# SPEECH EMOTION RECOGNITION USING MACHINE LEARNING

A PROJECT REPORT

submitted By

**NEETHU SATHEESH**
**TVE18MCA038**

**to**

the APJ Abdul Kalam Technological University
in partial fullfilment of the requirements for the award of the degree

**of**

Master of Computer Applications


**Department of Computer Applications**
College of Engineering
Trivandrum-695016

JULY 2021

# Declaration

I undersigned hereby declare that the project report titled **"Speech Emotion Recognition using Machine Learning"** submitted for partial fulfillment of the requirements for the award of degree of Master of Computer Applications of the APJ Abdul Kalam Technological University, Kerala is a bonafide work done by me under supervision of Smt.Pooja J P, Asst.Professor. This submission represents my ideas in my words and where ideas or words of others have been included. I have adequately and accurately cited and referenced the original sources. I also declare that I have adhered to ethics of academic honesty and integrity as directed in the ethics policy of the college and have not misrepresented or fabricated any data or idea or fact or source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the Institute and/or University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree, diploma or similar title.

Place : Trivandrum **Neethu Satheesh**

Date : 30/06/2021

**DEPARTMENT OF COMPUTER APPLICATIONS**

**COLLEGE OF ENGINEERING**
**TRIVANDRUM**


## CERTIFICATE

This is to certify that the report entitled **Speech Emotion Recognition using Machine Learning** submitted by **Neethu Satheesh** to the APJ Abdul Kalam Technological University in partial fulfillment of the requirements for the award of the Degree of Master of Computer Applications is a bonafide record of the project work carried out by her under my guidance and supervision. This report in any form has not been submitted to any University or Institute for any purpose.

Internal Supervisor External Supervisor

Head of the Dept

# Acknowledgement

First and for most I thank **GOD** almighty and to my parents for the success of this project. I owe a sincere gratitude and heart full thanks to everyone who shared their precious time and knowledge for the successful completion of my project.

I am extremely thankful to **Dr Jiji C V**, Principal, College of Engineering Trivandrum for providing me with the best facilities and atmosphere which was necessary for the successful completion of this project.

I am extremely grateful to **Dr. Sabitha S**, HOD, Dept of Computer Applications, for providing me with best facilities and atmosphere for the creative work guidance and encouragement.

I express our sincere thanks to **Smt. Pooja J P**, Asst. Professor, Department of Computer Applications, College of Engineering Trivandrum for her valuable guidance, support and advice that aided in the successful completion of my project.

I profusely thank other Asst. Professors in the department and all other staffs of CET, for their guidance and inspirations throughout my course of study.

I owe my thanks to my friends and all others who have directly or indirectly helped me in the successful completion of this project. No words can express my humble gratitude to my beloved parents and relatives who have been guiding me in all walks of my journey.

**Neethu Satheesh**

# Abstract

In the increase in man to machine interaction, speech analysis has become an integral part in reducing the gap between physical and digital world. Emotion recognition from audio signal requires feature extraction and classifier training. The feature vector consists of elements of the audio signal which characterise speaker specific features such as tone, pitch, energy, which is crucial to train the classifier model to recognise a particular emotion accurately. The North American English language open source dataset was divided into training and testing manually. Speaker vocal tract information, represented by Mel-frequency cepstral coefficients (MFCC), was extracted from the audio samples in training dataset. Pitch,Short Term Energy(STE), and MFCC coefficients of audio samples in emotions angry, happy, sad,clam,disgust,surprise,fear were obtained. These extracted feature vectors were sent to the classifier model. The test dataset will undergo the extraction procedure following which the classifier would make a decision regarding the underlying emotion in the test audio.The emotion of the human is predicted by using this system

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Emotions play a vital role in human communication. In order to extend its role towards the human-machine interaction, it is desirable for the computers to have some built-in abilities for recognizing the different emotional states of the user. With the advent of technology in the recent years, more intelligent interaction between humans and machines is desired. Hence, it is desirable for machines to have the ability to detect emotions in speech signals .Human speech conveys information and context through speech, tone, pitch and many such characteristics of the human vocal system.

