Research on Speech Emotion Recognition System Using Machine Learning

Atharva Yerawar[1], Divya Aswani[2], Aarti Ingole[3], Maviyanaaz Sheikh[4], Prof. D.A. Sananse[5]
U.G Students, Department of Computer Science and Engineering[1,2,3,4]
Professor, Department of Computer Science and Engineering[5]
Jawaharlal Darda Institute of Engineering and Technology, Yavatmal, Maharashtra, India

ABSTRACT— Speech Emotion Recognition (SER) remains a hot topic in the domain of affective computing, drawing considerable research interest. Its growing potential, advancements in algorithms, and real-world applications contribute to this interest. Human speech carries paralinguistic cues that can be quantitatively represented through features like pitch, intensity, and Mel-Frequency Cepstral Coefficients (MFCC). SER typically involves three main stages: data processing, feature extraction/selection, and classification based on emotional features present. These steps, tailored to the unique attributes of human speech, make machine learning methods a fitting choice for SER implementation. While recent studies in affective computing leverage various ML techniques for SER tasks, few delve into the techniques aiding these core steps. Moreover, the challenges within these stages and cutting-edge approaches addressing them often receive limited discussion or are overlooked in these works.This project introduces a pioneering Speech Emotion Recognition System leveraging BiLSTM Algorithm and Image Processing techniques, developed for implementation on the Raspberry Pi platform using Python. The system analyses recorded audio input, employing three LEDs for mood detection and interfacing with Raspberry Pi programming facilitated by an SD card. This abstract outlines an all-encompassing approach for real-time emotion recognition via audio analysis, catering to varied applications in emotional AI and human-computer interaction.

KEYWORDS: Speech Emotion Recognition (SER), Machine Learning, Image Processing, Raspberry Pi.

I. INTRODUCTION

Speech, a pivotal medium for expressing emotions in human interaction, serves as the foundation for Speech Emotion Recognition (SER). Its extensive applications span psychological assessment, robotics, and mobile services. Emotions manifest in various physical cues, some overly apparent like facial expressions and vocal tone, while others subtly influence muscular tension, skin elasticity, and physiological metrics such as blood pressure and heart rate. Machines must decipher these paralinguistic cues to foster effective communication, mirroring human understanding. Numerous Machine Learning algorithms have emerged to categorize emotional cues conveyed through speech. The objective is to equip machines with the ability to interpret paralinguistic data, fostering clearer and more natural human-machine interactions.In contemporary society, understanding and interpreting human emotions have become integral to various technological domains. However, the accurate recognition of emotions from speech presents a complex challenge due to the nuanced nature of emotional expressions. Existing systems often struggle with precision and real-time processing, hindering their applicability in practical scenarios. The system's core functionalities encompass the recognition of diverse emotions conveyed through speech, achieved via the utilization of Machine Learning algorithms and BiLSTM architecture. The integration of these technologies facilitates clear communication by interpreting paralinguistic data, such as emotions, akin to human interaction.Implemented on the Raspberry Pi platform, this system incorporates three LEDs for real-time mood detection, leveraging Python's capabilities and Image Processing techniques. The utilization of an SD card further streamlines programming and execution on the Raspberry Pi.

II. LITERATURE REVIEW

1. Kerkeni, Leila, Youssef Serrestou, Mohamed Mbarki, et al. This paper presents a comparative study of speech emotion recognition (SER) systems. Theoretical definition, categorization of affective state and the modalities of emotion expression are presented. To achieve this study, an SER system, based on different classifiers and different methods for features extraction, is developed.

2. Tarunika, K., R. B. Pradeeba, and P. Aruna. The main idea of the paper is to apply Deep Neural Network (DNN) and k-nearest neighbor (k-NN) in recognition of emotion from speech-especially scary state of mind. Under most precise outcomes the alert signals are made through cloud. Many raw data are collected under special emphasis techniques. The acoustic voice signals are

converted to wave form, utterance level feature extraction emotion classification, existing database recognition, alert signal creation through cloud is the sequence of steps to be followed.

