We got audio datasets with around 2000 audio files which were in the wav format from the following websites:

* <http://neuron.arts.ryerson.ca/ravdess/?f=3>,
* <http://kahlan.eps.surrey.ac.uk/savee/Download.html>

The first website contains speech data which is available in three different format.

1. Audio Visual – Video with speech
2. Speech – Audio only
3. Visual – Video only

We went with the Audio only zip file because we are dealing with finding emotions from speech. The zip file consisted of around 1500 audio files which were in wav format. The second website contains around 500 audio speeches from four different actors with different emotions.

We used Librosa library in Python to process and extract features from the audio files. Librosa is a python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems. Using the librosa library we were able to extract features i.e MFCC(Mel Frequency Cepstral Coefficient). MFCCs are a feature widely used in automatic speech and speaker recognition.

## 1.2. Motivation:

Detecting emotions is one of the most important marketing strategies in today’s world. You could personalize different things for an individual specifically to suit their emotional state. For this reason, we decided to do a project where we could detect a person’s emotions just by their voice which will let us manage many AI related applications.

## 1.3. Scope:

The emotion recognition system through speech system is a project which is beneficial for business and can also be combined with existing as well as upcoming technologies. The system would help business and technologies to learn and adjust to various human emotions and produce appropriate results. Following is such scenarios where speech characteristics can serve as a tool for identifying human emotion:

* + 1. Playing music and changing the ambient room’s lighting as per the tone of the conversation.
    2. Smart cars can slow down when a driver’s voice shows sign of angry or fear.
    3. Customer service centres can gather insights on their customer satisfaction by simply analysing the speech of their customers. Also, the scores received as a part of this analysis can be used to assess the overall opinion of company/product/services.

As a result this type of application has much potential in the world that would benefit companies and also even safety to consumers.

## 1.4. Organization of the report

* **Chapter 1(Introduction):-** The Emotion Recognition through Speech is a system that takes in input in the form of an audio signal and recognizes the emotional state of the person. Different scope for the emotion recognition system is also discussed.
* **Chapter 2(Background theory):** The Literature survey on the existing approaches of emotion recognition through speech was gone through. The advantages and disadvantages of these methods were compared and discussed and then at last the results are documented.
* **Chapter 3(Aim and objectives):** The project aims to design and develop the emotion recognition through speech System. The main objectives of the project is defined under certain steps which are followed which will help in the completing the project, starting from literature survey to listing the functional and non-functional requirements, from high and low level design, coding, testing and validation. Thus, the methods and methodologies help to accomplish the objectives of the project along with the resource used.
* **Chapter 4(Problem solving):** The problem-solving approach is defined by the literature survey for the existing systems. Major obstacles for the development of an emotion recognition system are discussed
* **Chapter 5(Software Implementation):** The functional as well non-functional requirements are also listed down along with high level and low-level designs. Based on the design the system is implemented using the desired software and platforms. The implemented system is tested under various conditions and results are tabulated.
* **Chapter 6(Results):** In this part the results of the planned system were discussed along with the data and its analysis. The data is analyzed then the expected results are displayed in the form of tables and figures. The results are tested, validated and verified and then alternative methods are also suggested so as to overcome the failure conditions.
* **Chapter 7(Project costing):** In this section the cost essential for making the project is explained which requires the working hours and the amount that each team member requires for the development of the website.

**Chapter 8(Conclusions):** The report is a well-defined document which involves a detailed description of the “Emotion Recognition through Speech System”. The report would explain the purpose and features of the system, the interfaces of the system, what and how the system would work, the restraints under which it must operate.

# 2. Background Theory

First, let us consider some works done in the past by other authors or groups that have provided motivation and ideas for our work and to preface our project and better understand the theory.

2.1. Logistic Regression using Pythonby Michael Galarnyk

* Utilizing logistic regression, via Python’s scikit-learn library, this work deals with predicting numerical digits from hand drawn images.
* Logistic Regression is a classifying technique used to model the probability of events.
* In this work, various hand drawn images are run through. The program recognizes and predicts the numeric digit that is comprised in that hand drawn image.

2.2. Speech Emotion Recognition with Convolutional Neural Networkby Reza Chu

* Author used Convolutional Neural network, which has also been applied in our work.

2.3. Analysis of CNN-Based Speech recognition system using raw speech as inputby Dimitri Palaz, Matthew Magimai and Ronan Collobert

* This system presents and ASR(Automatic Speech Recognizer)
* It aims at modelling the relationship that acoustic speech signal has with raw speech and use that to form a solid recognition system.

2.4. Emotion Detection through Speech by Derek Hung and Mitesh puthran

* Puts CNN to the test in detecting emotions in speech samples.
* Finds that CNN is more accurate than rest available models.

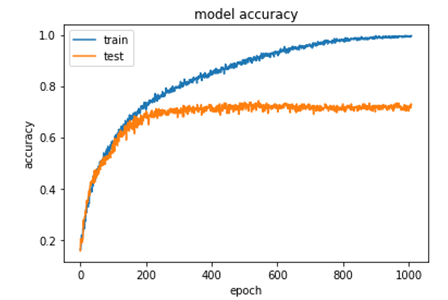


Figure i: Accuracy of CNN in Training vs Testing according to work by Derek Hung and Mitesh puthran.

Now, let’s understand some of the principles/theories that are heavily used in our project.

## 2.5. Convolutional Neural Network:

* Firstly, a neural network is a series of algorithm that aims to recognize relationships in a set of data through a process that mimics the way that the human brain operates. This means that, as if neurons interact inside the brain to send signals, a neural network aims to replicate a connection of that level and hence be excellently efficient at recognizing relationships and patterns.
* Therefore, a Convolutional neural network is a subclass of an NN, it has a minimum of one convolutional layer. It’s a deep learning algorithm that is capable of taking in an input media, assign importance in various aspects of that media and then be able to differentiate that media from other.
* There are 5 levels to a CNN:

1. Input layer: This layer contains the data
2. Convo layer: Features of the data are extracted here
3. Pooling layer: Used to reduce volume after convolution
4. Fully connected layer: Involves biases, weights and neurons.
5. Softmax/Logistic layer: Used for final classification

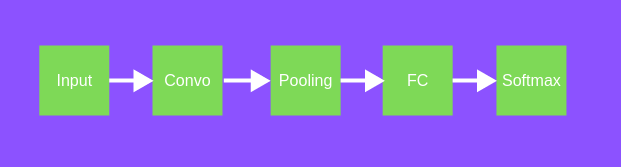


Figure ii: Layers of CNN.

## 2.6. Natural Language Processing:

* Natural language processing is the automatic manipulation of a natural language.
* Natural language is an umbrella term for all human communication with each other. It may be through languages, text etc.
* Processing a language is a simple task done daily by everyone. Text itself is a form of language processing.
* In terms of Machine learning, NLP is a field where the ability for the computer to understand, analyse, manipulate and possible generate a human speech is dealt with.
* We have put NLP to use in order to recognize the pattern with which the emotions in a human happen in context to speech. The changes in pitch, frequency etc is dealt with in this field.

Merits and Demerits:

* Emotion detection using speech recognition deals with quite a sensitive topic. Emotion detection has become one of the biggest marketing strategies, in which mood of the consumer plays an important role.
* Products’ demand can be increased by targeting them to users knowing their current emotions.
* Merits of using CNN: Quoting work by Derek Hung, Mitesh Puthran **“Emotion detection using speech”**

**“After building numerous different models, we have found our best CNN model for our emotion classification problem. We achieved a validation accuracy of 70% with our existing model”**

CNN has hence been proven to perform better than other models available.

* Our project model could perform better perhaps if more data was available to work on. Due to the ongoing COVID-19 crisis, gathering original raw data was rendered impossible
* By applying the model into hardware, the project would’ve been more hands-on, physical and practical. But due to the COVID crisis finding the right hardware and implementing the software into it was again rendered impossible.

# 3. Aim and Objectives

## 3.1. Title

* Detection of Emotional States by Speech Recognition

## 3.2. Aim

* Classify the given states of a person by using Speech Recognition

## 3.3. Problem statement

Currently, there are no major emotional state detecting system that can be deemed accurate. Emotion is a sensitive topic; they vary for everyone and are a crucial part of living beings.

Our aim is to create a deep learning model in order to analyse the emotion of a person through speech. This will be applicable in not only various ecommerce and IT sectors, but also in our day-to-day lives.

Now, lets talk about why this project idea was chosen.

Emotional state detection is one of the biggest marketing strategies in which the current mood of a customer plays a very vital role. Hence, to detect the mood of the person in the appropriate moment and suggest them a product to help them accordingly will increase the demand of the product or company. This is mainly what made us choose this topic.

## 3.4. Objectives

* Emotions appear whenever a perception of important changes occurs in the surroundings or in the human body. An emotion is a psychological state or a process that is designed to maintain a balance between the information processed by the brain and the most important goals that the brain must achieve. Therefore, the exploitation of this area is difficult, and the development of applications in this area is based mainly on well-known and extensively tested things.
* The most known example of emotions detection application is based on speech recognition. The study of emotions and their change has some major advantages in developing emotional intelligence by developing capabilities to help us better understand automatic reactions of the human body. Unlike gestures or language, the human speech is a universal language that reflects fluctuations of the mood in a person. Understanding how a person reacts helps us improve our social relationships, but also recognize and interpret their emotions better.
* Emotional speech recognition is an area of great interest for human computer interaction. The system must be able to recognize the user’s emotion and perform the actions specified accordingly.

## 3.5. The Main Objectives are: -

* To detect and analyse the emotional states of a person by using pitch and frequency features in speech recognition.
* Essential to have a framework that includes various modules performing actions like speech to text conversion, feature extraction, feature selection and classification of those features to identify the emotions.
* The pitch and Formants are first extracted from the speech signal and then their analysis is carried out to recognize different emotional states as Neutral, Happy, Sad etc.
* In taking Voice samples from user to detect the emotional state of the person
* Using pitch and frequency features in speech recognition.
* To analyze the features extracted from the voice samples.
* To determine the feature of the voice sample and select the state as per input.
* To classify the features of voice samples that taken as an input
* To store the voice samples in the database for efficient recognition output

## 3.6. Methods and Methodology/Approach to attain each objective

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Objective No.** | **Statement of the Objective** | **Method/ Methodology** | | **Resources Utilised** |
| 1 | *In taking Voice samples from user to detect the emotional state of the person.* | **Preprocessing:** -Converting the acoustic sound pressure wave into a digital signal which is suitable for voice processing. A microphone can be used to convert the acoustic wave into an analog signal. | Micro-Phone. | |
| 2 | *Using pitch and frequency features in speech recognition.* | **Feature extraction: -** The speech signal contains a large number of parameters that reflect the emotional characteristics. | | Python Librosa Library. |
| 3 | *To analyze the features extracted from the voice samples.* | **MFCC Features (Mel-frequency cepstral coefficient): -** These are the best for speech recognition as it takes human perception sensitivity with respect to frequencies into consideration. | | Python Librosa Library, Tensorflow & Keras library. |
| 4 | *To determine the feature of the voice sample and select the state as per input.* | **Feature selection: -** Feature selection (FS) aims to choose a subset of the relevant features from the original ones according to certain relevance evaluation criterion, which usually leads to higher recognition accuracy.  ***Techniques used:***  **LR-RFE (Linear Regression-Recursive feature elimination)** | | Build Five Layer CNN Model for train the datasets. |
| 5 | *To store the voice samples in the database for efficient recognition output.* | **Database: -** The performance and robustness of the recognition systems will be easily affected if it is not well trained with a suitable database. | | Excel Sheet in the .csv format and another file of model.json format. And for storing the user input we use output.wav file. |
| 6 | *To predict the outcome of a model.* | **Predictions: -** After tuning the model, tested it out by predicting the emotions for the test data. | | Jupyter Notebook in Python. |

Table 1: Method and Methodologies Table.

