Mini Project Report

On

"SPEECH EMOTION RECOGNITION USING MACHINE LEARNING"

Submitted in partial fulfillment of the

Requirements for the award of the degree of

Bachelor of Technology

In

Computer Science & Engineering

By

P. Harika Reddy – 17R21A05G3
Md Nizamuddin– 17R21A05F2
Md Abdul Mujeeb – 17R21A05F1

Under the guidance of

R. Anusha
Assistant Professor

Department of Computer Science & Engineering

MARRI LAXMAN REDDY GROUP OF INSTITUTIONS
MLR INSTITUTE OF TECHNOLOGY
(UGC AUTONOMOUS)
Affiliated to JNTUH, Approved by AICTE
Laxman Reddy Avenue, Dundigal, Hyderabad-500 043, Telangana, India
NBA NATIONAL BOARD OF ACCREDITATION
A NAAC ACCREDITED WITH GRADE

2021

MARRI LAXMAN REDDY GROUP OF INSTITUTIONS
MLR INSTITUTE OF TECHNOLOGY
(UGC AUTONOMOUS)
Affiliated to JNTUH,Approved by AICTE
Laxman Reddy Avenue, Dundigal, Hyderabad-500 043, Telangana, India
NBA NATIONAL BOARD ACCREDITATION
NAAC

Department of Computer Science & Engineering

CERTIFICATE

This is to certify that the project entitled "**SPEECH EMOTION RECOGNITION USING MACHINE LEARNING"** has been submitted by **P. Harika Reddy (17R21A05G3), Md Nizamuddin (17R21A05F2), Md Abdul Mujeeb (17R21A05F1)** in partial fulfillment of the requirements for the award of degree of Bachelor of Technology in Computer Science and Engineering from MLR Institute of Technology, Hyderabad. The results embodied in this project have not been submitted to any other University or Institution for the award of any degree or diploma.

Internal Guide **Head of the Department**

External Examiner

MARRI LAXMAN REDDY GROUP OF INSTITUTIONS
MLR INSTITUTE OF TECHNOLOGY
(UGC AUTONOMOUS)
Affiliated to JNTUH,Approved by AICTE
Laxman Reddy Avenue, Dundigal, Hyderabad-500 043, Telangana, India
NBA NATIONAL BOARD ACCREDITATION
NAAC

Department of Computer Science & Engineering

DECLARATION

We hereby declare that the project entitled **"SPEECH EMOTION USING MACHINE LEARNING"** is the work done during the period from **January 2020 to May 2020** and is submitted in partial fulfillment of the requirements for the award of degree of Bachelor of Technology in Computer Science and Engineering from MLR Institute of Technology, Hyderabad. The results embodied in this project have not been submitted to any other university or Institution for the award of any degree or diploma.

P. Harika Reddy	**17R21A05G3**
Md Nizamuddin	**17R21A05F2**
Md Abdul Mujeeb	**17R21A05F1**

Department of Computer Science & Engineering

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of people who made it possible, whose constant guidance and encouragement crowned our efforts with success. It is a pleasant aspect that I now have the opportunity to express my guidance for all of them.

First of all We would like to express my deep gratitude towards my internal guide **R.Anusha, Assistant Professor, Department of CSE** for her support in the completion of our dissertation. We wish to express our sincere thanks to **Dr. N. CHANDRASHEKHAR REDDY,**HOD, Dept. of CSE and also principal **Dr. K. SRINIVASA RAO** for providing the facilities to complete the dissertation.

We would like to thank all my faculty and friends for their help and constructive criticism during the project period. Finally, we are very much indebted to our parents for their moral support and encouragement to achieve goals.

P. Harika Reddy **17R21A05G3**

Md Nizamuddin **17R21A05F2**

Md Abdul Mujeeb **17R21A05F1**

Department of Computer Science & Engineering

ABSTRACT

Speech Emotion Recognition, abbreviated as SER, is the act of attempting to recognize human emotion and affective states from speech. This is capitalizing on the fact that voice often reflects underlying emotion through tone and pitch. This is also the phenomenon that animals like dogs and horses employ to be able to understand human emotion. SER is tough because emotions are subjective and annotating audio is challenging.In machine interaction with human being is yet challenging task that machine should be able to identify and react to human non-verbal communication such as emotions which makes the human computer interaction become more natural. In present research area automatic emotion recognition using speech is an essential task which paid close attention.Speech signal is a rich source of information and it is an attractive and efficient medium due to its numerous features of expressing approach & extracting emotions through speech is possible. In most of In this paper emotions is recognized through speech using spectral features such as Mel frequency spectrum co-efficient prosodic features like pitch , energy and were utilized & study is carried out using MLP classifiers which is used for detection of six basic emotional states of speaker's like anger ,happiness, sadness, fear, disgust and neutral using RAVDESS dataset.

INDEX

1. INTRODUCTION

1.1 Overview

As human beings speech is amongst the most natural way to express ourselves. We depend so much on it that we recognize its importance when resorting to other communication forms like emails and text messages where we often use emoji's to express the emotions associated with the messages. As emotions play a vital role in communication, the detection and analysis of the same is of vital importance in today's digital world of remote communication. Emotion detection is a challenging task, because emotions are subjective. There is no common consensus on how to measure or categorize them. We define a SER system as a collection of methodologies that process and classify speech signals to detect emotions embedded in them. Such a system can find use in a wide variety of application areas like interactive voice based-assistant or caller-agent conversation analysis. In this study we attempt to detect underlying emotions in recorded speech by analyzing the acoustic features of the audio data of recordings.

Emotion is a medium by which one expresses how a person feels and one's state of mind. Emotions play an important factor in sensitive job areas, like that of a surgeon, a Military Commander and many others where one has to maintain their emotions in check. Predicting emotions is a tough task as every individual has a different tone and intonation of speech. The elicit different types of emotions are happy, angry, neutral, sad and surprised. To classify these emotions from a given speech sample in the most appropriate method, is the goal of this paper. With different methods of predicting emotions we plan to use the multilayer perceptron.

1.2 Purpose of the project

Emotion recognition is the part of speech recognition which is gaining more popularity and need for it increases enormously. Although there are methods to recognize emotion using machine learning techniques, this project attempts to use deep learning to recognize the emotions from data.

SER(Speech Emotion Recognition) is used in call center for classifying calls according to emotions and can be used as the performance parameter for conversational analysis thus identifying the unsatisfied customer, customer satisfaction and so on …. for helping companies improving their services

It can also be used in-car board system based on information of the mental state of the driver can be provided to the system to initiate his/her safety preventing accidents to happen. It increases the human interaction.

