A

Major Project

On

SPEECH EMOTION RECOGNITION

(Submitted in partial fulfillment of the requirements for the award of Degree)

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

R. Mahesh Kumar (177R1A05A4)

S. Ambareesh (177R1A05B4)

CH. Jayanth (177R1A0572)

Under the Guidance of

M. Madhusudhan

(Assistant Professor)

CMR GROUP OF INSTITUTIONS

EXPLORE TO INVENT

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CMR TECHNICAL CAMPUS

UGC AUTONOMOUS

(Accredited by NAAC, NBA, Permanently Affiliated to JNTUH, Approved by AICTE, New Delhi)

Recognized Under Section 2(f) & 12(B) of the UGCAct.1956,

Kandlakoya (V), Medchal Road, Hyderabad-501401.

2017- 2021

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CMR
GROUP OF INSTITUTIONS
EXPLORE TO INVENT

CERTIFICATE

This is to certify that the project entitled "**SPEECH EMOTION RECOGNITION**" being submitted by **R. Mahesh Kumar (177R1A05A4), S. Ambareesh(177R1A05B4) & CH. Jayanth (177R1A0572)** in partial fulfillment of the requirements for the award of the degree of B.Tech in Computer Science and Engineering to the Jawaharlal Nehru Technological University Hyderabad, is a record of bonafide work carried out by him/her under our guidance and supervision during the year 2020-21.

The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Mr. M. Madhusudhan
Assistant Professor
INTERNAL GUIDE

Dr. A. Raji Reddy
DIRECTOR

Dr. K. Srujan Raju
HOD

EXTERNAL EXAMINER

Submitted for viva voice Examination held on ________________

ACKNOWLEGDEMENT

Apart from the efforts of us, the success of any project depends largely on the encouragement and guidelines of many others. We take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of this project. We take this opportunity to express my profound gratitude and deep regard to my guide

Mr. M. Madhusudhan, Assistant Professor for his exemplary guidance, monitoring and constant encouragement throughout the project work. The blessing, help and guidance given by him shall carry us a long way in the journey of life on which we are about to embark.

We also take this opportunity to express a deep sense of gratitude to Project Review Committee (PRC) Coordinators: **Mr. J. Narasimha Rao, Mr. B. P. Deepak Kumar, Mr. K. Murali, Dr. Suwarna Gothane and Mr. B. Ramji** for their cordial support, valuable information and guidance, which helped us in completing this task through various stages.

We are also thankful to **Dr. K. Srujan Raju,** Head, Department of Computer Science and Engineering for providing encouragement and support for completing this project successfully.

We are obliged to **Dr. A. Raji Reddy,** Director for being cooperative throughout the course of this project. We would like to express our sincere gratitude to Sri. **Ch. Gopal Reddy,** Chairman for providing excellent infrastructure and a nice atmosphere throughout the course of this project.

The guidance and support received from all the members of **CMR Technical Campus** who contributed to the completion of the project. We are grateful for their constant support and help.

Finally, we would like to take this opportunity to thank our family for their constant encouragement, without which this assignment would not be completed. We sincerely acknowledge and thank all those who gave support directly and indirectly in the completion of this project.

R. Mahesh Kumar (177R1A05A4)

S. Ambareesh (177R1A05B4)

CH. Jayanth (177R1A0572)

ABSTRACT

A real time Deep learning based system was built for the Emotion Recognition using speech that have been given as input with the help of a PC microphone. The main purpose of this project is to design a model that can record live speech from computer microphone, analyse the audio file, detect and recognize the particular emotion. After recognition is done, the particular user can view their emotion through an interface with appropriate emojis. Speech Emotion Recognition is an rapidly increasing research domain in recent years. In this paper eight basic emotions (Anger, Happy, Fear, Neutral, sad, calm, disgust, surprise) are analyzed from emotional speech signals. In Our project we have used Long Short Term Memory(LSTM) network which is an artificial Recurrent Neural Network. The reason we used LSTM is that it provides higher accuracy while dealing with emotion recognition using speech. Our model was trained using RAVDEES data set which contains 7356 audio files . I have took 2880 files for training my model, in order to increase our accuracy and also to detect and recognize emotion from different speech. Our model scored an accuracy of 93% for training data set.

LIST OF FIGURES

LIST OF SCREENSHOTS

TABLE OF CONTENTS

1.INTRODUCTION

1 INTRODUCTION

1.1 PROJECT SCOPE

Emotion plays a crucial role in daily human interactions. It helps us to know other people's feelings. It's been revealed through research that emotion plays a strong role in shaping human social interactions.In order to know a emotion of a particular person at any particular situation. This has opened a replacement research field called automatic emotion recognition, having basic goals to know and retrieve desired emotions. Our project is all about building a model to recognise emotion through Speech.

1.2 PROJECT PURPOSE

Human beings use speech to communicate with each other and emotion play an important role in direct communication, when we communicate through socialmedia platform like whatsapp, we use emojis to express our emotions, but while communicating through mobiles or while exchanging audio files to communicate we cannot detect the emotion of the person while conversation. So we are implementing a Deep learning model on Speech Emotion Recognition to recognize the emotion of the person by his audio.

1.3 PROJECT FEATURES

In our project we are using Long short-term memory (LSTM) classifier and Ravdess dataset, and by extracting main features like Mel-frequency cepstrum coefficients (MFCC) , pitches in audio and acoustic features (sound properties like tone, jitter, etc.).This features are extracted from the audio file and further used to train the classifier to classify emotional states: anger, happiness, sadness, surprise and a neutral state.

This speech emotion recognition system has been made by keeping the Users in mind and main purpose is to provide users an easy, simple and efficient system where users can recognize the emotion of a speaker through his/her speech. The user interface that is developed in this project helps users to upload audio data and get the emotion of speaker.

2. SYSTEM ANALYSIS

2. SYSTEM ANALYSIS

SYSTEM ANALYSIS

Speech signal is the most natural way to communicate among human beings. Researchers are constantly working to apply this mode in the domain of humanmachine interaction. However, it requires machines to interpret human spoken phrases intelligently and understand it semantically. Despite the great progress made in speech recognition, this process still requires a lot of efforts to make it a natural interaction between man and machine. One significant challenge in realizing this goal is the inability of machines to understand the emotional state hidden behind spoken words. In this context, speech emotion recognition (SER) refers to recognition of the emotional state of a speaker by analysis his/her speech. It is believed that SER can be used to extract useful semantics from speech, as well as improve the performance of speech emotion recognition systems.

2.1 PROBLEM DEFINITION

Emotion plays a crucial role in daily human interactions. It helps us to know other people's feelings. It's been revealed through research that emotion plays a strong role in shaping human social interactions.In order to know a emotion of a particular person at any particular situation. This has opened a replacement research field called automatic emotion recognition, having basic goals to know and retrieve desired emotions. Our project is all about building a model to recognize emotion through Speech.

2.2 EXISTING SYSTEM

Depending on the methods employed for feature extraction and classification, several speech emotion detection strategies are employed in the literature. Dellaert et al. used 17 features and compared three classifiers: maximum likelihood Bayes classification, kernel regression and k-NN (Nearest Neighbour) with four emotion categories. Scherer extracted 16 features and achieved overall accuracy 40.4% for fourteen emotional states

2.2.1 LIMITATIONS OF EXISTING SYSTEM

- Time consuming process
- Different models are used so no attention is paid
- Accuracy is low

2.3 PROPOSED SYSTEM

In our project we are using Long short-term memory (LSTM) classifier and Ravdess dataset, and by extracting main features like Mel-frequency cepstrum coefficients (MFCC) , pitches in audio and acoustic features (sound properties like tone, jitter, etc.).This features are extracted from the audio file and further used to train the classifier to classify emotional states: anger, happiness, sadness, surprise and a neutral state.

This speech emotion recognition system has been made by keeping the Users in mind and main purpose is to provide users an easy, simple and efficient system where users can recognize the emotion of a speaker through his/her speech. The user interface that is developed in this project helps users to upload audio data and get the emotion of speaker.

2.3.1 ADVANTAGES OF THE PROPOSED SYSTEM

- Accuracy is more
- Take less time
- User friendly (accessed easily)
- Shows output with appropriate emoji

2.4 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase andbusiness proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. Three key considerations involved in the feasibility analysis are

- Economic Feasibility
- Technical Feasibility
- Social Feasibility

2.4.1 ECONOMIC FEASIBILITY

The developing system must be justified by cost and benefit. Criteria to ensure that effort is concentrated on project, which will give best, return at the earliest. One of the factors, which affect the development of a new system, is the cost it would require.The following are some of the important financial questions asked during preliminary investigation:

- The costs conduct a full system investigation.
- The cost of the hardware and software.
- The benefits in the form of reduced costs or fewer costly errors.

Since the system is developed as part of project work, there is no manual cost to spend for the proposed system. Also all the resources are already available, it give an indication of the system is economically possible for development.

2.4.2 TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

2.4.3 BEHAVIORAL FEASIBILITY

This includes the following questions:

- Is there sufficient support for the users?
- Will the proposed system cause harm?

The project would be beneficial because it satisfies the objectives when developed and installed. All behavioral aspects are considered carefully and conclude that the project is behaviorally feasible.

2.5 HARDWARE & SOFTWARE REQUIREMENTS

2.5.1 HARDWARE REQUIREMENTS:

Hardware interfaces specifies the logical characteristics of each interface between the software product and the hardware components of the system. The following

- **Processor** : Intel Dual Core@ CPU 2.90GHz.
- **Hard disk** : 20 GB and Above.
- **RAM** : 4GB and Above.
- **Monitor** : 5 inches or above

2.5.2 SOFTWARE REQUIREMENTS:

Operating System : Windows/Linux.
Language used : Python 3
IDE : Anaconda, Jupyter notebook.

Libraries

1 .Numpy:

Numpy is a general-purpose array-processing package. It provides a high performance Multi dimensional array object, and tools for working with these arrays.

It is the fundamental package for scientific computing with Python.

2.Pandas:

Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. Python was majorly used for data munging and preparation. It had very little contribution towards data analysis. Pandas solved this problem. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data load, prepare, manipulate, model, and analyze. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc

3 Matplotlib:

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.

One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

4. Scikit – learn:

Scikit – learn is an open source Python library that implements a range of machine learning, pre-processing, cross-validation and visualization algorithms using a unified interface.

5 Keras:

Keras is a high-level neural networks API, capable of running on top of Tensorflow, Theano, and CNTK. It enables fast experimentation through a high level, user-friendly, modular and extensible API. Keras can also be run on both CPU and GPU

6 Librosa:

Librosa is a Python package for music and audio analysis. Librosa is basically used when we work with audio data like in music generation(using LSTM's), Automatic Speech Recognition.