# 4. Problem Solving

In this chapter we have approach towards problem solving in which we are going to evaluate the Design, Raw-Implementation & Data Collection.

## 4.1. Design of the Speech Emotion Recognition Model:

The approach of the speech emotion recognition model are as follows:

* First to take the input from user and stored it in the form of (.wav) file.
* Then import the datasets that are used (RAVDESS) the Ryerson Audio-Visual Database of Emotional Speech and Song and the (SAVEE) dataset.
* There-after, Extract the acoustic features over the full sample from dataset. To extract the features, pyAudio Analysis is used.
* After feature extraction, this data will be feed into training models Support Vector Machines (SVM), and Convolutional Neural Network (CNN) of 5 layers.
* To further analyse the dataset, the dataset will be splitted into male speech only samples and female speech only samples.
* Then, From the stored data we determine different types of emotion such as, happy, angry, sad, neutral, calm etc.

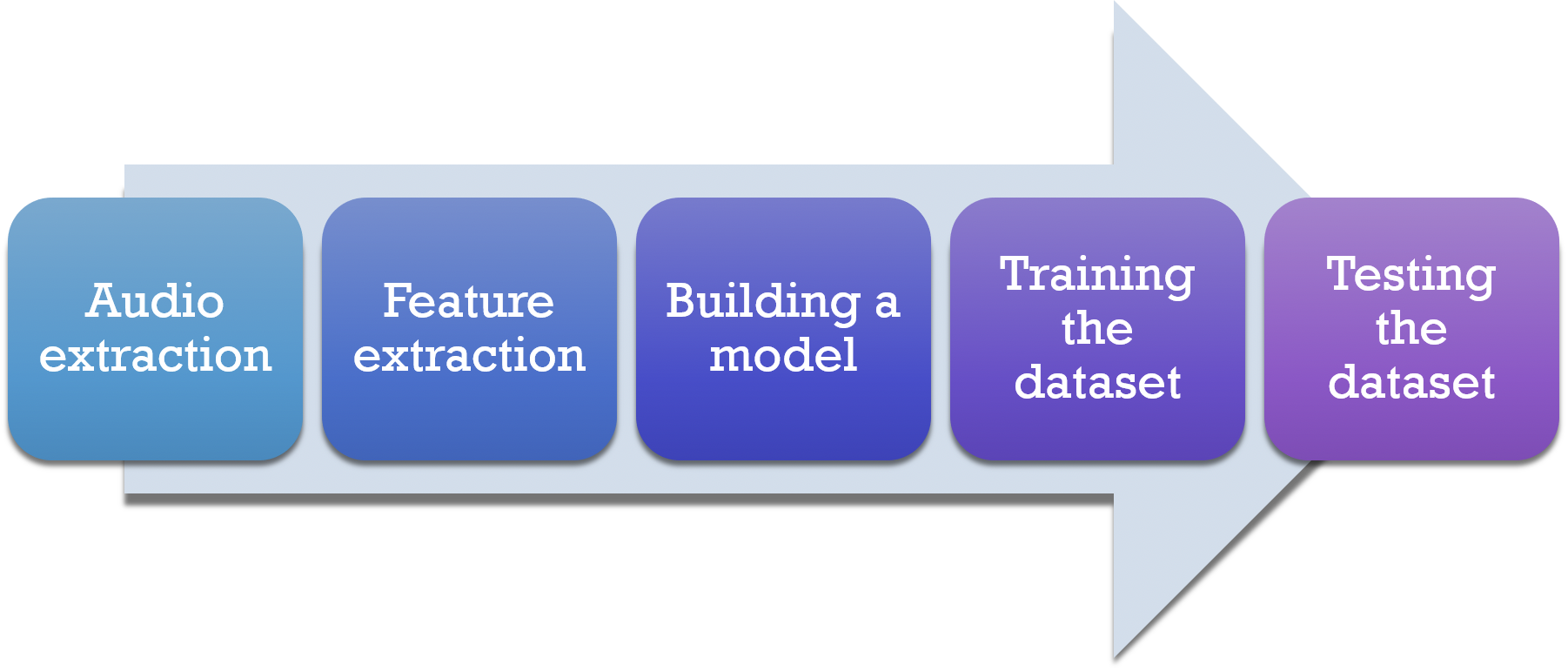


Figure iii: Raw Figure of design Approach

The Design we implement in our code are follow:

* Collected labelled data.
* Created a CNN model with 5 layers.
* Trained the model using 80% of available data
* Tested the model using 20% of data with a max accuracy of 60% - 70%.
* Setting parameters: like sampling rate, window size (choosing one audio from the file)
* Setting the labels of feelings.
* Getting the features of audio files using Librosa.
* Dividing the data into test and train.
* Building the model.
* Save model and evaluate it on test data.
* Test on real world samples.
* Detect the Emotion.

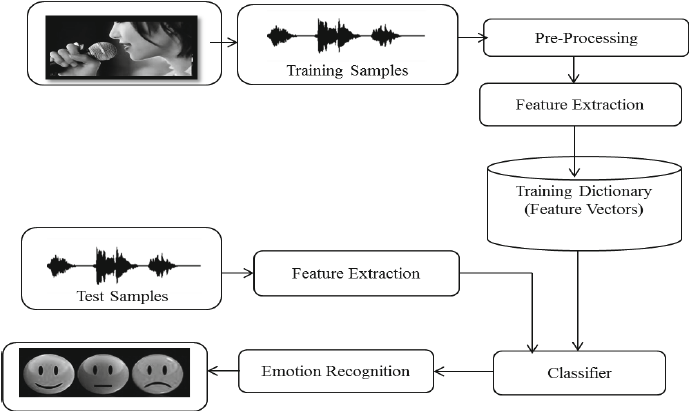
[](https://www.researchgate.net/figure/Architecture-of-Speech-Emotion-Recognition-System_fig1_270899631)

Figure iv: Architecture of Speech Emotion Recognition.

## 4.2. Major obstacles faces in developing the Speech Emotion Recognition Model:

* Emotions are subjective, people would interpret it differently. It is hard to define the notion of emotions.
* Annotating an audio recording is challenging. Should we label a single word, sentence or a whole conversation? How many emotions should we define to recognize.
* Collecting data is complex. There are lots of audio data can be achieved from films or news. However, both are biased since news reporting has to be neutral and actors’ emotions are imitated. It is hard to look for neutral audio recording without any bias.
* Labelling data require high human and time cost. Unlike drawing a bounding box on an image, it requires trained personnel to listen to the whole audio recording, analysis it and give an annotation. The annotation result must be evaluated by multiple individuals due to its subjectivity.

## 4.3. Data Collection for designing and implementing the Speech Emotion Recognition Model:

The data-sets we used to build the train models and for testing are as follow:

1. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). [4]
2. Surrey Audio-Visual Expressed Emotion (SAVEE) Dataset. [5]

* The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is used as the database. It is a database comprising of voice samples of 24 actors (12 male, 12 female). The speech of these are in various emotions and all the actors speak in North American English accent. The dataset also contains 7,356 high-quality video recordings which correspond to the audio dataset. The dataset voice samples for eight emotional expressions viz., neutral, calm, happy, sad, angry, fearful, surprise, and disgust. The data set for song samples for six emotional expressions viz., neutral, calm, happy, sad, angry, and fearful. All emotion expressions, except neutral, are expressed at two levels of emotional intensity normal and strong. The database has been validated in by 297 participants.
* SAVEE (Surrey Audio-Visual Expressed Emotion) is an emotion recognition dataset. It consists of recordings from 4 male actors in 7 different emotions, 480 British English utterances in total. The sentences were chosen from the standard TIMIT corpus and phonetically-balanced for each emotion. This release contains only the audio stream from the original audio-visual recording. The data is split so that the training set consists of 2 speakers, and both the validation and test set consists of samples from 1 speaker, respectively.[6]
* Note: The datasets which we used have speech voice upto 3 – 5 seconds because we use large amount of files so the data-sets must have smaller memory size.
* Google Drive Link of Dataset: <https://drive.google.com/drive/folders/1mlZKNEPD5iATFSX5uIaqCOfTG-IZnsOo?usp=sharing>.

|  |  |  |  |
| --- | --- | --- | --- |
| Emotion | Speech Sample Count | Song Sample Count | Summed Count |
| Neutral | 60 | 80 | 140 |
| Calm | 60 | 160 | 220 |
| Happy | 60 | 160 | 220 |
| Sad | 60 | 160 | 220 |
| Angry | 60 | 160 | 220 |
| Fearful | 40 | 160 | 220 |
| Disgust | 40 | 100 | 140 |
| Surprised | 40 | 100 | 140 |
| Total | 420 | 1080 | 1500 |

Table 2: Total Number of Data Taken to Train the Model.

# 5. Problem Solving & Software Implementation

In this chapter we have approach towards problem solving & Software Implementation in which we are going to evaluate the Requirements, Diagram, Implementation with explanation and Test Cases.

## 5.1. Functional and Non- Functional Requirements:

Functional requirements and non – functional requirements are as follow:

**Functional Requirements:**

**FR1:** The application should allow the user to log in or sign up if he/she is a new user. The application should allow the user to log in with his/her email id or phone number along with a password.

**FR2:** The application should allow the user to reset the password if he/she forgets it, by sending an OTP to the phone number.

**FR3:** The application should ask for the details of the user such as name, age, gender, profession, hobbies and the description of the user’s mood like what is going on his/her mind.

**FR4:** The application should allow user to proceed after filling all the details.

**FR5:** the application should ask for the voice node. The user can select option of uploading the voice node or recording it.

**FR6:** The application should ask the user to wait for a minute or two.

**FR7:** The application should display the emotional state of the user.

**FR8:** The application should display the graph of the user’s voice node which shows the pitch and frequency.

**FR9:** The application should ask the user to save and if he/she wants to share it in WhatsApp or message.

**FR10:** The application should create a link if user selects to share and then send the link with the selected method.

**FR11:** The application should ask the user to select if he wants to test another voice node or want to exit.

**FR12:** The application should move back to asking the voice node if the user selects to test another voice node.

**FR13:** The application should exit i.e. close application if the user selects to exit.

**Non-Functional Requirements:**

**NFR1:** The application should have a safe and secure database.

**NFR2:** The application should be efficient in terms of performance and response time.

**NFR3:** The application should be simple and user friendly.

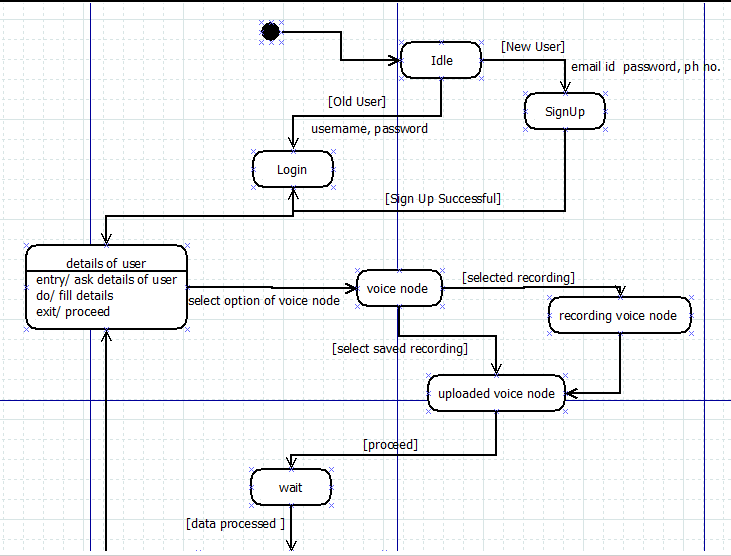
**NFR4:** The application should have a modern and appealing UI.

