

# **CAPSTONE PROJECT REPORT**

(Project Term January-May 2021)

## **Emotion Recognition Through Speech**

Submitted by

**Gundreddy Sohith Naidu**

**Registration Number: 11713077**

**Ashokraj K**

**Registration Number: 11705871**

**Garine Sai kiran**

**Registration Number: 11705768**

**Sharon Bino**

**Registration Number: 11704249**

**Edara Raj Kumar**

**Registration Number: 11700979**

**Project Group Number KC345**

**Course Code CSE445**

Under the Guidance of

**Krishan Bansal (Assistant Professor)**

**School of Computer Science and Engineering**



## TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering (SCSE)

**Program :** P132::B.Tech. (Computer Science & Engineering)

**COURSE CODE :** CSE445

**REGULAR/BACKLOG** Regular

**GROUP NUMBER :** CSERGC0345

**Supervisor Name :** Krishan Bansal

**UID :** 16348

**Designation :** Assistant Professor

**Qualification :** \_\_\_\_\_ **Research Experience :** \_\_\_\_\_

SR.NO.	NAME OF STUDENT	Prov. Regd. No.	BATCH	SECTION	CONTACT NUMBER
1	Gundreddy Sohith Naidu	11713077	2017	K17QS	9877458337
2	Ashokraj K	11705871	2017	K17GU	7708275675
3	Garine Sai Kiran	11705768	2017	K17CF	8985679284
4	Sharon Bino	11704249	2017	K17KH	9877497172
5	Edara Raj Kumar	11700979	2017	K17GU	9182936277

**SPECIALIZATION AREA :** Software Engineering

**Supervisor Signature:** \_\_\_\_\_

**PROPOSED TOPIC :** SPEECH EMOTION RECOGNITION

Qualitative Assessment of Proposed Topic by PAC		
Sr.No.	Parameter	Rating (out of 10)
1	Project Novelty: Potential of the project to create new knowledge	6.91
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	7.23
3	Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program.	7.45
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	8.00
5	Social Applicability: Project work intends to solve a practical problem.	7.45
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	6.68
PAC Committee Members		
PAC Member (HOD/Chairperson) Name: Dalwinder Singh	UID: 11265	Recommended (Y/N): Yes
PAC Member (Allied) Name: Asha Rani	UID: 11332	Recommended (Y/N): Yes
PAC Member 3 Name: Dr. Parampreet Kaur	UID: 18758	Recommended (Y/N): Yes

**Final Topic Approved by PAC:** SPEECH EMOTION RECOGNITION

**Overall Remarks:** Approved

**PAC CHAIRPERSON Name:** 14307::Raj Karan Singh

**Approval Date:** 11

Mar 2021

4/22/2021 3:23:46 PM

## DECLARATION

We hereby declare that the project work entitled “**Emotion Recognition Through Speech**” is an authentic record of our own work carried out as requirements of Capstone Project for the award of B. Tech degree in “**Computer Science & Engineering**” from Lovely Professional University, Phagwara, under the guidance of “**Krishnan Bansal**”, during August to November 2020. All the information furnished in this capstone project report is based on our own intensive work and is genuine.

Project Group Number: KC345

Name of Student 1: Gundreddy Sohith Naidu

Registration Number: 11713077

Name of Student 2: Ashokraj K

Registration Number: 11705871

Name of Student 3: Garine Sai Kiran

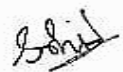
Registration Number: 1170576

Name of Student 4: Sharon Bino

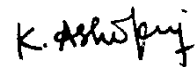
Registration Number: 11704249

Name of the Student: Edara Raj Kumar

Registration Number: 11700979



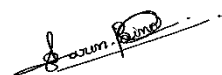
Date:24-04-2021



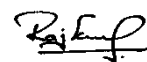
Date:24-04-2021



Date:24-04-2021



Date:24-04-2021



Date:24-04-2021

## **CERTIFICATE**

This is to certify that the declaration statement made by this group of students is correct to the best of my knowledge and belief. They have completed this Capstone Project under my guidance and supervision. The present work is the result of their original investigation, effort, and study. No part of the work has ever been submitted for any other degree at any University. The Capstone Project is fit for the submission and partial fulfillment of the conditions for the award of B. Tech degree in “**Computer Science & Engineering**” from Lovely Professional University, Phagwara.



(Krishnan Bansal)

**Signature and Name of the Mentor**

**Designation** – Assistant Professor

**School of Computer Science and Engineering,**  
Lovely Professional University,  
Phagwara, Punjab.

Date: 24-04-2021

## **ACKNOWLEDGEMENT**

Here by declaring that the case study entitled "Speech Emotion Recognition through speech" submitted at Lovely Professional University, Phagwara, Punjab is an authentic work and has not been submitted elsewhere.

Furthermore, understanding that the work presented here with is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of all group members knowledge, the content of this case study represents authentic and honest effort conducted, in its entirety, by all of us. Therefore, all the members of the group fully responsible for the contents the case study report.

Thankfully by gaining support from the university and our faculty member who is An Assistant Professor "Krishnan Bansal". We had completed the case study report as soon as possible within timeline from week to week within in a one and half month by conducting the online classes and describing the work gradually.

Ashokraj K

Edara Rajkumar

Garine Saikiran

Gundreddy Sohith Naidu

Sharon Bino

## Contents

1. INTRODUCTION -----	8
1.1 OBJECTIVE-----	8
2. PROFILE OF THE PROBLEM -----	8
3. Existing System-----	9
3.1 Introduction-----	9
3.2 Existing Software-----	10
3.3 DFD For Present System -----	11
3.4 What's New in the System To be Developed-----	12
4. Problem analysis-----	12
4.1 Product definition -----	12
4.2 Feasibility Analysis -----	13
<b>4.2.1 Linear Prediction Cepstral Coefficients (LPCC):</b> -----	13
<b>4.2.2 Mel-frequency spectrum coefficients (MFCC):</b> -----	14
4.3. Project Plan -----	14
5. Software Requirement Analysis -----	15
5.1 Introduction -----	15
<b>5.1.1. List of Libraries</b> -----	15
5.2 General Requirements-----	16
<b>5.2.1 Software and Hardware Requirements</b> -----	16
<b>5.2.2 Hardware Requirements</b> -----	17
<b>5.2.3 Software Requirements</b> -----	17
5.3 SPECIFIC REQUIREMENTS: -----	17
6. DESIGN -----	18
6.1. System Design-----	18
6.2. Flowchart -----	19
7. TESTING -----	20
7.1. Testing the Project-----	21
7.2. Levels of Testing -----	22
7.3. Manual testing to improve design -----	23
8. Implementation of project -----	24
8.1. Conversion Plan-----	25
8.2. LIMITATIONS AND FURTHER WORKS -----	27
8.3. Limitations -----	27
8.4. Future Enhancements -----	27
8.5. Software Maintenance-----	27
9. PROJECT LEGACY -----	29
9.1. Current Status of the Project -----	29

9.2. Remaining Areas of Concern -----	29
9.3. Technical and Managerial Lessons Learnt -----	30
10. User Manual -----	30
11. Source Code-----	32
12. Bibliography-----	37

# **1. INTRODUCTION**

Emotion plays an important role in Human beings. Interaction between human and computer has been increasingly rapidly. To make this interaction better and efficient, natural communication between them is needed. Aim of computer should be able to respond based on the emotion perceived from the user. To achieve these criteria, computer need to predict our emotion through facial expression or by our speech. In Human interaction, speech plays a major role. In this project we tried to detect the emotion of a person through their speech.

Detecting emotion from speech makes the interaction efficient. It will make improvements in voice recognition systems when we implement this kind of models in it. When we try this model with native language dataset of the users will be more effective as well. Detecting emotion through speech can be benefitted in many ways which all given in this report.