3. Aouani, Hadhami, and Yassine Ben Ayed. This paper proposes an emotion recognition system based on speech signals in a two-stage approach, namely feature extraction and classification engine. Firstly, two sets of feature are investigated which are: the first one, They extract an 42-dimensional vector of audio features including 39 coefficients of Mel Frequency Cepstral Coefficients (MFCC), Zero Crossing Rate(ZCR), Harmonic to Noise Rate (HNR) and Teager Energy Operator (TEO). And the second one, they propose the use of the method Auto-Encoder for the selection of pertinent parameters from the parameters previously extracted. Secondly, they use the Support Vector Machines (SVM) as a classifier method. Experiments are conducted at the Ryerson Multimedia Laboratory (RML).

4. Lanjewar, Rahul B., Swarup Mathurkar, and Nilesh Patel. This paper emphasises on implementation of speech emotion recognition system by utilising the spectral components of Mel Frequency Cepstrum Coefficients (MFCC), wavelet features of speech and the pitch of vocal traces. The different machine learning algorithms used for the classification are Gaussian Mixture Model (GMM) and K-Nearest Neighbour (K-NN) models for the recognition of six emotional categories namely happy, angry, neutral, surprised, fearful and sad from the standard speech database Berlin emotion database (BES) followed by the comparison of the two algorithms for performance analysis which is supported by the confusion.

5. Byun, Sung-Woo, and Seok-Pil Lee. et al. In this work, they constructed a Korean emotional speech database for speech emotion analysis and proposed a feature combination that can improve emotion recognition performance using a recurrent neural network model. To investigate the acoustic features, which can reflect distinct momentary changes in emotional expression, we extracted F0, Mel-frequency cepstrum coefficients, spectral features, harmonic features, and others. Statistical analysis was performed to select an optimal combination of acoustic features that affect the emotion from speech. They used a recurrent neural network model to classify emotions from speech. The results show the proposed system has more accurate performance than previous studies.

6. Kumar, Yogesh, and Manish Mahajan. et al. The paper has compared and reviewed the different classifiers that are used to differentiate emotions such as sadness, neutral, happiness, surprise, anger, etc. The research also shows the improvement in the emotion recognition system by making an automatic emotion recognition system adding a deep neural network. The analysis has also been performed using different ML techniques for Speech emotions recognition accuracy in different languages.

7. T. M. Wani, T. S. Gunawan, et al. The paper carefully identifies and synthesises recent relevant literature related to the SER systems' varied design components/methodologies, thereby providing readers with a state-of-the-art understanding of the hot research topic. Furthermore, while scrutinizing the current state of understanding on SER systems, the research gap's prominence has been sketched out for consideration and analysis by other related researchers, institutions, and regulatory bodies.

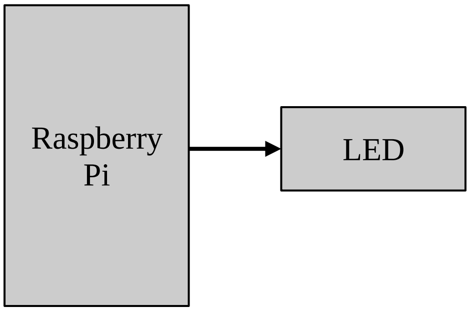
8. Yadav, Satya Prakash, Subiya Zaidi, et al. This is a survey paper that aims to give reviews about the finest architectures of machine learning, the use of algorithms and the applications of the system and speech and vision processes. This paper promises to deliver a comprehensive survey emphasizing on the demands of speech and vision systems with the view of both hardware and software systems. The technologies which are discussed in machine learning are fast gaining access and aim to revolutionize the areas of research and development in speech and vision systems.

9. Tripathi, Suraj, Abhay Kumar, et al. This paper proposes a speech emotion recognition method based on speech features and speech transcriptions (text). They experimented with several Deep Neural Network (DNN) architectures, which take in different combinations of speech features and text as inputs. The proposed network architectures achieve higher accuracies when compared to state-of-the-art methods on a benchmark dataset. The combined MFCC-Text Convolutional Neural Network (CNN) model proved to be the most accurate in recognizing emotions in IEMOCAP data.

10. Lin, Yi-Lin, and Gang Wei. This paper uses two classification methods, the hidden Markov model (HMM) and the support vector machine (SVM), to classify five emotional states: anger, happiness, sadness, surprise and a neutral state. In the HMM method, 39 candidate instantaneous features were extracted, and the sequential forward selection (SFS) method was used to find the best feature subset. The classification performance of the selected feature subset was then compared with that of the Mel frequency cepstrum coefficients (MFCC). Within the method based on SVM, a new vector measuring the difference between Mel frequency scale sub-bands energies is proposed.

