SPEECH EMOTION RECOGNITION

Mini Project Report

Submitted by

Ann Sarah Babu

Reg No: FIT20MCA-2029

Submitted in partial fulfillment of the requirements for the award of the degree of

Master of Computer Applications
Of
A P J Abdul Kalam Technological University

FEDERAL INSTITUTE OF SCIENCE AND TECHNOLOGY (FISAT)®
ANGAMALY-683577, ERNAKULAM(DIST)
FEBRUARY 2022

DECLARATION

I, **Ann Sarah Babu** hereby declare that the report of this project work, submitted to the Department of Computer Applications, Federal Institute of Science and Technology (**FISAT**), Angamaly in partial fulfillment of the award of the degree of Master of Computer Application is an authentic record of our original work.

The report has not been submitted for the award of any degree of this university or any other university.

Date :

Place: Angamaly

FEDERAL INSTITUTE OF SCIENCE AND TECHNOLOGY (FISAT)®

ANGAMALY, ERNAKULAM-683577

DEPARTMENT OF COMPUTER APPLICATIONS

CERTIFICATE

This is to certify that the project report titled **"Speech Emotion Recognition"** submitted by **Ann Sarah Babu** towards partial fulfillment of the requirements for the award of the degree of Master of Computer Applications is a record of bonafide work carried out by them during the year 2022.

Project Guide **Head of the Department**

Submitted for the viva-voice held on ..... at .....

ACKNOWLEDGEMENT

I am extremely glad to present my mini project which I did as a part of our curriculum. I take this opportunity to express my sincere thanks to those who helped me in bringing out the report of my project.

I am deeply grateful to **Dr. Manoj George** , Principal, FISAT, Angamaly and **Dr. C. Sheela** ,Vice Principal FISAT, Angamaly.

My sincere thanks to **Dr. Deepa Mary Mathews**, Head of the department of MCA, FISAT, who had been a source of inspiration. I express heartiest thanks to **Dr. Deepa Mary Mathews** my project guide for her encouragement and valuable suggestions. I express my heartiest gratitude to my scrum master **Ms. Rosemary Mathew** and the faculty members in our department for their constant encouragement and never ending support throughout the project.I would also like to express my sincere gratitude to the lab faculty members for their guidance.

Finally I express my thanks to all my friends who gave me wealth of suggestion for successful completion of this project.

ABSTRACT

Human communications have come a long way from cave drawings to sophisticated internet media. Emotions are conveyed by all humans in one way or another but speech remains most crucial of all as it is the most natural and convenient method to convey our feelings. Speech usually contains two types of information, paralinguistic and linguistic. Linguistic information conveys language, accent, dialect and paralinguistic information conveys emotional state, context, gender, environment and attitude thus using paralinguistic features is beneficial as it really helps to understand the emotional state of the speaker but each language has its way of expressing feeling irrespective of the speaker and its interpretation is also subjective.

This project use Neural Networks to classify the emotions from a given speech, known as Speech Emotion Recognition (SER). It is based on the fact that voice often reflects underlying emotion through tone and pitch. Speech Emotion Recognition helps to classify elicit specific types of emotions. The MLP-Classifier is used to classify the emotions from the given wave signal, which makes the choice of learning rate to be adaptive.

Contents

Chapter 1

Introduction

Speech is a complex signal which contains information about the message, speaker, language and emotions. There are various kinds of emotions which can be articulated using speech.Emotional speech recognition is a system which basically identifies the emotional state of human being from his or her voice; speech is very misleading even for humans to judge the emotion of the speaker.The necessity to develop emotionally intelligent systems is exceptionally important for the modern society because such systems have great impact on decision making, social communication and smart connectivity.

Traditional emotional feature extraction concentrates on the analysis of the emotional features in the speech from time construction, amplitude construction, and fundamental frequency construction and signal feature. In this proposed system we use the deep neural networks, to extract the emotional characteristic parameter from emotional speech signal automatically.

This project aims to study how to improve the recognition rate of speech emotion recognition. It recognize the underlying emotional state of a speaker from his/her voice. Practical applications of emotion recognition systems can be found in many domains such as audio/video surveillance, web-based learning, commercial applications, clinical studies, entertainment, banking, call centers, computer games and psychiatric diagnosis.