## **1.1 OBJECTIVE**

In the emotion recognition from speech system, different feature extraction techniques are taken into consideration and made SVM classification and MLP classifiers for a better accuracy. Machine learning model is created from this classifier. It predicts the human emotions by its previous trained and test sets. The data used for trained the model is taken form The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) which contains audio files of different emotions. Currently, it can detect 7 emotions : Anger, Calm, Happy, Disgust, Surprise, Sad and Fear. If the model can have an improvement with additional research, more emotions can be generated. The project is in a ready-go state to be used by a user to detect whether his tone of speech in any of these seven emotions. We discussed the complete process and steps we can in this project here.

## **2. PROFILE OF THE PROBLEM**

Predicting the emotion through user speech is the task. For that we made a webpage in order to record the speech of the user. When we drag that audio file into this model, it detects the emotion of that speech. This project has immense possibilities in the areas like in Call Centre to detect customer's emotion, voice based virtual assistance like Siri & Alexa etc. Since the emotions are calculated as distribution on gender basis, the



people who are not categorised as male and female voice can be identified if an aggressive research happens behind this project. Even we could detect emotion of animals if we worked on the datasets related to that in further research study. Making this project into a product can also help as an emergency system through voice, for an example, in a home, if a person is happened to cause a terrible injury where he is not able to walk, he can scream at the product asking help. The product would detect his emotion & emergency and would proceed with further steps like calling nearby hospital, ambulance or even police if a robbery happens. These were the scenarios where we can implement this model.

### **3. Existing System**

#### **3.1 Introduction**

Emotion recognition through Speech is a particular type of technology which has been used in now today's world. Such that this technology is used various applications like conducting a crime Investigation of observing the victim words which has been said for truth or false statements.

In this case emotion plays a crucial role for a person to understand the victim words and conclude the victim is telling the officers truth or not. Therefore, for the machine it will be an easy role for the taking the process of the victim such that the machine will find out the truth behind the words which has been spoken by the victim. Further it will identify the clarity of vocals of speech of mid pitch, high pitch and low pitch. Based on the pitches the machine will find out the following feelings of the particular person like anger, fear, happy, sad, distress, etc. They are three types of particular features which are used for extraction of emotion detection such as follows:

- 1) Elicited features
- 2) Prosodic features
- 3) Spectral features

Thus, therefore for different features different types of technology used such as for the spectral type of features they have been used MFCC, LPCC and MEDC. For the prosodic features they have been used by based on following pitch, fundamental frequency, global parameters, loudness.

There are other kinds of classification for classifying emotions are used such as Artificial neural network, Support vector machine, Gaussian Mixture model, Hidden Markov Model.

In this Project will be able to show you the way to acknowledge totally different emotions from pre-recorded audio recordings. we all know that voice-controlled personal assistants like Amazon Alexa, Apple Siri, and Google Assistant and plenty of additional became additional powerful and still evolving. we have a tendency to begin to visualize them integrated into phones, laptops, room gadgets, cars, primarily on nearly something we have a tendency to use daily. We can feel the convenience of use is that the primary key that produces this field grow magnificently.

When we have ascertained regarding the Speech feeling Recognition project on Emotional speech audio dataset. We feel this is often associate degree exciting and fun project. As we have a tendency to use additional voice-controlled gadgets, we feel feeling recognition are a part of these devices within the following years. the unreal intelligence behind these devices is good enough to grasp our emotions once we speak to them and provides additional customized responses.

For example, before more matured the road, one of our group raised Siri to “play music from the Music app,” and so it starts to play the broad combine. However, imagine if we tend to add the feeling recognition power into that command. This way, it will play varieties of music counting on our mood. several music apps area unit already giving classes with totally different needs, therefore why not play that blend with simply an easy “play music” command.

### **3.2 Existing Software**

The various classifier used for the classification of the options of the speech. There square measure numerous classifiers used like **Gaussian Matrix Model, Hidden Markov Model, Support Vector Machine**. B. Yang, M. Lugar projected a piece wherever the emotion detected by the sound options.

It absolutely was supported the music theory. It took the 2 different pitch intervals. It took the 2 totally different pitch intervals. and so found the occurrences that square measure the explanation behind of a consonant or dissonant impression.

They can evaluate these harmony options during a lot of realistic manner. Yashpal sing Chavan, M. L. Dhore, Pallavi Yes aware projected the speech options like, **Mel Frequency ceptrum coefficients** and **Mel Energy Spectrum Dynamic Coefficients**.

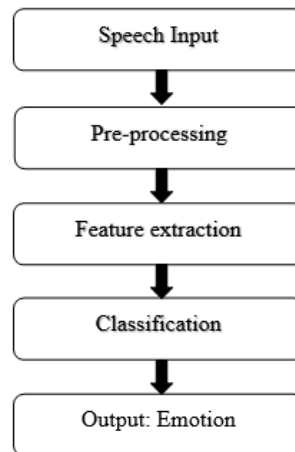
They took the utterances of speech of somebody's voice as associate degree input and so numerous options extracted from those utterances. The classifier used has been used for classifying emotions was the “**support vector machine**” or we can say **SVM**. The LIBSVM was used for classification of feeling. J. Sirisha Devi, Y. Srinivas and Shiva Prasad Nandyala introduced text dependent speaker recognition with associate degree sweetening of sleuthing the feeling of the speaker previous victimisation the hybrid FFBN and GMM ways.

Lingli Yu, Kaiju Chou dynasty, Yishao Huang projected humans' emotional speeches recognition contributes a lot of to created harmonious human machine interaction, conjointly with several potential applications. In projected system they used 3 ways to extend binary “**support vector machines**” square measure compared for recognizing emotions from speech by the Chinese and therefore the Berlin speech information.

One was customary SVM schemes, and 2 alternative ways square measure DAG and UDT that might type a binary of call tree classifiers. Meanwhile, a ranked classification technique of feature driven ranked SVMs classifiers square measure designed, whose structure is analogous with DAG, it used totally different feature parameters to drive every layer, and therefore the feeling will be divided layer by layer. Finally, analysis of the classification rate of these 3 extend binary SVMs, DAG performed the most effective for testing information, and customary SVM wasn't so much behind, the UDT was the poorest thanks to depend on its higher layer classification accuracy.

### **3.3 DFD For Present System**

The speech samples square measure taken as input. the primary factor to be finished the speech samples is that the pre-processing wherever noise from the sample is removed. currently from the noise free samples desired options square measure extracted. These options square measure then more expire to the classifier a. The classifier so classifies the emotions consequently and outputs the emotions.



**Figure 1: Speech Emotion Recognition System [1]**

### **3.4 What's New in the System To be Developed**

The new in the system to be developed to be adding some of the Phycological events in the project such that if the physiological signals associated with involuntary systema nervosum enable to assess objectively emotions. These embody such as vital sign, cardiac, respiration, force per unit area, myogram, skin electrical phenomenon, blood volume pulse, and skin temperature. Further victimization physiological signals to acknowledge emotions is additionally useful to those folks that suffer from physical or psychopathy so exhibit issues with facial expressions or tone of voice.

## **4. Problem analysis**

### **4.1 Product definition**

A definition is each necessary and troublesome as a result of the everyday word “emotion” may be a notoriously fluid term in that means. feeling is one amongst the foremost troublesome ideas to outline in scientific discipline. In fact, there are a unit completely different definitions of emotions within the scientific literature.

In everyday speech, feeling is any comparatively transient aware expertise characterised by intense mental activity and a high degree of delight or annoyance. Scientific discourse has drifted to alternative meanings and there's no agreement on a definition. feeling is commonly entwined with temperament, mood, temperament, motivation, and disposition. In scientific discipline, feeling is usually outlined as a fancy state of feeling that leads to physical and psychological changes.