**NFR5:** The application should be free from bugs and technical errors.

## 5.2. Functional Diagram of the Software Model:

The State Diagram and DFD diagram of the software are given below:

**State Diagram:**



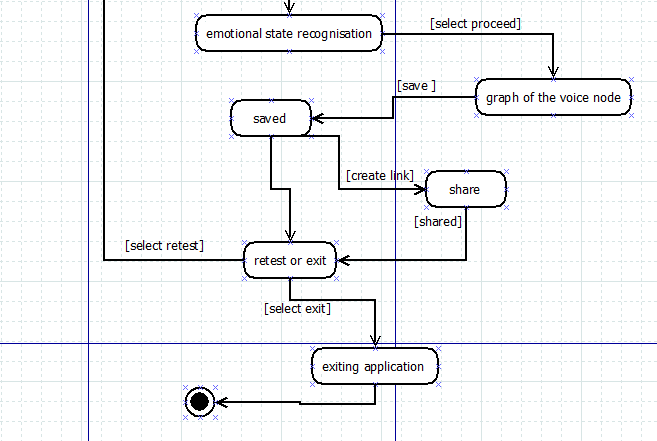


Figure v: State Diagram of the Speech Emotion Recognition Model.

**DFD Diagram Level 0 & 1:**

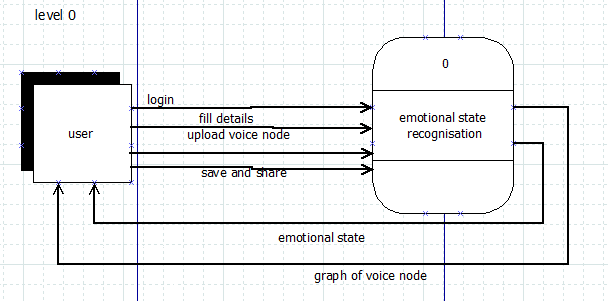
****

Figure vi: DFD Level 0 diagram for Speech Emotion Recognition Model.

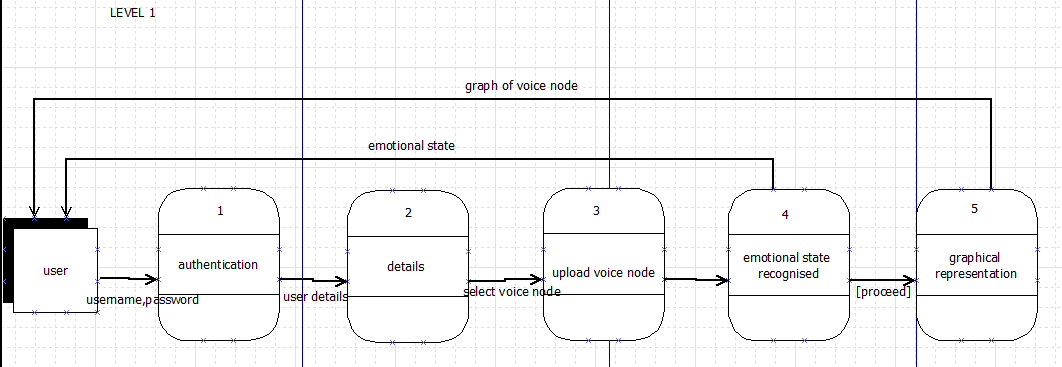


Figure vii: DFD Level-1 Diagram for Speech Emotion Recognition Model.

## 5.3. Implementation of the Software Model:

First, we import all the necessary python library which we going to use for training the model and then we call the raw data file which are using for feature extraction:

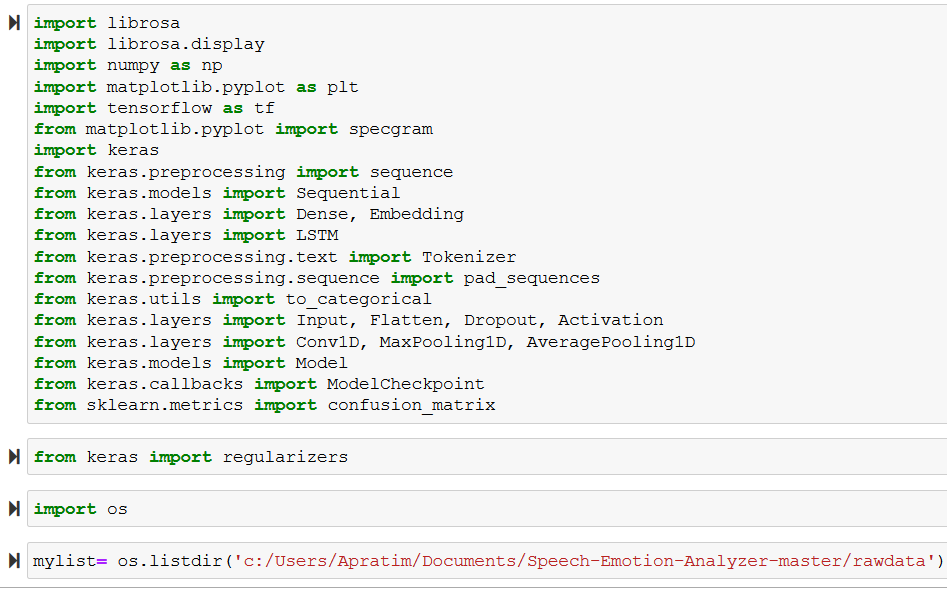


Figure viii: Import the necessary python library for training the model.

Now, we Tested one of the audio files from dataset by plotting out the waveform and a spectrogram for knowing the features.

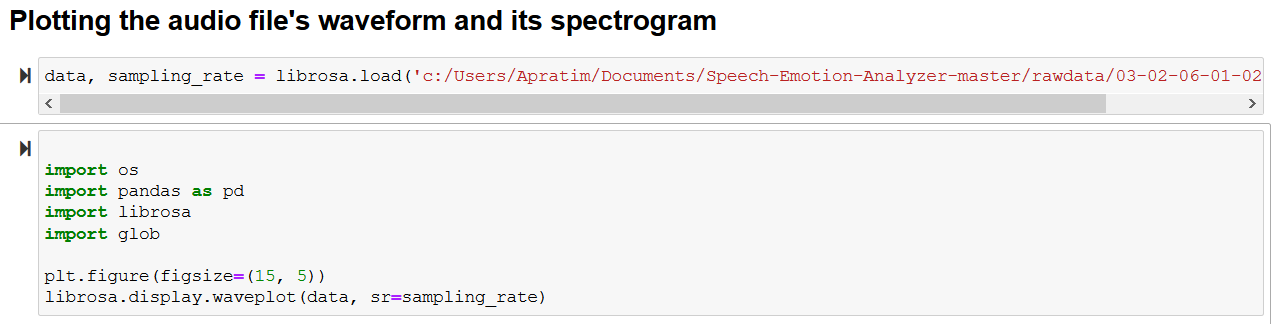


Figure ix: Code implementation for knowing the waveform of one of audio files.

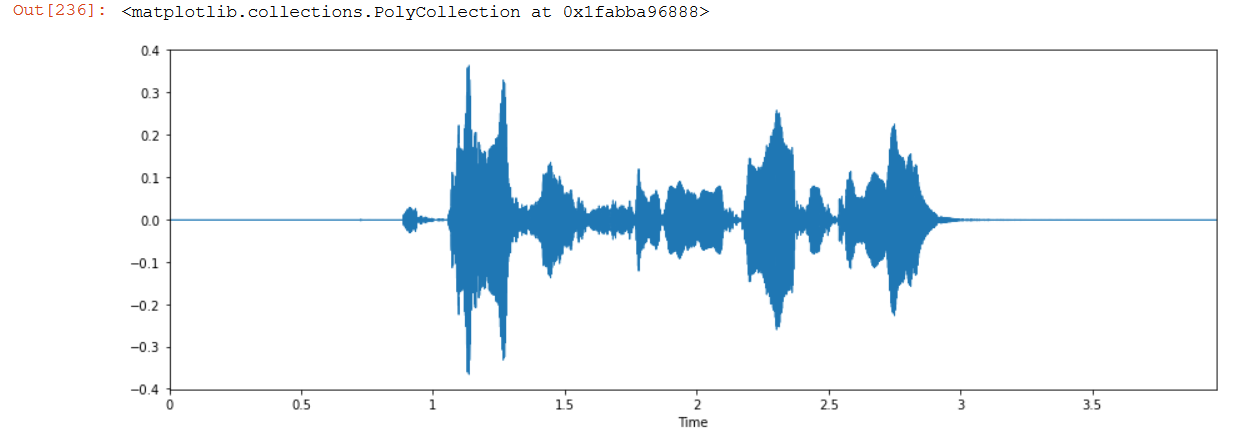
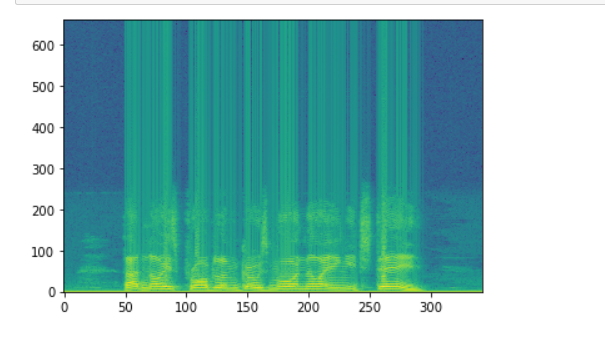


Figure x: Implementation of the waveform of the audio file(.wav).



Figure xi: Code implementaion of showing the spectogram of the audio file.

Figure xii: Getting the Spectrogram of the audio file.

Now, The next step involves organizing the audio files. Each audio file has a unique identifier at the 6th position of the file name which can be used to determine the emotion the audio file consists. We have 5 different emotions in our dataset.

Figure xiii: Code Implementation to organizing the audio files and setting labels.

1. Calm
2. Happy
3. Sad
4. Angry
5. Fearful
6. Neutral
7. Surprise

Now, We used Librosa library in Python to process and extract features from the audio files. Librosa is a python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems. Using the librosa library we were able to extract features

i.e MFCC(Mel Frequency Cepstral Coefficient). MFCCs are a feature widely used in automatic speech and speaker recognition. We also separated out the females and males voice by the using the identifiers provided in the website. This was because as experiment we found out that separating male and female voices increased by 15%. It could be because of the pitch of the voice was affecting the results.

Each audio file gave us many features which were basically array of many values. These features were then appended by the labels which we created in the previous step.

The next step involved dealing with the missing features for some audio files which were shorter in length. We increased the sampling rate by twice to get the unique features of each emotional speech. We didn’t increase the sampling frequency even more since it might collect noise thus affecting the results.