Emotion recognition from audio signal requires feature extraction and classifier training. The feature vector consists of elements of the audio signal which characterise speaker specific features such as tone, pitch, energy, which is crucial to train the classifier model to recognise a particular emotion accurately. The North American English language open source dataset was divided into training and testing manually.Speaker vocal information, represented by Mel-frequency cepstral coefficients (MFCC), was extracted from the audio samples in training dataset. Pitch, Short Term Energy(STE), and MFCC coefficients of audio samples in emotions angry, happy, sad,clam,disgust,surprise and fear were obtained.

# Chapter 2

# Problem Defnition and Motivation

Speech is the most common and efficient method of human communication.This reality motivate many researchers to consider speech signal as a quick and effective process to interact between computer and human. It means the computer should have enough knowledge to identify human voice and speech.Speech emotion recognition is mostly beneficial for applications, which need human-computer interaction such as speech synthesis, customer service, education, forensics and medical analysis

For a natural and effective human- machine interaction , it is important to recognize ,analyze ,and respond to the emotional state of user. Development of speech emotion recognition helps to Business Marketing, Customer care support, Digital Assistance.

# Chapter 3

# Literature Review

In today's applications, identifying the emotion exhibited in a spoken percept has various applications. Human-Computer Interaction is a branch of study that looks at how humans and computers interact with each other. As part of my literature research, I looked at a number of publications and briefings on the subject. This chapter is a concise summary of my findings.

## 3.1 Using Support Vector Machine

Emotion recognition from audio signal requires feature extraction and classifier training. Mel-frequency cepstral coefficients (MFCC) was extracted from the audio samples in training dataset. These extracted feature vectors were sent to the classifier model SVM(Support Vector Machine). SVM is a supervised learning algorithm used most widely for pattern recognition applications. The algorithm is simple to use and provides good results even when trained on limited size training dataset.Three features are extracted(Happy ,Sad ,Anger )

## 3.2 Using Random Decision Forest

Recognize emotions in speech and classify them in 7 emotion output classes namely, anger, boredom, disgust, anxiety, happiness, sadness and neutral . The proposed approach is based upon the Mel Frequency Cepstral coefficients (MFCC) and energy of the speech signals as feature inputs and uses Berlin database of emotional speech.The features extracted from speech are converted into a feature vector, which in turn are used to train different classification algorithms namely, Support Vector Machine (SVM) , Random Decision Forest and Gradient Boosting .The random

forests classifier was implemented using 15 forests. Gini impurity was used to measure the quality of the split of a node. Testing for different max depths and decided to split nodes until each leaf had samples belonging to a single class only.

## 3.3 Using Deep Belief Network

This paper analyzes different characteristics to make a better description of speech emotion. The main works of this paper are the selection of the database, the extraction of emotion features, and the selection of classificationalgorithm. Then, two methods are used to evaluate the result, including overall and average recognition rate. This paper uses contrastive divergence algorithm on emotion feature extraction. Compared with the traditional algorithms, such as support vector machine (SVM) and artificial neural network (ANN), the accuracy of test emotion sample has a better performance after feature extraction by DBN, to about 5traditional classification algorithm.

# Chapter 4

# Requirement Analysis

## 4.1 Overall Description

Emotion recognition from audio signal requires feature extraction and classifier training. The feature vector consists of elements of the audio signal which characterise speaker specific features such as tone, pitch, energy, which is crucial to train the classifier model to recognise a particular emotion accurately. The North American English language open source dataset was divided into training and testing manually. Speaker vocal tract information, represented by Mel-frequency cepstral coefficients (MFCC), was extracted from the audio samples in training dataset. Pitch, Short Term Energy(STE), and MFCC coefficients of audio samples in emotions anger, happiness, and sadness were obtained. These extracted feature vectors were sent to the classifier model. The test dataset will undergo the extraction procedure following which the classifier would make a decision regarding the underlying emotion in the test audio.

### 4.1.1 Product Functions

- Preprocessing of audio signal
- Feature extraction of audio signal
- Training and testing of model
- Save Model
- connect the UI with the model.