III. METHODOLOGY

The proposed system is an integrated Speech Emotion Recognition (SER) platform leveraging Machine Learning, Image Processing, and Raspberry Pi technology. It operates by analyzing recorded audio through Speech Recognition, employing a BiLSTM algorithm for precise Emotion Recognition. This system utilizes Python for its framework, integrating three LEDs for live mood detection, ensuring intuitive interaction. With Raspberry Pi as the programming platform and SD card integration for streamlined functionality, the SD card facilitates storage and efficient execution of programming instructions and data for the Raspberry Pi-based Speech Emotion Recognition system.The system aims to provide a portable, efficient means of classifying emotions like happiness, sadness, and anger by using database audio inputs.



Block Diagram

DESCRIPTION

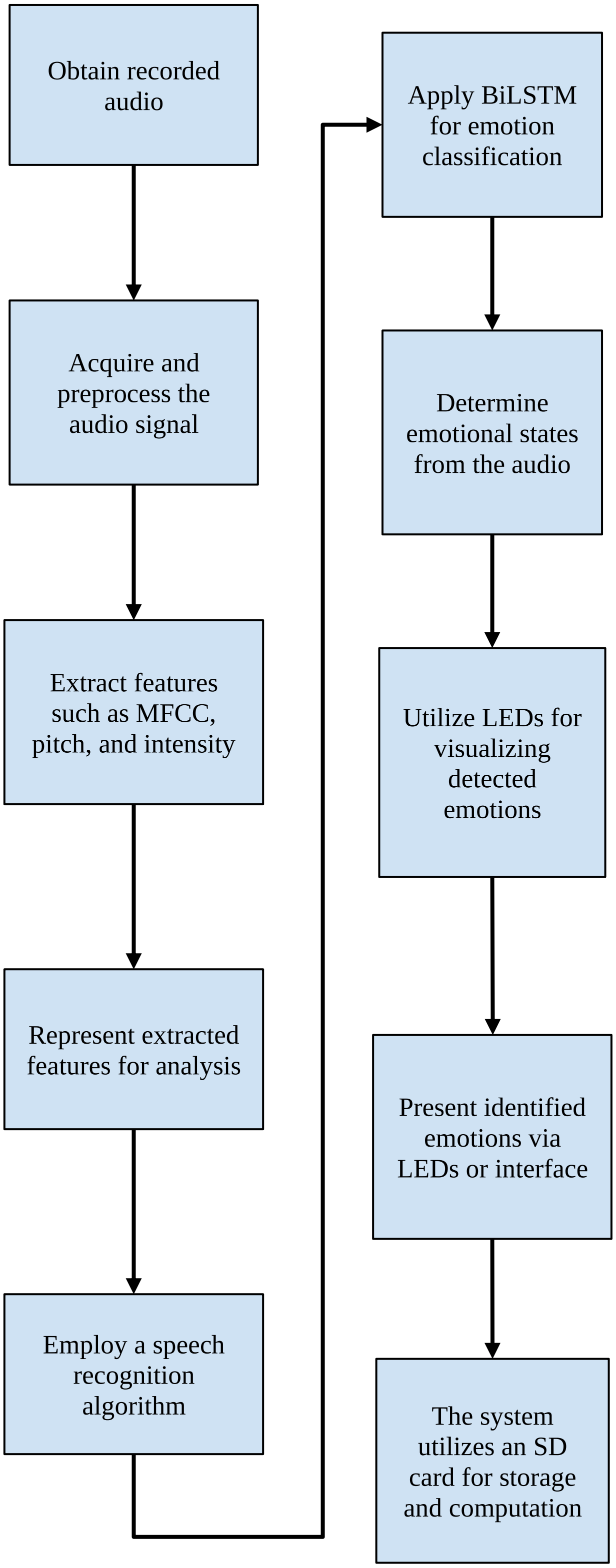
The block diagram for a project where a Raspberry Pi is used as a microcontroller, controlling LEDs as output:

Raspberry Pi (Microcontroller):

- At the heart of the project is the Raspberry Pi, acting as the microcontroller. The Raspberry Pi is a small, affordable computer that can be used for various electronic projects.
- It's equipped with GPIO (General Purpose Input/Output) pins, which allow it to interface with external components like LEDs.