It provides the building blocks necessary to create the music information retrieval systems. Librosa helps to visualize the audio signals and also do the feature extractions in it using different signal processing techniques.

7 TensorFlow :

TensorFlow is a free and open-source software library for machine learning. It can be used across a range of tasks but has a particular focus on training and inference of deep neural networks. Tensorflow is a symbolic math library based on dataflow and differentiable programming.

3. ARCHITECTURE

3.ARCHITECTURE

3.1 PROJECT ARCHITECTURE

This project architecture shows the procedure followed for Speech detection using machine learning, starting from input to final prediction.

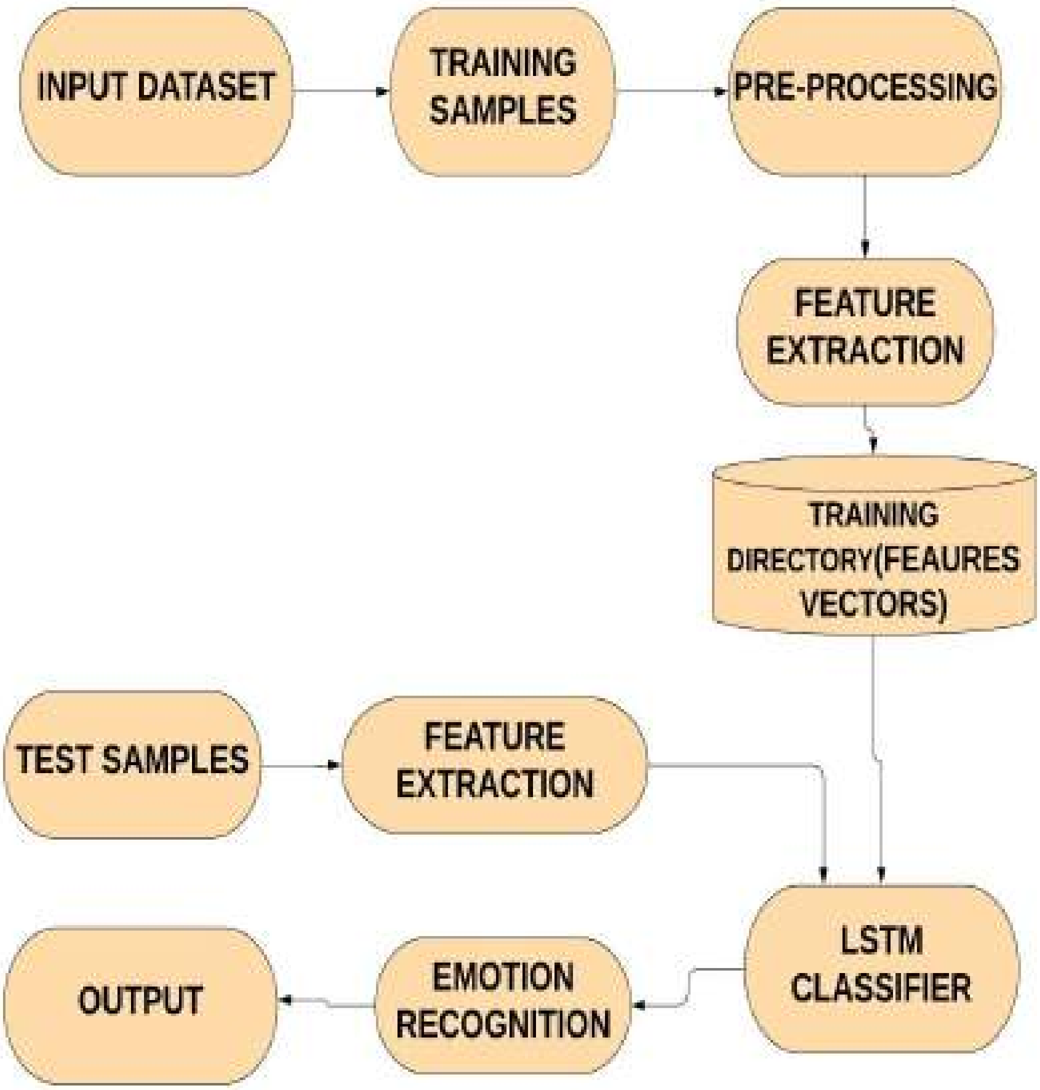

Figure 3.1 Project Architecture of Speech Emotion Recognition

3.2 DESCRIPTION

Input Dataset:

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains 7356 audio files from that we used 1GB of audio files which is 2880 audio files for data training with LSTM

Preprocessing:

Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

Feature Extraction:

Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing. A characteristic of these large data sets is a large number of variables that require a lot of computing resources to process.

Training:

Training Data: The part of data we use to train our model. This is the data which your model actually sees(both input and output) and learn from. ... This is how we evaluate and see how much our model has learned from the experiences feed in as training data, set at the time of training

LSTM Classifier:

Long Short Term Memory is a kind of recurrent neural network. In **RNN** output from the last step is fed as input in the current step. **LSTM** can by default retain the information for long period of time. It is used for processing, predicting and **classifying** on the basis of time series data.

Output:

Emotion of the voice is given as output and that is visible on console with respective emojies for better understanding

3.3 USE CASE DIAGRAM

Here we have two actors in this use case diagram, one is a user and the other is the model. The audio file to which the emotion has to be recognised is being made and saved by the user. This audio file is inputted to the model. The emotion of the speaker in the audio file is predicted and the output emotion is displayed. This emotion output can be viewed by the user and the new audio file given by the user is updated to the existing dataset by the admin

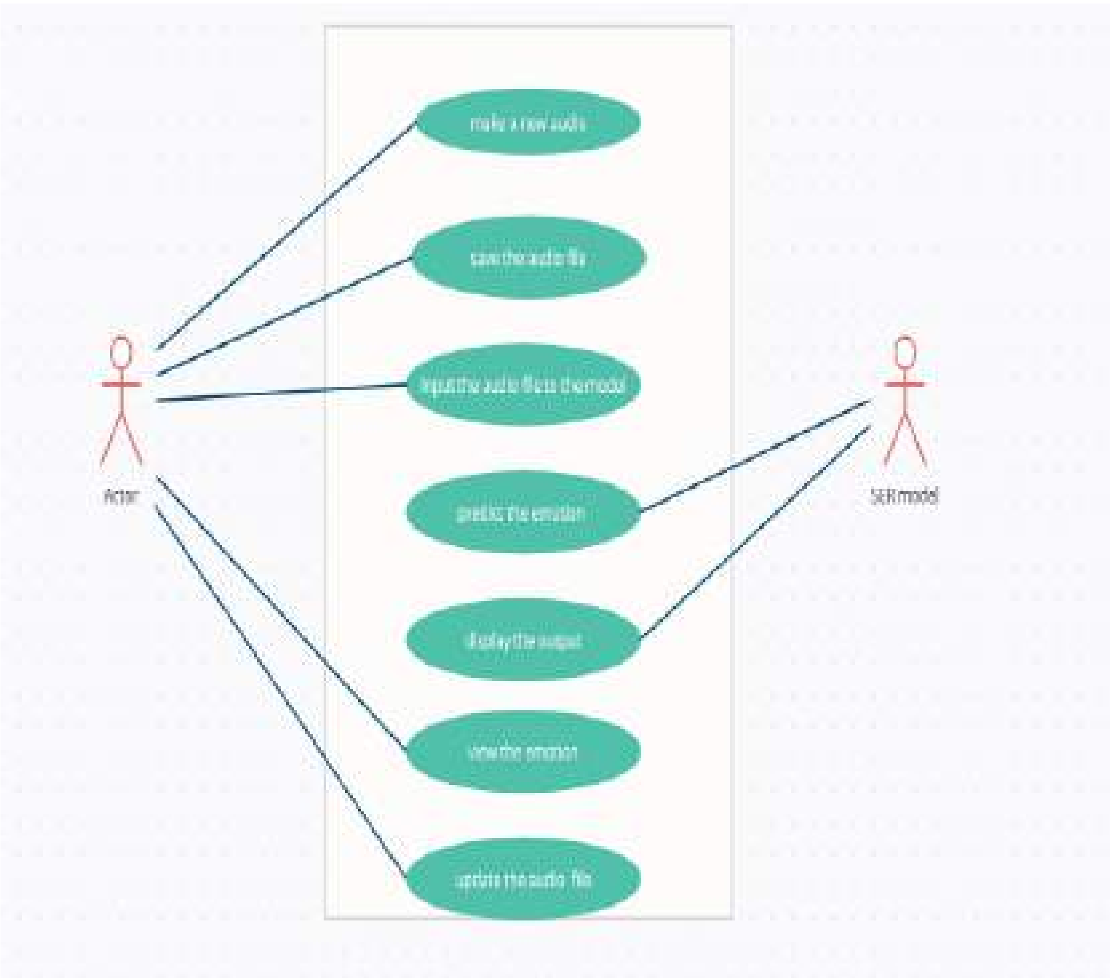

Figure 3..3 Use Case Diagram

3.4 CLASS DIAGRAM

A Class Diagram describes the structure of a system by showing it's classes, their attributes, methods, and the relations between them. A class represents an entity, a noun, of the system.