### 4.1.2 Hardware Requirements

- Processor : Intel Core i3
- Storage : 512 GB Hard Disk space
- Memory : 4 GB RAM

### 4.1.3 Software Requirements

- Operating System : Linux/Windows
- Platform : Python
- Librarie used : librosa,matplotlib, numpy, sklearn,keras, tensorflow

## 4.2 Functional Requirements

The functional requirements includes all the activities or processes that should be achieved by the proposed system. It includes

- **librosa:**Librosa is a robust Python package for working with and analysing audio. It's the first step toward working with audio data at scale for a variety of applications, from identifying a person's voice to extracting personal features from audio.It provides the necessarybuilding elements for the development of music information retrieval systems. Librosa aids in the visualisation of audio signals as well as feature extractions utilising various signal processing techniques.
- **sklearn:**By using this library,implements various regression, classification and clustering algorithms such as random forest, support vector machine, k-means and DBSCAN. And the sk learn library is built in a way that it can work with various scientific and numeric libraries of python such as scipy and numpy.
- **matplotlib:** It's used for the visualisation of data in python programming language. It's implemented to work with the wider scipy stack and it's built on numpy arrays. It's a multi platform data visualization technique. It was developed in 2002 by John Hunter. Visualization is the most efficient way to understand the data. Using this library, It can represent our data in various plots such as line, bar, histogram, scatter etc.

- **keras:** Keras is a lightweight Python deep learning package that may be used with Theano or Tensor Flow. It was created to implement deep learning models for research and development as simple and quick as feasible. It operates on Python 2.7 or 3.5 and, thanks to the underlying frameworks, can run on both GPUs and CPUs.

## 4.3 Non Functional Requirements

### 4.3.1 Performance Requirements

- Accuracy : Accuracy in functioning and the nature of user-friendly should be maintained by the system.
- Speed : The system must be capable of offering speed.
- Low cost: This system is very cheap to implement and is also user-friendly.
- Less Time consuming: It uses very less time comparing to the existing sysytem .
- User Friendly: This proposed system is highly user friendly they enables to create a good environment.

### 4.3.2 Quality Requirements

- Scalability : The software will meet all of the functional requirements.
- Maintainability : The system should be maintainable. It should keep backups to atone for system failures, and should log its activities periodically.
- Reliability : The acceptable threshold for down-time should be large as possible. i.e. mean time between failures should be large as possible. And if the system is broken, time required to get the system backup again should be minimum.
- Availability: This system is easily available as the core equiments in building the sofware is easily obtained.
- High- Functionality: This system is highly functional in all environment since, They are highly adaptable.

# Chapter 5

# Design And Implementation

The proposed system is used to Identify Emotions of audio by using a pre-trained model. The model is trained using MLP Classifiers upon the features extracted by MFCC.

## 5.1 Overall Design

The proposed system follows client server architecture. That is the predict the emotion has a client part and a server part as well. The client part is used by the user to input the audio which is to be prdicted. The input is passed to server and the evaluated result is given back to the client. The server side is developed in Python and the client side is built using HTML and Python.

### 5.1.1 System Design

The system is web based. The input is taken from the user through a web page and the input is passed to the python program running in the server side. The server program perform tasks such as pre processing and feature extraction on the input data. The results of these processes are used to evaluate the input using the pre trained model.

The model is created using the data obtained from Kaggle.com. They provide approximately 1500 audio in English language. These audio are divided into 24 folderes - each set of folder contain different context - to ensure variability of the domain. It also have emotion as label.

### 5.1.2 Methodology

There are two parts in this project. The first part is the creation of the model and the second one is the creation of user program which will work with the pre-trained model.

The main process of the automated essay scoring is the creation of the trained model. The major steps in the model creation Feature extraction, training, testing and model evaluation. The major steps in the model creation are mentioned below.