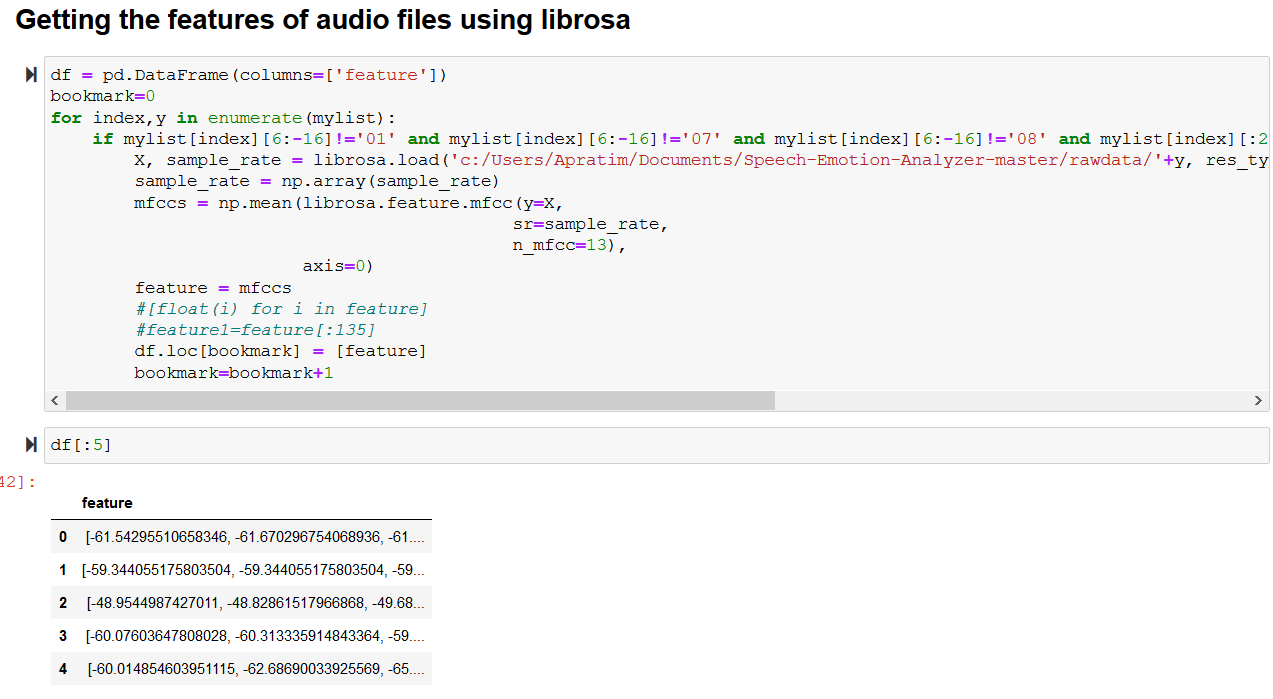


Figure xiv: Coding implementation of Feature Extraction using librosa library.

Note:

* While extracting the features, all the audio files have been timed for 3 seconds to get equal number of features.
* The sampling rate of each file is doubled keeping sampling frequency constant to get more features which will help classify the audio file when the size of dataset is small.

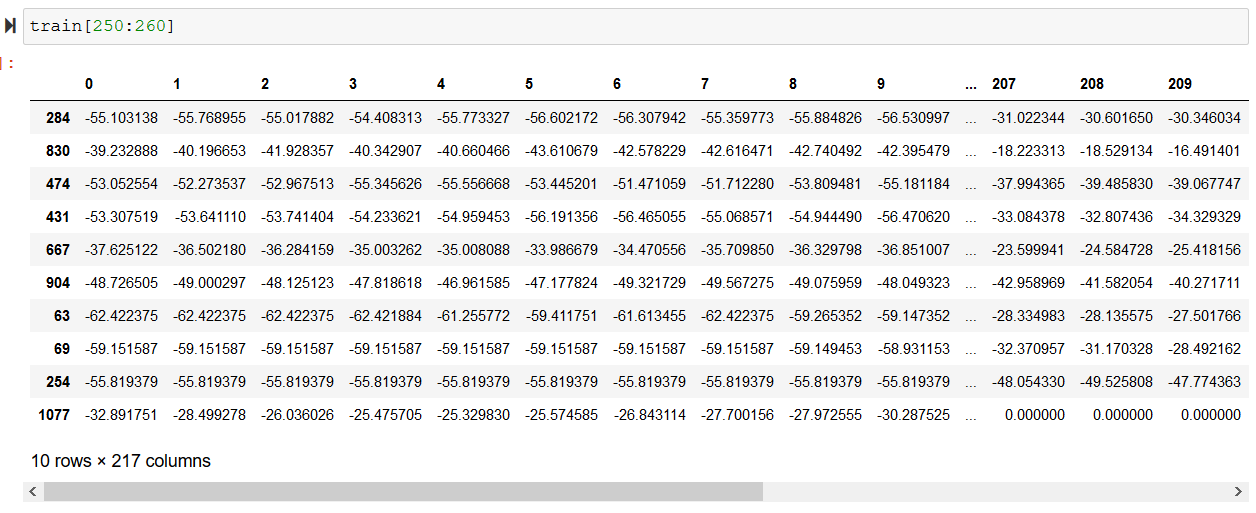


Figure xv: Training the datasets after extracting its features.

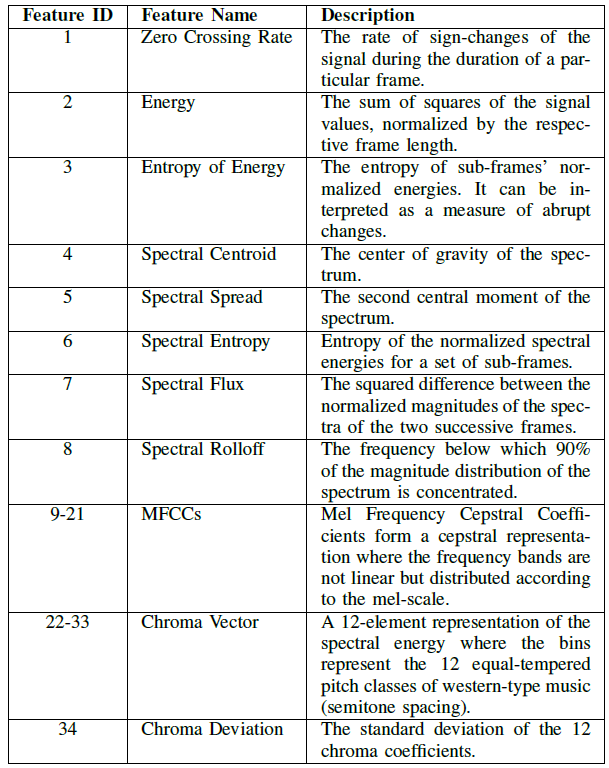


Table 3: List of Features Implemented.[7]

The next steps involve shuffling the data, splitting into train and test and then building a model to train our data.

Since the project is a classification problem, Convolution Neural Network seems the obivious choice. We also built Multilayer perceptrons and Long Short Term Memory models but they under-performed with very low accuracies which couldn't pass the test while predicting the right emotions.

We built a LSTM model and CNN models. The MLP and LSTM were not suitable as it gave us low accuracy. As our project is a classification problem where were categorize the different emotions, CNN worked best for us.

So, The Five Layer CNN model which we used to get the accuracy of 70% are as follow:

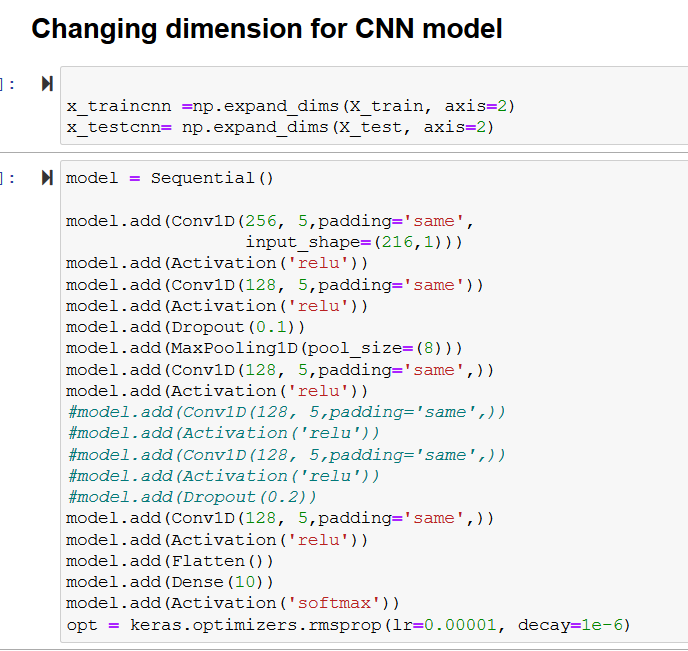


Figure xvi: Coding Implementation of CNN Model.

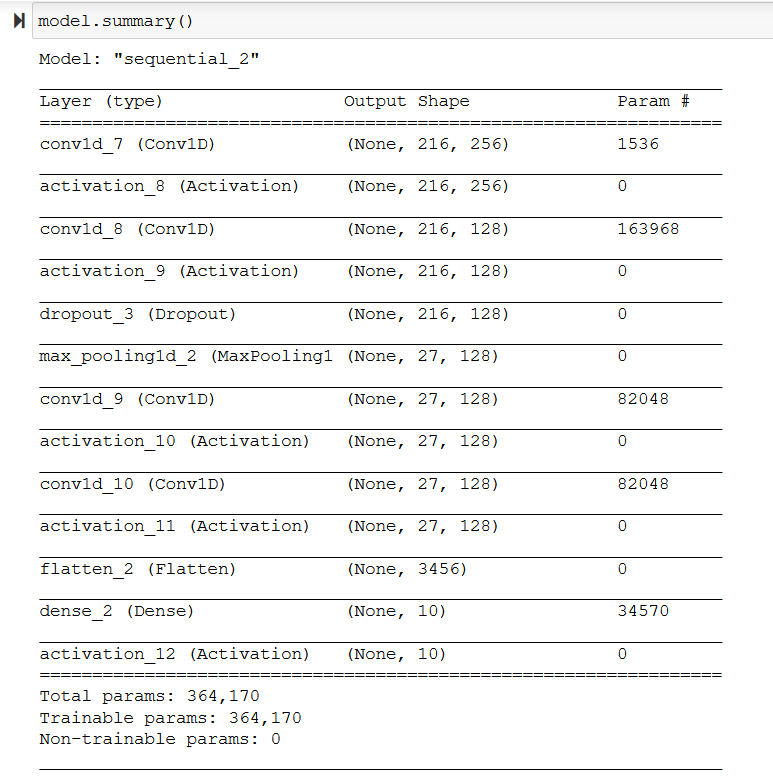


Figure xvii: Model Summary of the build CNN model.