Chapter 2

PROOF OF CONCEPT

As humans, speech is among the maximum herbal manner to express ourselves. As feelings play a critical position in communication, the detection and evaluation of the same is of critical significance in today's world. We outline a SER system as a group of methodologies that operate and classify speech signals to detect emotions in them. Applications of speech emotion recognition system are wide-ranging from your home to industries, it's already in use in our assistants and smart speaker. Humanmachine interface will reach new heights as SER systems will increase accessibility, useability and help people from all sections of life.

2.1 Objectives

- The primary objective of SER is to improve man-machine interface.
- To choose a good speech database.
- To extract effective features, and to design reliable classifiers using machine learning algorithms.
- To predict the emotion from speech.

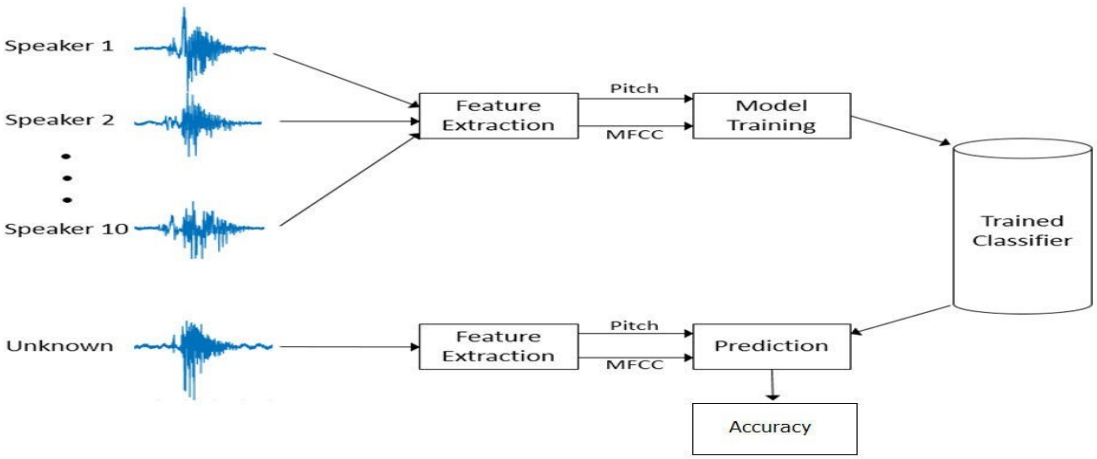
Chapter 3

IMPLEMENTATION

In the Speech Emotion Recognition System (SER), the audio files are given as the input. The data sets travels through a number of blocks of processes which makes it executable to help for the analysis of the speech parameters. The data is preprocessed to change it to the suitable format and the respective features from the audio files are extracted using various steps such as framing, hamming, windowing, etc. This process helps in breaking down the audio files into the numerical values which represents the frequency, time, amplitude or any other such parameters which can help in the analysis of the audio files.

After the extraction of the required features from the audio files, the model is trained. We have used the RAVDESS dataset of audio files which has speeches of 24 people with variations in parameters. For the training, we store the numerical values of emotions and their respective features correspondingly in different arrays. These arrays are given as an input to the MLP Classifier that has been initialized.

The Classifier identifies different categories in the datasets and classifies them into different emotions. The model will now be able to understand the ranges of values of the speech parameters that fall into specific emotions. For testing the performance of the model, if we enter the unknown test dataset as an input, it will retrieve the parameters and predict the emotion as per training dataset values. The accuracy of the system is displayed in the form of percentage.



3.1 Dataset

The dataset used is Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). It's a Speech audio-only files (16bit, 48kHz .wav). RAVDESS contains 1440 files: 60 trials per actor x 24 actors = 1440. The file consists of 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent.Speech emotions includes calm, happy, sad, neutral, surprise and angry expressions. Each expression is produced at two levels of emotional intensity (normal, strong).