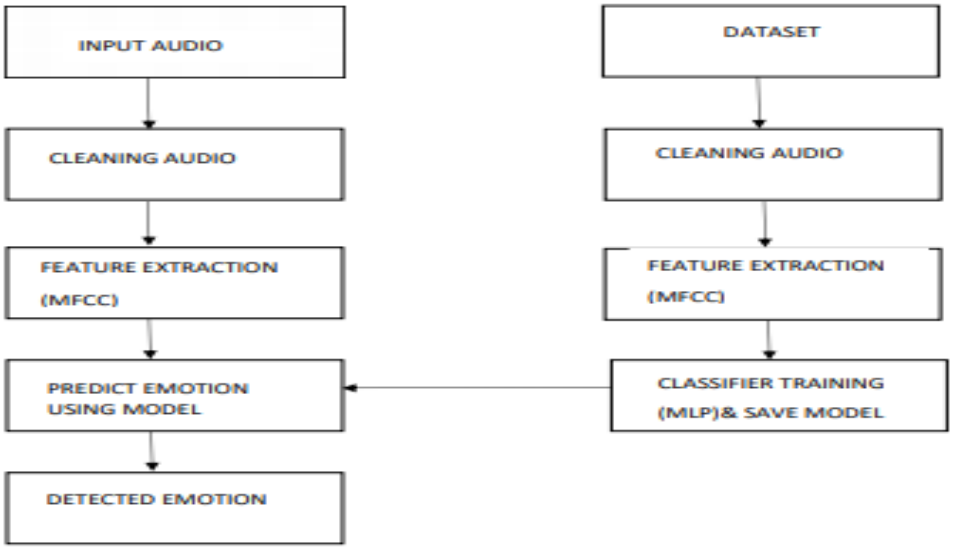


Figure 5.1: Architecture of model Creation

- **Data preprocessing:**Data preprocessing is process of removing unwanted noices from audio.

- **feature extraction:**Feature extraction is the most important part of any machine learning task and so is the case with us. Here specify the MFCC features of the audio and they are extracted using MFCC. Extract the features such as MFCC,mel,chroma etc. as the features .

- **Training:**Multi Layer Perceptron Classifier is the technique used for the learning of the system.It's one of the widely used supervised learning techniques It uses the labelled dataset along with the extracted features to generate the model. A portion of our dataset is used for training. The remaining portion is for testing the model.

- **Testing:** In testing phase test the generated model with the remaining portion of the

dataset. The data set is fed into the generated model and their results are recorded for the next stage which is the model evaluation and error analysis.

- **Model evaluation and error analysis:** The results of the testing data along with their original values are used for the error calculation. Various statistical measures can be adopted for calculating the efficiency of the model. The various measures are accuracy, precision, recall, kappa score etc. If the results of these quantitative analyses are acceptable, then move forward with the generated model. If the results are poor, the model is to be regenerated with more features so that the best result is obtained.

The second part of the project is to build the user interface. The user interface is build using HTML and Python. This is the part of project which deals with the user. The input audio is fed into the server through this. And the results returned are also displayed in the user program. The interface is built in a way such that it is easy and understandable for the person who uses it. For that uses responsible HTML designs which uses CSS and JavaScript also to provide the better user experience. Python Flask is also used for the development of user interface.

## 5.2 Data Flow Diagram

DFD is one of the graphical representation techniques used in a project to show the flow of the data through a project. DFD helps us to obtain an idea about the input, output, and process involved. The things absent in a DFD are control flow, decision rules, and loops. It can be described as a representation of functions, processes that capture, manipulate, store, and distribute data between a system and the surrounding and between the components of the system. The visual representation helps for good communication.

It shows the journey of the data and how will it be stored in the last. It does not provide details about the process timings or if the process shall have a parallel or sequential operation. It is very different from a traditional flow chart or a UML that shows the control flow or the data flow.

In level 0 the basic data flow of the application is showcased. It does not show the flow of data much deeper. It will be evaluated in the higher levels of Data Flow Diagram. The Data Flow Diagram of Automated essay scoring system is shown below.