Now, The Raw figure of Five layer CNN are as follow:

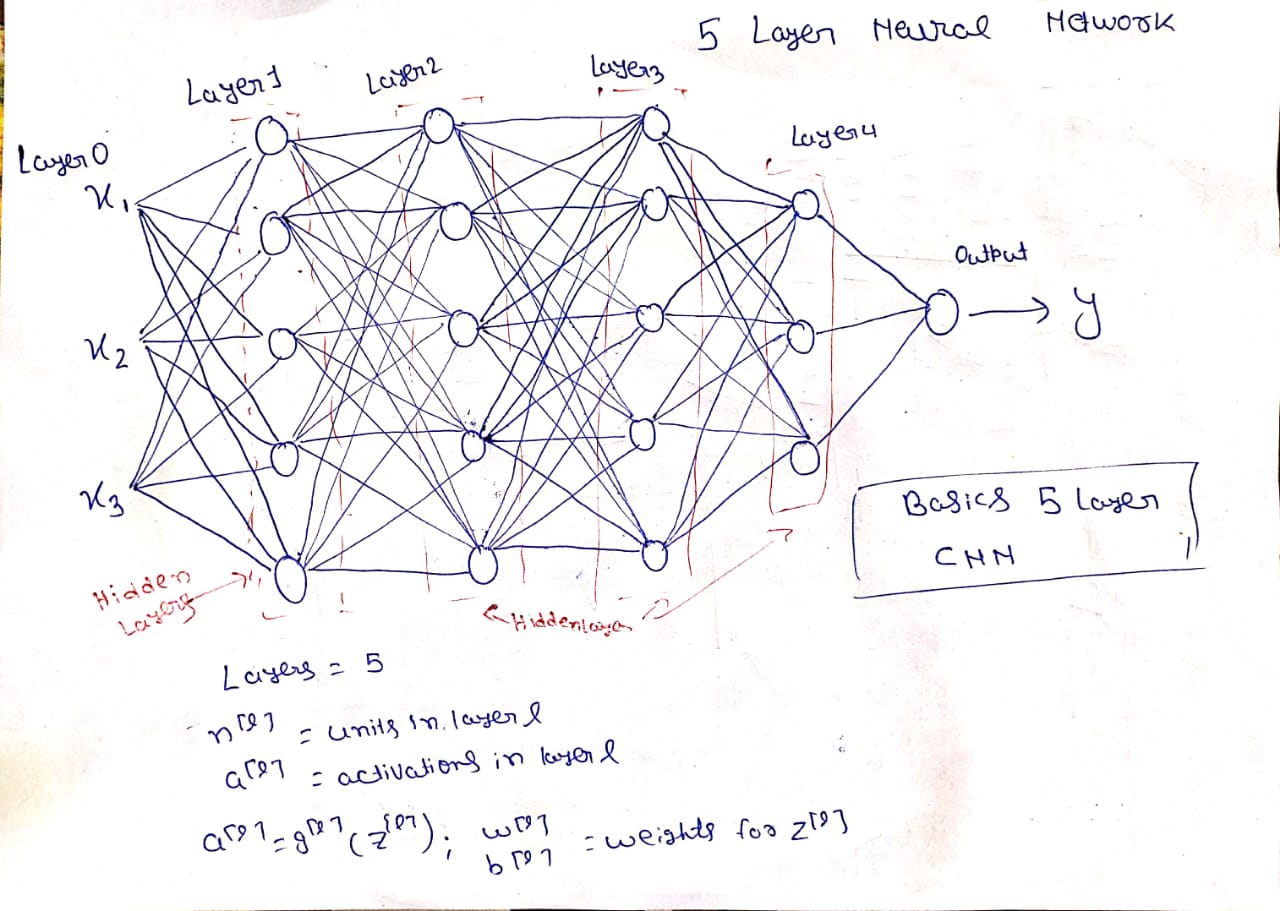


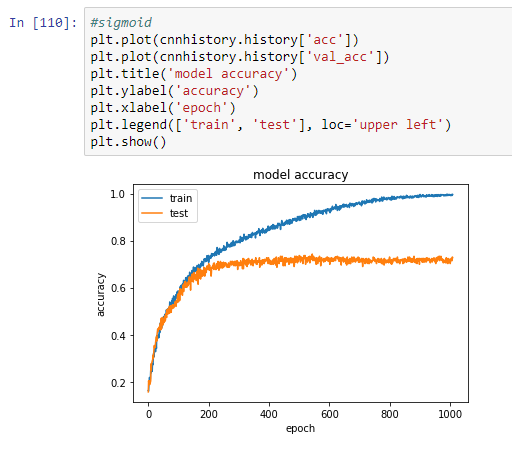
Figure xviii: Five Layer CNN Network.

Now, after the 5 layer CNN then we have to train the model with batch\_size=16 and 700 epochs without any learning rate schedule, etc.



Figure xix: here we compile and fit the model to get the accuracy after training.

Now, it get trained upto 700 epochs with batch\_size=16 and then we have to load the train data into model.json file and then run it to get the accuracy.

Figure xx: Graph which shown the model accuracy.

After training the model we had to save the model and load the model to get the accuracy then we implement the testing part.



Figure xxi: Implementation after saving and loading the model in which the model achieved 72.73% accuracy.

After building numerous different models, we have found our best CNN model for our emotion classification problem. We achieved a validation accuracy of 72.23% with our existing model. Our model could perform better if we have more data to work on. What’s more surprised is that the model performed excellent when distinguishing between a males and females voice. We can also see above how the model predicted against the actual values.

## 5.4. Testing and Analysis of the Software Model:

Now, after train the model and store the dataset in model.json file then we must do the live testing of the speech emotion detection model. The test case table are as follow with explanation.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Test Case ID | Test Case Description | INPUT | Expected Output | Obtained Output | Pass/Fail |
| TC 01 | Register into the application | Username and password | Registered  Successfully | Figure: xxiii | PASS |
| TC 02 | Login into the application. | Username & Password | Welcome | Figure: xxiv | PASS |
| TC 03 | Login into the application | Username & password | Welcome | Incorrect username or password  Figure: xxv | Fail |
| TC 04 | Live prediction of voice using microphone. | Hello I am fine. | Detect the same input speech and neutral emotion. | Figure: xxvi & xxvii | PASS |
| TC 05 | Live prediction of voice using microphone | Oh! That officer | Detect the same speech and anger emotion. | Figure: xxviii & xxix | PASS |
| TC 06 | Live prediction of voice using microphone. | Why are you so sad. | Detect the same input speech and sad emotion. | Figure: xxx & xxxi | Pass |
| TC 07 | Live prediction of voice using microphone. | I am Very Delighted today. | Detect the same input speech and sad emotion. | Figure: xxxii & xxxiii | Pass |
| TC 08 | Live prediction of voice using microphone. | This Coffee Sucks. | Detect the same input speech and sad emotion. | Figure: xxxiv & xxxv | Fail |

Table 4: Test Case of Different Input and Output.

**ANALYSIS OF TEST CASES:**

**TC01:**

Register into the application:

User should register into the application by entering username and password and also confirm the password.

**Source code for this Test case is in: Appendix-A**

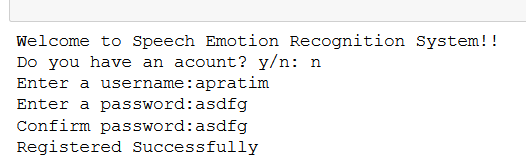


Figure xxii: User Registered Successfully.

**TC 02:**

Login into the application

User should login into the application by entering username and password if he/ she is an old user and has already registered before.

**Source Code for this Test Case is in: Appendix-A**

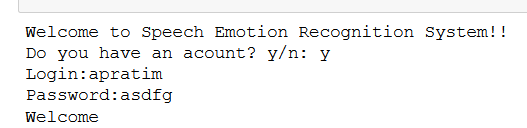


Figure xxiii: User Login Successfully.

**TC 03:**

Login into the application

User should login into the application by entering username and password if he/ she is an old user and has already registered before.

**Source Code for this Test Case is in: Appendix-A**

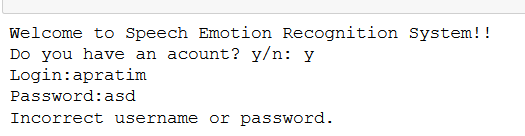


Figure xxiv: User Unable to Login

**TC 04:**

Live prediction of voice using microphone.

User should ask to record the voice of his/her and then after recording it print the speech, Waveform of the output file and detect the emotion.

**Source Code for this Test Case is in: Appendix (B, C & D)**

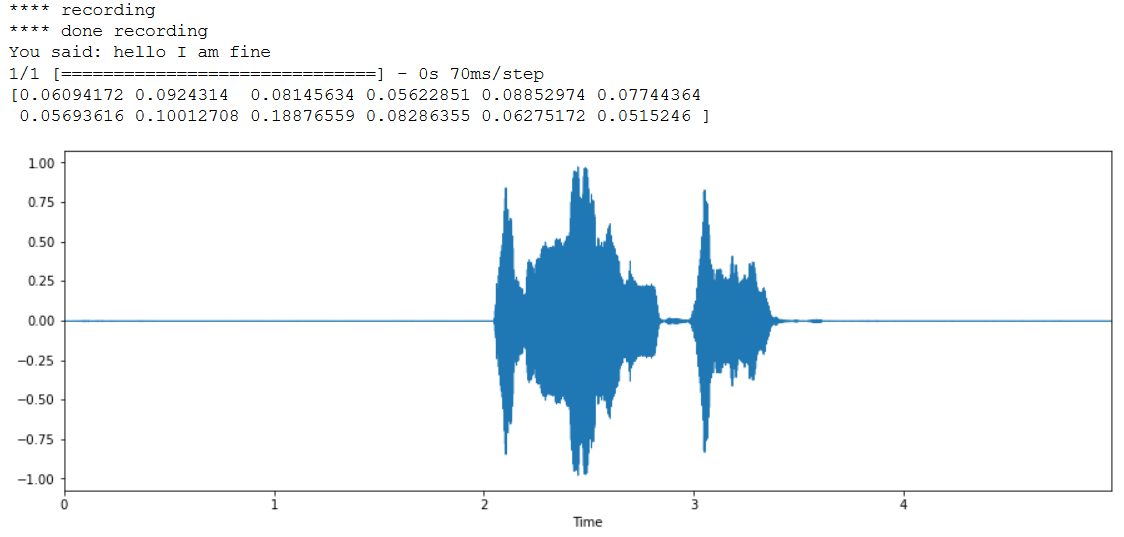


Figure xxv: Recording and waveform of the output file.



Figure xxvi: Neutral Emotion Detected Successfully.

**TC 05:**

Live prediction of voice using microphone.

User should ask to record the voice of his/her and then after recording it print the speech, Waveform of the output file and detect the emotion.

**Source Code for this Test Case is in: Appendix (B, C & D)**

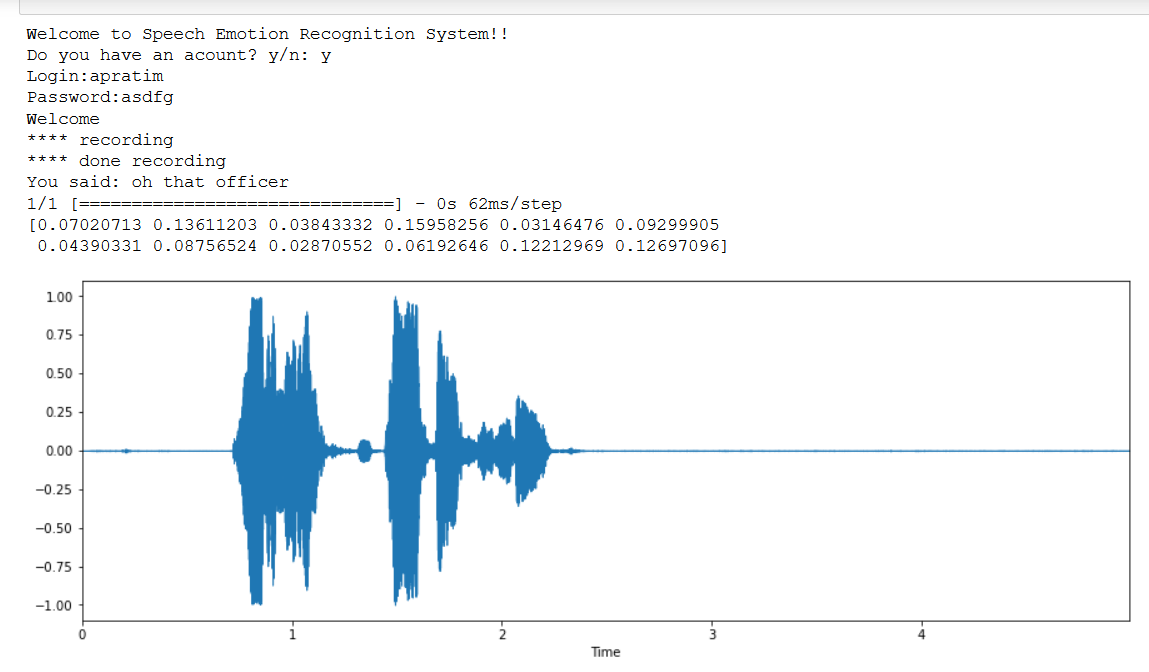


Figure xxvii: Recording and the waveform of the output file.