These changes influence thought and behaviour in line with alternative theories, emotions aren't causative forces however merely syndromes of elements like motivation, feeling, behaviour, and physiological changes. In 1884, In what's Associate a Nursing emotion? yank scientist and thinker James planned a theory of feeling whose influence was considerable. in line with his thesis, the sensation of intense feeling corresponds to the perception of specific bodily changes. This approach is found in several current theories: the bodily reaction is that the cause and not the consequence of the feeling. The scope of this theory is measured by the numerous debates it provokes. This illustrates the issue of agreeing on a definition of this dynamic and sophisticated development that we tend to decision feeling.

“Emotion” refers to a good vary of emotive processes like moods, feelings, affects, and well-being. The term “emotion” in has been additionally observed a particularly advanced state related to a good sort of mental, physiological, and physical events.

## 4.2 Feasibility Analysis

Human speech consists of many parameters which shows the emotions comprise in it. As there is change in emotions these parameters also get changed. It is necessary to select proper feature vector to identify the emotions.

Features are categorized as excitation source features, spectral features, and prosodic features. Excitation source features are achieved by suppressing characteristics of **vocal tract**. Prosodic features used for emotion recognition are pitch, energy, intensity. Statistical measurements are also used to distinguish emotions like minimum, maximum, standard deviation, range, mean, median, variance, skewness, kurtosis etc. of features.

Spectral features used for emotion recognition are **Linear prediction coefficients, Perceptual linear prediction coefficients, Mel-frequency spectrum coefficients, Linear prediction cepstrum coefficients, perceptual linear prediction**. The accuracy of differentiating different emotions can be achieved by using MFCC, LFPC, LPC, PLP, and RASTA-PLP.

### 4.2.1 Linear Prediction Cepstral Coefficients (LPCC):

The cepstral could be a common remodel accustomed gain data from associate degree EEG signal. It may be accustomed separate the excitation signal (which contains the

words and also the pitch) and also the transfer operates (which contains the voice quality). The cepstrum may be seen as data concerning rate of modification within the totally different spectrum bands.

#### 4.2.2 Mel-frequency spectrum coefficients (MFCC):

Feature extraction is employed to extract the feature from the speech signal. **Mel Frequency Cepstral constant** is that the most significant and effective methodology for the feature extraction. Feature extraction aims for knowledge reduction by changing the signal into a compact set of parameters whereas protective spectral and/or temporal characteristics of the speech signal info. The diagram of feature extraction is given below.

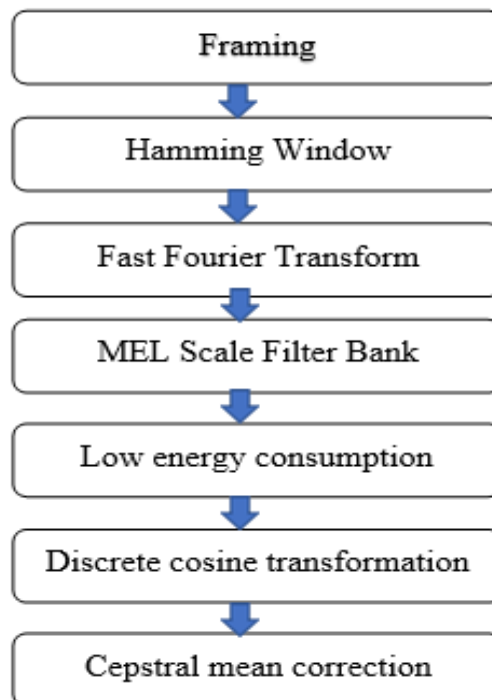


Figure 2: Feature extraction using MFCC [2]

### 4.3. Project Plan

In our speech emotion recognition system consists of 4 main steps. Initial is that the sample of the recorded voice will be collected or recorded. After that we gone analyze the voice sample with different feature methods and organize it further then we going to detect the emotion from the speech using the MFCC method.

Therefore, analysing the emotion detected by the software by using the different kinds of data sets we collected like happy, disgust, sad etc. and applying classifiers & filters to form an accurate pitch of the emotion to be displayed.

## 5. Software Requirement Analysis

### 5.1 Introduction

These are the technologies we are used: The programming language to be dominantly used is “**Python**”, probably version 3 with various libraries as listed in below.

For Website to take the input from the user we have used Web Languages like **HTML**, **CSS**, **JavaScript**.

**Python:** It a general-purpose programing language. Hence, you'll use the programming language for developing both desktop and web applications. Also, you'll use Python for developing complex scientific and numeric applications. Python is supposed with features to facilitate data analysis and visualization. It is easy to implement machine learning techniques using python.

**HTML:** HTML or Hyper Text Mark-up Language is the standard mark-up language used to create web pages. A web browser can read HTML files and comprise them into visible or audible web pages. Internet browser is able to read through HTML data and compose them in to audible or visible pages. The browser doesn't exhibit the HTML tags, but makes use of them to understand the information in the page. HTML details the framework of a site semantically together with cues for presentation, making it a markup language as opposed to a programming language.

**CSS:** Cascading Style Sheets (CSS) is a style sheet language used for describing the look and formatting of a document written in a markup language.

**Java Script:** “JavaScript is a cross-platform, object-oriented scripting language used to make web pages interactive”.

#### 5.1.1. List of Libraries

**Librosa:** A Python package for audio and music signal processing. At a high level, librosa provides implementations of a variety of common functions used throughout the field of music and audio information retrieval.

**Soundfile:** SoundFile module can read and write the sound files in the project . File reading is supported through [libsndfile](#), which is a free, cross-platform, open-source (LGPL) library for reading many different sampled sound file formats that runs on many platforms including Windows, OS X, and Unix.

**NumPy:** Numerical python It is a open source python library used for working with multi-dimensional arrays and matrices. It also has functions for working in domain of linear algebra, and matrices.

**Sklearn:** Scikit-learn is perhaps the foremost useful library for machine learning in Python. The SkLearn library contains tons of efficient tools for machine learning and statistical modeling including classification, regression, clustering, and dimensionality reduction

**Pyaudio:** To get started dealing with playback and recording audio on Windows, Linux, and MacOS in a **Python** environment you should consider using the library **PyAudio**.

**Pickle:** Python pickle module is used for serializing and de-serializing a Python object structure. Any object in Python can be pickled so that it can be saved on disk.

**Os:** The OS module in Python provides functions for interacting with the operating system. OS comes under Python's standard utility modules. This module provides a portable way of using operating system-dependent functionality.

**Glob:** Glob (short for global) is used to return all file paths that match a specific pattern. We can use glob to search for a specific file pattern, or perhaps more usefully, search for files where the filename matches a certain pattern by using wildcard characters.