3.2 Modules

3.2.1 DATA PREPARATION

The initial step is to prepare the data for training on a neural network. Data must be numerical. If you have categorical data, such as emotion attributes such as "happy", "sad", "angry" etc, we can convert it to a real-valued representation.

3.2.2 FEATURE EXTRACTION

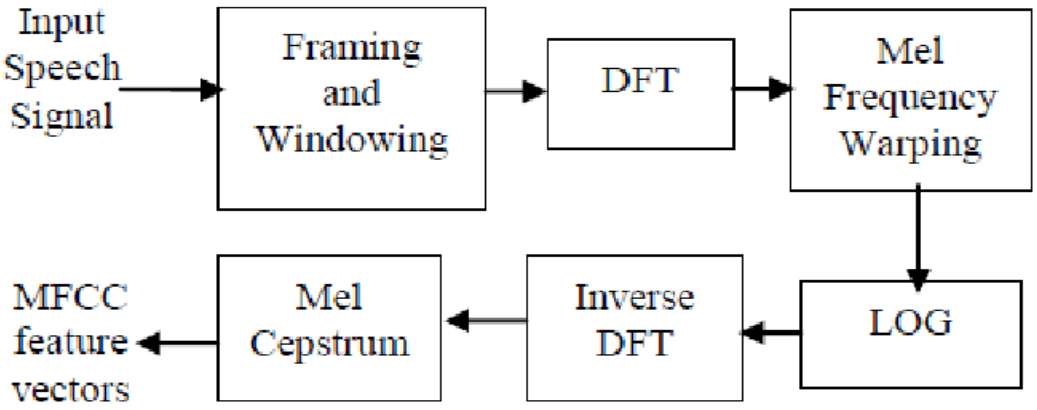
Voice frequently reflects hidden feeling through tone and pitch. The objective of feature extraction is to reveal applicable feature from discourse signals as for feelings. Five feature are extracted from the discourse signals given as information. The five features are, MFCC, Contrast, Mels Spectrograph Frequency, Chroma and Tonnetz.

(a) MFCC

Mel Frequency Cepstral Coefficients(MFCC), are the final features used in machine learning models trained on audio data.The MFCC feature extraction technique basically includes windowing the signal, applying the DFT, taking the log of the magnitude, and then warping the frequencies on a Mel scale, followed by applying the inverse DCT.

Steps for calculating MFCCs for a given audio sample:

1. Slice the signal into short frames (of time).
2. Compute the periodogram estimate of the power spectrum for each frame.
3. Apply the mel filterbank to the power spectra and sum the energy in each filter.
4. Take the discrete cosine transform (DCT) of the log filterbank energies.



(b) Chroma

The chroma feature, which are also referred to as ”pitch class profiles” is a descriptor which represents the tonal content of a audio signal in a condensed form. A better quality of the extracted chroma feature enables much better results in these high-level tasks. Short Time Fourier Transforms and Constant Q Transforms are used for chroma feature extraction.

(c) MEL Spectrogram Frequency

A mel spectrogram is a spectrogram where the frequencies are converted to the mel scale.The Mel scale relates evident repeat, or pitch, of an unadulterated tone to its real assessed recurrence. Individuals are incredibly improved at perceiving little changes in pitch at low frequencies than they are at high frequencies. Solidifying this scale makes our features arrange even more eagerly what individuals listen.

(d) Contrast
In an audio signal, the spectral contrast is the measure of the energy of frequency at each timestamp. Since most of the audio files contain the frequency whose energy is changing with time. It becomes difficult to measure the level of energy. Spectral contrast is a way to measure that energy variation.

(e) Tonnetz
Computes the tonal centroid features. This is also feature which detects the changes in the harmonic content of the audio signals. A peak in the detection function represents that a transition was made from one harmonically stable region to another.