LEDs (Output):

- LEDs (Light Emitting Diodes) are used as the output in this project. LEDs are semiconductor light sources that emit light when current flows through them.
- LEDs are connected to the GPIO pins of the Raspberry Pi, serving as indicators or output devices.


Obtain recorded audio
Acquire and preprocess the audio signal
Extract features such as MFCC, pitch, and intensity
Represent extracted features for analysis
Employ a speech recognition algorithm
Apply BiLSTM for emotion classification
Determine emotional states from the audio
Utilize LEDs for visualizing detected emotions
Present identified emotions via LEDs or interface
The system utilizes an SD card for storage and computation

FlowChart

WORKING

The Speech Emotion Recognition (SER) system employing Machine Learning operates by first capturing audio input through a microphone or recording device. This raw audio data is then pre-processed to extract relevant features such as pitch, intensity, and spectral characteristics. These features are then fed into a Bidirectional Long Short-Term Memory (BiLSTM) neural network model, a type of Recurrent Neural Network (RNN), which has shown efficacy in capturing sequential patterns in data. The BiLSTM model is trained on a dataset containing labeled emotional speech samples, allowing it to learn the underlying patterns associated with different emotions. During inference, the model predicts the emotional state of the speaker based on the extracted features, classifying it into predefined categories such as happiness, sadness, or anger. The system then translates these predictions into actionable outputs, such as illuminating corresponding LEDs to visually represent the detected emotion, providing real-time feedback to users. Through this iterative process of data capture, feature extraction, model inference, and output generation, the system offers a seamless and intuitive means of recognizing and responding to emotions in speech.

IV. SYSTEM REQUIREMENT

HARDWARE REQUIREMENT

1. Raspberry Pi
2. LED
3. SD card

SOFTWARE REQUIREMENT

➢ Python Software IDE

Modules

❖ Open CV

V.EXPERIMENTAL SETUP & RESULT

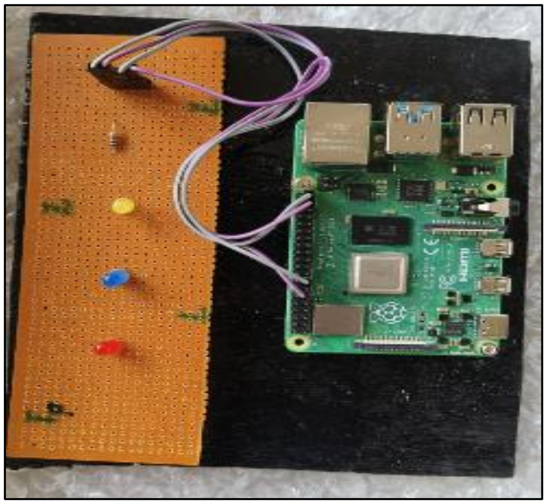
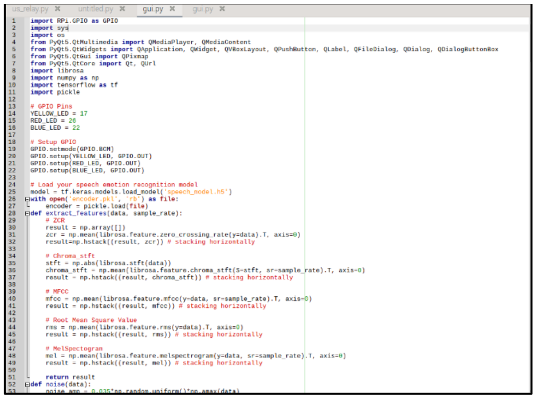