## **5.2 General Requirements**

### **5.2.1 Software and Hardware Requirements Environment:**

**Operating System:** - Microsoft Windows 2000 or Higher



**Offline view:** Microsoft Internet Explorer, Google chrome

**Tools:** Python ide 3.7 or above, Notepad, Microsoft office

**Dataset:** RAVDESS Dataset.

**User Interface:** Html, CSS, JavaScript

### 5.2.2 Hardware Requirements

**Table 1: Hardware Requirement**

Number	Description
1	Pc with 10 GB hard-disk and 4 GB RAM

### 5.2.3 Software Requirements

**Table 2: Software Requirement**

Number	Description
1	Windows XP or Higher
2	Python IDE 3.7 or Higher

## 5.3 SPECIFIC REQUIREMENTS:

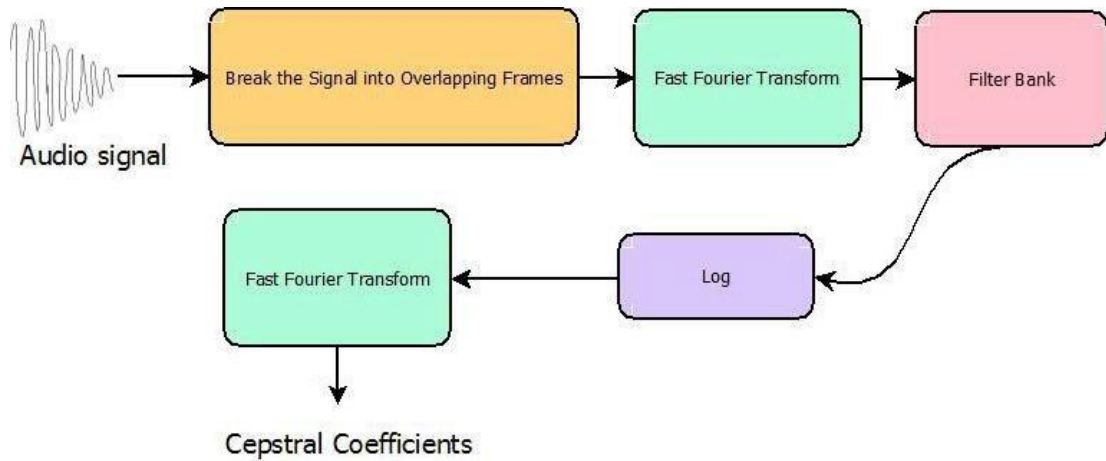
A demand could be a statement that identifies a necessary characteristic or quality of a system so as for it to possess worth and utility to a user. They're used as associate degree input into the look stage of a system as per utilized in coming up with and implementation of the solution. necessities computing a really important input into the testing part, because the take a glance at need to turn out output of a speech emotion as expressed within the demand. Hence, it's vital for necessities to be easy and specific.

As we are ready to see, the aim of the project is to find out the emotion through speech, the requirements of the emotion classier, justifying some of the implementation choices and showing evidence that software engineering concepts are thoroughly applied. It outlines the audio features used for classification, including measures such as signal energy, pitch and voice quality. It also justices the choice of labels for emotion classification (Neutral, Clam, Happy, Sad, Angry, Fearful, Disgust, Surprised) and for speech quality assessment.

## 6. DESIGN

### 6.1. System Design

MLP classifiers is used to predict the emotion. Here is the diagram to understand the systematic design of MLP.



**Figure 3: MLP classifiers [3].**

Using the training and testing functions, loaded data would be split according to the features in each audio. Here, libraries and packages such as librosa and numpy are being used. First training of the model will be taken and then testing happens, this pattern is followed in every audio files.

MLP Classifiers are used to optimise the loss functions. Minimum the loss function, better the results will be. MLP Classifier is functioned in hidden layers using the MLP supervised learning Algorithm. The function of MLP Classifier in this project is to provide the hidden layer value, alpha, batch size, learning rate and maximum iteration. After the initialisation of MLP Classifier the model is being trained

## 6.2. Flowchart

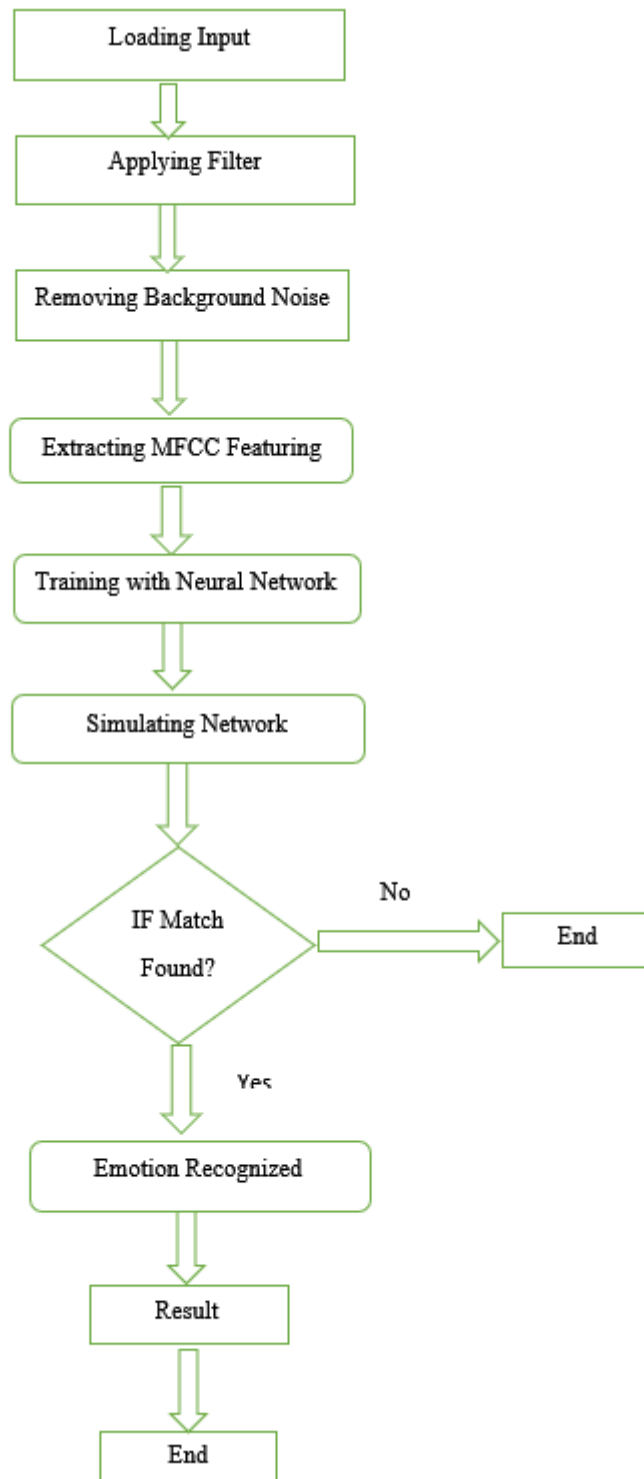


Figure 4: Flowchart of the model

## 7.TESTING

In the Speech Emotion Recognition System (SER), the sound documents are given as the information. The informational indexes goes through a number of squares of cycles which makes it executable to help for the examination of the discourse boundaries.

The information is preprocessed to transform it to the reasonable organization and the separate highlights from the sound records are removed utilizing different advances like outlining, hamming, windowing, and so forth This measure helps in separating the sound records into the mathematical qualities which addresses the recurrence, time, adequacy or whatever other such boundaries which can help in the examination of the sound records.

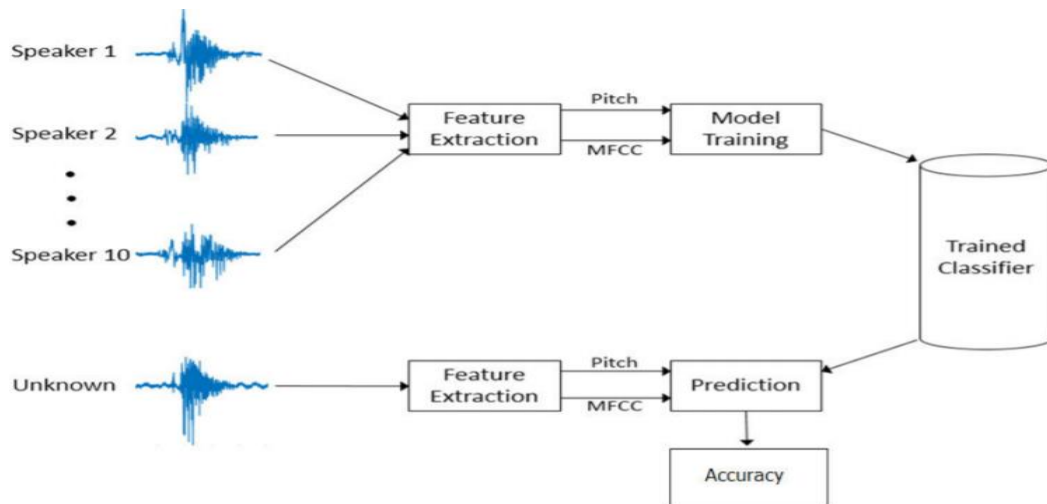
After the extraction of the required highlights from the sound documents, the model is prepared. We have utilized the RAVDESS dataset of sound documents which has addresses of 24 individuals with varieties in boundaries.