3.2.3 TRAINING & TESTING THE MODEL

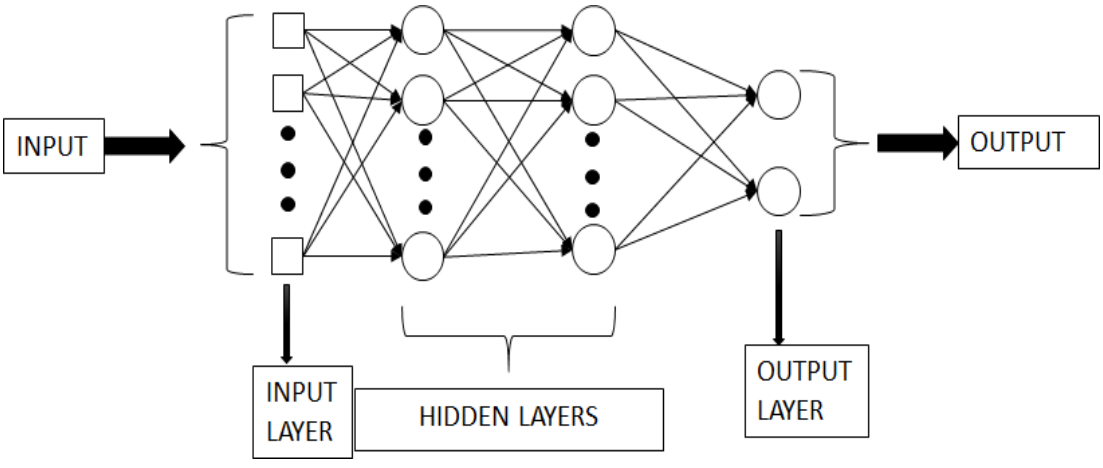
The input to the model should be the features extracted along with the emotion category that it belongs to, classifier will be able to identify and then classify the data. This training helps the model to understand, which emotions have what range of the respective features. So, when an unseen data is given as an input, it will be able to correlate and predict the emotion.
The Ravdess dataset is passed to the MLP Classifier to train the model, we split the dataset into a 75:25 ratio, i.e.; the training and testing dataset.

3.2.4 DEPLOYMENT

Model deployment is simply the engineering task of exposing an ML model to real use. The term is often used quite synonymously with making a model available via real-time APIs.

3.3 Algorithm

MLPClassifier stands for Multi-layer Perceptron classifier which in the name itself connects to a Neural Network. MLPClassifier relies on an underlying Neural Network to perform the task of classification. A Multi-Layer Perceptron (MLP) is a network made up of perceptron. It has an input layer that receives the input signal, an output layer that makes predictions or decisions for a given input, and the layers present in between the input layer and output layer is called hidden layer. There can be many hidden layers, the number of hidden layers can be changed as per requirement. In the proposed methodology for Speech Emotion Recognition, the Multi-Layer Perceptron Network will have one input layer, hidden layers and one output layer. The input layer will take as input, the five features, that are extracted from the audio file. The extracted five features being, Mel Frequency Cepstral Coefficients, Mel Spectrogram Frequency, Chroma, Tonnetz and Contrast. The hidden layer uses an activation function to act upon the input data and to process the data. The activation function used is logistic activation function. The output layer brings out the information learned by the network as output. This layer classifies and gives output of the predicted emotion, according to the computation performed by the hidden layer. The figure below illustrates Multilayer Perceptron.



Chapter 4

RESULT ANALYSIS

The result of the proposed project Speech Emotion Detection using MLP-Classifier lies in developing a model that can successfully predict the underlying emotional state of a speaker from his/her voice, depending on the features identified from the given input(voice). The proposed system takes speech as input and predict the corresponding emotion as output.

Chapter 5

CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

Our ultimate aim is to study how to improve the recognition rate of speech emotion recognition.Human–machine interaction is becoming popular day by day; to interact with machine, speech emotion recognition is as important as human to human interaction.Through this project, we demonstrate a speech emotion recognition system which takes speech as input and classify emotions that the speech contains. We choose multilayer perceptron (MLP) classifier to do this task.

This project, showed how we can leverage Machine learning to obtain the underlying emotion from speech audio data and some insights on the human expression of emotion through voice. This system can be employed in voice-based virtual assistants or chatbots, in linguistic research, etc.In addition, other real-time applications like remote tracking of persons in a distressed phase, communication between human and robots, customer care services, where emotion is perpetually expressed.