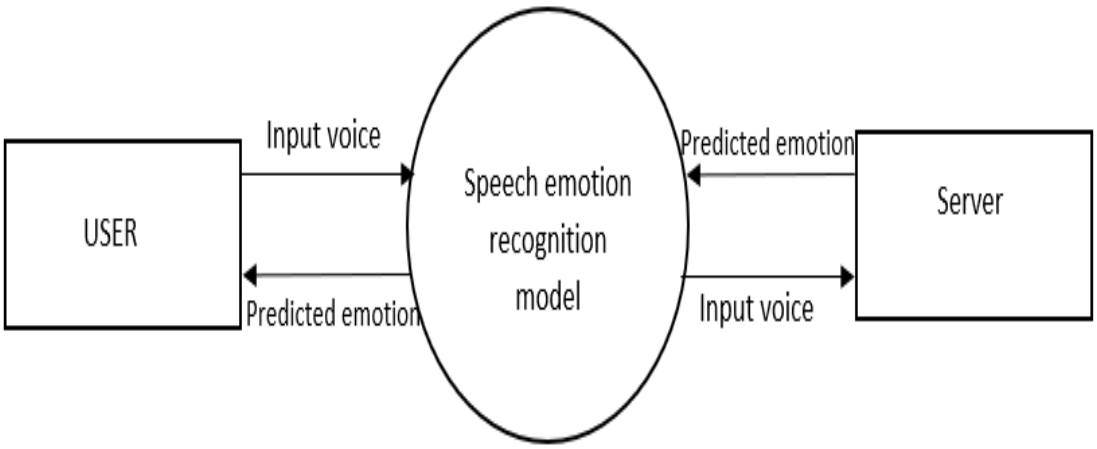


Figure 5.2: Level 0 DFD

The diagram shows Level 0 Data flow diagram of the Speech emotion recognition. As the diagram indicates there is a user part and an Server. The input of the project is the voice by the user and which is given to the Speech emotion recognition.The input is transferred to Server side program. The feature extraction and emotion prediction is occurred in the Server side. The predicted emotion is passed back to the user through the application. This is how the data flows through the application. Since there is no database in the application, the data is not stored anywhere. The data is lost after the prediction.