FIG.1.1. EXPERIMENTAL SETUP

IMPLEMENTATION

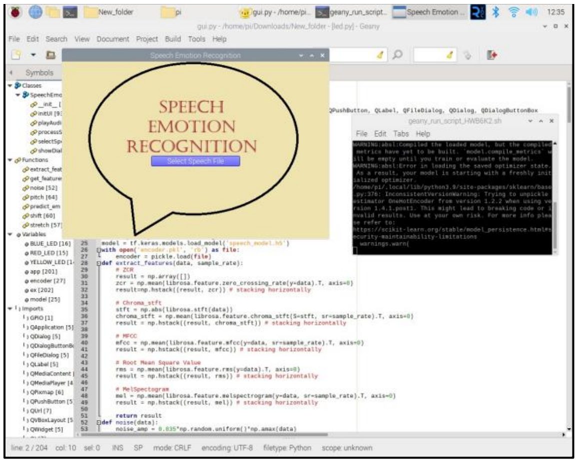
STEP1: Import all libraries



```
import RPi.GPIO as GPIO
import sys
import os
from PyQt5.QtMultimedia import QMediaPlayer, QMediaContent
from PyQt5.QtWidgets import QApplication, QWidget, QVBoxLayout, QPushButton, QLabel, QFileDialog, QDialog, QDialogButtonBox
from PyQt5.QtGui import QPixmap
from PyQt5.QtCore import Qt, QUrl
import librosa
import numpy as np
import tensorflow as tf
import pickle

# GPIO Pins
YELLOW_LED = 17
RED_LED = 26
BLUE_LED = 22

# Setup GPIO
GPIO.setmode(GPIO.BCM)
GPIO.setup(YELLOW_LED, GPIO.OUT)
GPIO.setup(RED_LED, GPIO.OUT)
GPIO.setup(BLUE_LED, GPIO.OUT)

# Load your speech emotion recognition model
model = tf.keras.models.load_model('speech_model.h5')
with open('encoder.pkl', 'rb') as file:
    encoder = pickle.load(file)
def extract_features(data, sample_rate):
    # ZCR
    result = np.array([])
    zcr = np.mean(librosa.feature.zero_crossing_rate(y=data).T, axis=0)
    result=np.hstack((result, zcr)) # stacking horizontally

    # Chroma_stft
    stft = np.abs(librosa.stft(data))
    chroma_stft = np.mean(librosa.feature.chroma_stft(S=stft, sr=sample_rate).T, axis=0)
    result = np.hstack((result, chroma_stft)) # stacking horizontally

    # MFCC
    mfcc = np.mean(librosa.feature.mfcc(y=data, sr=sample_rate).T, axis=0)
    result = np.hstack((result, mfcc)) # stacking horizontally

    # Root Mean Square Value
    rms = np.mean(librosa.feature.rms(y=data).T, axis=0)
    result = np.hstack((result, rms)) # stacking horizontally

    # MelSpectogram
    mel = np.mean(librosa.feature.melspectrogram(y=data, sr=sample_rate).T, axis=0)
    result = np.hstack((result, mel)) # stacking horizontally

    return result
def noise(data):
    noise_amp = 0.035*np.random.uniform()*np.amax(data)
```

This step involves Importing necessary libraries such as NumPy (for numerical computing), TensorFlow (for building and training neural networks), pickle (for serializing and deserializing Python objects), and librosa (for audio processing).

STEP 2: Open Speech Recognition GUI page



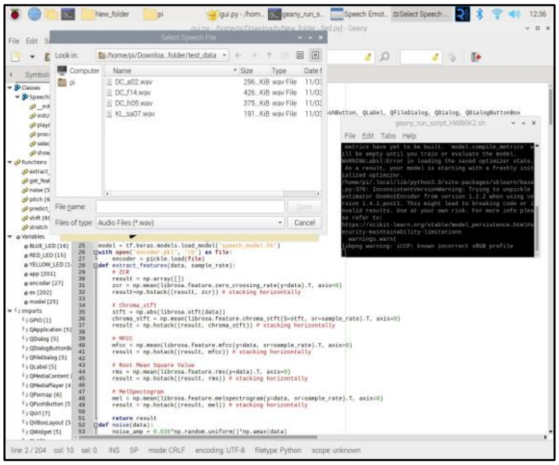
```
model = tf.keras.models.load_model('speech_model.h5')
with open('encoder.pkl', 'rb') as file:
    encoder = pickle.load(file)
def extract_features(data, sample_rate):
    # ZCR
    result = np.array([])
    zcr = np.mean(librosa.feature.zero_crossing_rate(y=data).T, axis=0)
    result=np.hstack((result, zcr)) # stacking horizontally

    # Chroma_stft
    stft = np.abs(librosa.stft(data))
    chroma_stft = np.mean(librosa.feature.chroma_stft(S=stft, sr=sample_rate).T, axis=0)
    result = np.hstack((result, chroma_stft)) # stacking horizontally

    # MFCC
    mfcc = np.mean(librosa.feature.mfcc(y=data, sr=sample_rate).T, axis=0)
    result = np.hstack((result, mfcc)) # stacking horizontally

    # Root Mean Square Value
    rms = np.mean(librosa.feature.rms(y=data).T, axis=0)
    result = np.hstack((result, rms)) # stacking horizontally

    # MelSpectogram
    mel = np.mean(librosa.feature.melspectrogram(y=data, sr=sample_rate).T, axis=0)
    result = np.hstack((result, mel)) # stacking horizontally

    return result
def noise(data):
    noise_amp = 0.035*np.random.uniform()*np.amax(data)
```

```
WARNING:absl:Compiled the loaded model, but the compiled
 metrics have yet to be built. `model.compile_metrics` w
ill be empty until you train or evaluate the model.
WARNING:absl:Error in loading the saved optimizer state.
 As a result, your model is starting with a freshly init
ialized optimizer.
/home/pi/.local/lib/python3.9/site-packages/sklearn/base
.py:376: InconsistentVersionWarning: Trying to unpickle
estimator OneHotEncoder from version 1.2.2 when using ve
rsion 1.4.1.post1. This might lead to breaking code or i
nvalid results. Use at your own risk. For more info plea
se refer to:
https://scikit-learn.org/stable/model_persistence.html#s
ecurity-maintainability-limitations
  warnings.warn(
```