5.2 Future Scope

In future work, we will continue to further study speech emotion recognition and further expand the training data set. Our ultimate aim is to study how to improve the recognition rate of speech emotion recognition.

This proposed model is only able to detect emotions by using voice as input; further, we can to try to detect emotions by using video, image, text and also by combining all these four videos,image,text, and voice as inputs so that it will be advantageous in applications where emotion detection is essential for acting or responding according to that particular person emotions. In the future, we can also try with multiple emotions from a single input file since, in real-time, there may be various emotions when people speak. Currently, this model can detect only single emotion from the given input speech signal.

.

Chapter 6

APPENDIX

6.1 Coding

6.1.1 utils.py

```
import soundfile
import numpy as np
import librosa
import glob
import os
from sklearn.model_selection import train_test_
# all emotions on RAVDESS dataset
int2emotion = {
"01": "neutral",
"02": "calm",
"03": "happy",
"04": "sad",
"05": "angry",
```

```
”06”: ”fearful”,
”07”: ”disgust”,
”08”: ”surprised”
}

# we allow only these emotions
AVAILABLE_EMOTIONS = {
”angry”,
”sad”,
”neutral”,
”happy”
}

def extract_feature(file_name, **kwargs):
”””
Extract feature from audio file ‘file_name
‘ Features supported:
- MFCC (mfcc)
- Chroma (chroma)
- MEL Spectrogram Frequency (mel)
- Contrast (contrast)
- Tonnetz (tonnetz)
e.g:
‘features = extract_feature(path, mel=True, mfcc=True)
‘ ”””
mfcc = kwargs.get(”mfcc”)
chroma = kwargs.get(”chroma”)
mel = kwargs.get(”mel”)
contrast = kwargs.get(”contrast”)
tonnetz = kwargs.get(”tonnetz”)
```

```
with soundfile.SoundFile(file_name) as sound_file:
X = sound_file.read(dtype="float32")
sample_rate = sound_file.samplerate
if chroma or contrast:
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.me(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))
if contrast:
contrast =
np.mean(librosa.feature.spectral_contrast(S=stft,
sr=sample_rate).T,axis=0)
result = np.hstack((result, contrast))
if tonnetz:
tonnetz =
np.mean(librosa.feature.tonnetz(y=librosa.effects.harmoni
c(X), sr=sample_rate).T,axis=0)
result = np.hstack((result, tonnetz))
return result
```

```
def load_data(test_size=0.2):
X, y = [], []
for file in glob.glob("data/Actor_*/*.wav"):
# get the base name of the audio file
basename = os.path.basename(file)
# get the emotion label
emotion = int2emotion[basename.split("-")[2]]
# we allow only AVAILABLE_EMOTIONS we set
if emotion not in AVAILABLE_EMOTIONS:
continue
# extract speech features
features = extract_feature(file, mfcc=True, chroma=True,
mel=True)
# add to data
X.append(features)
y.append(emotion)
# split the data to training and testing and return it
return train_test_split(np.array(X), y, test_size=test_size,
random_state=7)
```

6.1.2 ser.py

```
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import accuracy_score
from utils import load_data
import os
import pickle
# load RAVDESS dataset
X_train, X_test, y_train, y_test = load_data(test_size=0.25)
# print some details
# number of samples in training data
print("[+] Number of training samples:", X_train.shape[0])
# number of samples in testing data
print("[+] Number of testing samples:", X_test.shape[0])
# number of features used
# this is a vector of features extracted
# using utils.extract_features() method
print("[+] Number of features:", X_train.shape[1])
# best model, determined by a grid search
model_params = {
'alpha': 0.01,
'batch_size': 256,
'epsilon': 1e-08,
'hidden_layer_sizes': (300,),
'learning_rate': 'adaptive',
'max_iter': 500,
}
# initialize Multi Layer Perceptron classifier
# with best parameters ( so far )
model = MLPClassifier(**model_params)
# train the model
```

```
print("[*] Training the model...")
model.fit(X_train, y_train)
# predict 25y_pred = model.predict(X_test)
# calculate the accuracy
accuracy = accuracy_score(y_true=y_test, y_pred=y_pred)
print("Accuracy: :.2f

# now we save the model
# make result directory if doesn't exist yet
if not os.path.isdir("result"):
os.mkdir("result")


pickle.dump(model, open("result/mlp_classifier.model", "wb"))
```