1.3 Motivation

Speech Emotion Analysis refers to the use of various methods to analyze vocal behavior as a marker of affect (e.g., emotions, moods, and stress), focusing on the nonverbal aspects of speech. The basic assumption is that there is a set of objectively measurable voice parameters that reflect the affective state a person is currently experiencing (or expressing for strategic purposes in social interaction). This assumption appears reasonable given that most affective states involve physiological reactions (e.g., changes in the autonomic and somatic nervous systems), which in turn modify different aspects of the voice production process. For example, the sympathetic arousal associated with an anger state often produce changes in respiration and an increase in muscle tension, which influence the vibration of the vocal folds and vocal tract shape, affecting the acoustic characteristics of the speech, which in turn can be used by the listener to infer the respective state. This helps in various aspects, and also can be used in many different applications.

2. Literature Survey

2.1 Existing System

Speech Emotion Recognition is a subject under research. Speech emotion recognition abbreviated as SER. It creates a natural Human Computer interaction. There are various kinds of methods used to identify the emotion from the speech, such as using support vector machine (SVM), Recurrent Neural Network, K-nearest neighbour, Hidden Markov Model (HMM).

Support Vector Machine (SVM)

Support Vector Machine approach computes the audio parameters to identify the emotion and has high accuracy in predicting the emotion from the speech. But this approach can only classify the dataset into 2 classes only. That means we can only identify among the 2 emotions trained to the classifier. The other disadvantages of this approach are long processing time, background noise leading to error and it has low accuracy.

K-nearest neighbor

The other classifier is k nearest neighbour classifier. This is the simplest classification algorithm which identifies the emotion of speech. This classifier uses pitch and energy of the audio to predict emotion and has accuracy of 64% for 4 emotions audio.

Hidden Markov Model (HMM)

HMM models temporal sequencing from the audio. This modelling is useful in predicting the emotion from the speech. The main limitation of this classifier is feature selection process. As the features don’t carry the complete information of the emotion of the speech. But it has good classification accuracy compare to other classifiers.

2.2 Disadvantages of Existing System

1. It can only tell only limited number of emotions for instance only two emotions.

2. The model trained heavily depends on the language used, words used in that particular language rather than depending on features of the language such as pitch, tone, pauses etc.

3 .There are less number of features extracted from the test data which makes it difficult to predict accurate emotion on test data

4. Audio Visual enhancements are not considered in the Existing system to predict the emotion

5. Song and Speech are not differentiated distinctly in Existing system, which leads to the problem Song is confused with Speech Emotion and vice versa.

3. Proposed System

3.1 Proposed System

- The proposed system extracts features such as pitch, tone, frequency from the input audio and plot the features against properties such as MFCC, Mel and Chroma to generate a feature vector of float type 32-bit floating point integers.
- The feature vector represents the features of the audio, then those features are used to train a Multi-layer perceptron model which internally makes use of a Artificial Neural Network with binary Inputs and Outputs.
- It has 3 layers Input Layer, hidden Layer and the Output Layer, there might ne N number of layers under the hidden layer but Input and Output layers are just one, the feature vector goes through input of Artificial Neural Network and it gives us output as the observed emotion which matches the audio's features most or the closest correct emotion
- The proposed system can identify can understand seven emotions
    1. Neutral
    2. Calm
    3. Happy
    4. Sad
    5. Angry
    6. Fearful
    7. Disgust
    8. Surprised
- The Proposed System can observe and predict four emotions
    1. Calm
    2. Happy
    3. Fearful

4. Disgust

- If any emotion does not match any of the observable emotions then the closest observed emotion is predicted, emotions are recognized through speech using spectral features such as Mel frequency cepstrum coefficient
- Prosodic features like pitch , energy and were utilized & study is carried out using MLP classifiers which is used for detection of six basic emotional states of speaker's such as anger ,happiness , sadness , fear , disgust and neutral using RAVDESS dataset.
- RAVDESS dataset has recordings of 24 actors, 12 male actors and 12 female actors, the actors are numbered from 01 to 24. The male actors are odd numbered and female actors are even numbered. The emotions contained in the dataset are as sad, happy, neutral, angry, disgust, surprised, fearful and calm expressions. The dataset contains all expressions in three formats, which are: Only Audio, Audio-Video and Only Video.

3.2 Advantages of Proposed System

- The Human computer interaction will be improved and will be more natural.
- We can detect the emotion of the speaker irrespective of the language.
- Proposed system can detect 8 different emotions (calm, happy, sad, angry, fearful, surprise, and disgust expressions)
- We can detect the correct emotion of the speaker approximately since the accuracy of the model high.