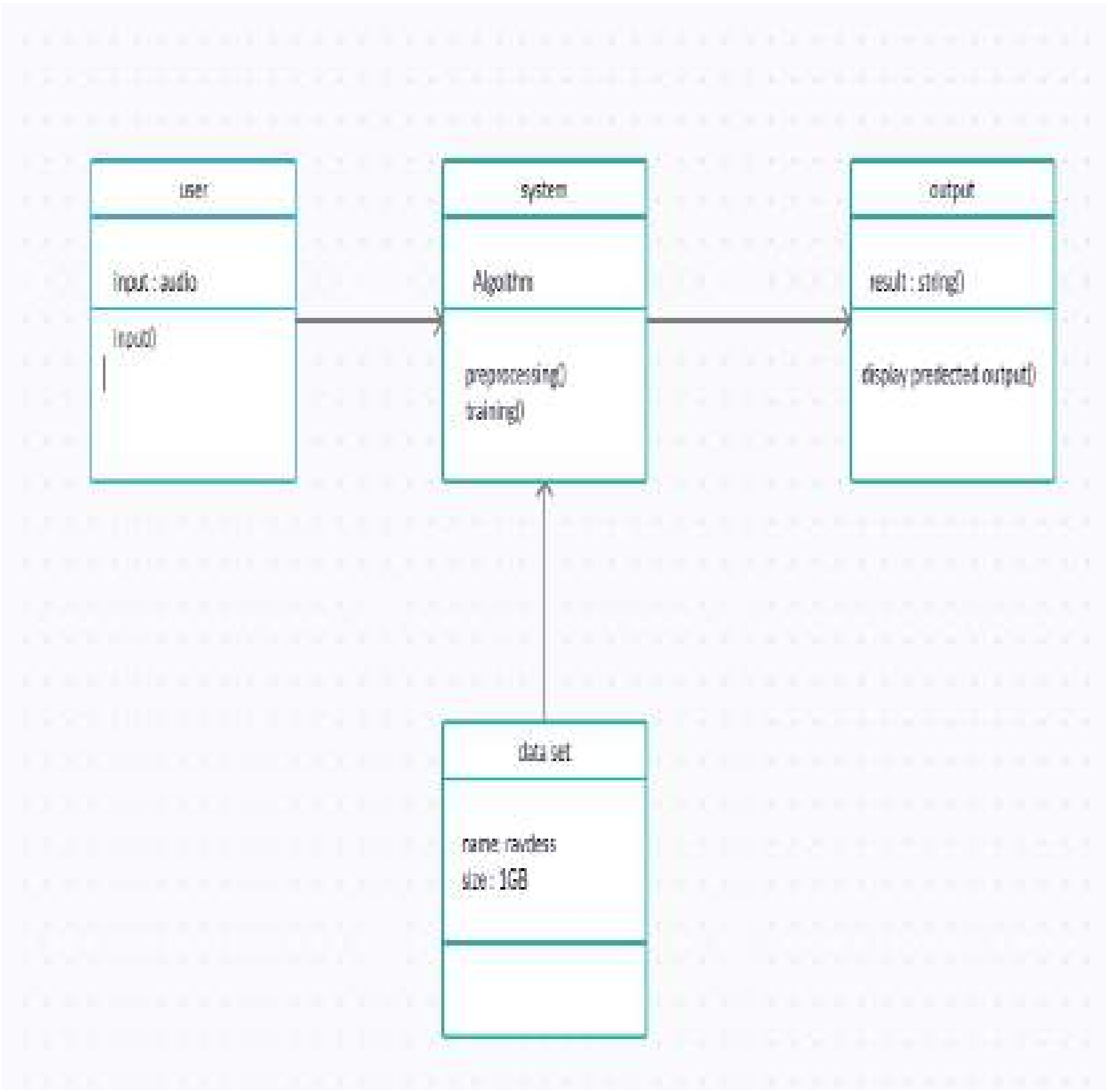

Figure 3.4Class Diagram

3.5 SEQUENCE DIAGRAM

A sequence diagram simply depicts interaction between objects in a sequential order i.e. the order in which these interactions take place. We can also use the terms event diagrams or event scenarios to refer to a sequence diagram. Sequence diagrams describe how and in what order the objects in a system function. These diagrams are widely used by businessmen and software developers to document and understand requirements for new and existing systems.

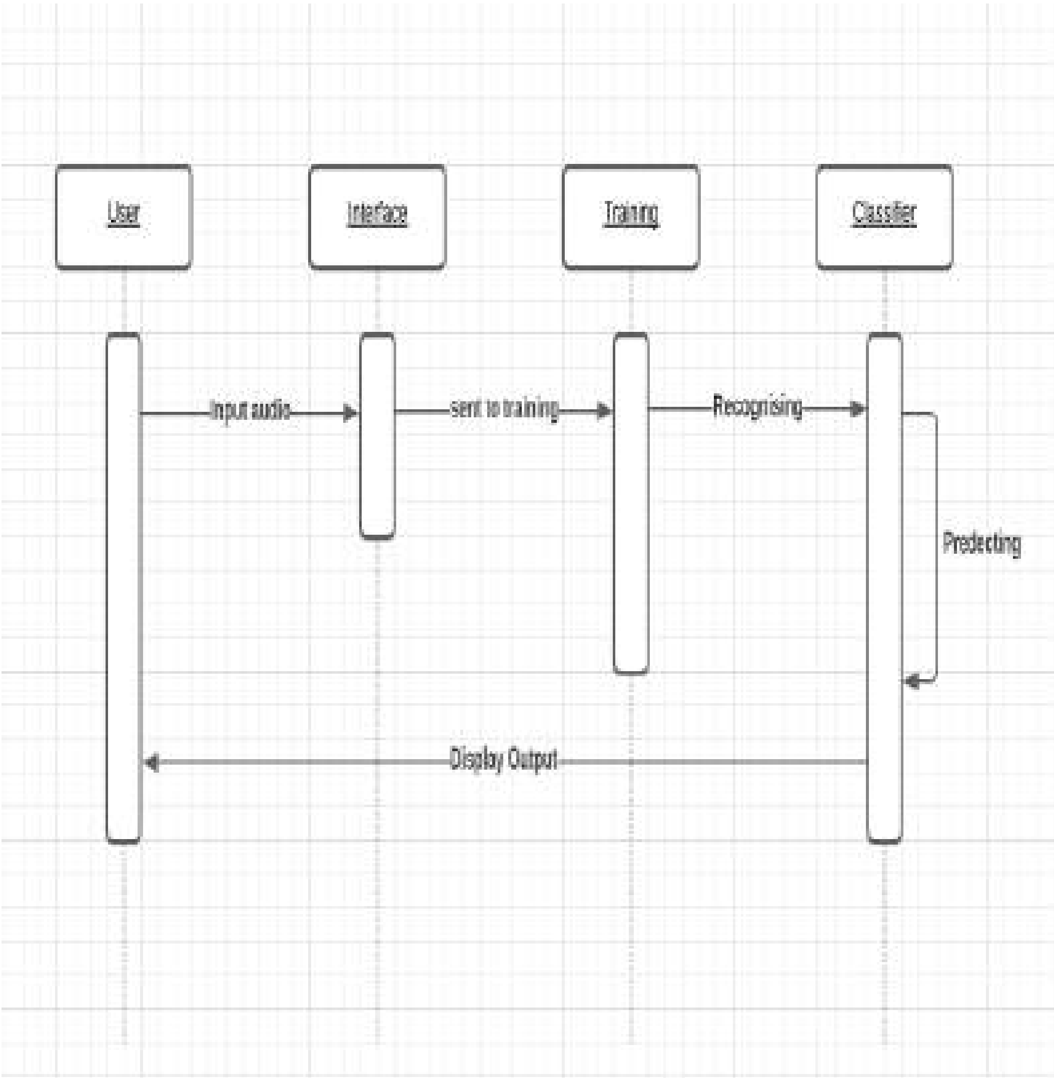

Figure 3.5 Sequence Diagram

3.6 ACTIVITY DIAGRAM

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency.