6.1.3 test.py

```
import pyaudio
import os
import wave
import pickle
from sys import byteorder
from array import array
from struct import pack
from sklearn.neural_network import MLPClassifier
from utils import extract_feature


# Flask utils
from flask import Flask, redirect, url_for, request,
render_template
from werkzeug.utils import secure_filename
#from gevent.pywsgi import WSGIServer


# Define a flask app
app = Flask(__name__)


@app.route('/', methods=['GET'])
def index():
# Main page
return render_template('index.html')


@app.route('/predict', methods=['GET'])
def upload():
# if request.method == 'POST':
```

```
# f = request.files['file']
# basepath = os.path.dirname(__file__)
# if not os.path.exists('uploads'):
# os.mkdir('uploads')
# file_path = os.path.join(basepath, 'uploads',
secure_filename(f.filename))
# f.save(file_path)


#Prediction


# load the saved model (after training)
try:
model = pickle.load(open("result/mlp_classifier.model",
"rb"))
# extract features and reshape it
features = extract_feature("test.wav", mfcc=True,
chroma=True, mel=True).reshape(1, -1)
# predict
result = 'you are '+ model.predict(features)[0]
# show the result !
print("result:", result)
return render_template('pred.html',emo=result)
except:
return render_template('pred.html',emo='No file found')
# return None
if __name__ == '__main__':
app.run(debug=True)
```

Chapter 7

SCREEN SHOTS

Here I add some sample screenshots of the proposed system:

- Home Screen
- Emotion ”Happy” is predicted
- Emotion ”Sad” is predicted
- Emotion ”Angry” is predicted