3.3 System Requirements

➢ **Software Requirements**

1. Python 3 or above (with IDE)
2. Pip to install python packages
3. AI ML Libraries like librosa, sklearn

4. RAVDESS dataset (Audio-Visual Database)

➢ Hardware Requirements

1. Minimum 4GB RAM (8GB recommended)
2. Windows/UNIX based OS
3. JupyterLab (Cloud).

4. System Design

4.1 Proposed System Architecture

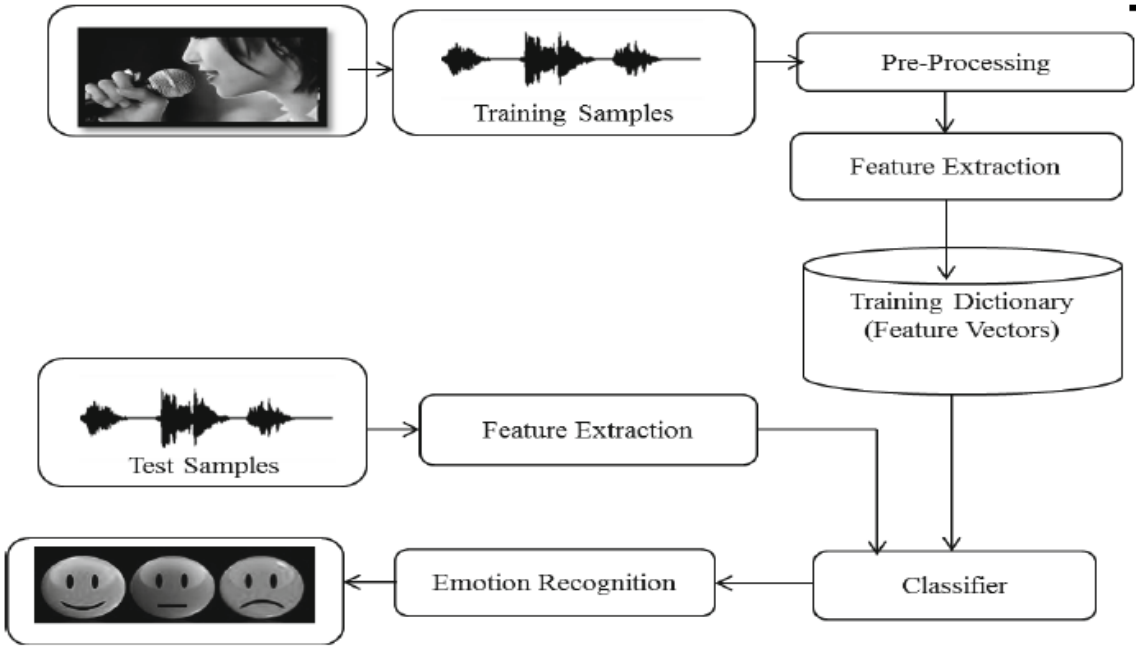

Fig 4.1.1 Architecture

In proposed system, we are using RAVDESS dataset as a data for the system. The data present in the dataset is pre-processed to clean the audio and remove the disturbance from the audio to reduce the error in the output. The audio is divided into equal time intervals frames. Then the dataset is divided into 2 parts as training data and testing data. Training data is 80% of the dataset and testing data is 20% of the dataset.The features are extracted from the audio and given to the classifier to predict the emotion.The model is created by training data inputs to the classifier then this model is tested with the testing data inputs.

We get the accuracy by calculating the output of the model and the actual emotion in dataset.

4.2 Modules

Based on the methodology, we divided the project into two modules. They are:

(1) Speech processing module

(2) Classification module.

Speech Processing:

Speech processing consists of two phases. They are:

(a) Preprocessing phase

(b) Feature extraction phase.

Let's discuss these two phases.

(a)Preprocessing

Preprocessing is the initial phase which is there after followed by feature extraction. Preprocessing includes:

Silence removal

There are many different types of silences present in a audio. To get clear audio of the person or the speaker, we need clear those silence present in the audio. So, this silent is removed by "silence removal".There are many methods to remove the silence , few of them are STE(short time energy)and ZCR(zero crossing rate). Below equation represents, short time energy method to remove the silence[6]:

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2$$

PRE-EMPHASIS

The most vital steps of preprocessing at high frequency is the pre-emphasis of the speech signal. Pre-emphasis is implemented by the following below equation[6].

$$x'(n) = x(n) - \propto x(n-1)$$

Where α is the pre-emphasis parameter whose valuelies between (0.9) and 1.

Normalization

The signal sequence division of the highest value of the signal to ensure that the each sentence has a comparable volume level is done by normalization[6].

Windowing

$$y_1(n) = x_1(n)w(n), \quad 0 \le n \le N-1 \quad N-1$$

Hamming window is one among the most popular windowing techniques. The below formula describes hamming window function [6].

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1$$

:

Feature extraction

Feature extraction is an important step in the whole process. The processing of the audio signal takes place here in feature extraction. In this module, extraction of feature vectors is carried out. Here, we have extracted three main features namely MFCC, mel Spectrogram and chroma.

(b)Classification

In our system, we are using MLP classifier for classification of the features. After completing the training process our model is created. After completion the testing, the accuracy of the model is calculated by the data given which is created by the classifier [7].

feedforward artificial neural network which is well known as ANN is super class of MLP which stands for multi-level perceptron. It has minimum of three layers which are one input layer, an

$$y = \varphi(\sum_{i=1}^{n} w_i x_i + b) = \varphi(\mathbf{w}^T \mathbf{x} + b)$$

output layer and other hidden layers [8].

Where, x represents input vector

W represents weights vector

B represents bias

Φ represents activation function.

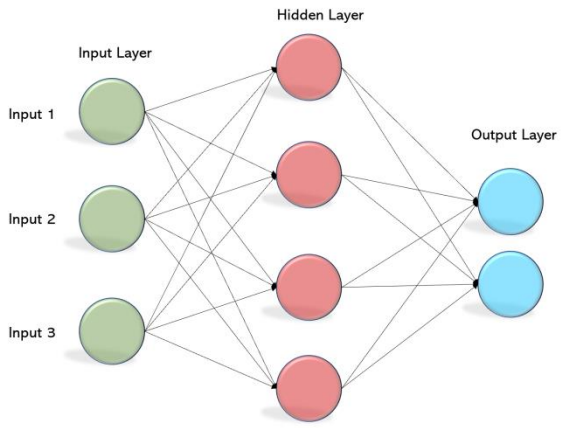

Fig 4.2.1 MLP Classifier

The classification is the main process in the system. We divided the extracted features into classes which predicts the emotion of the speech. In our system we are using the MLP (MultiLayer Perceptron) classifier [9].

We initialize the MLP classifier by defining the required parameters. After that we give the data to the network to train the model. The model is created to predict the emotion of the speech. The accuracy of the model is calculated by comparing the emotion predicted by the model and the actual emotion in the dataset [10]