For the preparation, we store the mathematical upsides of feelings and their particular includes correspondingly in various clusters. These clusters are given as a contribution to the MLP Classifier that has been introduced.

The Classifier recognizes various classifications in the datasets and orders them into various feelings. The model can now comprehend the scopes of upsides of the discourse boundaries that fall into explicit feelings. For testing the exhibition of the model, in the event that we enter the obscure test dataset as an info, it will recover the boundaries and foresee the feeling according to preparing dataset values.

The precision of the framework is shown as rate which is the end-product of our undertaking.

Which we mentioned in the figure 5.



**Figure 5: Speech Emotion Recognition System [3]**

Ensuing work with multi-facet perceptron has shown that they are fit for approximating a XOR administrator just as numerous other non-straight capacities. Multi-facet perceptron are regularly applied to administered learning issues.

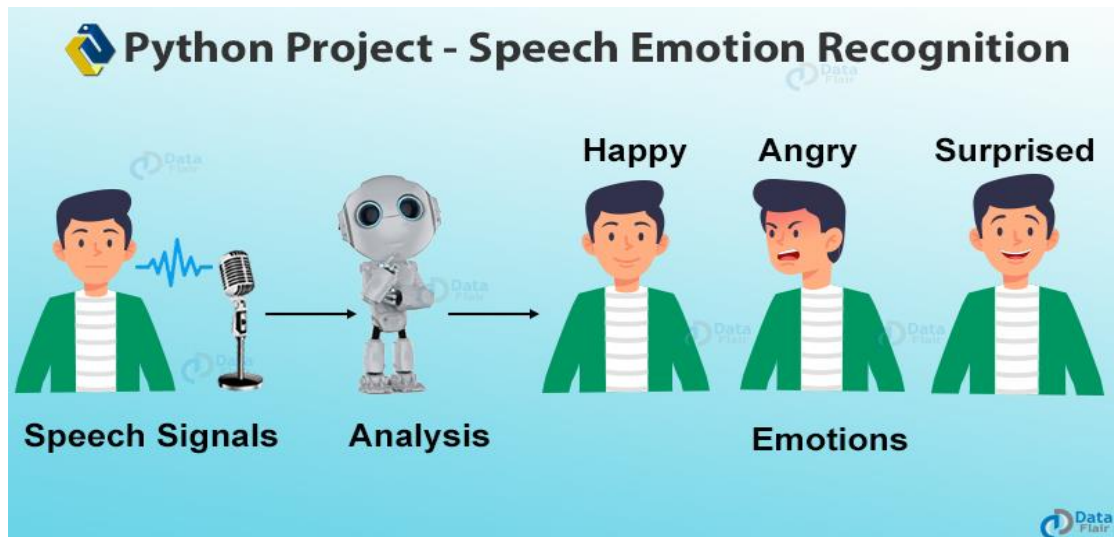
They train on a set of info yield combines and figure out how to show the relationship (or conditions) between those inputs and outputs, Important issues in MLP plan incorporate particular of the quantity of covered up layers and the number of units in these layers.

The quantity of covered up units to use is a long way from clear. As great a beginning stage as any is to utilize one secret layer, with the quantity of units equivalent to a large portion of the amount of the quantity of info and yield units

## 7.1. Testing the Project

Discourse gathering tests are generally directed by physically scoring the oral reaction of the subject. This requires a test manager to be persistently present. To stay away from this, a subject can type the reaction, after which it very well may be scored naturally. Be that as it may, spelling blunders may then be considered acknowledgment mistakes, impacting the test outcomes. We exhibit an autocorrection approach dependent on two scoring calculations to adapt to spelling mistakes. The primary calculation manages sentences and depends on word scores. The subsequent calculation manages single words and depends on phoneme scores. The two calculations were assessed with a corpus of composed answers dependent on three diverse Dutch

discourse materials. The level of contrasts among programmed and manual scoring was resolved, notwithstanding the mean distinction in discourse acknowledgment edge. The sentence remedy calculation performed at a higher exactness than usually got with these discourse materials. The word adjustment calculation performed better compared to the human administrator. The two calculations can be utilized by and by and permit discourse gathering tests with open set discourse materials over the web.



**Figure 6: Pictural representation of the project from "Data Flair"**

The contribution to the model ought to be the highlights removed along with the feeling classification that it has a place with, put away correspondingly into individual exhibits so that, classifier will be ready to recognize the examples, connections and afterward group the information.

This preparation assists the model with getting, which feelings have what scope of the particular highlights. Thus, when a concealed information is given as an info, it will actually want to connect.

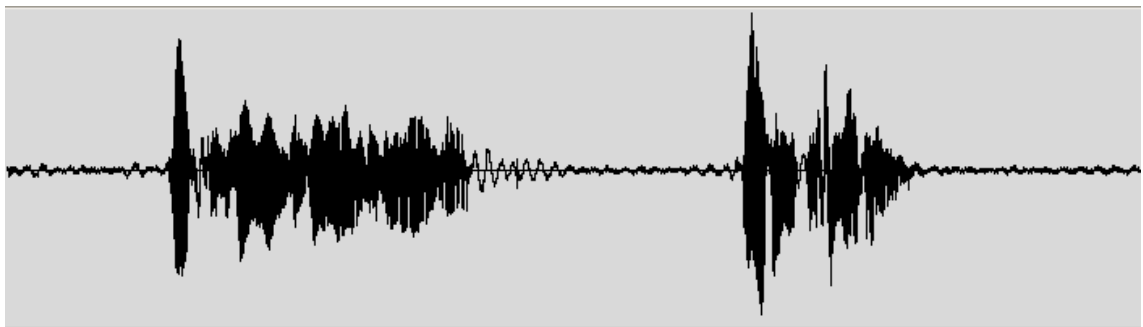
## **7.2. Levels of Testing**

The current testing interaction of a voice application is completed physically. Test designs presently need to utilize experimentation techniques to track down the most appropriate test systems. This makes bottlenecks in the pipeline.

Test engineers should be effective and mindful to get the sound reactions from the gadget as that is the sole technique for correspondence for greater part of the voice

gadgets. For a couple more up to date gadgets, test engineers have focus on sound as well as observing showcases on specific gadgets (like the Amazon Echo Show). This is indispensable in situations where the sound stream breaks or when an obscure ability is summoned.

Focusing on such subtleties guarantees a smooth and pleasant client experience and adds to the nature of the application made by the task group. Nonetheless, this interaction is amazingly tedious, particularly when the quantity of experiments is extremely huge.



**Figure 7: Waveform from audio**

### **7.3. Manual testing to improve design**

Even though manual testing can be time-consuming, it can be reliable and useful to help improve the design and product experience. This is because it replicates the user experience. Sometimes, when testing for our project, we realized that some responses could be better framed, or that we could improve the flow of certain invocations. We wouldn't have known this without experiencing the flow as a user, so we tweaked the design to accommodate these changes.