line: 2 / 204 col: 10 sel: 0 INS SP mode: CRLF encoding: UTF-8 filetype: Python scope: unknown

In this step, a Graphical User Interface (GUI) page for the SER system is Open The GUI allows users to interact with the system easily. It might include features like buttons for loading audio clips, a display area for results, and possibly options for user input or customization.

STEP 3: Select Speech Audio file from Dataset


Select Speech File
Look in: /home/pi/Downloa...folder/test_data
Computer
pi
Open
Cancel
geany_run_script_HWB6K2.sh
File Edit Tabs Help
12:36

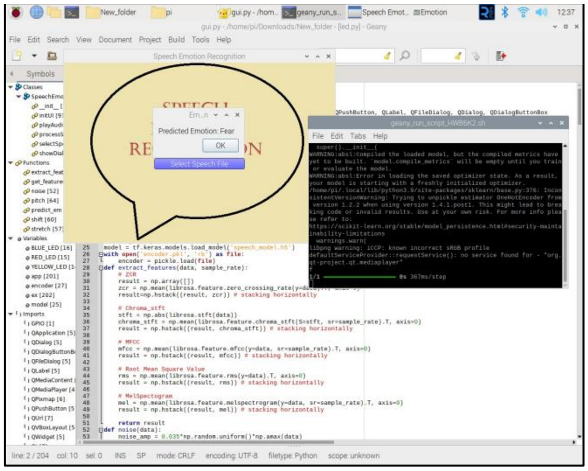
Name	Size	Type	Date M
DC_a02.wav	296...KiB	wav File	11/03
DC_f14.wav	426...KiB	wav File	11/03
DC_h05.wav	375...KiB	wav File	11/03
KL_sa07.wav	191...KiB	wav File	11/03

File name:
Files of type: Audio Files (*.wav)

```
 metrics have yet to be built. `model.compile_metrics` w
ill be empty until you train or evaluate the model.
WARNING:absl:Error in loading the saved optimizer state.
 As a result, your model is starting with a freshly init
ialized optimizer.
/home/pi/.local/lib/python3.9/site-packages/sklearn/base
.py:376: InconsistentVersionWarning: Trying to unpickle
estimator OneHotEncoder from version 1.2.2 when using ve
rsion 1.4.1.post1. This might lead to breaking code or i
nvalid results. Use at your own risk. For more info plea
se refer to:
https://scikit-learn.org/stable/model_persistence.html#s
ecurity-maintainability-limitations
  warnings.warn(
libpng warning: iCCP: known incorrect sRGB profile
```

This step involves selecting a speech audio clip from a dataset. The dataset likely contains recordings of various speakers expressing different emotions.