4.3 UML Diagrams

a) Activity Diagram

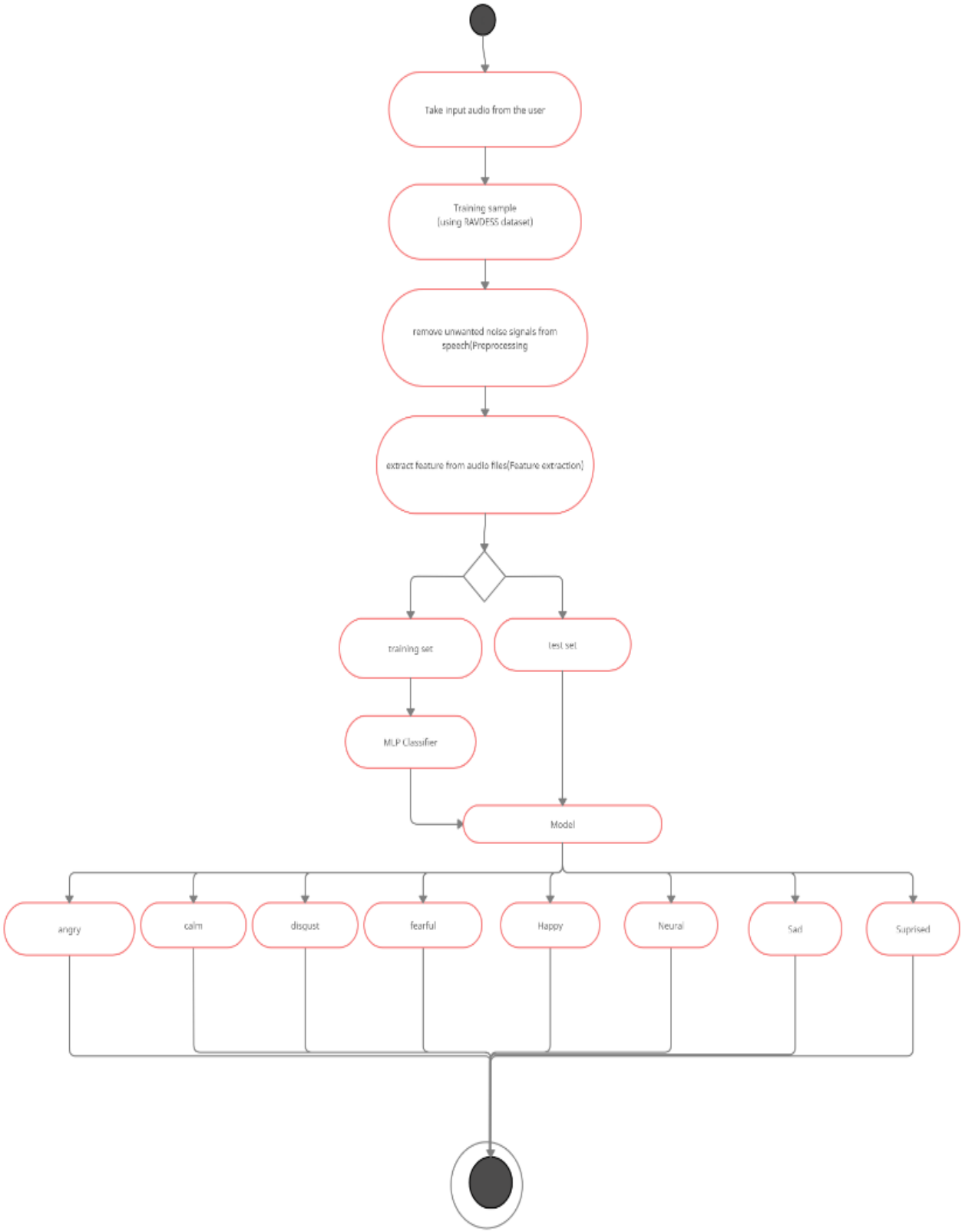

Fig 4.3.1 Activity Diagram

b) Class Diagram

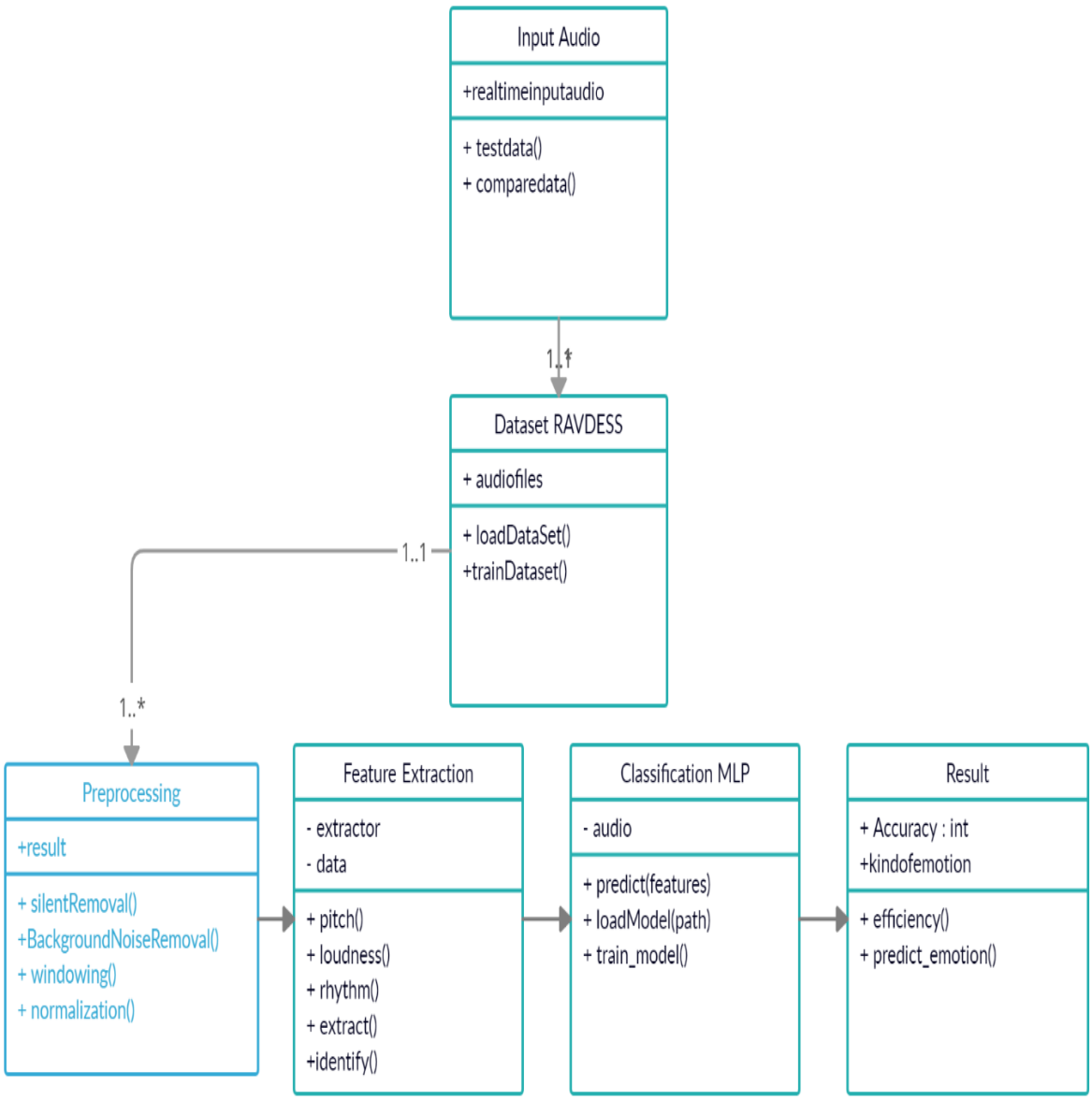

Fig 4.3.2 Class Diagram

5. Implementation

5.1 Algorithms

Extracting Features from the audio files using -

Fig 5.1.1 Extracting features

Important features

1. MFCC

2. Chroma

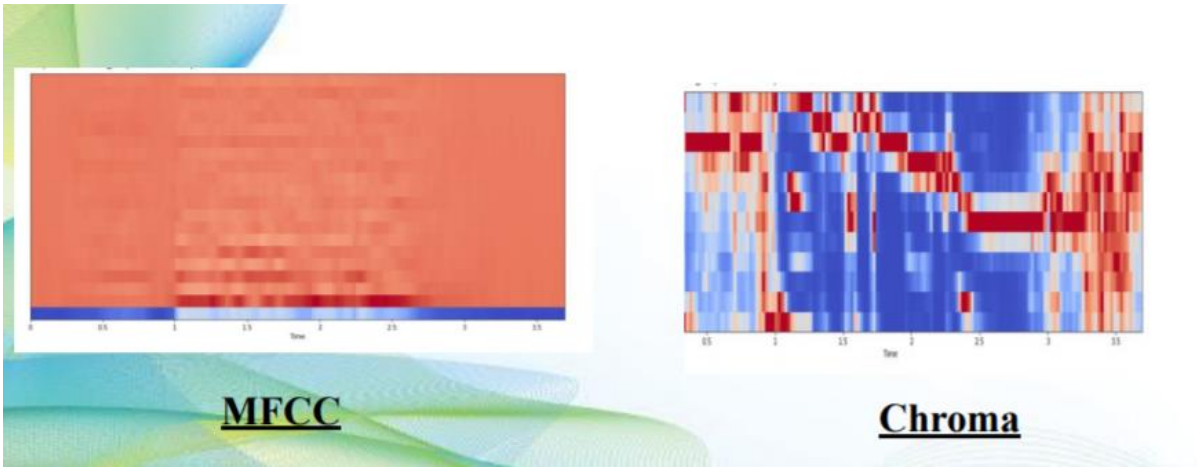

Fig 5.1.2 MFCC and Chroma

Making of Multi-Layer Perceptron Classifier

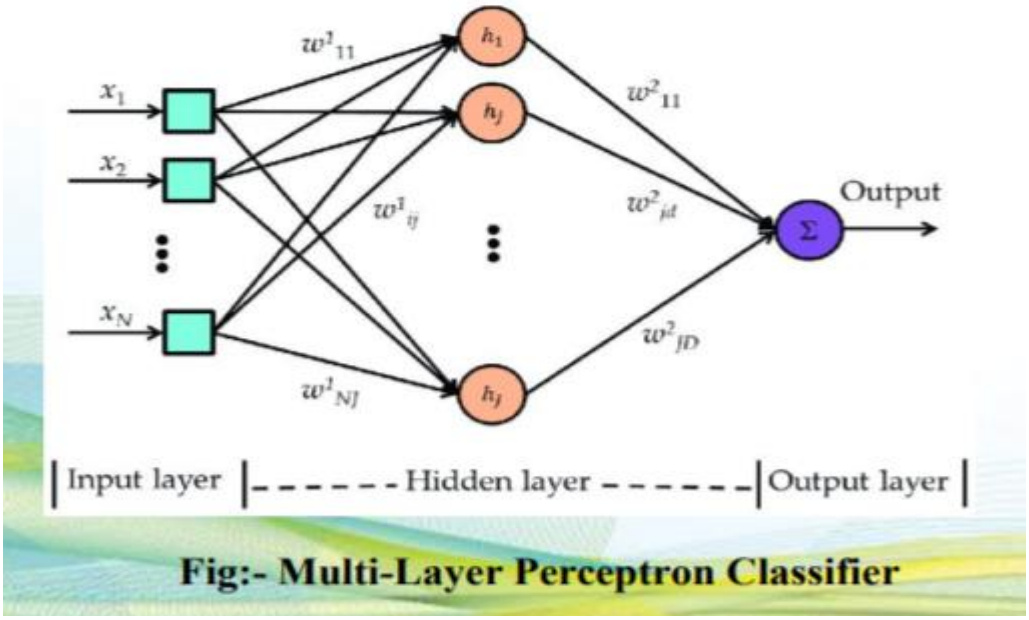

Fig 5.1.3 MLP Classifier layers

5.2 Implementation Steps

- **Step 1** –

Dataset Creation , gather data for dataset , using RAVDESS Audio Visual Database dataset.

- **Step 2 –**

Data pre-processing is done to further refine our dataset, data divided into test and train datasets.

- **Step 3 –**

Create a Model for Speech Recognition using Multi-Layer perceptron Classifier (Neural Network based)

- **Step 4 –**

Training the model using Train Dataset, which we created in Step 2 (using Train Dataset)

- **Step 5 –**

Testing the model using Test Dataset, which we created in Step 2 (using Test Dataset)

- **Step 6 –**

Evaluating model, Improving Accuracy, Using the model for Speech emotion Recognition