#### **Exploratory testing**

Exploratory testing is key when testing a voice project. It is vital for a test engineer to make sure that no flow breaks or crashes the application. For example, we not only checked that each skill was working, but we also checked to ensure that the user would be able to jump from one skill type to another without any errors. Testing on a voice stage is an incredible learning experience for a test engineer because of its exceptional testing measure and the sudden discoveries that accompany it. Each task increases the value of the stage and making testing procedures. Testing front line innovation requires

tolerance, various degrees of imagination and procedures to ensure the expertise is prepared for the most exceptional client.

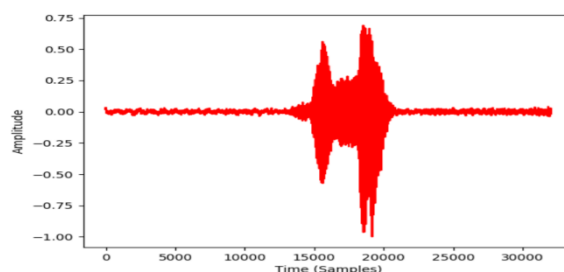
Voice acknowledgment innovation is the future and by discovering and tackling issues with through testing different voice applications, we will empower this innovation to be valuable and reliable for its clients.

## 8. Implementation of project

Speech Recognition and, comprehension of unconstrained discourse has been an objective of examination since 1970. It is an interaction of transformation of discourse to message. The object of human discourse isn't simply an approach to pass on words starting with one individual then onto the next yet additionally to make the other individual to comprehend the profundity of the expressed words.

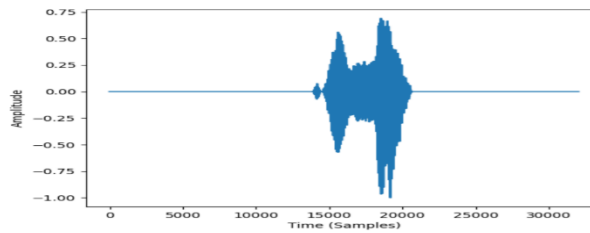
For understanding discourse human not just consider for data passed to the ears yet in addition judge the data by the setting of the data. That is the reason human can undoubtedly comprehend the communicated in language pass on to them indeed, even in loud climate. Perceiving discourse by machine is so hard for the dynamic attributes of communicated in dialects. Individuals utilized various methodologies for robotized discourse acknowledgment framework.

For perceiving discourse individuals consistently favor English as a large portion of the research and executed for them. It is a region where a great deal to contribute for our language to build up in PC field. We used Spectral Subtraction method to reduce the additive noise from the speech signal. Based on testing the various level of threshold, we we able to remove the additive noise from the signal and have a mostly clean signal.



**Figure 8: Noisy signal**





**Figure 9: Refined Sound Signal**

During the preparation test age and real testing, we cut back the quiet part furthermore, extricated just the voiced district. By breaking the example into pieces and dependent on the edge level, we separated just the voiced sound.

```
In [7]: def load_data(test_size=0.2):
        x,y=[],[]
        for file in glob.glob('C:/Users/hp/Desktop/SER/Speech_Emotion_Detection-master/speech-emotion-recognition-ravdess-data/Actor_
            file_name=os.path.basename(file)
            emotion=emotions[file_name.split("-")[2]]
            if emotion not in observed_emotions:
                continue
            feature=extract_feature(file, mfcc=True, chroma=True, mel=True)
            x.append(feature)
            y.append(emotion)
        return train_test_split(np.array(x), y, test_size=test_size, train_size= 0.75,random_state=9)

In [8]: import time
        x_train,x_test,y_train,y_test=load_data(test_size=0.25)

In [9]: print((x_train.shape[0], x_test.shape[0]))

(1838, 613)
```

**Figure 10: Training the dataset**

```
In [21]: new_feature= extract_feature("C:/Users/hp/Desktop/SER/Speech_Emotion_Detection-master/03-01-02-01-02-06.wav", mfcc=True, chrom
        new_feature.shape
        Emotion_Recognition_Through_Speech.predict([new_feature])

Out[21]: array(['calm'], dtype='<U9')
```

**Figure 11: Emotion of given speech is displayed**

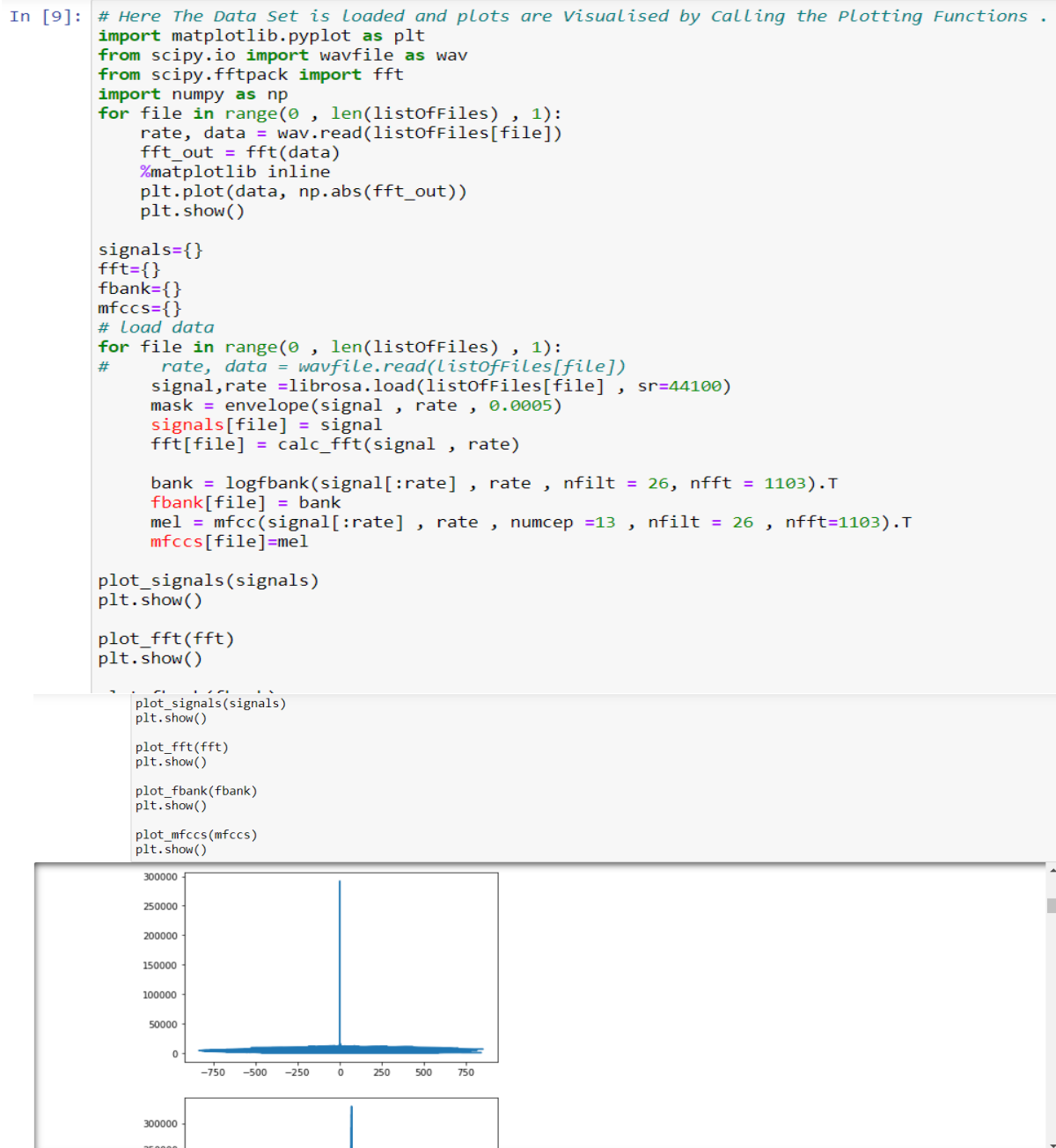
Result will be shown as a output in which emotion of that audio will be shown.