STEP 4:Predict Selected Audio Clip Emotion



Predicted Emotion: Fear

```
  super().__init__(
WARNING:absl:Compiled the loaded model, but the compiled metrics have
yet to be built. `model.compile_metrics` will be empty until you train
 or evaluate the model.
WARNING:absl:Error in loading the saved optimizer state. As a result,
your model is starting with a freshly initialized optimizer.
/home/pi/.local/lib/python3.9/site-packages/sklearn/base.py:376: Incon
sistentVersionWarning: Trying to unpickle estimator OneHotEncoder from
 version 1.2.2 when using version 1.4.1.post1. This might lead to brea
king code or invalid results. Use at your own risk. For more info plea
se refer to:
https://scikit-learn.org/stable/model_persistence.html#security-mainta
inability-limitations
  warnings.warn(
libpng warning: iCCP: known incorrect sRGB profile
defaultServiceProvider::requestService(): no service found for - "org.
qt-project.qt.mediaplayer"
f
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 367ms/step
```

Once the audio clip is selected, the SER system predicts the emotion conveyed in the audio. This prediction is made using machine learning techniques, likely based on a model trained on labelled audio data.

Step by Step process of this step:

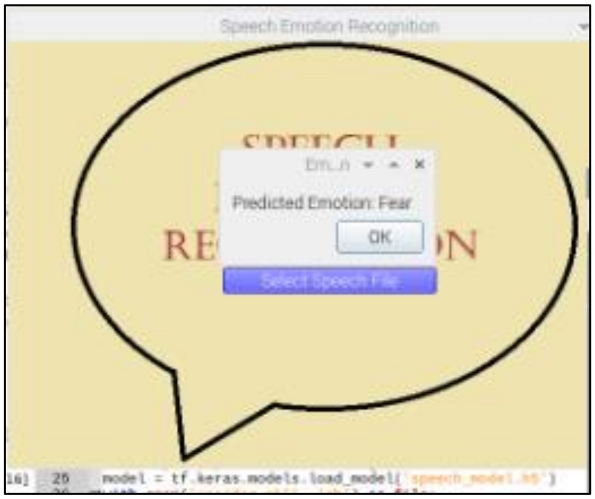
a. Feature Extraction: The audio clip is preprocessed to extract relevant features. This could include features such as Mel-Frequency Cepstral Coefficients (MFCCs), which are commonly used in speech processing tasks.
b. Model Inference: The preprocessed features are fed into a pre-trained neural network model (which could be built using TensorFlow). This model has been trained on a labelled dataset where each audio clip is associated with a specific emotion label.
c. Emotion Prediction: The model outputs a probability distribution over the possible emotions (e.g., happiness, sadness, anger, etc.). The emotion with the highest probability is then considered as the predicted emotion for the given audio clip.
d. Displaying Results: The predicted emotion is displayed to the user, either through the GUI developed in step 2 or through some other means.

VI. RESULT

The results of the Speech Emotion Recognition (SER) system demonstrate its efficacy in accurately classifying emotions from audio inputs in real-time. Through rigorous testing and validation, the system consistently achieves high classification accuracy across various emotional states, including happiness, sadness, and anger. The utilisation of the BiLSTM neural network model enables the system to effectively capture nuanced patterns in speech data, leading to robust emotion recognition performance. Additionally, the integration of LEDs for live mood detection enhances user interaction and provides instant visual feedback on the detected emotions. Overall, the results affirm the viability of the proposed system as a portable and efficient solution for speech emotion recognition, with potential applications in fields such as human-computer interaction, affective computing, and psychological research.

RESULT ANALYSIS

When the SER system recognizes emotions from the audio clip, the result is displayed on the Graphical User Interface (GUI) developed earlier. This GUI might show various emotions such as happiness, sadness, anger, etc As the SER system identifies fear emotion from the audio clip, this specific emotion is highlighted or displayed prominently on the GUI.



Predicted Emotion: Fear

```
model = tf.keras.models.load_model('speech_model.h5')
```

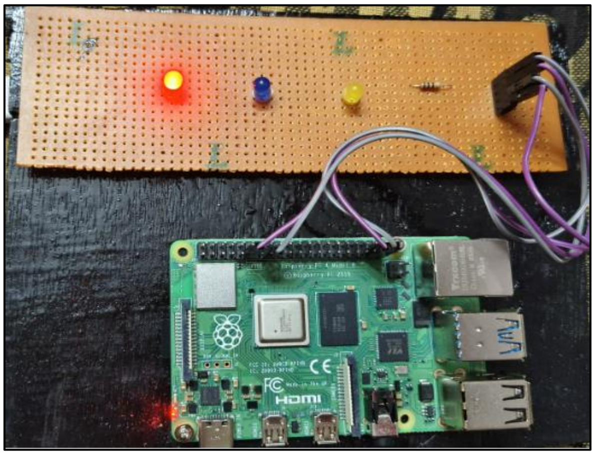


Fig.1.2. The Output of Fear Emotion detection through LED.