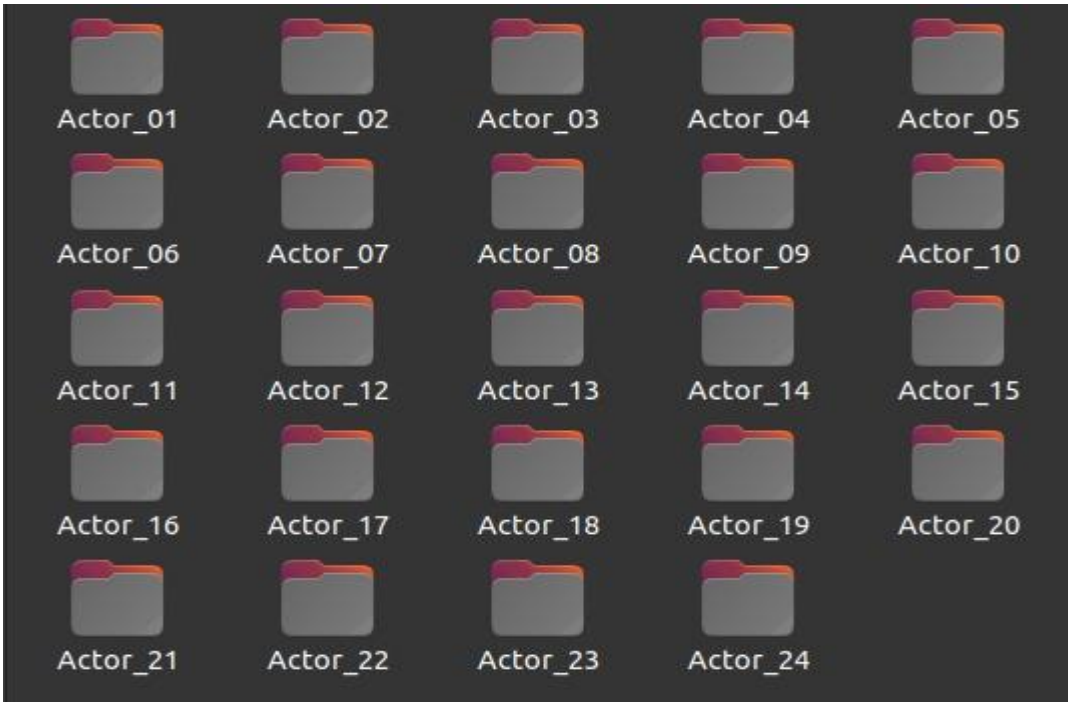

Fig 5.2.1 Sounds folders

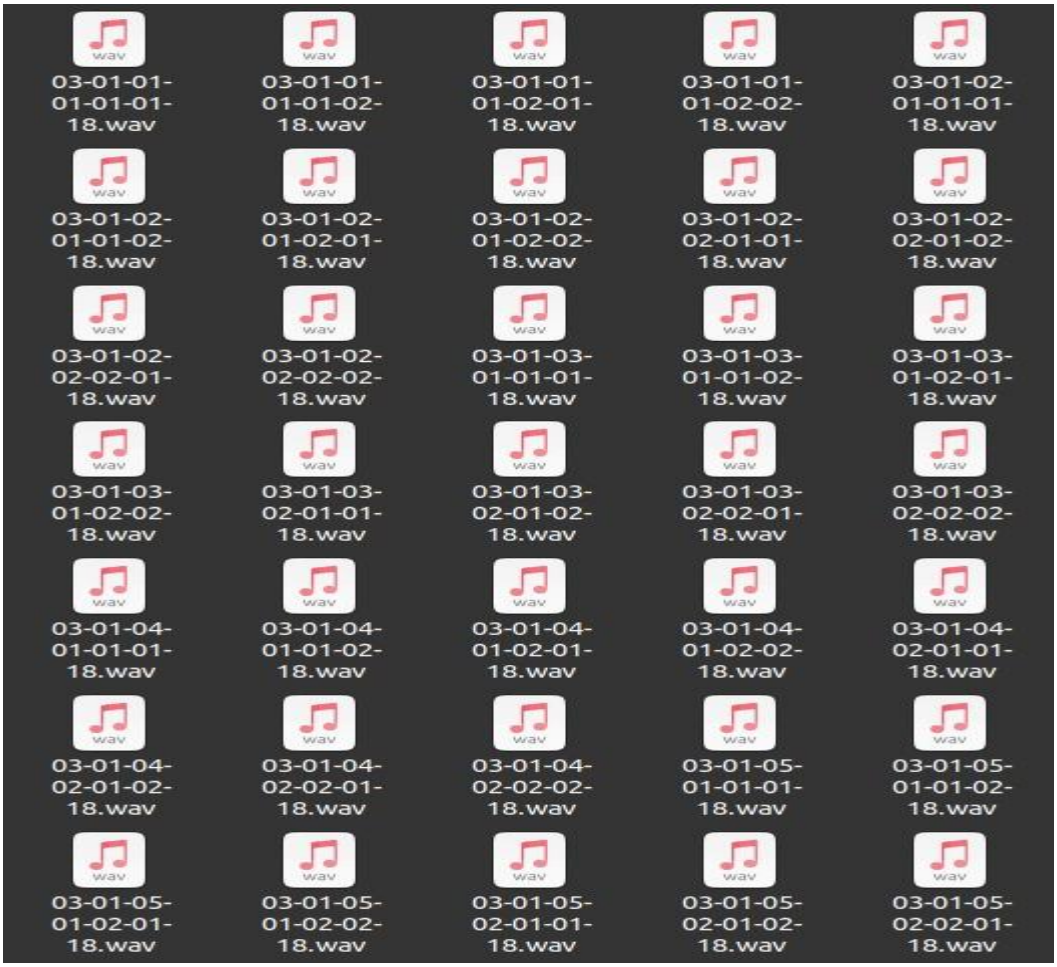

Fig 5.2.2 Sound files

5.3 Source Code

```
# Extract features (mfcc, chroma, mel) from a sound file
def extract_feature(file_name, mfcc, chroma, mel):
    with soundfile.SoundFile(file_name) as sound_file:
        X = sound_file.read(dtype="float32")
        sample_rate=sound_file.samplerate
        if chroma:
            stft=np.abs(librosa.stft(X))
        result=np.array([])
        if mfcc:
            mfccs=np.mean(librosa.feature.mfcc(y=X, sr=sample_rate, n_mfcc=40).T,
axis=0)
            result=np.hstack((result, mfccs))
        if chroma:
            chroma=np.mean(librosa.feature.chroma_stft(S=stft, sr=sample_rate).T,axis=0)
            result=np.hstack((result, chroma))
        if mel:
            mel=np.mean(librosa.feature.melspectrogram(X, sr=sample_rate).T,axis=0)
            result=np.hstack((result, mel))
    return result
 # Emotions in the RAVDESS dataset
emotions={
'01':'neutral',
'02':'calm',
'03':'happy',
'04':'sad',
```

```
'05':'angry',
'06':'fearful',
'07':'disgust',
'08':'surprised'
}
# Emotions to observe
observed_emotions=['calm', 'happy', 'fearful', 'disgust']
# Load the data and extract features for each sound file
def load_data(test_size=0.2):
    x,y=[],[]
    files = glob.glob("Dataset/all_audio/*.wav")
    #print(*files)
    for file in files:
        file_name=os.path.basename(file)
        emotion=emotions[file_name.split("-")[2]]
        if emotion not in observed_emotions:
            continue
        feature=extract_feature(file, mfcc=True, chroma=True, mel=True)
        x.append(feature)
        y.append(emotion)
    return train_test_split(np.array(x), y, test_size=test_size, random_state=9)
# main function
if __name__ == "__main__":
    maxx=0.00000000000000000000000000
    while 1:
        # Split the dataset
```

```
        x_train,x_test,y_train,y_test=load_data(test_size=0.25)
        # Initialize the Multi Layer Perceptron Classifier
        model = MLPClassifier(alpha=0.01, batch_size=256, epsilon=1e-08,
hidden_layer_sizes=(300,), learning_rate='adaptive', max_iter=500)
        # Train the model
        model.fit(x_train,y_train)
        # Predict for the test set
        y_pred=model.predict(x_test)
        # Calculate the accuracy of our model
        accuracy=accuracy_score(y_true=y_test, y_pred=y_pred)
        if accuracy>maxx:
            maxx=accuracy
            # Print the accuracy
            print('Shape of test and train datsets respectively', (x_train.shape[0],
x_test.shape[0]))
            # Get the number of features extracted
            print(f'Features extracted: {x_train.shape[1]}')
            now = datetime.now()
            current_time = now.strftime("%H:%M:%S")
            print("Accuracy: {:.2f}%".format(accuracy*100),'\t',current_time)
            # Pkl_Filename = "Pickle_sorted_Model.pkl"
            # with open(Pkl_Filename, 'wb') as file:
            #     pickle.dump(model, file)
                for input in argv[1:]:
                    inputs.append(extract_feature(input))
                predicted_emotions = model.predict(inputs)
```

```
for file, emotion in zip(argv[1:], predicted_emotions):
    print('File {} predicted to have a {} emotion.'.format(file, emotion))
print(model)
```