## 8.1. Conversion Plan

Speech Recognition has gotten vital in this day and age. With the progressions in innovation and upgrades in acknowledgment calculations, discourse has gotten one of

the essential wellsprings of contribution for some applications. Discourse is the most effective and common method of correspondence. Along these lines, it is natural that discourse acknowledgment frameworks have discovered applications in different fields

Here down we mentioned the code which helps in converting the audio files to plots and Visualization



**Figure 12,13: Implementation of code**

## **8.2. LIMITATIONS AND FURTHER WORKS**

It has been an exceptionally intriguing field with regards to research and technology. Many specialized groups all throughout the planet are cooperating to get the agreeable result. Several explores are continuous in this field and because of progression in innovation and effective new models it has made conceivable to make further improvement in this field to get more precise results. Like numerous different ventures our undertaking has additionally got restrictions and the upgrades that can be made in future.

### **8.3. Limitations**

Narrow Recognition Domain : Currently the IVR framework with discourse acknowledgment framework chips away at exceptionally slender area. In view of time impediment and trouble in gathering the information tests at present we are centred around utilizing just Nepali numbers from zero to nine in the system.

The precision level of acknowledgment is reliant upon accessible number of preparing tests yet because of inaccessibility of preparing information as it where less information are being prepared and perceived.

Offline operation: Another impediment of the current framework is that it is planned as it were for the disconnected activity for example accessible just on work area climate however not on the web. Narrow application domain: Our current framework is centered uniquely in execution discourse acknowledgment on mechanizing a basic assignment in desktop environment.

### **8.4. Future Enhancements**

By increasing the training data samples using effective data collection mechanism the domain of recognition can be increased. The system may be enhanced to make work for online mode by integrating it in web applications. The system can be enhanced to apply on the real time applications using telephone. For this further research on domain is necessary.

### **8.5. Software Maintenance**

Advances in discourse-based innovations have arisen to furnish PCs with the capacity to cost-adequately perceive and integrate discourse. Furthermore, remote interchanges have climbed to where the quantity of cell phones will obscure land-based telephones

and the Internet has become a typical correspondence instrument for organizations. The juncture of these advances forecasts fascinating freedoms for upkeep the executives.

Upkeep, by its actual nature, is an exceptionally versatile action. This portability necessity compels an expert's capacity to get and give data that can improve efficiency, decrease costs, and improve generally the executives of the upkeep interaction. When the laborer adventures past their wired climate, their alternatives to access data assets lessen.

An extra factor is that upkeep laborers, alongside other "talented" specialists like doctors, regularly see PC innovation as unessential to the current task. Their fondness towards, just as openness to PC applications is regularly not exactly ideal for keeping up and completely using data assets.

Associations understand that administration of the upkeep interaction can prompt tremendous expense investment funds just as efficiency enhancements. Patterns towards arranged and booked upkeep projects to improve productivity and adequacy have incited support associations to convey electronic upkeep the board frameworks (CMMS). While a CMMS is the spine for robotizing the administration approach, proficiently catching information and giving simple admittance to framework clients are key achievement factors for the framework.

A portion of the significant discourse innovation programming sellers in the market incorporate IBM, Nuance Communications, Speech Works International, and Motorola. Data on the ROI benefits identified with discourse-based applications can be downloaded from their sites.

Significant associations in the aircraft, monetary assistance, and transportation markets have effectively sent discourse-based applications to give more elevated levels of client support just as save a great many dollars in working expenses. A portion of the business fragments that could profit by discourse empowered upkeep the executives incorporate Utilities, Municipal Transportation Authorities, Food and Beverage, Academic foundations, Medical offices, and Retail chains. As associations hope to make their labor forces more productive by carrying out a CMMS, they need to address the openness of these applications by a versatile labor force. While equipping them with

PCs, and other hand-held gadgets can help, the ease, convenience, and universality of voice correspondences presents an elective that will be difficult to overlook.

Associations that need to guarantee that their interests in CMMS applications are completely used should start taking a gander at how discourse empowered support the board can stretch out current applications to laborers that are exceptionally portable and need to remain as such.

## **9. PROJECT LEGACY**

### **9.1. Current Status of the Project**

The Project has undergone various test with many numbers of samples and the result provided are almost accurate. More the number of samples provided more accurate result would be given as an output. The project is able to make predictions based on the RAVDESS dataset. Currently, it can detect 7 emotions : Anger, Calm, Happy, Disgust, Surprise, Sad and Fear. If the model can have an improvement with additional research, more emotions can be generated. The project is in a ready-go state to be used by a user to detect whether his tone of speech in any of these seven emotions.

### **9.2. Remaining Areas of Concern**

The Project comprises of efficient model to detect the emotions already so if the model can be researched more, then more emotions can play a role in it. If further improved, this project has immense possibilities in the areas like in Call Centre to detect customer's emotion, voice based virtual assistance like Siri & Alexa etc. Since the emotions are calculated as distribution on gender basis, the people who are not categorised as male and female (Transgender's) voice can be identified if an aggressive research happens behind this project. Making this project into a product can also help as an emergency system through voice, for an example, in a home, if a person is happened to cause a terrible injury where he is not able to walk, he can scream at the product asking help. The product would detect his emotion & emergency and would proceed with further steps like calling nearby hospital, ambulance or even police if a robbery happens.

### 9.3. Technical and Managerial Lessons Learnt

Through this project, we were able to study how speech emotion can play a vital role in the advanced technology of this era. Project helped in understanding the emotion distribution by gender, variation in energy across each emotion, the role of pitch in an emotion etc. We were able to learn about features like MFCC, Mel and Chroma. Representation of emotions was an important study as we are classifying emotions in labels like Happy, Sad, Anger etc. Furthermore, we were able to acquire knowledge about the libraries used in this project, about the RAVDESS dataset, how to load data, produce the texts of a particular audio file, masking an audio to reduce unnecessary voices, splitting the data, plotting graphs for the audios, how to do feature extraction in an audio, Training and testing the model etc.

## 10. User Manual

The Project can be divided into various steps of activities so a user would feel easy to navigate :

*1. Installation of Libraries and Packages :* Libraries and Packages like os, glob, tqdm, pandas, numpy, matplotlib.pyplot, scipy, python\_speech\_features, librosa, noisereduce, tensorflow, keras, sklearn were being used.

*2. Getting the list of data files and identifying the Path :* The directories of the datasets are being discovered with the help of functions in libraries like os and glob. We would get the list of all files in the directory ( def getListOfFiles(dirName) : ). Proving a list of files and sub directories and iterate over all the entries.

*3. Speech Recognition to Texts :* With the help of speech\_recognition, which is a library to perform speech recognition, we could get the text of each audios given in the dataset. Some accurate and some inaccurate results will be given but the speeches which were not able to recognise would be shown as an error.

*4. Mask/ Cleaning of Data :* The down sampling of the voices would be done and the masking process would be undergone as to clean the data. Here, Cleaning the data means to reduce the unwanted noises in the audio file. This method is termed as Mask. After the Masking is done, the data would be directed into a clean folder for later use.

5. *Plot the Audios as Graphs* : Plotting the audio files as graphs helps the user to identify the energy distribution that happens in each emotion. As an example, in an anger voice, the pitch and energy would be high compared to calm/neutral state. This process can be done by loading the file and then assigning the matplotlib.pyplot library.