Figure xxviii: Surprise Emotion Detected Successfully.

**TC 06:**

Live prediction of voice using microphone.

User should ask to record the voice of his/her and then after recording it print the speech, Waveform of the output file and detect the emotion.

**Source Code for this Test Case is in: Appendix (B, C & D)**

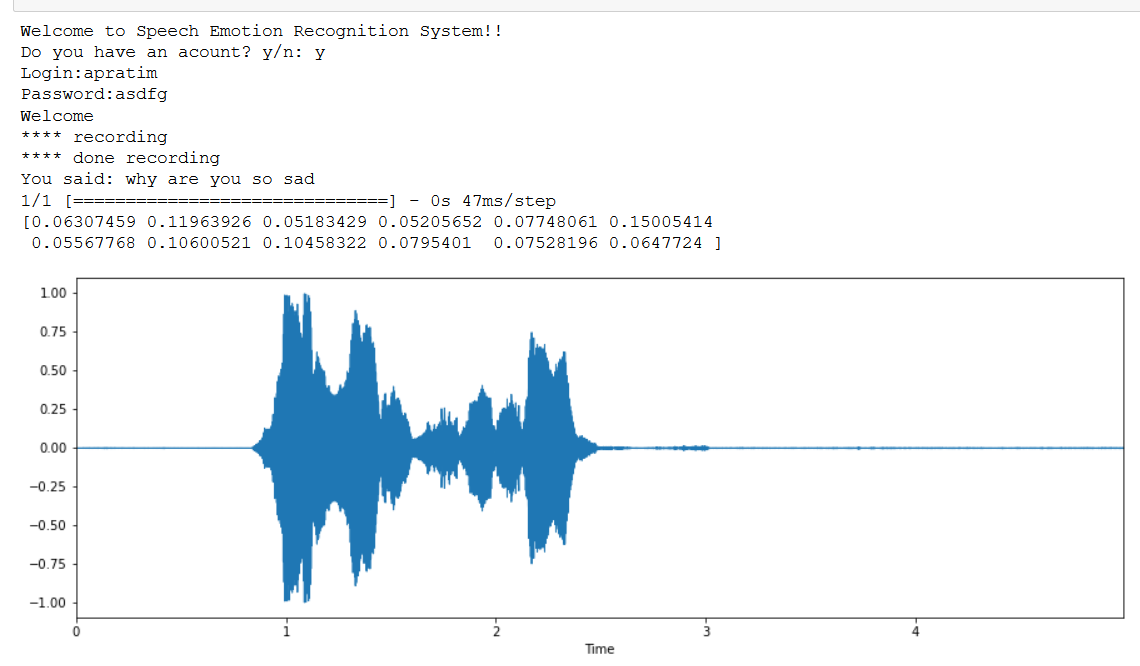


Figure xxix: Recording and the Waveform of the output file.



Figure xxx: Sad Emotion Detected Successfully.

**TC 07:**

Live prediction of voice using microphone.

User should ask to record the voice of his/her and then after recording it print the speech, Waveform of the output file and detect the emotion.

**Source Code for this Test Case is in: Appendix (B, C & D)**

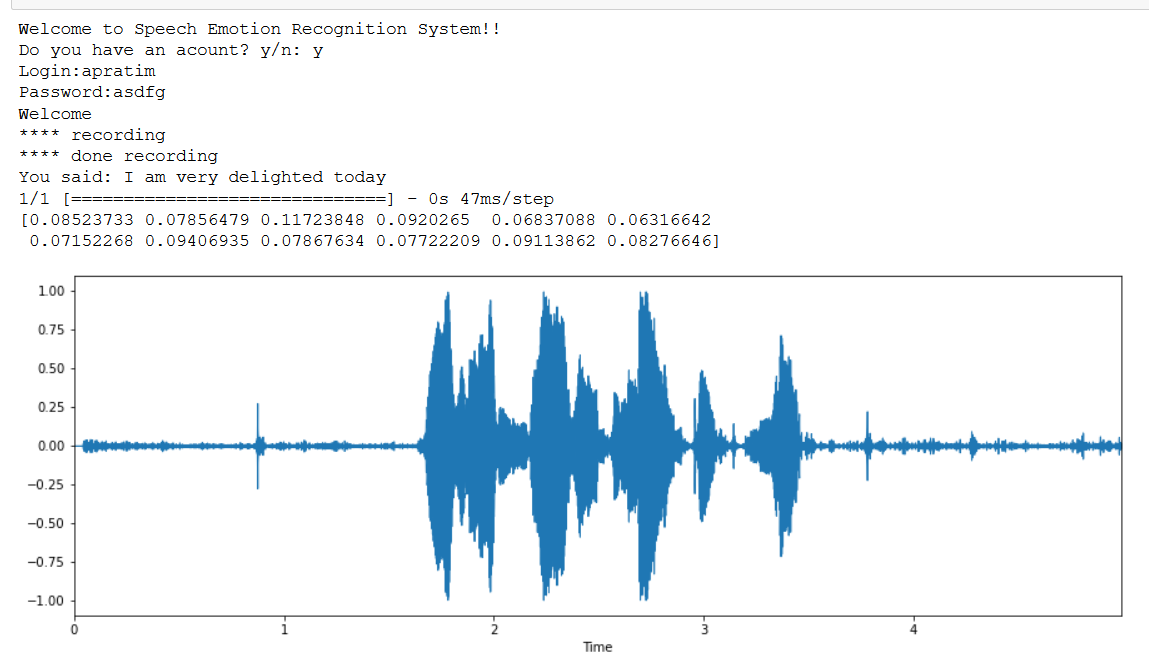


Figure xxxi: Recording and The waveform of the output file.



Figure xxxii: Happy Emotion Detected Successfully.

**TC 08:**

Live prediction of voice using microphone.

User should ask to record the voice of his/her and then after recording it print the speech, Waveform of the output file and detect the emotion.

**Source Code for this Test Case is in: Appendix (B, C & D)**

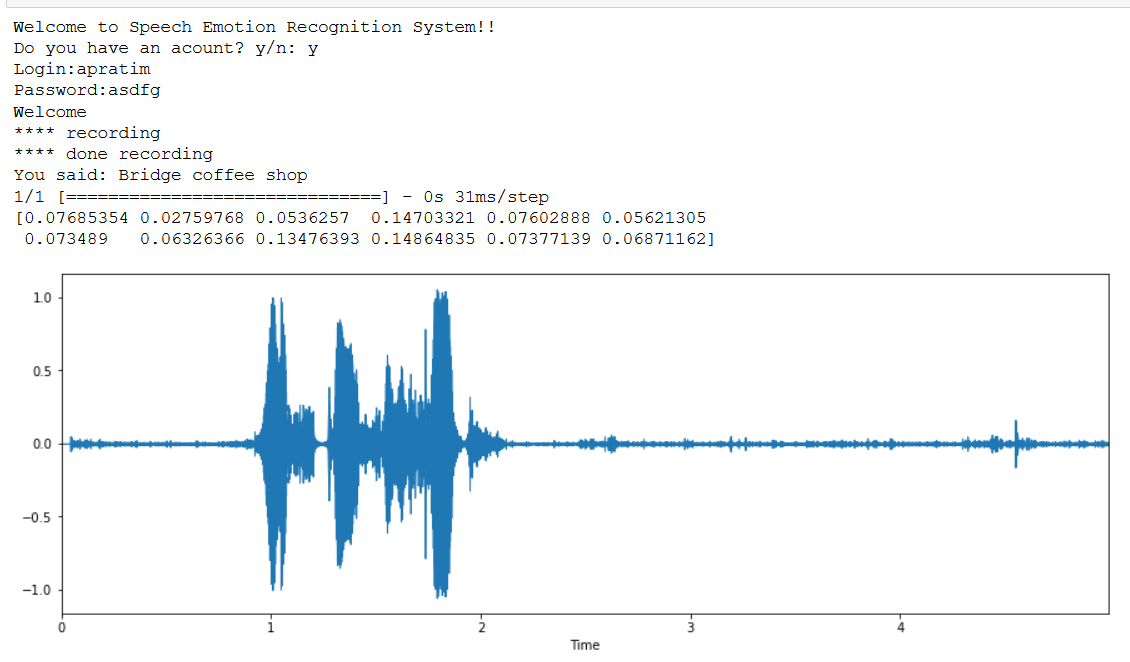


Figure xxxiii: Recording and the Waveform.

In the above **TC 08** User try to detect the anger emotion with the speech ‘this coffee Sucks’ But the software unable to get the speech correctly and print ‘Bridge Coffee Shops’. And the Test case Fail.



Figure xxxiv: Failed to detect the anger emotion.

# 6. Results

In this chapter we tried to show the whole output by recording the voice of the user to get their emotion. As we know from the software implementation chapter that by using CNN (Convolutional Neural Network) of layer 5 we train our data set and achieve an accuracy of 70.23% which is quite good enough to replicate the emotion. But Sometime the train data model failed to recognize the correct emotion, which generally occurs:

• Due to the bad quality of microphone or

• Due to the python speech recognition system which sometime not able to detect the correct voice.

• Due to less datasets because if we train a large number of datasets then we can achieve 90 % of accuracy.

Therefore, we are going to test the number of emotions that are detect by our system and for showing this we made a list of tables to replicate this and then analyse the emotions.

|  |  |  |  |
| --- | --- | --- | --- |
| **SNO:** | **User Voice** | **Output File** | **Emotion Detected** |
| **1.** | Hello I am Fine. | Output.wav | Neutral |
| **2.** | I am very Delighted Today | Outout1.wav | Happy |
| **3.** | I am going to kill you | Output2.wav | Worry or Sad |
| **4.** | Oh! that officer | Output3.wav | Surprised |
| **5.** | Why are you so Sad | Output4.wav | Sad |

Table 5: Result of the Output File with Correct Emotion Detected.

Analysis:

**Note: For Source Code see the appendix page.**

SN0: 1

Here, In this ouput.wav file user recorded the voice Hello I am Fine and the system correctly detected the Speech and correctly detect the emotion is Neutral.