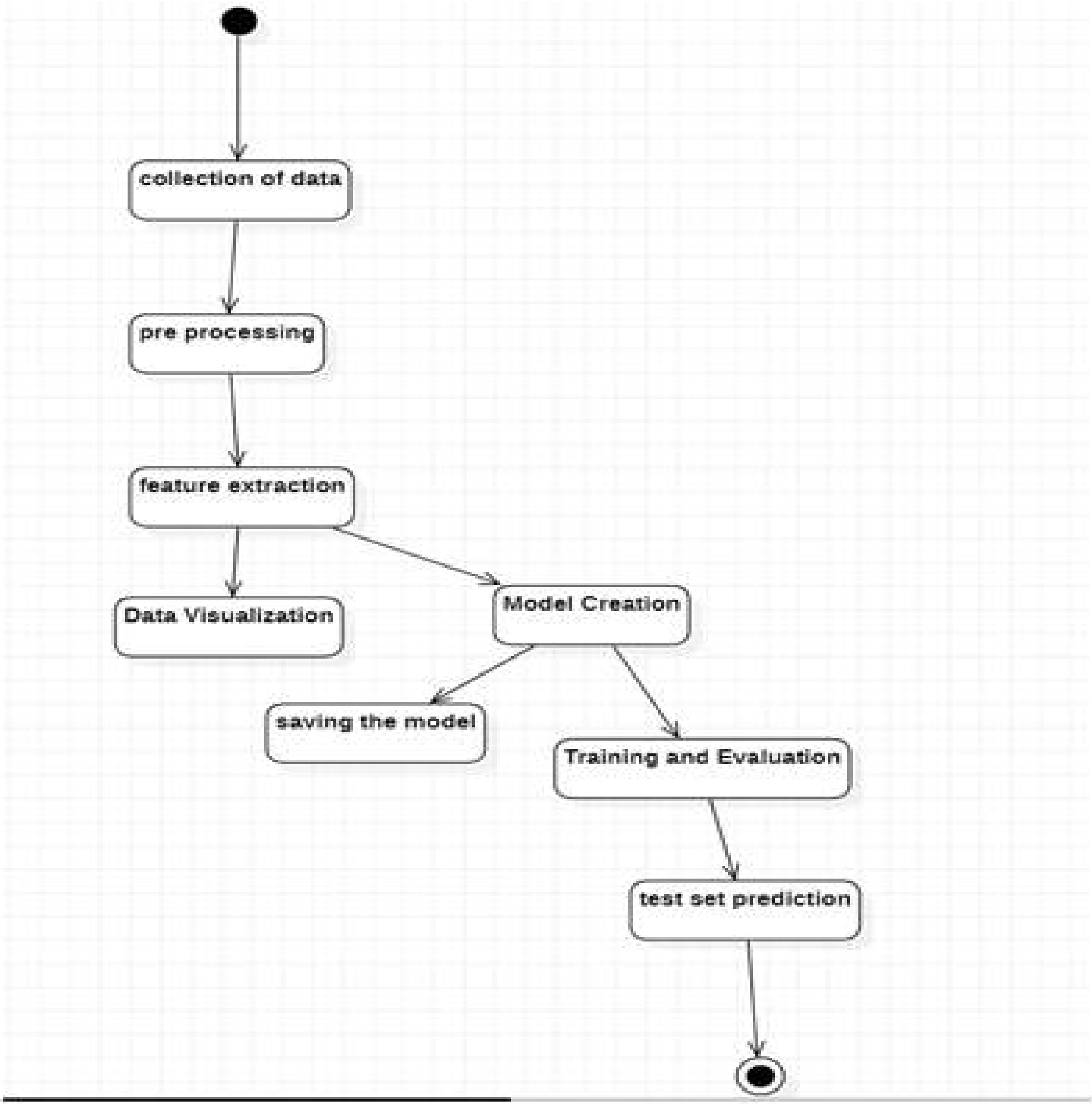

Figure 3.6 Activity Diagram

4.IMPLIMENTATION

4 IMPLEMENTATION

4.1 SAMPLE CODE

```
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import os # to use operating system dependent functionality
import librosa # to extract speech features
import wave # read and write WAV files
import matplotlib.pyplot as plt # to generate the visualizations
import numpy as np

# LSTM Classifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn import preprocessing
lb = preprocessing.LabelBinarizer()
import keras
from keras.utils import to_categorical
from keras.models import Sequential
from keras.layers import *
from keras.optimizers import rmsprop


def extract_mfcc(wav_file_name):
    #This function extracts mfcc features and obtain the mean of each dimension
    #Input : path_to_wav_file
    #Output: mfcc_features'''
    y, sr = librosa.load(wav_file_name,duration=3
                        ,offset=0.5)
    mfccs = np.mean(librosa.feature.mfcc(y=y, sr=sr, n_mfcc=40).T,axis=0)

    return mfccs


##### load radvess speech data #####
radvess_speech_labels = [] # to save extracted label/file
ravdess_speech_data = [] # to save extracted features/file
for dirname, _, filenames in
os.walk('C:/Users/Dell/Downloads/MaheshMajor/Dataset'):
    print(dirname)
    for filename in filenames:
        #print(os.path.join(dirname, filename))
        radvess_speech_labels.append(int(filename[7:8]) - 1) # the index 7 and 8 of the
file name represent the emotion label
        wav_file_name = os.path.join(dirname, filename)
        ravdess_speech_data.append(extract_mfcc(wav_file_name)) # extract MFCC
features/file
```

```
print("Finish Loading the Dataset")



#### convert data and label to array
ravdess_speech_data_array = np.asarray(ravdess_speech_data) # convert the input to
an array
ravdess_speech_label_array = np.array(radvess_speech_labels)
ravdess_speech_label_array.shape # get tuple of array dimensions

#### make categorical labels
labels_categorical = to_categorical(ravdess_speech_label_array) # converts a class
vector (integers) to binary class matrix
labels_categorical.shape


ravdess_speech_data_array.shape


x_train,x_test,y_train,y_test=
train_test_split(np.array(ravdess_speech_data_array),labels_categorical,
test_size=0.20, random_state=9)

len(x_train)
len(y_train)

len(x_test)
len(y_test)


# Split the training, validating, and testing sets
number_of_samples = ravdess_speech_data_array.shape[0]
training_samples = int(number_of_samples * 0.8)
validation_samples = int(number_of_samples * 0.1)
test_samples = int(number_of_samples * 0.1)


# Define the LSTM model
def create_model_LSTM():
    model = Sequential()
    model.add(LSTM(128, return_sequences=False, input_shape=(40, 1)))
    model.add(Dense(64))
    model.add(Dropout(0.4))
    model.add(Activation('relu'))
    model.add(Dense(32))
    model.add(Dropout(0.4))
    model.add(Activation('relu'))
    model.add(Dense(8))
    model.add(Activation('softmax'))
```

```
    # Configures the model for training
    model.compile(loss='categorical_crossentropy', optimizer='Adam',
metrics=['accuracy'])
    return model


w = np.expand_dims(ravdess_speech_data_array[:training_samples],-1)


w.shape

model_A = create_model_LSTM()
history =
model_A.fit(np.expand_dims(ravdess_speech_data_array[:training_samples],-1),
labels_categorical[:training_samples],
validation_data=(np.expand_dims(ravdess_speech_data_array[training_samples:traini
ng_samples+validation_samples], -1),
labels_categorical[training_samples:training_samples+validation_samples]),
epochs=100, shuffle=True)


### loss plots using LSTM model
loss = history.history['loss']
val_loss = history.history['val_loss']

epochs = range(1, len(loss) + 1)

plt.plot(epochs, loss, 'ro', label='Training loss')
plt.plot(epochs, val_loss, 'b', label='Validation loss')
plt.title('Training and validation loss')
plt.xlabel('Epochs')
plt.ylabel('Loss')
plt.legend()

plt.show()


plt.clf()

acc = history.history['accuracy']
val_acc = history.history['val_accuracy']

plt.plot(epochs, acc, 'ro', label='Training acc')
plt.plot(epochs, val_acc, 'b', label='Validation acc')
plt.title('Training and validation accuracy')
plt.xlabel('Epochs')
plt.ylabel('Loss')
plt.legend()
```

```
plt.show()


### evaluate using model A
model_A.evaluate(np.expand_dims(ravdess_speech_data_array[training_samples +
validation_samples:], -1), labels_categorical[training_samples + validation_samples:])


model_A.save_weights("Modeled2_2LSTM.h5")

path_=input()
#path_ = 'C:/Users/Dell/Downloads/MaheshMajor/Dataset/audio_speech_actors_01-
24/Actor_12/03-01-02-01-02-01-12.wav'
#path_ ='C:/Users/Dell/Downloads/MaheshMajor/my_Audio_file.wav'

import IPython.display as ipd
ipd.Audio(path_)

a = extract_mfcc(path_)

a.  Shape
a1 = np.asarray(a)
a1.shape

q = np.expand_dims(a1,-1)
qq = np.expand_dims(q,0)
qq.shape

pred = model_A.predict(qq)
Pred

preds=pred.argmax(axis=1)
Preds
```