## 5.3 Screenshots of user interface

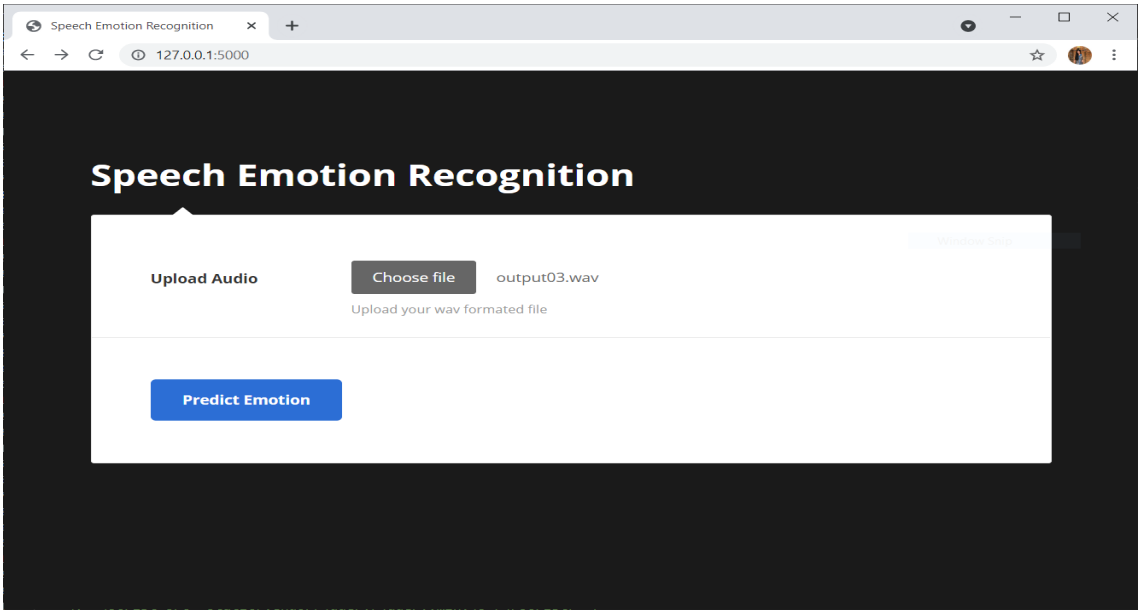


Figure 5.3: input - uploading audio

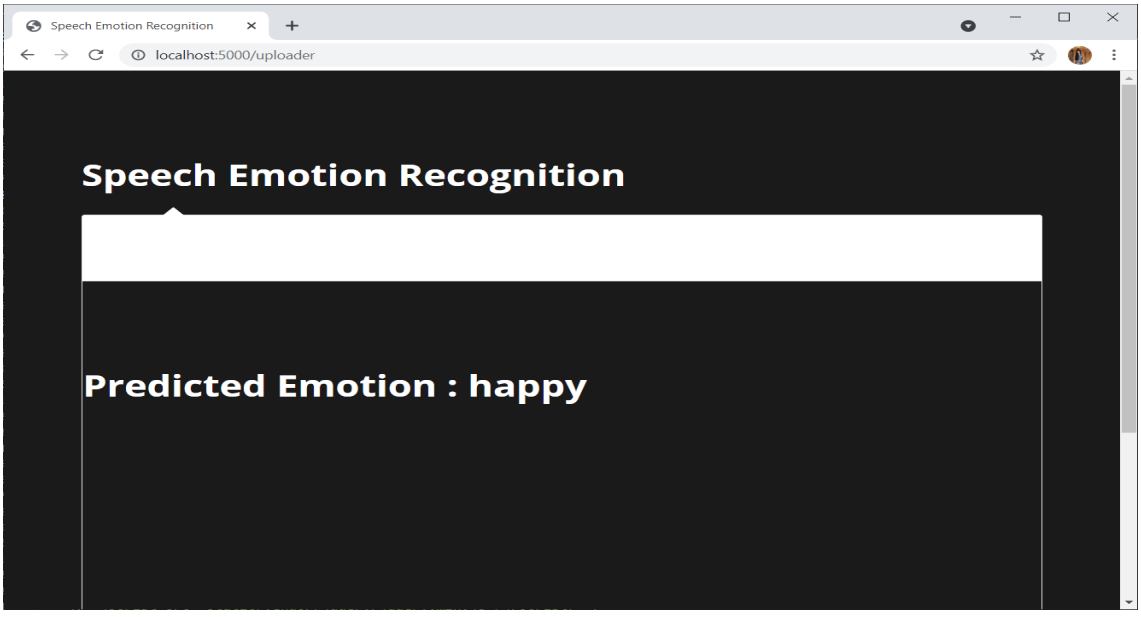


Figure 5.4: output - predicted emotion

# Chapter 6

# Coding

**Algorithm 1** Algorithm for Creating the model:

1: Split the data set into training data set and testing dataset. 80% of the dataset is used for training and the remaining 20% is used for testing to obtain better result.

2: The training dataset is used for the preprocessing stage and the preprocessed data is further used to extract the features. The features of the input voice is obtained using librosa libraries.

3: The extracted features are used to create the model.

4: The model is created using MLP Classifier with the selected features. The MLP Classifier model evaluates how much the dependent values depend upon the independent values.

5: The testing dataset is feeded into the created model and their results are noted down.

6: The result of testing dataset evaluated using the created model is then compared with the actual values of the testing dataset to evaluate the efficiency of the model. Various statistical measures such as Accuracy, Precision, Recall, Kappa score etc. can be used to evaluate the model.

7: Further tuning is performed upon the created model to improve the efficiency of the model.

**Algorithm 2** Algorithm for web Application and Speech Emotion Recognition:

1: Read the input voice from the user through the user interface.

2: On button click the value in the web page is passed to the server program for the Prediction of voice.

3: From the server program, access the input Voice and perform the preprocessing tasks on it.

4: The preprocessed voice is used to extract the required features from it using librosa libraries.

5: Using the pretrained model,evaluate the input voice using the extracted features.

6: The Emotion is predicted by the results of the model and the Emotion is passed to the web page.

7: The Emotion is displayed in the web application.

# Chapter 7

# Testing and Implementation

## 7.1 Testing and various types of testing used.

Once a software is developed, the major activity is to test whether the actual results match with the experimental results. This process is called testing. It's used to make sure that the developed system is defect free. The main aim of testing is to find the errors and missing operations by executing the program. It also ensure that all of the objectivs of the project are met by the developer. The objective of testing is not only to evaluate the bugs in the created software but also finding the ways to improve the efficiency, usability and accuracy of it. It aims to measure the functionality, specification and performance of a software program. Tests are performed on the created software and their results are compared with the expected documentation. When there are too much errors occurred, debugging is performed. And the result after debugging is tested again to make sure that the software is error free. The major testing processes applied to this project are unit testing, integration testing and system testing. In unit testing, our aim is to test all individual units of the software. It makes sure that all of the units of the software works as it intended. In integration testing, the combined individual units are tested to check whether it met the intended function or not. It helps us to find out the faults that may arise when the units are combined. In system testing the entire software is tested to make sure that it satisfies all of the requirements. The tables shown below describes the testing process occurred during the development of this project "Speech Emotion Recognition". This defines the various steps took to create the project error free.