6 Testing

Training was done using the below Flow-chart

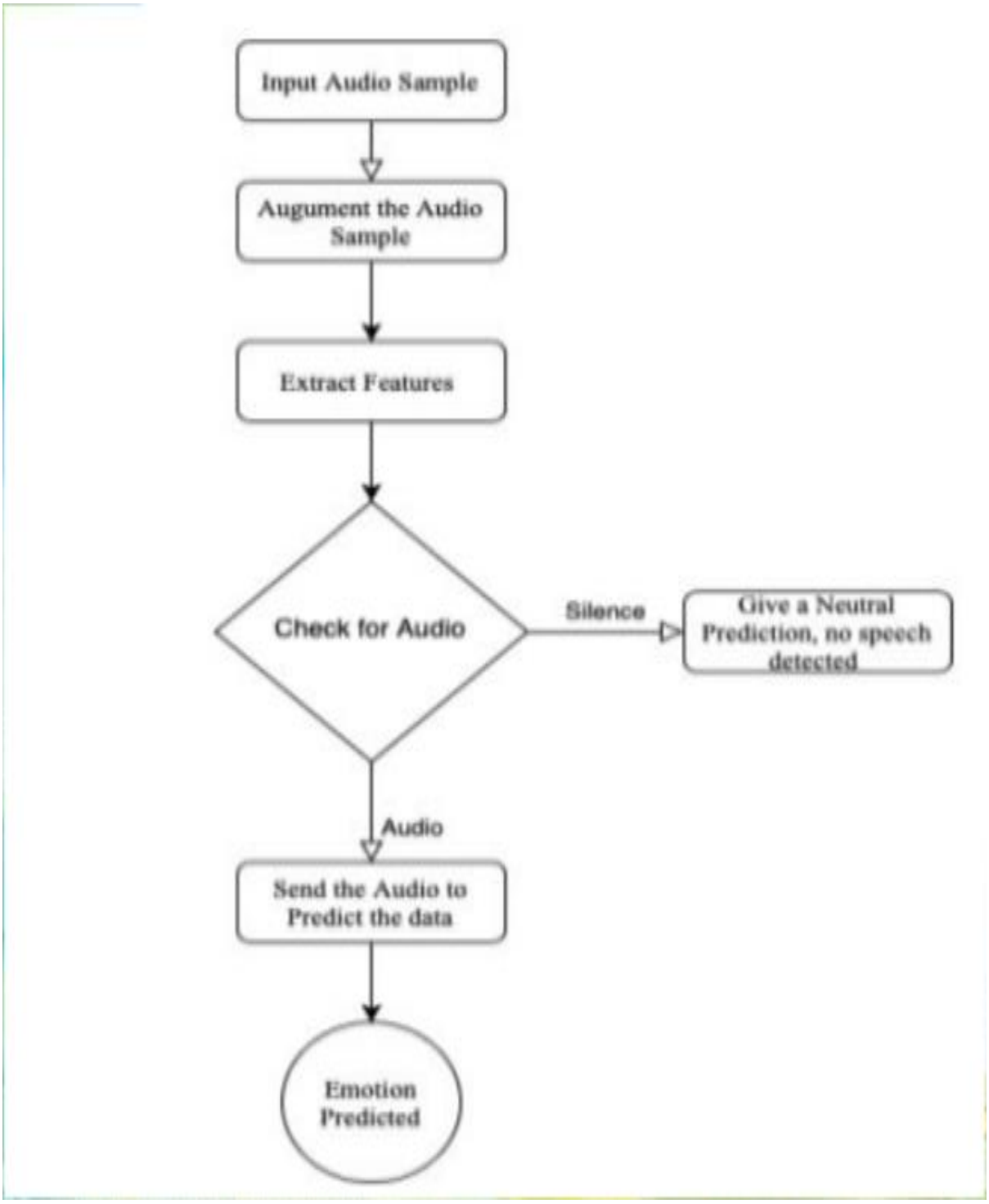

Fig 6.1 Training data

Testing has been done using the below flow chart.

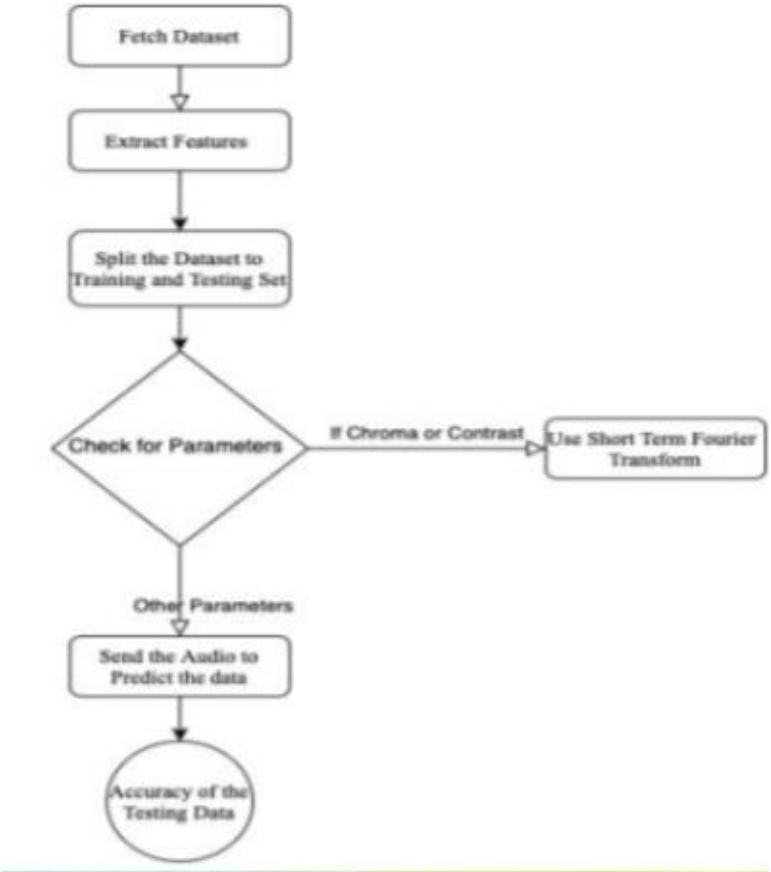

Fig 6.2 Testing data

7. Result

Our project is successfully executed by predicting the emotion of the speaker of 84.38% accuracy. The Project describes