5. SCREENSHOTS

5. SCREENSHOTS

5.1 BEFORE INPUT

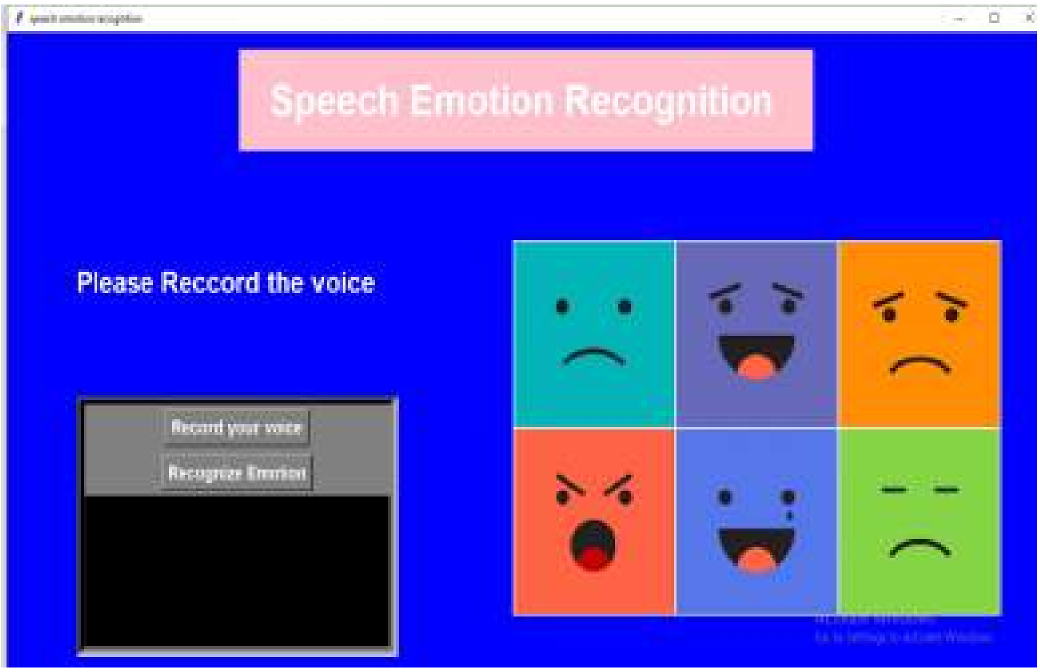

Figure 5.1.1 Before Input

5.2 AFTER OUTPUT

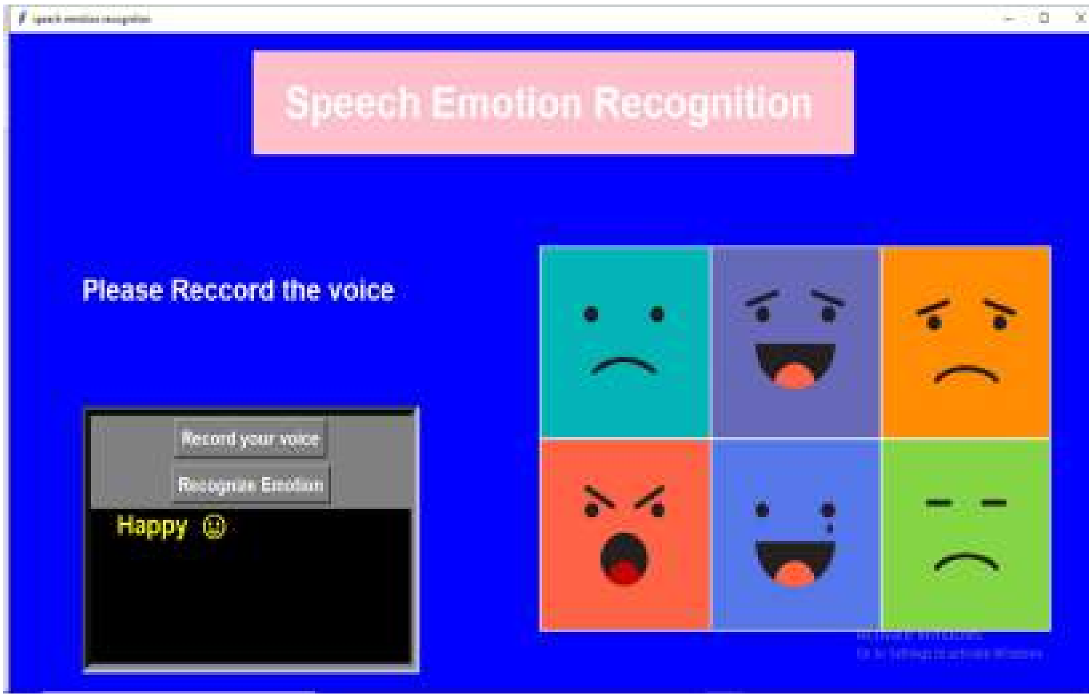

Figure 5.2.1 Output Of Happy

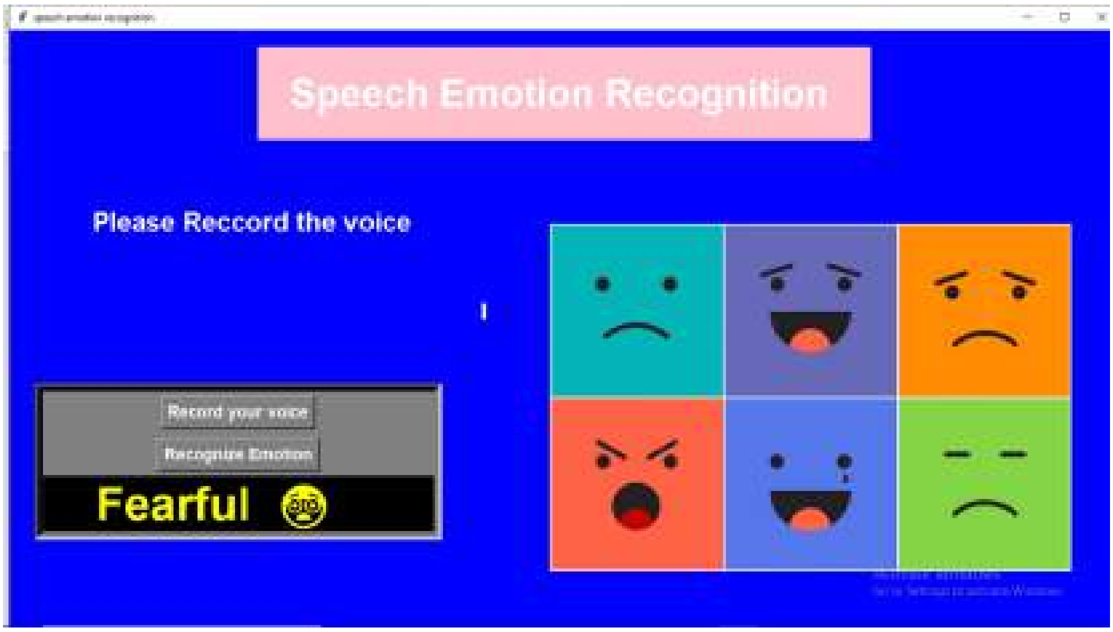

Figure 5.2.2 Output Of Fearful

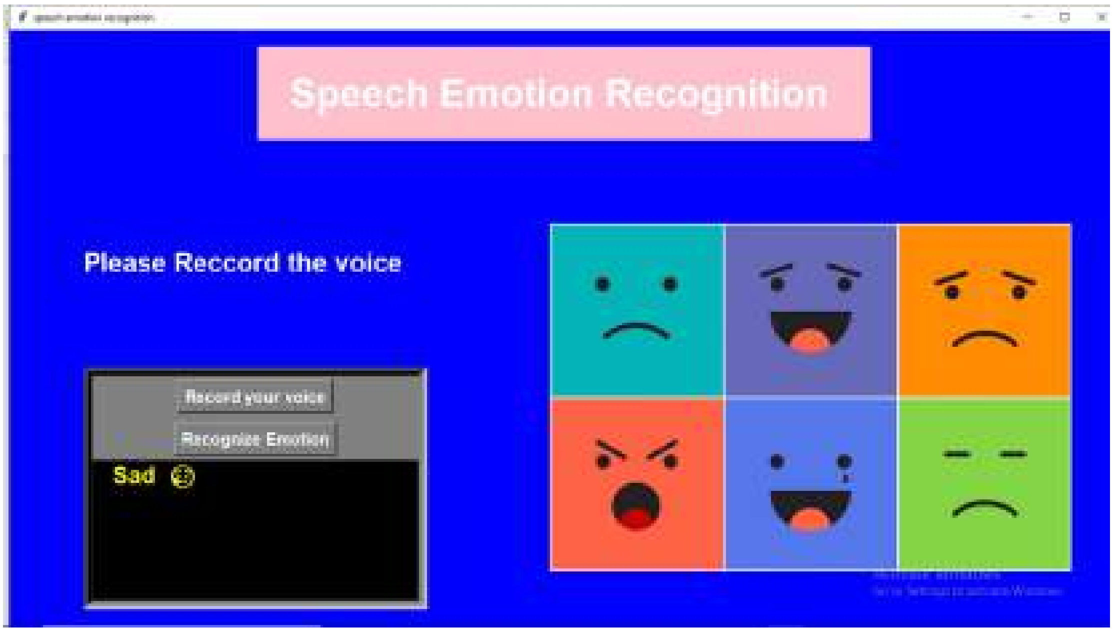

Figure 5.2.3 Output Of Sad

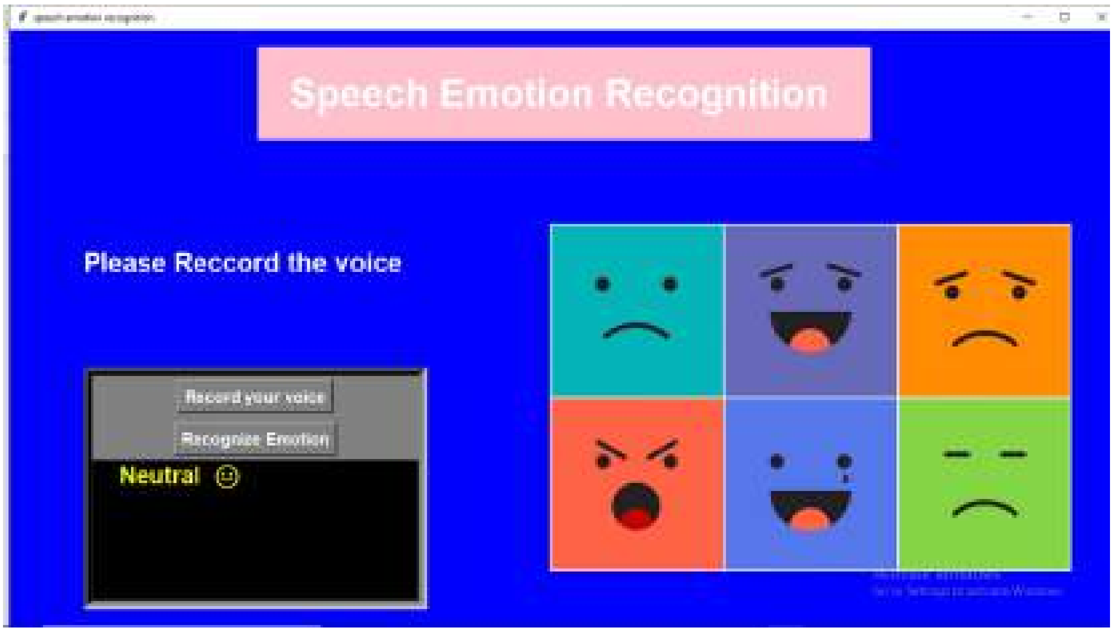

Figure 5.2.4 Output Of Neutral

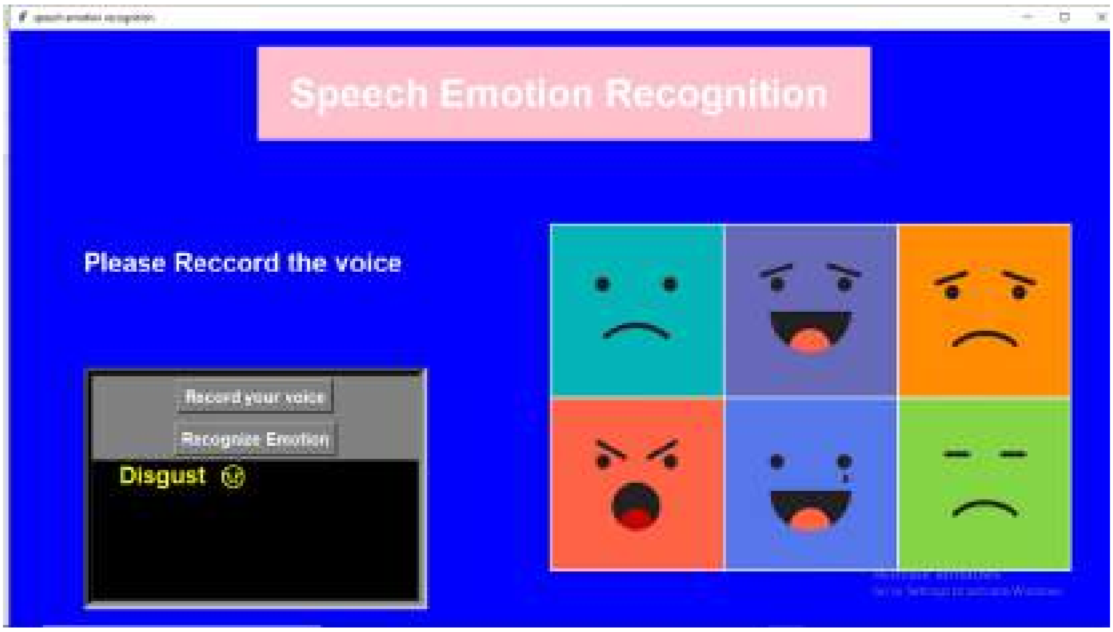

Figure 5.2.5 Output Of Disgust

6. TESTING

6. TESTING

6.1 INTRODUCTION TO TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

6.2 TYPES OF TESTING

6.2.1 UNIT TESTING

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

6.2.2 INTEGRATION TESTING

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

6.2.3 FUNCTIONAL TESTING

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.
Invalid Input : identified classes of invalid input must be rejected
Functions :identified functions must be exercised.
Output : identified classes of application outputs must be exercised.

Systems/Procedures: interfacing systems or procedures must be invoked. Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes.

6.3 TEST CASES

6.3.1 INPUT AUDIO

Test case ID	Test case name	Purpose	Test Case	Output
1	User speaks out some sentence as input	Use it to detect the emotion of the audio	The user with speaks angry voice	The audio file get saved in the file location
2	User speaks out some sentence as input	Use it to detect the emotion of that audio	The user speaks with calm and relaxed audio voice	The audio file get saved in the file location

6.3.2 CLASSIFICATION

Test case ID	Test case name	Purpose	Input	Output
1	Classification test 1	To check if the classifier performs its task	Angry voice is given as input	Angry is the output
2	Classification test 2	To check if the classifier performs its task	Calm voice is given as input	Calm is the output
3	Classification test 3	To check if the classifier performs its task	Sad voice is given as input	Sad is the output

6.4 TRAINING AND VALIDATION LOSS

The loss function we used is “Categorical_Crossentropy”. This loss function is used when there are more than two label classes in our dataset. The formula related to it is as follows

$$\mathcal{L}(\hat{y}, y) = -\frac{1}{N}\sum_{i}^{N}[y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)]$$