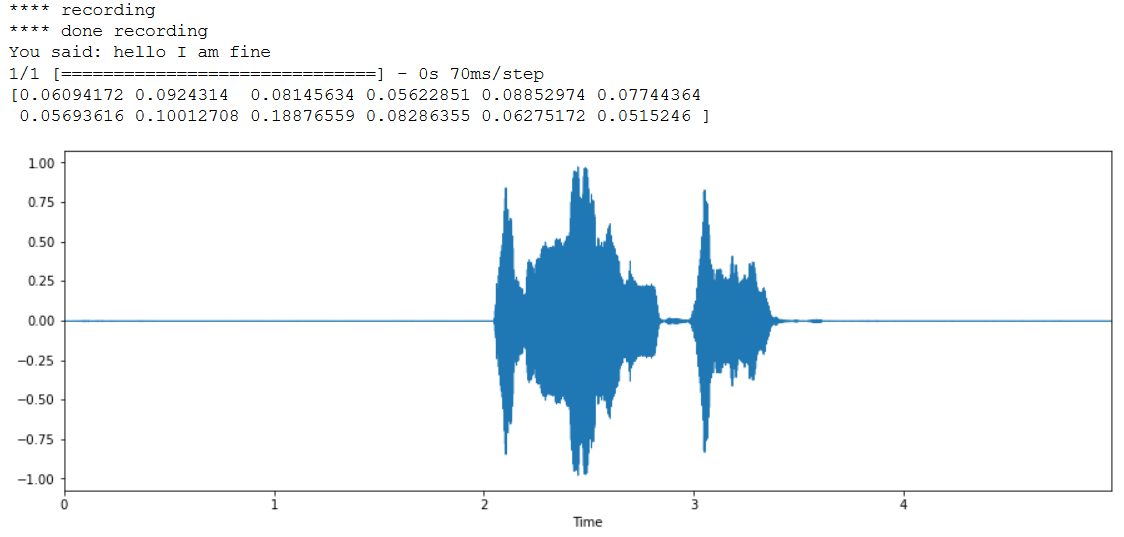


Figure xxxv: Recording and waveform of the output.wav file.



Figure xxxvi: Neutral Emotion Detected Successfully.



Figure xxxvii: The output.wav file

SN0: 2

Here, In this ouput1.wav file user recorded the voice I am very Delighted Today and the system correctly detected the Speech and correctly detect the emotion is Happy.

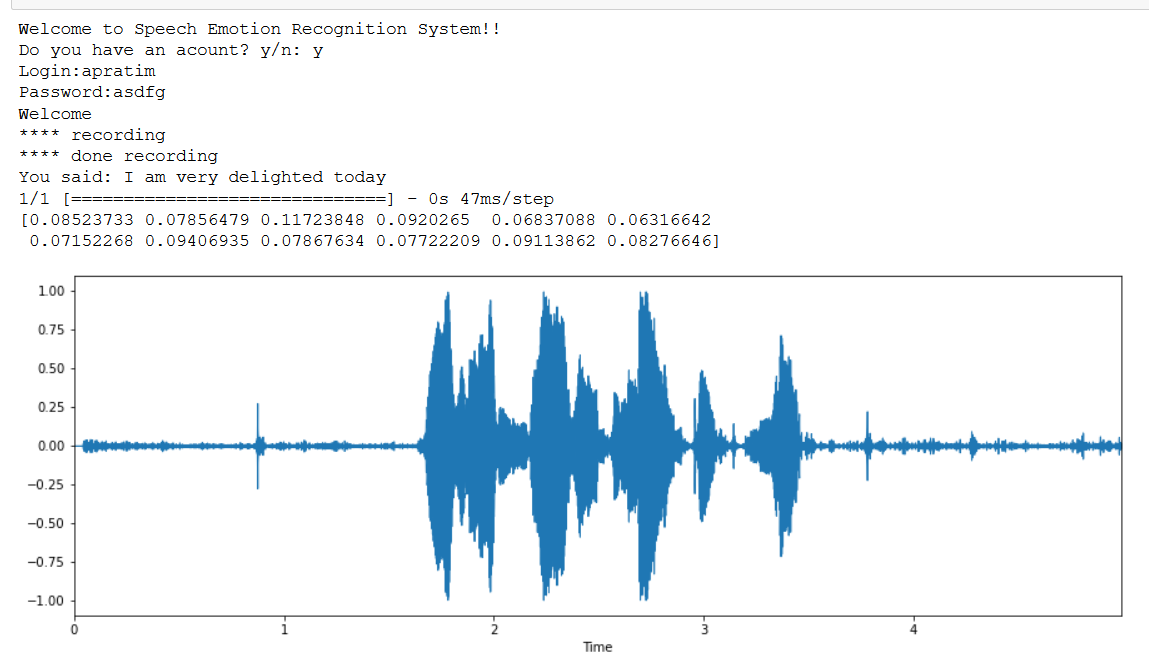


Figure xxxviii: Recording and The waveform of the output1.wav file.



Figure xxxix: Happy Emotion Detected Successfully.



Figure xl: The output1.wav file

SN0: 3

Here, In this ouput2.wav file user recorded the voice I am going to kill you and the system correctly detected the Speech and correctly detect the emotion is Sad or Worry.

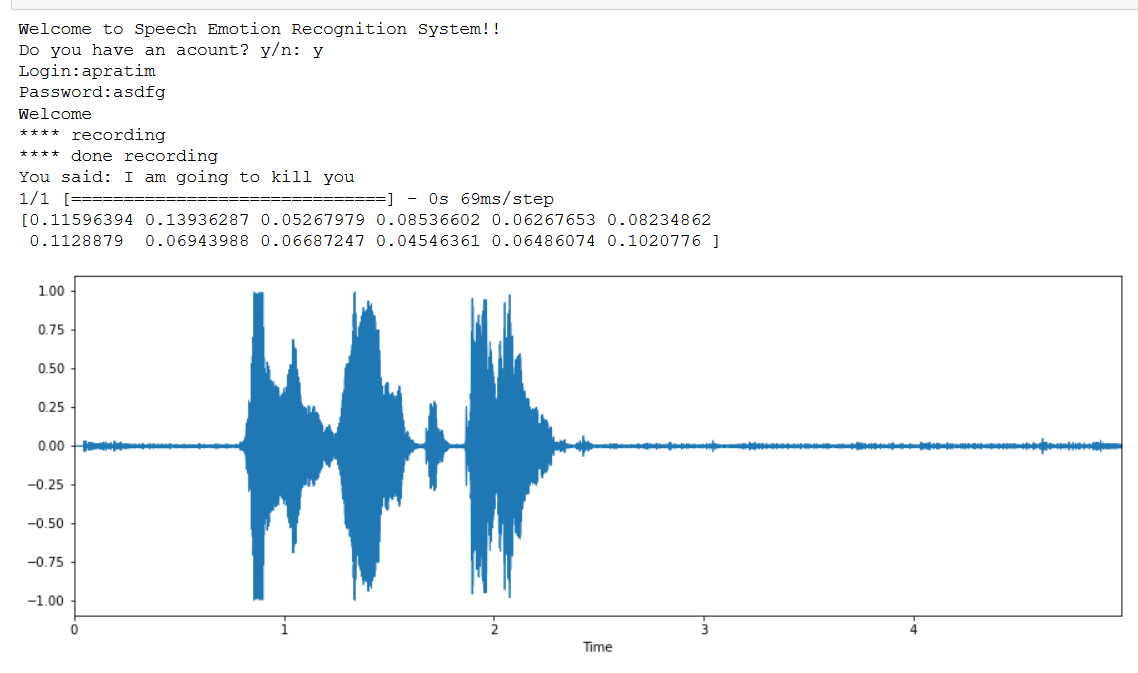


Figure xli: Recording and the waveform of output2.wav file



Figure xlii: Sad or Worry emotion detected.



Figure xliii: The Output2.wav file

SN0: 4

Here, In this ouput3.wav file user recorded the voice Oh! that officer and the system correctly detected the Speech and correctly detect the emotion is Surprised.

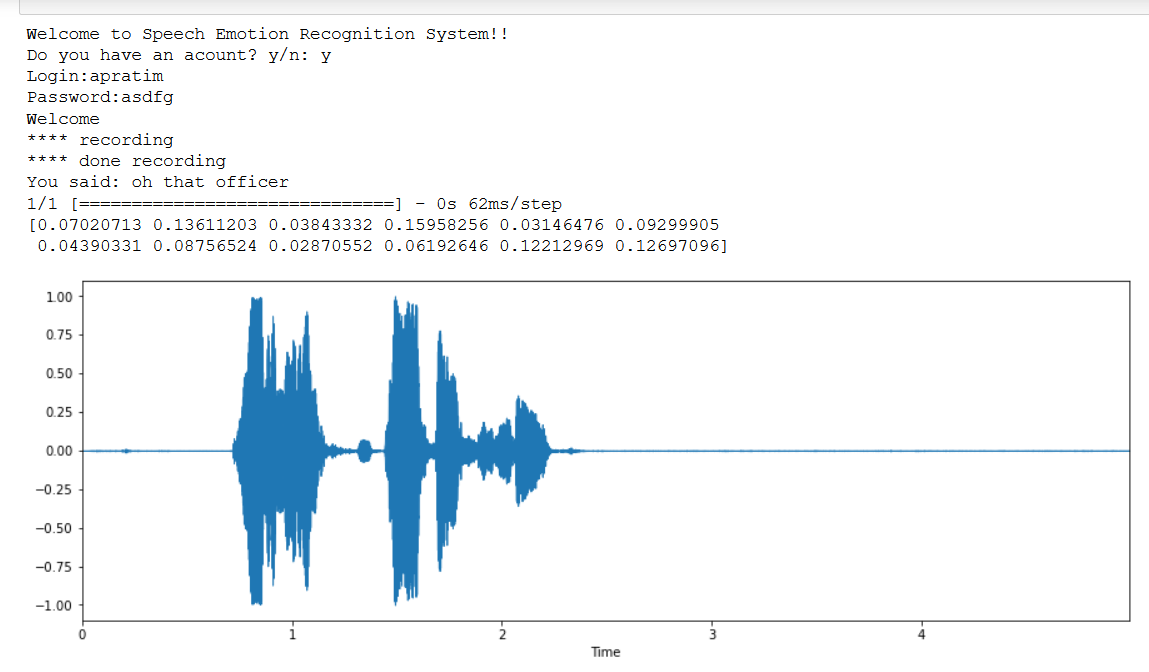


Figure xliv: Recording and the waveform of the output3.wav file.



Figure xlv: Surprise Emotion Detected Successfully.

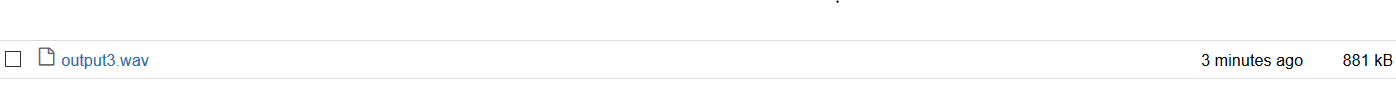


Figure xlvi: The Output3.wav

SN0: 5

Here, In this ouput4.wav file user recorded the voice Why are you so Sad and the system correctly detected the Speech and correctly detect the emotion is Sad.