### 7.1.1 Unit Testing

**Test Cases and Result**

| Sl No | Procedures | Expected result | Actual result | Pass or Fail |
|---|---|---|---|---|
| 1 | pre-processing | clean the dataset for feature extraction | same as expected | Pass |
| 2 | extract features from dataset | extract various features like MFCC,mel,chroma | same as expected | Pass |
| 3 | training and testing of model | create the model and store it in a pickle file | pickle file generated | Pass |
| 4 | prediction | predict the result accurately | same as expected. | Pass |
| 5 | python server program | set up a python flask server to run the program | Same as expected | Pass |
| 6 | create the user interface | To load the web page with required fields | Same as expected | Pass |

Table 7.1: Unit test cases and results

### 7.1.2 Integration Testing

**Test Cases and Result**

| Sl No | Procedures | Expected result | Actual result | Pass or Fail |
|---|---|---|---|---|
| 1 | load the user interface from python | the user interface is loaded when the flask program is run | Same as expected | Pass |
| 2 | pass input voice from web page to server | To pass the input voice entered by the user to the python program to and receive it there. | Same as expected | Pass |
| 3 | Emotion prediction | load the previously generated pickle file to the server and predict the emotion with it and extracted features. | Same as expected | Pass |
| 4 | display results | pass the result to web page and display it there | Same as expected | Pass |

Table 7.2: Integration cases and result

### 7.1.3 System Testing

**Test Cases and Result**

| Sl No | Procedures | Expected result | Actual result | Pass or Fail |
|---|---|---|---|---|
| 1 | to run python server | Server program executed successfully, hence the entire program worked without any crash | Same as expected | Pass |
| 2 | Emotion prediction | allow user to input voice and output generated according to the input voice. | Same as expected | Pass |

Table 7.3: System test cases and results

# Chapter 8

# Results and Discussion

The main aim of the project was to predict the emotion of the voice with a machine learning model. And it is observed that the system performs all the functionalities as expected.

## 8.1 Accuracy

Accuracy of system is calculated from comparing expected and predicted Emotions.

```
#from sklearn.matrics import accuracy_score
print("Accuracy Score :",accuracy_score(y_test,y_pred))
```

```
Accuracy Score : 0.8364779874213837
```

Figure 8.1: accuracy

## 8.2 Advantages and Limitations

The proposed system is a machine learning model to evaluate the input voice and predict it's emotion. The proposed system posses more advantages over the existing system. Like every other system, this system also have it's own disadvantages. But they are negligible while comparing with the advantages and they can be overcame in future.

### 8.2.1 Advantages

- Speech emotion of humans are detected with high accuracy

- The Human resource needed for the evaluation can be saved.
- Can save time needed for the prediction of emotion.

### 8.2.2 Limitations

- The current dataset is comparatively small. hence the results are neither too good nor too bad. It can be improved by improving dataset

# Chapter 9

# Conclusion and Future Scope

Speech emotion recognition is a very use full machine learning application to the current environment. It can be done in various methods. Through this project, I tried to develop a machine learning model to evaluate and predict the Emotion of human voice. The speech emotion recognition predict 7 emotion namely, angry,calm, disgust, surprised, happy, sad and neutral

The results obtained by the created model seems encouraging and can be improved in future. The rate of errors in the machine learning model is very minimum . The majority of the project was built in python. It uses a flask server to connect to the user interface built using HTML, JavaScript and CSS. The project was built with the help of various python libraries such as librosa,soundfile,matplotlib, sklearn, pandas, numpy etc.

The feature scope of this particular machine learning model can be extended to multiple dimensions.Predict more emotions and change the language of speech by adding additional datasets.And also change the model to predict real-time speech emotions as well.

# Bibliography

[1] Girija Deshmukh, Apurva Gaonkar, Gauri Golwalkar, Sukanya Kulkarni ,"Speech based Emotion Recognition using Machine Learning" Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019).

[2] Mohan Ghai, Shamit Lal, Shivam Dugga l and Shrey Manik, "Emotion Recognition On Speech Signals Using Machine Learning" 2017 International Conference On Big Data Analytics and computational Intelligence (ICBDACI).

[3] Ruhul Amin Khalil , "Speech Emotion Recognition using Deep Learning Techniques",2018

[4] Ayadi M E, Kamel M S, Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases[J]. Pattern Recognition, 2011.