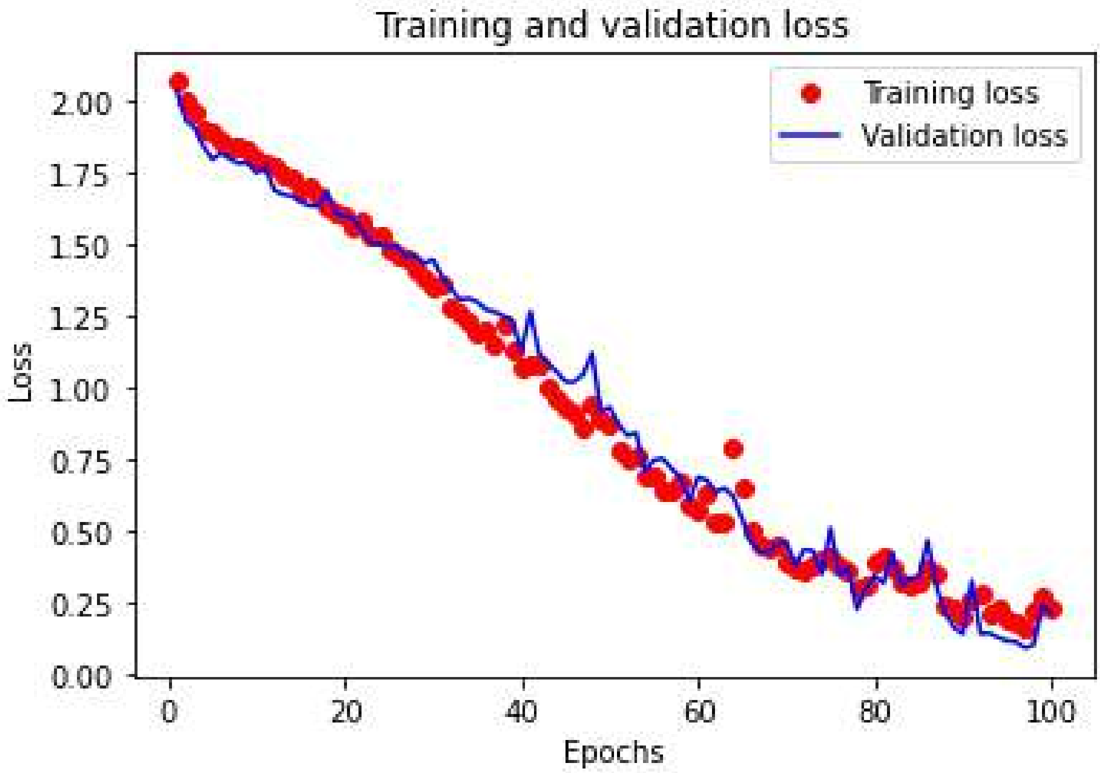

Figure 6.4 Training and validation loss

6.5 TRAINING AND VALIDATION ACCURACY

The mathematical formula behind 'Accuracy metric' is as follows.

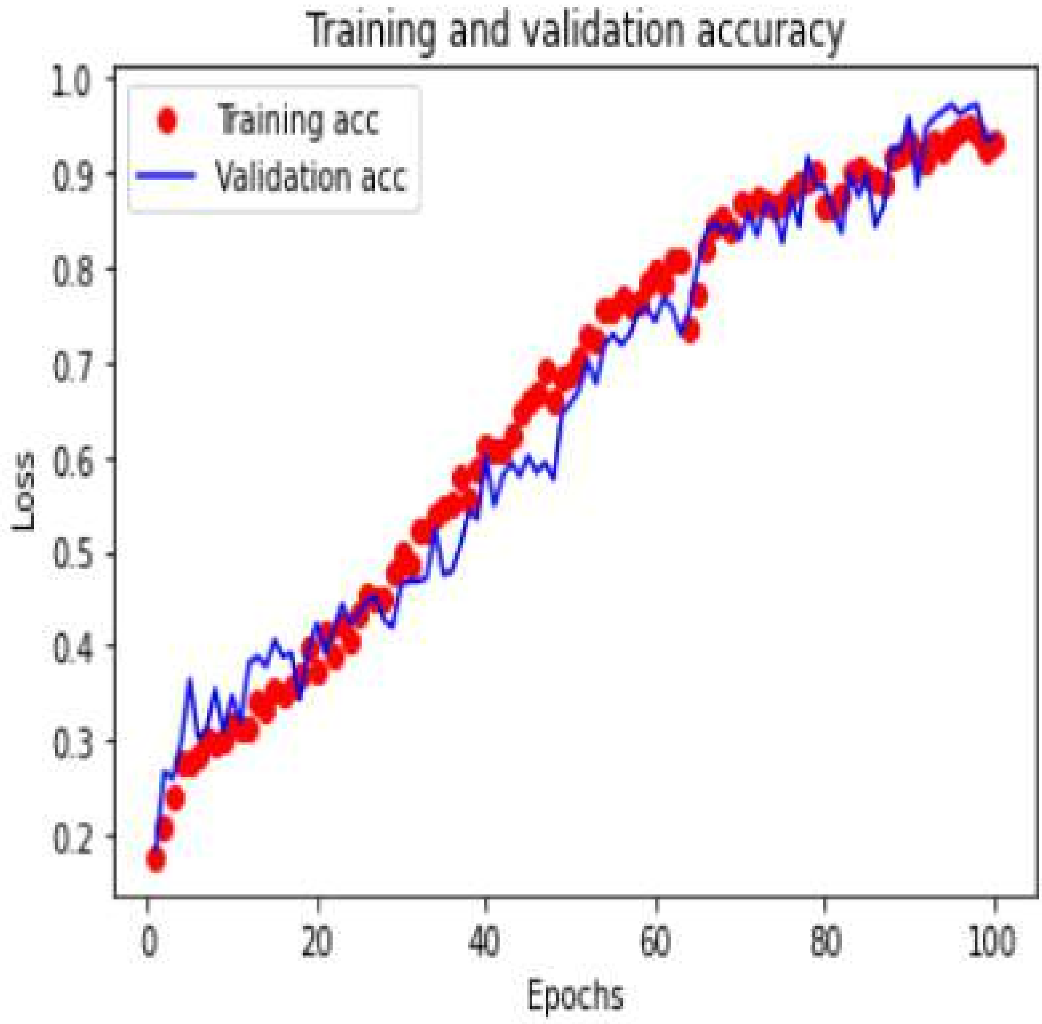

Figure 6.5 Training and validation Accuracy

6.6 SUMMARY OF LSTM MODEL

Model: "sequential_1"

Layer (type)	Output Shape	Param #
lstm_1 (LSTM)	(None, 128)	66560
dense_1 (Dense)	(None, 64)	8256
dropout_1 (Dropout)	(None, 64)	0
activation_1 (Activation)	(None, 64)	0
dense_2 (Dense)	(None, 32)	2080
dropout_2 (Dropout)	(None, 32)	0
activation_2 (Activation)	(None, 32)	0
dense_3 (Dense)	(None, 8)	264
activation_3 (Activation)	(None, 8)	0

Figure 6.6 Summary of LSTM Model

7 . GITHUB LINK

7 GITHUB

7.1 GITHUB LINK

https://github.com/Ambareesh899/Speech-emotion-recognition

8 . CONCLUSION

8 CONCLUSION & FUTURE SCOPE

8.1 PROJECT CONCLUSION

Our project presents a new way to give the ability to a machine to determine the emotion of the human. It will give the machine the ability to have a better approach towards having better and seamless conversations like humans do.SER systems most commonly use prosodic and spectral features since they support a wider range of emotion and yield better results. The results can further be improved by adding features from other modalities, such as the ones that depend on visual or linguistic features. We believe that, as SER systems become more part of our daily lives, there will be more data available to learn from, which will improve their performance, even when at times humans can fail. The subtle differences which may not be registered by humans can be picked up by these networks that will improve the areas where emotion recognition is applicable, such as human computer interaction, healthcare, and alike.

8.2 FUTURE SCOPE

If we can merge speech emotion detection and facial emotion detection in the future, we can get more precise emotional response and we can build a perfect project for emotion detection

9 BIBILOGRAPHY

9.BIBILOGRAPHY

9.1 REFERENCES

- Mirsamadi, E. Barsoum and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 2227-2231, doi: 10.1109/ICASSP.2017.7952552.

- Abdelwahab and C. Busso, "Domain Adversarial for Acoustic Emotion Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 12, pp. 2423-2435, Dec. 2018, doi: 10.1109/TASLP.2018.2867099.

- McFee, Brian, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. “librosa: Audio and music signal analysis

9.2 WEBSITES

- HTTP://librosa.org/doc/latest/index.HTML#broadcasting
- HTTP://WWW.TensorFlow.org/guide
- HTTP://WWW.TensorFlow.org/tutorials/images/CNN
- HTTP://flask.pallets projects.com/en/1.1.x/
- HTTP://devolves.Io/HTML/
- HTTP://bootstrap.com/docs/3.4/CSS/
- https://numpy.org/doc/stable/

SPEECH EMOTION RECOGNITION

M.Madhusudhan[1], R.Mahesh kumar[2], S.Ambareesh[3], CH.Jayanth[4]
[1]*AssistantProfessor,CSE,CMR Technical campus,Hyderabad, India*
[2]*Student, CSE, CMR Technical Campus, Hyderabad, India*
[3]*Student, CSE, CMR Technical Campus, Hyderabad, India*
[4]*Student, CSE, CMR Technical Campus, Hyderabad, India*

***ABSTRACT*: In this paper, a real time Deep learning based system was built for the Emotion Recognition using speech that have been given as input with the help of a PC microphone. The main purpose of this project is to design a model that can record live speech from computer microphone, analyse the audio file, detect and recognize the particular emotion. After recognition is done, the particular user can view their emotion through an interface with appropriate emojis. Speech Emotion Recognition is an rapidly increasing research domain in recent years. In this paper eight basic emotions (Anger, Happy, Fear, Neutral, sad, calm, disgust, surprise) are analyzed from emotional speech signals. In Our project we have used Long Short Term Memory(LSTM) network which is an artificial Recurrent Neural Network. The reason we used LSTM is that it provides higher accuracy while dealing with emotion recognition using speech. Our model was trained using RAVDEES data set which contains 7356 audio files . I have took 2880 files for training my model, in order to increase our accuracy and also to detect and recognize emotion from different speech. Our model scored an accuracy of 93% for training data set.**

***Keywords*: Long Short Term Memory (LSTM), Recurrent neural network, RAVDEES data set.**

1. INTRODUCTION

As we know Emotion plays a crucial role in daily human interactions. It helps us to know other people's feelings while we communicate with them through directly or indirectly.

when we communicate through social media platform like whats app, face book etc we use emojis to express our emotions, but while communicating through mobiles or while exchanging audio files to communicate we cannot detect the emotion of the person while conversation.

To overcome this problem, we developed a deep learning model which is capable recognizing emotions by just through the audio of the speaker.This speech emotion recognition is very useful in various fields like call centers, entertainment, voice assistance , good human computer interactions and education systems.

We have achieved speech emotion recognition using Recurrent neural networks. Our model detects certain emotions from the sound and shows users through the Tinker window. Therefore, it will reduce the problem of not being able to see emotions during a communication through audio exchange only.

This process of emotional processing and speech recognition usually consists of three parts, which were the selection of the emotional expression database, the feature extract, and the emotional recognition.

2. LITERATURE SURVEY

This section provides an overview of the major speech recognition techniques developed in recent years. Because of the importance of emotion recognition in human communication with the computer and the construction of artificial intelligence systems, there are many other recent publications and surveys in which it is conducted with the SER. In this section, we review the most recent studies related to current work.