Figure 7.1: Home Screen

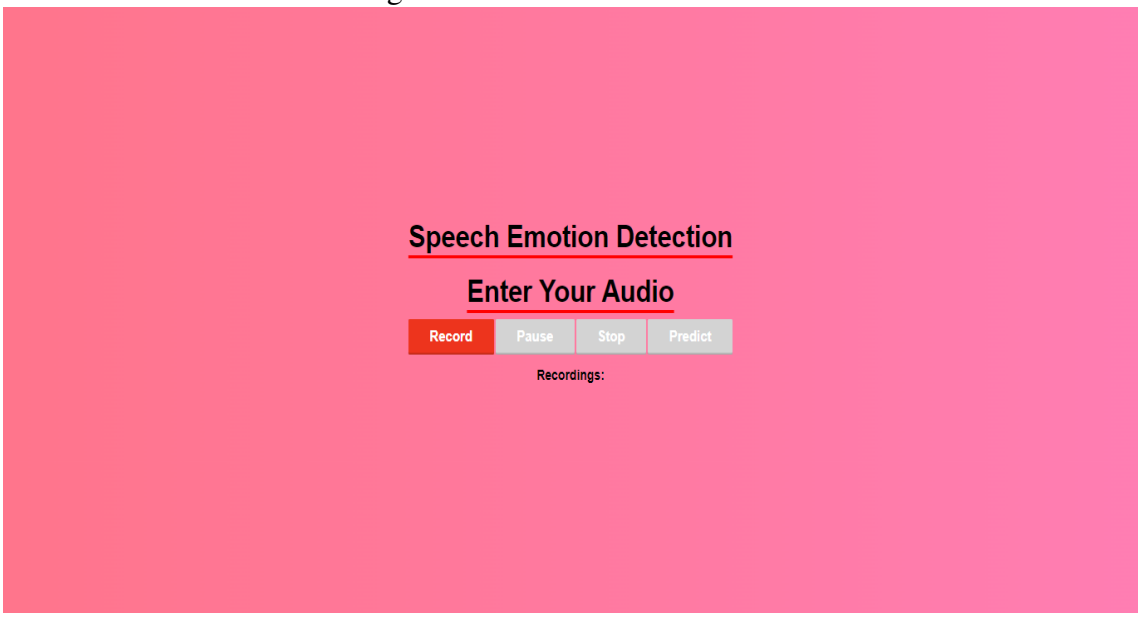


Figure 7.2: Prediction1

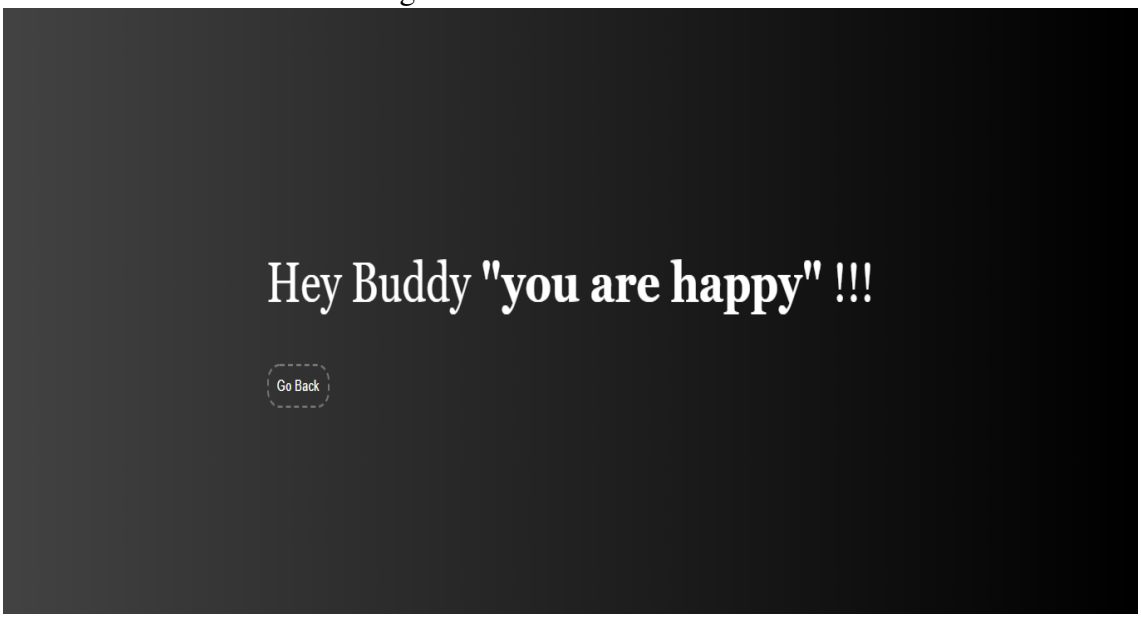


Figure 7.3: Prediction2

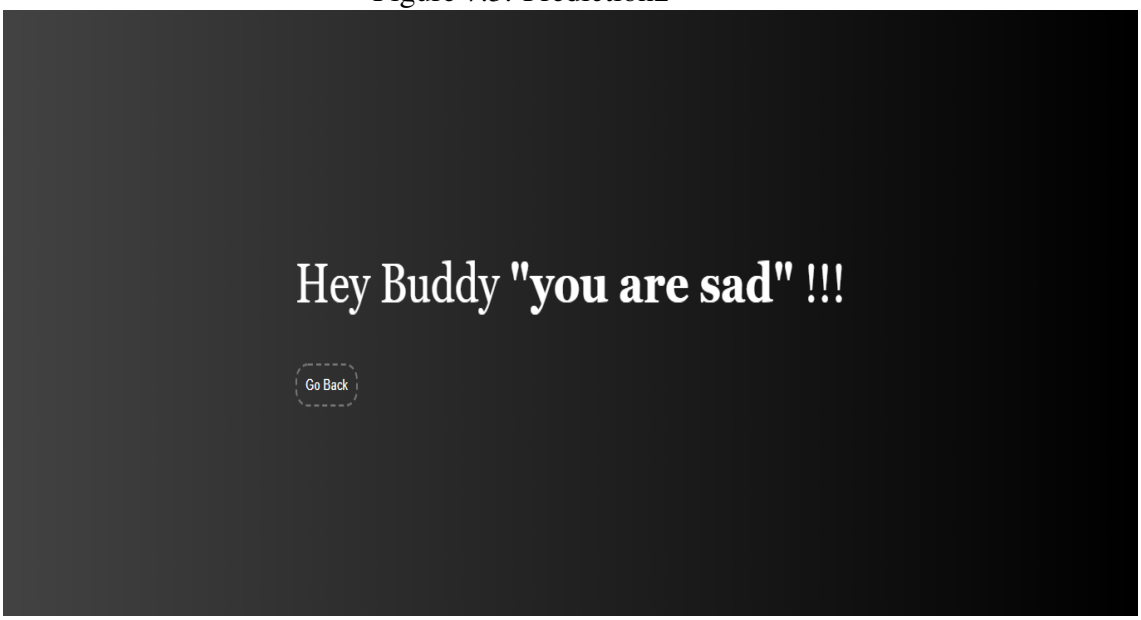
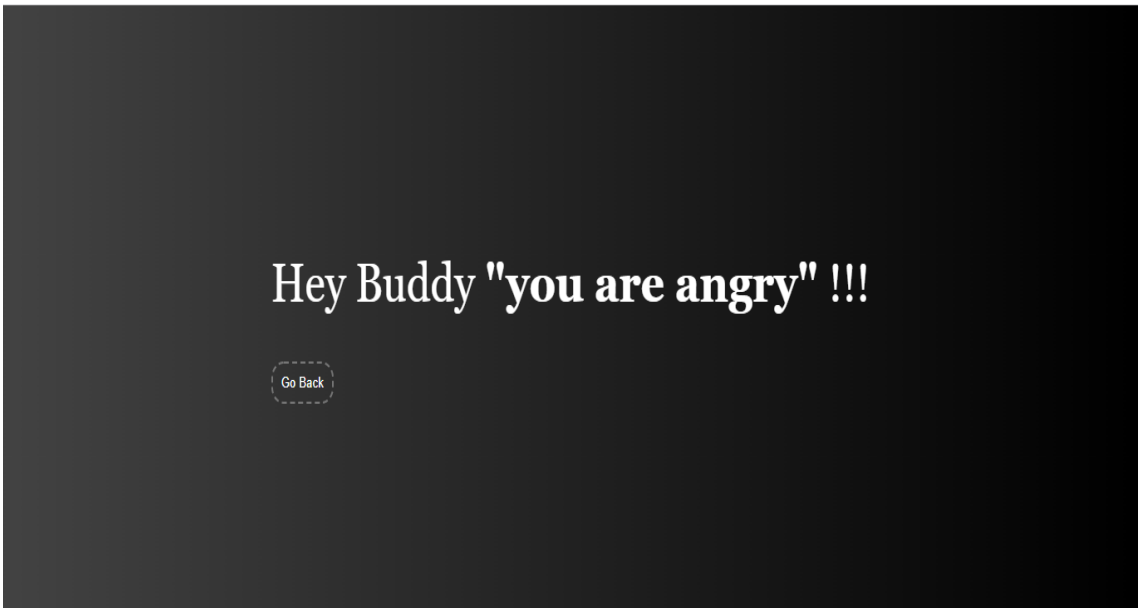


Figure 7.4: Prediction3



Chapter 8

REFERENCES

(a) https://youtu.be/aircAruvnKk

(b) https://ieeexplore.ieee.org

(c) https://www.analyticsinsight.net/speech-emotion-recognition-ser-through-machine-learning/

(d) https://www.kaggle.com/kuntaldas599/emotional-speech-classification2d

(e) H.K. Palo, Mihir Narayana Mohanty and Mahesh Chandra. Use of different features for Emotion Recognition using MLP network. Springer India 2015, Computational Vision and Robotics, Advances in Intelligent Systems and Computing

(f) HTTP://PRACTICALCRYPTOGRAPHY.COM/MISCEL LANEOUS/MACHINE-LEARNING/GUIDE-MELFREQUENCY-CEPSTRAL-COEFFICIENTS-MFCCS/

(g) Ayush Kumar Shah ,Mansi Kattel,Araju Nepal. Chroma Feature Extraction using Fourier Transform. Chroma_ Feature_ xtraction. January 2019