6. *In-Depth Visualisation of Data* : Feature extractions like Mel Frequency Cepstral Coefficients (MFCC) would be plotted, which is to extract the features of the data (voices). Calculations are carried out on the basis of signals provided in each audio files. Plotting functions would be loaded and signal rates would be identified using librosa package. Here, mask function would also be applied as to clean the data.

7. *Redirecting of Cleaned Data* : The clean data which was produced after the step above would be redirected to a Clean Audio Folder. Signal rate using Librosa package would be once again performed at this step, just like the previous step.

8. *Feature Extraction* : The features of the Audio Files would be taken out using the MFCC, Mel and Chroma of the Audio data. Audio library such as Soundfile would be used to read and write the sound files. Soundfile shows the audio files as a NumPy array. Chroma, MFCC and Mel are all features that comes under librosa package. So, the features are being checked and the results are declared.

9. *RAVDESS Dataset and Classification*: RAVDESS is a dataset which comprises hundreds of Vocal datas. It stands for Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). It is the dataset which we used in our project. Emotions in this dataset are being classified as Calm, Sad, Happy, Fearful, Surprised, Disgust and Angry based on the energy of the each audio files.

10. *Split the data* : Using the training and testing functions, loaded data would be split according to the features in each audio. Here, libraries and packages such as librosa and numpy are being used. First training of the model will be taken and then testing happens, this pattern is followed in every audio files.

11. *Multi-Layer Perceptron Classifier or MLP Classifier* : MLP Classifiers are used to optimise the loss functions. Minimum the loss function, better the results will be. MLP Classifier is functioned in hidden layers using the MLP supervised learning Algorithm. The function of MLP Classifier in this project is to provide the hidden layer value,

alpha, batch size, learning rate and maximum iteration. After the initialisation of MLP Classifier the model is being trained.

*12. Saving, Loading and Predicting the Model* : Pickle module is used to Save the model once the training of the model is done. Loading the model from the file can be done by 'open' function. After the model is being loaded, it is then predicted. Prediction shows the result in arrays with each emotions listed.

*13. Store to CSV file* : The probability results that we had found through prediction are now being saved into a CSV format file. In this file, the file name with the prediction results assigned to serial number are being given. Libraries like NumPy and Pandas are being used here.

*14. Recording the User voice* : In order to get a live recording of the voice and provide the output as speech emotion detection, the audio is being reordered in a file with a file name. Library like Pyaudio which is used to take the input audio for the recording and the wave module for a convenient interface for WAV format are being used.

*15. Live Voice with Prediction Result* : After the recording, the live voice of the user would be saved and the system starts functioning from step 1 to 13 purely based on this live audio provided with the training history from RAVDESS dataset. Based on the Accuracy, the accurate result will be published.

## **11. Source Code**

```
pip install librosa soundfile numpy sklearn pyaudio

import librosa

import soundfile

import os, glob, pickle

import numpy as np

from sklearn.model_selection import train_test_split

from sklearn.neural_network import MLPClassifier

from sklearn.metrics import accuracy_score
```



```

def extract_feature(file_name, mfcc, chroma, mel):

    X, sample_rate = librosa.load(os.path.join(file_name), res_type='kaiser_fast')

    if chroma:

        stft=np.abs(librosa.stft(X))

        result=np.array([])

    if mfcc:

        mfccs=np.mean(librosa.feature.mfcc(y=X, sr=sample_rate, n_mfcc=40).T,
axis=0)

        result=np.hstack((result, mfccs))

    if chroma:

        chroma=np.mean(librosa.feature.chroma_stft(S=stft, sr=sample_rate).T,axis=0)

        result=np.hstack((result, chroma))

    if mel:

        mel=np.mean(librosa.feature.melspectrogram(X, sr=sample_rate).T,axis=0)

        result=np.hstack((result, mel))

    return result

emotions={

    '01':'neutral',

    '02':'calm',

    '03':'happy',

    '04':'sad',

    '05':'angry',

    '06':'fearful',

```

```

'07': 'disgust',

'08': 'surprised'

}

# Emotions to observe

observed_emotions=['neutral','calm','happy','sad','angry','fearful','disgust','surprised']

def load_data(test_size=0.2):

    x,y=[],[]

    for file in glob.glob('C:/Users/hp/Desktop/SER/Speech_Emotion_Detection-master/speech-emotion-recognition-ravdess-data/Actor_*/*.wav'):

        file_name=os.path.basename(file)

        emotion=emotions[file_name.split("-")[2]]

        if emotion not in observed_emotions:

            continue

        feature=extract_feature(file, mfcc=True, chroma=True, mel=True)

        x.append(feature)

        y.append(emotion)

    return train_test_split(np.array(x), y, test_size=test_size, train_size=0.75,random_state=9)

import time

x_train,x_test,y_train,y_test=load_data(test_size=0.25)

model=MLPClassifier(alpha=0.01, batch_size=256, epsilon=1e-08,
hidden_layer_sizes=(300,), learning_rate='adaptive',

max_iter=500)

model.fit(x_train,y_train)

```

```
MLPClassifier(activation='relu', alpha=0.01, batch_size=256, beta_1=0.9,
              beta_2=0.999, early_stopping=False, epsilon=1e-08,
              hidden_layer_sizes=(300,), learning_rate='adaptive',
              learning_rate_init=0.001, max_fun=15000, max_iter=500,
              momentum=0.9, n_iter_no_change=10, nesterovs_momentum=True,
              power_t=0.5, random_state=None, shuffle=True, solver='adam',
              tol=0.0001, validation_fraction=0.1, verbose=False,
              warm_start=False)
```

```
y_pred=model.predict(x_test)
```

```
accuracy=accuracy_score(y_true=y_test, y_pred=y_pred)
```

```
# Print the accuracy
```

```
print("Accuracy: {:.2f}%".format(accuracy*100))
```

```
from sklearn.metrics import classification_report
```

```
print(classification_report(y_test,y_pred))
```

```
from sklearn.metrics import confusion_matrix
```

```
matrix = confusion_matrix(y_test,y_pred)
```

```
print (matrix)
```

```
import pickle
```

```
# Save the Model to file in the current working directory
```

```
#For any new testing data other than the data in dataset
```

```
Pkl_Filename = "Emotion_Recognition_Through_Speech.pkl"
```

```
with open(Pkl_Filename, 'wb') as file:
```

```
    pickle.dump(model, file)
```

```
with open(Pkl_Filename, 'rb') as file:
```

```
    Emotion_Recognition_Through_Speech = pickle.load(file)
```

```
Emotion_Recognition_Through_Speech
```

```
A=Emotion_Recognition_Through_Speech.predict(x_test)
```

```
A
```

```
new_feature=
```

```
extract_feature("C:/Users/hp/Desktop/SER/Speech_Emotion_Detection-master/03-01-02-01-01-02-06.wav", mfcc=True, chroma=True, mel=True)
```

```
new_feature.shape
```

```
Emotion_Recognition_Through_Speech.predict([new_feature])
```

## 12. Bibliography

- [1] R. A. S. Nilofer, R. P. Gadhe, R. Deshmukh and P. V. B. Wasghmare, "Automatic emotion recognition from speech signals," *International Journal of Scientific and Engineering Research*, vol. 6, no. 4, p. 4, April 2015.
- [2] A. Rawat and P. K. Mishra, "Emotion Recognition through Speech Using Neural Network," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 5, no. 5, May 2015.
- [3] S. B. R, N. A and R. Desai, "Speech Emotion Recognition using MLP Classifier," *International Journal of Engineering Science and Computing*, vol. 10, no. 5, May 2020.