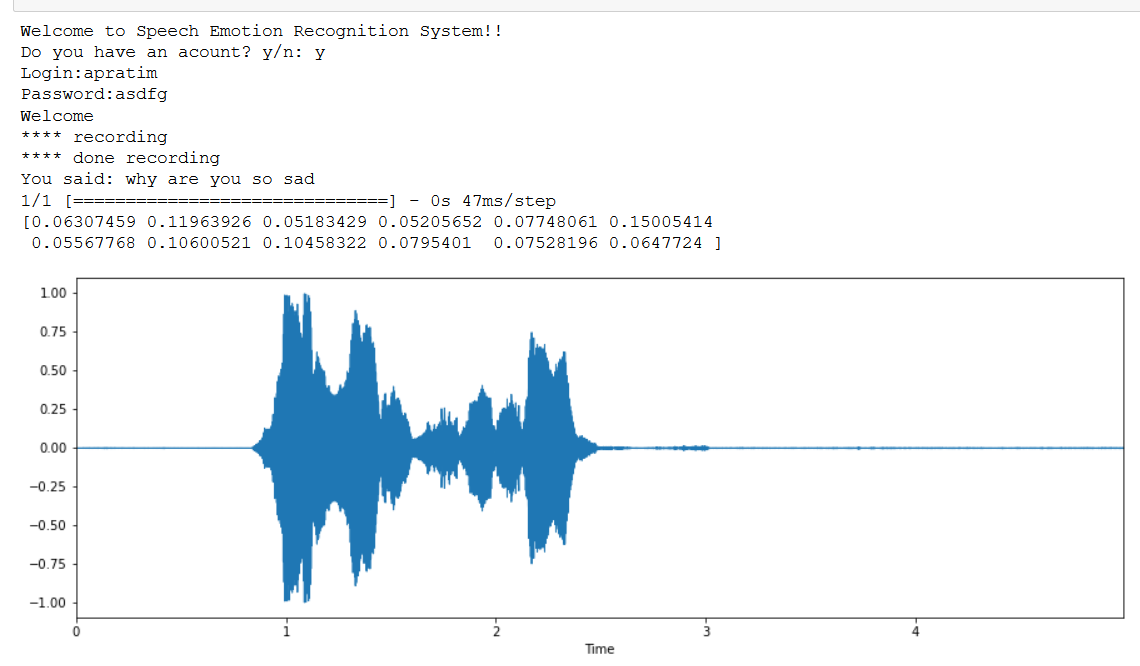


Figure xlvii: Recording and the Waveform of the output4.wav file.



Figure xlviii: Sad Emotion Detected Successfully.



Figure xlix: The Output4.wav file.

Hence, we show all the emotions that are detected by our system and in future if we train the model with large dataset and build a more multilayer CNN to train the model than we get more accuracy and able to get detected more emotion.

# 7. Project Costing

In this Chapter we Show the total Cost that are required to build the Speech Emotion Detection Model.

|  |  |  |
| --- | --- | --- |
| **Component** | **Quantity** | **Cost** |
| Software | NA | Rs 0 |
| Hardware | NA | Rs 0 |
| Man Hours | (22 hours \* 11 weeks \* 5members) = 1210 hours | 1210 man hours |

Table 6: Overall Cost Table.

# 8. Conclusions and Suggestions for Future Work

In this Chapter we are going to talk about the Conclusion of our Software Model and In Future where it would be use.

## 7.1. Conclusion

In this project, the emotional State recognition by speech recognition a model is developed which detects the human emotions by just recognizing their voice. The dataset (voice samples) for this model are taken mainly from the two sources first the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and Surrey Audio-Visual Expressed Emotion (SAVEE) Dataset. These datasets consist of British English and contains different emotions (as explained in chapter 4 . Problem solving under the section 4.3 data collection). After building numerous different models, the best CNN model for emotion classification problem was found. This model was trained successfully with different datasets. The labels were also set for the different emotions. The model was also tested on real world samples.

Thus, after testing on different sample datasets it is found that the model here developed can detect the emotional state of a person with almost 70.23% of accuracy. The total of 7 different emotions can be detected with such accuracy in this model. The accuracy obtained in this study cannot be the compared to other studies as there is a difference in the datasets. Also, the dataset used here was originally intended to incorporate visual cues such as facial features which were recorded in video samples and contained songs sung by various actors. The accuracies obtained for data from female speakers is more than with the male speakers. This is in accordance with other studies which found that accuracies on a data consisting of only female speakers are higher than the data consisting of only male speakers.

This project is mainly to help people to know the emotional state of a person by just listening their voice and thus, understating the mental condition of a person. This model can be used by psychiatrists to easily understand their patients and help them out. Thus this project is to meet the requirements of people and to help them out in their mental conditions. It is hoped that this model can help to shift the current scenario of people dealing with depression to a scenario of mentally happy society.

## 7.2 Suggestions for future work:

The model here developed detects the emotional state by training the voice samples and the accuracy given is 70.23%. Some further improvements can be made on this model in order to increase its reliability and effectiveness.

In the future we could build a sequence to sequence model to generate voice based on different emotions. E.g. A happy voice, A surprised one etc. Here, majority of the datasets available for Speech Sentiment Analysis have prompted emotions. That is, it involves speech samples which are produced by identical utterances of a speech in a given emotion. As these speeches are a deliberate effort it may not always be like an unprompted speech which are more natural in nature. However, the major drawback in obtaining unprompted speech samples is that it would require more human effort and time. It would also involve recording speech samples all the time which may lead to privacy concerns.

Incorporating a mechanism which would in real time update the models could also be looked into. However, this approach would require the active interaction of a participant to annotate the data in real time. Supplementing the audio dataset with visual cues like facial features may help in predicting the broader spectrum of emotion. This can be taken further as a future work in the research to improve the accuracy of the predictions. Also, tuning the parameters of the models further can also lead to better accuracy that can be taken into future work.

Thus, with greater accuracy this model can be used in the major medical field of mental health. The psychiatrist can use this model to detect the conditions of their patients and help them out. With the passing time there is increase in completions and at the same time there is increased rate of mental pressure. In today’s world a person deals with lot of emotions at the same time and thus, understanding the correct emotion and working towards a happier mind is of huge importance. This model can thus, help to build a happy society in the future by helping people understand each other better by knowing the emotional states.

# References

**[Referring a Journal Paper]**

1. Nakayama, W. and Nakajima, T. (1982) Effects of Pore Diameters and System Pressure on Saturated Pool Nucleate Boiling Heat Transfer from Porous Surfaces, *J. Enhanced Heat Trans.*, 104(2), pp. 286-291.

**[Referring a Book]**

1. Nield, D.A. and Bejan, A. (2005) *Convection in Porous Media*, New York: Springer-Verlag.

**[Referring a Edited Book]**

1. Yang, W. and Kim, J.H. (1992) *Rotating Machinery*, New York: Begell House.

**[Referring a Chapters of Edited Book]**

1. Nishikawa, K., Ito, T. and Eds (1982) Augmentation of Nucleate Boiling Heat Transfer by Prepared Surfaces, In Mizushima T. and Yang W.J., Eds., *Heat Transfer in Energy Problems*, Washington, D.C.: Hemisphere Publishing, pp. 1111-1182.

**[Referring a E-Book]**

1. Buyevich Yu, A. and Alexandrov, D.V. (2005) *Heat Transfer in Dispersons*, Connecticut: Begell House, Available at http://www.edata-center.com/ebooks/b7f98f1e271b3e77a.html/ E-Books [accessed 5 May 2005].

**[Referring a Conference Proceedings]**

1. Ma, T.M. (1987) Effects of Geometrical Shapes of Reentrant Grooves on Boiling Heat Transfer from Porous Surfaces, Heat Transfer 1986, *Proc. of 8th Intl. Heat Trans. Conf.*, Vol. 4, pp. 2013-2018.

**[Referring a Dissertation]**

1. Richmond, J. (2004) *Steady State Thermal Conductivity*, PhD Thesis, University of Connecticut.

**[Referring a White Paper]**

1. Bettencourt, S. and Anderson, J.L. (1990) Pen-Reared Salmonid Aquaculture in the Northeastern United States, U.S. Department of Agriculture, Northeast Regional Aquaculture Center Report 100. Kingston, RI.

**[Referring Standards]**

1. ASME, (1996) Scheme for the Identification of Piping Systems, ASME Standard A13.1.

**[Referring a Manual]**

1. Ricardo Software Company, (2005) VECTIS Help Manual, VECTIS Help Centre v3.9.2.

**[Referring Webpages]**

1. Nonperiodical Web Document - Author A.A., Author B.B., (Year of publication) *Title of document*. Retrieved month day, year, from http://Web address.
2. Chapter or Section of a Web document - Author A.A., Author B.B., (Year of publication) *Title of article, In Title of book or larger document* (chapter or section number), Retrieved month day, year, from <http://www.someaddress.com/full/url/>
3. Online Scholarly Journal Article - Author A.A., Author B.B., (Year of publication) Title of Article, *Title of Journal*, volume number, Retrieved month day, year, from <http://www.someaddress.com/full/url/>

**[Referring a Patent]**

1. Vladimir, E. S. (1996) Light Duty Box Wing Aeroplane, US Patent No. 5503352, available at www.freepatentsonline.com/5503352.pdf, April 2, 1996.

List of references are arranged in Harvard referencing style:

* [1]<https://towardsdatascience.com/speech-emotion-recognition-with-convolution-neural-network-1e6bb7130ce3>
* [2]<https://towardsdatascience.com/logistic-regression-using-python-sklearn-numpy-mnist-handwriting-recognition-matplotlib-a6b31e2b166a>
* [3] T. Giannakopoulos (2015) pyAudioAnalysis: An Open-Source Python

Library for Audio Signal Analysis. PLoS ONE 10(12): e0144610.

https://doi.org/10.1371/journal.pone.0144610

* [4] <http://neuron.arts.ryerson.ca/ravdess/?f=3>
* [5] <http://kahlan.eps.surrey.ac.uk/savee/Download.html>
* [6] <https://www.tensorflow.org/datasets/catalog/savee>
* [7] T. Giannakopoulos (2015) pyAudioAnalysis: An Open-Source Python

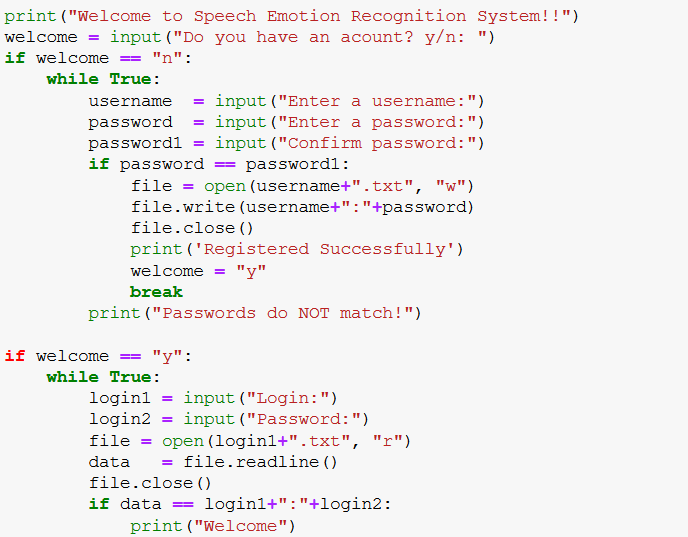
Library for Audio Signal Analysis. PLoS ONE 10(12): e0144610.

https://doi.org/10.1371/journal.pone.0144610

# Appendix

## Appendix-A

**Source Code for Login and register into the system are given below:**



## Appendix-B

**Source Code for recording the voice and store it in output.wav file and also print the speech into text.**



## Appendix-C

**Source Code for load the model.json file, output.wav file and get the feature of the output.wav file by print the waveform.**



## Appendix-D

**Source Code for live emotion testing of the output.wav file by using the training dataset model.**