1. **Accuracy**

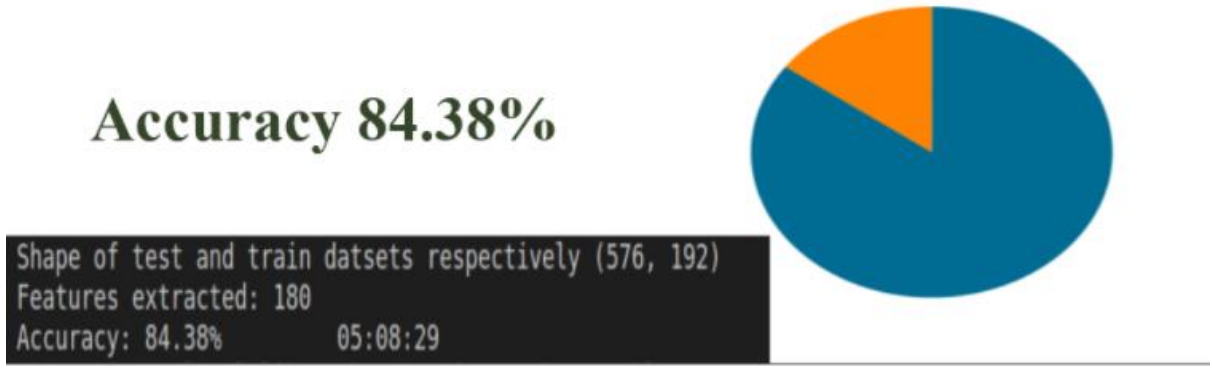

Fig 7.1 Accuracy of model

2. **Emotions predicted through the audio**

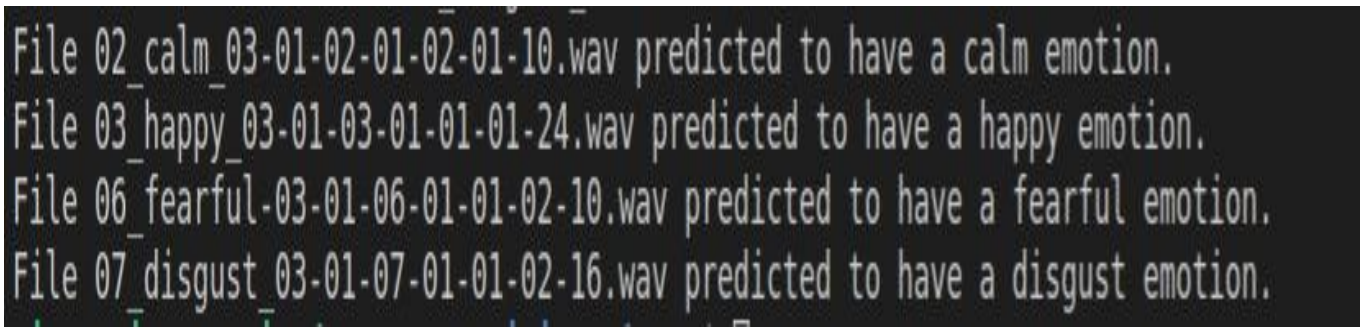

Fig 7.2 Predicted Emotions

8. Conclusion

It can be used in virtual assistants such as google assistant, siri, alexa etc.

Machine can also understand the emotion of the human and can respond in corresponding way.

We conclude that our project is improving the human computer interaction.

The proposed model achieved an accuracy of 66.67%.

Calm was the best identified emotion.

The model gets confused between similar emotions like calm-neutral, happy-surprised.

We tested the model on our own voice file for the sentence “Dogs are sitting by the door” and it identified the emotion correctly.

The system could take into consideration multiple speakers from different geographic locations speaking with different accents.

Though standard feed forward MLP is powerful tool for classification problems, we can use CNN, RNN models with larger data sets and high computational power machines and compare between them.

Study shows that people suffering with autism have difficulty expressing their emotions explicitly. Image based speech processing in real time can prove to be of great assistance.

9. Future Enhancement

Some of the drawbacks can be resolved in the future to make the model more accurate and efficient.

- We can improve the model by training the model a variety of datasets which increases the accuracy of the model.
- Removing the disturbance from the input audio which deviates from correct prediction.
- Adding more emotions to the system since this system can identify only 8 emotions.
- Extracting more features from the speech to improve the classification process.

10. References

[1] https://data-flair.training/blogs/python-mini-project-speech-emotion-recognition/

[2] Mehmet Berkehan Akçay, Kaya Oğuz, 2000. www.sciencedirect.com/science/article/abs/pii/S0167639319302262

[3] Akalpita Das,Laba Kr. Thakuria,Purnendu Acharjee,Prof. P.H. Talukdarhttps://www.ijser.org/researchpaper/A-Brief-Study-on-Speech-Emotion-Recognition.pdf

[4]YashpalsingChavhan,ManikraoDhorehttps://www.researchgate.net/publication/43785303_Speech_Emotion_Recognition_Using_Support_Vector_Machines

[5] Department of Electrical and Computer Engineering, National University of Singapore, 4 Engineering Drive 3, Singapore 117576, Singaporehttps://www.sciencedirect.com/science/article/abs/pii/S0167639303000992

[6] Bashar M. Nema, Ahmed A. Abdul-Kareem Department of Computer Science, College of Science, Mustansiriyah University, IRAQ

https://www.iasj.net/iasj/download/62b38cc626857948

[7] Prabhakar Reddy G,Arun Dalton, Sai Prasad K ”Fuzzy logics associated with neural networks in intelligent control for better world”. International Journal of Reasoning based Intlligent Systems.

[8] https://en.wikipedia.org/wiki/Multilayer_perceptron

[9] R. Anusha, Boggula.Lakshmi,Spruthikankala ,T. Mounika, Deep Stock Prediction" International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-4, November 2019

[10] R.Anusha, T.Nirmala, N.Thulasichitra “System for smart protection of crop” .International Journal of Grid and Distributed Computing Vol. 13, No. 2, (2020), pp. 1707-1715

[11] https://smartlaboratory.org/ravdess/

[12] Subhashini Peneti (MLR Institute of Technology), Hemalatha E(Jawaharlal Nehru Technological University)-DDOS Attack Identification using Machine Learning Techniques

[13]ThomasWood,https://deepai.org/machine-learning-glossary-and-terms/precision-and-recall

[14] TeemuKanstrén, https://towardsdatascience.com/a-look-at-precision-recall-and-f1-score-36b5fd0dd3ec