In 2018, Swain et al. revised studies between 2000 and 2017 in the SER programs according to the three retained methods, input domain, and separators. An important phase of data research research and feature releases; however, only traditional machine learning methods such as CNN, KNN, SVM etc are considered a distinguishing tool, and the authors feel remorse for the neural networks and deep learning methods.

A year later, Khalil et al. reviewed comprehensible approaches to the SER using in-depth reading. Many in-depth, updated learning methods include deep neural network (DNN), convolution neural network (CNN), repetitive neural network (RNN), and auto encoder, spoken and some of their issues and strengths in the study. However, research cannot address accessible ways to overcome weaknesses.

Recently, Anjali et al. has published a review as a summary of how to identify speech emotions. A comprehensive discussion of various factors used in the sensitivity of speech and in the review of the various approaches used for this purpose from 2009 to 2018 is provided in the review. The retrieval of this paper was the depth of the analysis. Still, it can be considered a start.

In 2020, Bas et al. published a brief review of the importance of data sets and features of speech recognition, audio removal; finally, they analyzed the importance of differentiated approaches involving SVM and HMM. The power of the study was to identify a number of factors related to the recognition of speech emotions; however, its weakness is the leak of modern research methods and is briefly mentioned in the interaction of repetitive neural networks as an in-depth learning method.

In our model, the data set we needed to train was the RAVDEES data set. In this training process, we were able to achieve 93% accuracy. In our project we have chosen the LSTM RNN network because it is best suited to solve speech prediction problems.

3. EMOTIONAL SPEECH DATA

The effectiveness and robustness of emotion recognition systems will be easily affected if they are not properly trained with the appropriate database. Therefore, it is imperative to have sufficient and appropriate clauses in the database to train the speech emotion recognition system and to evaluate and verify its effectiveness. In this section, we provide detailed data about the database we used which is the RAVDEES data set.

3.1 RAVDEES DATA SET

The full form of RAVDEES is the Ryerson Audio-Visual Database of Emotional Speech and Song. It is a data set with eight different emotions such as joy, sadness, anger, fear, surprise, disgust, calmness, and neutrality performed by 24 characters (12 male actors and 12 female actors). The total size of this data set is 24.8 GB where we spent about 1.10 GB on our project.

The total audio files set by this data are 7356 audio files. These files are a mixture of both standard sentences and songs. RAVDESS is very rich in a variety of samples; and each feeling is made to performed in two different intensity and both with a normal voice and singing voice. RAVDEES data set consists of North American English accent.

4. FEATURE EXTRACTION

The speech signal contains of many number of parameters that reflects and are related to the emotional characteristics. One of the difficult task in emotion recognition is to decide what features should be used for building model. In recent research, there are many common features are extracted, such as energy, pitch, tone, and some spectrum features such as Linear Prediction Coefficient (LPC), Mel-Frequency Cep-strum Coefficient (MFCC) and Modulation spectral features. In this work, we have selected Mel Frequency Cep-strum Coefficient (MFCC), to extract the emotional features.

4.1 MFCC FEATURES

The Mel-Frequency Cep-strum Coefficient is widely used representation of the spectral property of the voice signals.These are the best in terms of speech recognition as it takes sensitivity to human perceptions of frequency into consideration. In each frame, the Fourier transform and power spectrum are estimated and map on Mel-frequency scale . There are various techniques used for MFCC features extraction some windowing the signal, applying the DFT, taking the log of the magnitude, and then warping the frequencies on a Mel scale.

In our project LSTM classifier will extract all the MFCC features from the given data set and then all the feature vectors are used for training the classifier.

5. SYSTEM ARCHITECTURE

The next phase after completing feature extraction is creating our LSTM mode land training our model and saving it.

The training of our LSTM Network model is done using TensorFlow and Keras. Since our model involves recognition of audio data, we used Recurrent Neural Networks. Recurrent Neural Networks is an extension to Neural Networks which was developed to deal with speech data.

Once the data set is imported and pre-processing of data is completed, we trained our model using TensorFlow and Keras. To avoid the problem of training our data set every time we want to use our model, we saved our trained LSTM model using Keras. Any deep learing model should be well trained in order to get accurate outputs, so in our project we are training our LSTM model with 100 numbers of epochs so as to get more accuracy for our model.

Our LSTM model consists of one hidden LSTM layer with 128 neurons and next we have three dense layers with continuous dropouts. First dense layer consists of 64 neurons with RELU as Activation function and second dense layer consists of 32 neurons with RELU as Activation function and the final dense layer which is an output layer consists of 8 neurons which denotes 8 different emotion states and here we have used SoftMax as an Activation function.

The below figure shows the visual diagram of LSTM classifier used in our project.

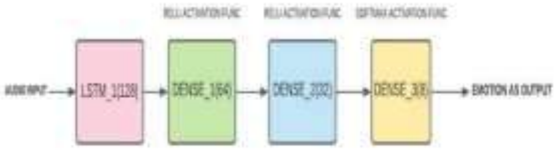

Fig. 1. visual diagram of LSTM

The below figure depicts the system architecture of our Deep learning project.

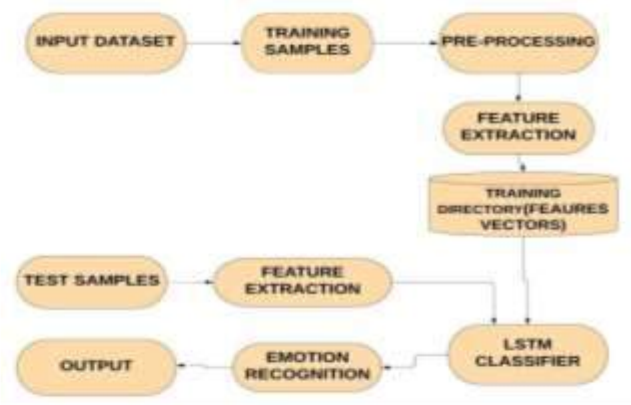

Fig. 2. System Architecture

In order to make our model to recognize the emotion of audios present in our dataset, we need to first train our model. We use TensorFlow module to train.

TensorFlow is an open-source library for numerical computation and large-scale machine learning that ease Google Brain TensorFlow, the process of acquiring data, training models, serving predictions, and refining future results.

TensorFlow bundles together both Machine Learning and Deep Learning models and algorithms. It uses Python as a convenient front-end and runs it efficiently in optimized C++. TensorFlow allows developers to create a graph of computations to perform. Each node in the graph represents a mathematical operations and each connection represents data. Hence, instead of dealing with low-details like figuring out proper ways to hitch TensorFlow allows developers to create a graph of computations to perform. Each node in the graph represents a mathematical operations and each connection represents data. Hence, instead of dealing with low-details like figuring out proper ways to hitch the exact output of one function to the input of another, the developer can simply focus on the overall logic of the application. TensorFlow is for the backend of keras. Thus this tensorFlow plays an major role in almost all deep learning based projects.

A loss function is used to optimize the deep learning algorithm. The loss is calculated on training and testing of model and its interpretation is based on how well the model is performing in these two sets. It is the sum of errors made for each example in training or testing sets. Loss value implies how inefficently or well a model behaves after each iteration of optimization. The loss at each iteration of our deep learning model has been decreasing which indicates a better accuracy of model for detection.

The loss of our model is shown in the below figure.

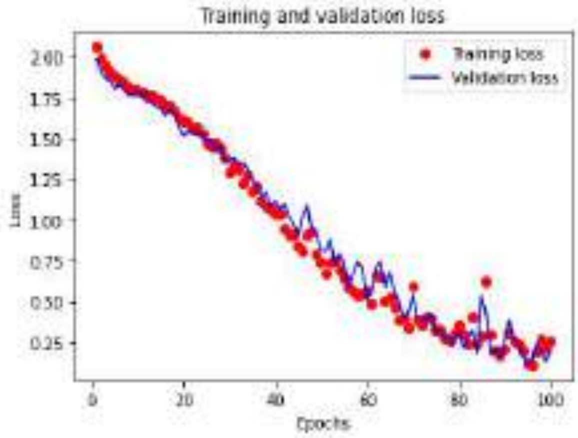

Fig. 3. Loss Graph of our model

The loss function we used is "Categorical_Crossentropy". This loss function is used when there are more than two label classes in our dataset. The formula related to it is as follows

$$\mathcal{L}(\hat{y}, y) = -\frac{1}{N}\sum_{i}^{N}[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)]$$

In our project we have trained our LSTM model with 100 number of epochs so as to get more accurate results as we know that any deep learning models accuracy depends upon how many number of times it is being trained. We can set epochs to any number of times depending upon our requirement. In first epoch the loss will be more and accuracy will be less but when it comes to last epoch loss and accuracy are vice versa; loss is less and accuracy is more.

The below figure shows the loss after each epoch in our training process.

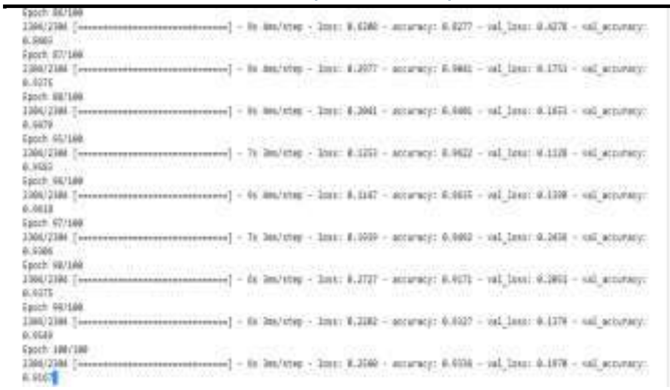

Fig. 4. Loss after each epoch of our model

Since our model is a predictive algorithm as it predicts the emotions of human speech, we must need to evaluate the model before using in real time. Evaluation is nothing but checking the extent up to which we can use our developed model. There are many evaluation criteria present in Keras. We chose our evaluation metric as 'Accuracy'

The reason we chose accuracy is that it will give accurate results when all the classes in our dataset has same number of records in each folder. Since our dataset contains equal number of records in each folder, we proceeded using this as our evaluation metric. The accuracy graph of our model is as shown in below figure.