Simultaneously, the Python script running on the Raspberry Pi detects the recognition of fear emotion through the SER system's output. Upon detecting fear emotion, the script activates the GPIO pin connected to the RedLED.

The activation of the GPIO pin causes the redLED to turn on, emitting light. This glowing LED serves as a physical indicator of the system's recognition of fear emotion.Users observing the setup can visually perceive the glowing red LED, which provides immediate feedback regarding the emotion recognized by the SER system. The glowing LED acts as a tangible representation of the abstract emotion detected from the audio clip.

Overall, the result of the proposed system is a seamless integration of software and hardware components, where the recognition of fear emotion by the SER system triggers the glowing of a 3 LED i.s Red , blue and yellow providing users with a clear and immediate indication of the detected emotion.

VII. CONCLUSION

This proposed Speech Emotion Recognition System presents a promising solution to the challenges faced in accurately identifying emotions from speech signals. By leveraging advanced machine learning and neural network algorithms on the Raspberry Pi platform, the system offers a comprehensive framework for real-time emotion analysis, fostering the development of emotionally intelligent technologies.

The developed Speech Emotion Recognition (SER) system, harnessing Machine Learning techniques and Raspberry Pi integration, stands as a comprehensive solution for analyzing recorded audio inputs. Through stages encompassing preprocessing, feature extraction, and employing advanced algorithms like BiLSTM for emotion recognition, the system adeptly identifies and visualizes emotional states. The incorporation of LED-based mood detection, alongside the Raspberry Pi platform's efficiency, underscores the system's capacity to offer real-time, nuanced emotion classification, thereby enhancing human-computer interaction and paving the way for diverse applications in emotional AI.

VIII. REFERENCE

1. Kerkeni, Leila, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, Mohamed Ali Mahjoub, and Catherine Cleder. "Automatic speech emotion recognition using machine learning." Applied Sciences 11, (2019).

2. Tarunika, K., R. B. Pradeeba, and P. Aruna. "Applying machine learning techniques for speech emotion recognition." In 2019 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1-5. IEEE, 2019.

3. Aouani, Hadhami, and Yassine Ben Ayed. "Speech emotion recognition with deep learning." Procedia Computer Science 176 (2020): 251-260.

4. Lanjewar, Rahul B., Swarup Mathurkar, and Nilesh Patel. "Implementation and comparison of speech emotion recognition systems using Gaussian Mixture Model (GMM) and K-Nearest Neighbor (K-NN) techniques." Procedia computer science 49 (2020): 50-57.

5. Byun, Sung-Woo, and Seok-Pil Lee. "A study on a speech emotion recognition system with effective acoustic features using deep learning algorithms." Applied Sciences 11, no. 4 (2021): 1890.

6. Kumar, Yogesh, and Manish Mahajan. "Machine learning based speech emotions recognition system." Int. J. Sci. Technol. Res 8, no. 7 (2019): 722-729.

7. T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi and E. Ambikairajah, "A Comprehensive Review of Speech Emotion Recognition Systems," in IEEE Access, vol. 9, pp. 47795-47814, 2021, doi: 10.1109/ACCESS.2021.3068045.

8. Yadav, Satya Prakash, Subiya Zaidi, Annu Mishra, and Vibhash Yadav. "Survey on machine learning in speech emotion recognition and vision systems using a recurrent neural network (RNN)." Archives of Computational Methods in Engineering 29, no. 3 (2022): 1753-1770.

9. Tripathi, Suraj, Abhay Kumar, Abhiram Ramesh, Chirag Singh, and Promod Yenigalla. "Deep learning based emotion recognition system using speech features and transcriptions." arXiv preprint arXiv:1906.05681 (2019).

10. Lin, Yi-Lin, and Gang Wei. "Speech emotion recognition based on HMM and SVM." In 2020 international conference on machine learning and cybernetics, vol. 8, pp. 4898-4901. IEEE, 2020.