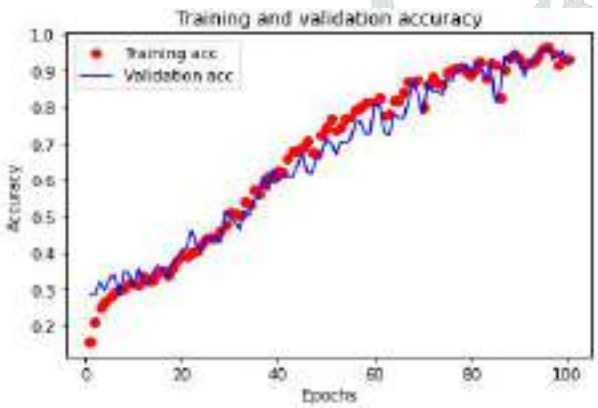

Fig. 5. Accuracy Graph of our model

The mathematical formula behind 'Accuracy metric' is as follows.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

The activation functions we used in our two Dense layers is 'ReLU'. ReLU stands for rectified linear unit, and it is a type of activation function. Mathematically, it is defined as *y = max (0, x)*. Visually, it looks like the following:

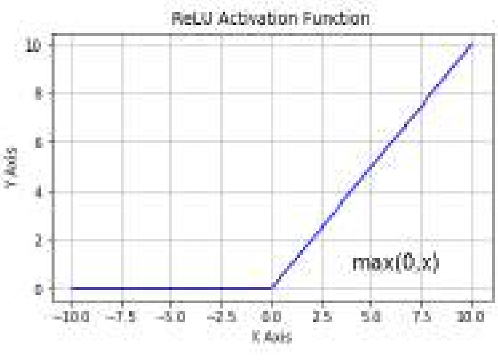

Fig. 6. ReLU Activation Function Graph

The activation function we used in our last layer i.e.; Dense layer is 'SoftMax'. SoftMax activation function is used when there are more than two classes present in our dataset or when our model should predict more than two class labels and our model need to predict eight emotion so we have used SoftMax Activation Function. The mathematical formula related to 'SoftMax' is as follows.

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

6. REAL TIME EMOTION RECOGNITION

The next phase after training our model is recognizing the emotions using real time Audio feed from computer microphone. Emotion recognition from real time data is done by importing the trained model using Keras. We need to access the microphone of our laptop/computer with the use of Sounddevice. The live audio recording using sounddevice module is sent for pre- processing of audio. Our system microphone is capable of recording 44,100 audio samples per second. However it is not unusual to see 96,000 samples **a second** with some digital audio formats. So, our algorithm sends around 48,000 samples per second for pre-processing. Usually, the audio extracted from computer microphone is of size 44.1 kHZ with 3 channels. Initially while training our model, we set the size of our audio to 48,000 multiplied by 0.8 sec . So, we need to resize the raw feed in order to sent it to Detection algorithm.

After pre-processing of audio samples is completed, then it will extract the mfcc and all other related features from audio, then the detection algorithm will be able to analyse the features and then finally it will detect the exact emotion of the speaker in real time.

7. RESULTS AND DISCUSSIONS

A real-time Speech Emotion Recognition using Neural Networks named LSTM is introduced. In this paper,Speech Emotion recognition is done to help easily find out the emotion of customers in call centers and many other wide applications. This system showed good results in recognizing Emotions of particular person and shows results with appropriate Emojis. With the help of Neural Networks, in particular Recurrent Neural Networks we were able to develop a LSTM model and thereby easing the work of emotion detection.

Neural Networks results in more accurate recognition predictions compared to other methods of speech emotionl recognition. I strongly believe neural networks as an ideal way to solve this problem.

Below figure shows the execution of our Speech emotion recognition model.

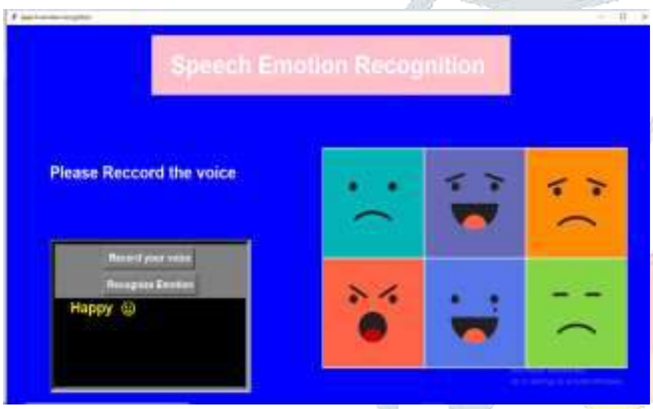

Fig. 7. Emotion recognition of real time audio

As soon as the recognition of emotion is done, the user will be able to check his/ her emotion through an tkinter interface along with related emojis.

8 CONCLUSION

In this paper, a real-time Deep learning based Speech Emotion Recognition system was built. The use of Recurrent Neural Networks that is an LSTM network helps our model to recognize the emotion of an audio file or emotion from particular persons speech.

The system achieved a maximal accuracy of about 93% for training and 91% for the validation set. Our model will also recognize emotions of speech files that aren't trained by our model.

Through this project we have shown that how to implement Speech Emotion Recognition through deep learning. Our project give ability to a machine to determine the emotions of human through his speech so that there will be a better communication between humans and machines.

If we merge speech emotion recognition and facial emotion recognition in the future, then we can get more precise emotional response and we can build a perfect project for emotion detection.

speech emotion recognition and understanding will eventually show the way to true artificial intelligence.

8. REFERENCES

1. Booth, P.A. An Introduction to Human-Computer Interaction; Psychology Press: Hove, UK, 1989.

2. Harper, E.R.; Rodden, T.; Rogers, Y.; Sellen, A. Being Human: Human-Computer Interaction in the Year 2020; Microsoft Research: Redmond, WA, USA, 2008; ISBN 0955476119.

3. Cambria, E.; Hussain, A.; Havasi, C.; Eckl, C. Sentic computing: Exploitation of common sense for the development of emotionsensitive systems. In Development of Multimodal Interfaces: Active Listening and Synchrony; Springer: Berlin/Heidelberg, Germany, 2010; pp. 148–156.

4. Patil, K.J.; Zope, P.H.; Suralkar, S.R. Emotion Detection From Speech Using Mfcc and Gmm. Int. J. Eng. Res. Technol. (IJERT) 2012, 1, 9.

5. Hassan, A.; Damper, R.I. Multi-class and hierarchical SVMs for emotion recognition. In Proceedings of the INTERSPEECH 2010, Makuhari, Japan, 26–30 September 2010; pp. 2354–2357.

6. Lin, Y.L.; Wei, G. Speech emotion recognition based on HMM and SVM. In Proceedings of the 2005 International Conference on Machine Learning and Cybernetics, Guangzhou, China, 18–21 August 2005; Volume 8, pp. 4898–4901.

7. Nicholson, J.; Takahashi, K.; Nakatsu, R. Emotion Recognition in Speech Using Neural Networks. In Proceedings of the 6th International Conference on Neural Information Processing (ICONIP '99), Perth, Australia, 16–20 November 1999.

8. Schüller, B.; Rigoll, G.; Lang, M. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, QC, Canada, 17–21 May 2004.

9. France, D.J.; Shiavi, R.G.; Silverman, S.; Silverman, M.; Wilkes, D.M. Acoustical Properties of Speech as Indicators of Depression and Suicidal Risk. IEEE Trans. Biomed. Eng. 2000, 47, 829–837. [CrossRef] [PubMed]

10. Hansen, J.H.; Cairns, D.A. ICARUS: Source generator based real-time recognition of speech in noisy stressful and Lombard effect environments. Speech Commun. 1995, 16, 391–422. [CrossRef]

JETIR
Journal of Emerging Technologies and Innovative Research
An International Open Access Journal Peer-reviewed, Refereed Journal
www.jetir.org | editor@jetir.org An International Scholarly Indexed Journal

Certificate of Publication

The Board of

Journal of Emerging Technologies and Innovative Research (ISSN : 2349-5162)

Is hereby awarding this certificate to

R.Mahesh Kumar

In recognition of the publication of the paper entitled

SPEECH EMOTION RECOGNITION

Published In JETIR (www.jetir.org) ISSN UGC Approved (Journal No: 63975) & 7.95 Impact Factor

Published in Volume 8 Issue 6 , June-2021 | Date of Publication: 2021-06-10

Parin P

EDITOR

EDITOR IN CHIEF

ISSN 2349-5162

JETIR2106163 **Research Paper Weblink http://www.jetir.org/view?paper=JETIR2106163** **Registration ID : 310474**

Journal of Emerging Technologies and Innovative Research
An International Open Access Journal Peer-reviewed, Refereed Journal
www.jetir.org | editor@jetir.org An International Scholarly Indexed Journal

Certificate of Publication

The Board of

Journal of Emerging Technologies and Innovative Research (ISSN : 2349-5162)

Is hereby awarding this certificate to

S.Ambareesh

In recognition of the publication of the paper entitled

SPEECH EMOTION RECOGNITION

Published In JETIR (www.jetir.org) ISSN UGC Approved (Journal No: 63975) & 7.95 Impact Factor

Published in Volume 8 Issue 6 , June-2021 | Date of Publication: 2021-06-10

Parin P

EDITOR

EDITOR IN CHIEF

ISSN 2349-5162

JETIR2106163 **Research Paper Weblink http://www.jetir.org/view?paper=JETIR2106163** **Registration ID : 310474**

Journal of Emerging Technologies and Innovative Research
An International Open Access Journal Peer-reviewed, Refereed Journal
www.jetir.org | editor@jetir.org An International Scholarly Indexed Journal

Certificate of Publication

The Board of

Journal of Emerging Technologies and Innovative Research (ISSN : 2349-5162)

Is hereby awarding this certificate to

Ch.Jayanth

In recognition of the publication of the paper entitled

SPEECH EMOTION RECOGNITION

Published In JETIR (www.jetir.org) ISSN UGC Approved (Journal No: 63975) & 7.95 Impact Factor

Published in Volume 8 Issue 6 , June-2021 | Date of Publication: 2021-06-10

Parin P

EDITOR

EDITOR IN CHIEF

ISSN 2349-5162

JETIR2106163 **Research Paper Weblink http://www.jetir.org/view?paper=JETIR2106163** **Registration ID : 310474**

Journal of Emerging Technologies and Innovative Research
An International Open Access Journal Peer-reviewed, Refereed Journal
www.jetir.org | editor@jetir.org An International Scholarly Indexed Journal

Certificate of Publication

The Board of

Journal of Emerging Technologies and Innovative Research (ISSN : 2349-5162)

Is hereby awarding this certificate to

M.Madhusudhan

In recognition of the publication of the paper entitled

SPEECH EMOTION RECOGNITION

Published In JETIR (www.jetir.org) ISSN UGC Approved (Journal No: 63975) & 7.95 Impact Factor

Published in Volume 8 Issue 6 , June-2021 | Date of Publication: 2021-06-10

Pariu P
EDITOR

EDITOR IN CHIEF

ISSN 2349-5162

JETIR2106163 Research Paper Weblink http://www.jetir.org/view?paper=JETIR2106163 Registration ID : 310474