KIET
GROUP OF INSTITUTIONS
Connecting Life with Learning

A

Project Report

on

Speech Emotion Recognition using Deep Learning

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2023-24

in

Computer Science and Engineering

By

Divya Garg (2000290100058)

Amrita Singh (2000290100019)

Anshika Gupta (2000290100025)

Under the supervision of

Dr. Sanjiv Sharma

KIET Group of Institutions, Ghaziabad

Affiliated to

Dr. A.P.J. Abdul Kalam Technical University, Lucknow

(Formerly UPTU)

May, 2024

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Signature

Name: Divya Garg

Roll No.: 2000290100058

Signature

Name: Amrita Singh

Roll No.: 2000290100019

Signature

Name: Anshika Gupta

Roll No.: 2000290100025

Date:

CERTIFICATE

This is to certify that Project Report entitled "**Speech emotion Recognition using Deep Learning**" which is submitted by "**Divya Garg, Amrita Singh, Anshika Gupta**" in partial fulfillment of the requirement for the award of degree B. Tech. in Department of Computer Science & Engineering of Dr. A.P.J. Abdul Kalam Technical University, Lucknow is a record of the candidates own work carried out by them under my supervision. The matter embodied in this report is original and has not been submitted for the award of any other degree.

.

Dr. Sanjiv Sharma **Dr. Vineet Sharma**

(Associate Professor) **(HoD-CSE)**

Date:

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the B. Tech Project undertaken during B. Tech. Final Year. We owe special debt of gratitude to Dr. Sanjiv Sharma, Department of Computer Science & Engineering, KIET, Ghaziabad, for his constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only his cognizant efforts that our endeavors have seen light of the day.

We also take the opportunity to acknowledge the contribution of Dr. Vineet Sharma, Head of the Department of Computer Science & Engineering, KIET, Ghaziabad, for his full support and assistance during the development of the project. We also do not like to miss the opportunity to acknowledge the contribution of all the faculty members of the department for their kind assistance and cooperation during the development of our project.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members, especially faculty/industry person/any person, of the department for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

Date:

Signature:
Name: Divya Garg
Roll No.: 2000290100058

Signature:
Name: Amrita Singh
Roll No.: 2000290100019

Signature:
Name: Anshika Gupta
Roll No.: 2000290100025

ABSTRACT

Speech Emotion Recognition (SER) is an emerging field that involves recognizing emotions conveyed in speech. Emotions expressed through speech can greatly impact decision-making. This paper delves into the topic of speech emotion recognition (SER) and its focus on interpreting emotions conveyed through spoken language. The importance of SER lies in its potential to improve human-computer interaction, cognitive analysis, and psychiatric assessment. The study combines and preprocesses audio data from various datasets, such as RAVDESS, CREMA-D, TESS, and SAVEE, and uses log mel spectrograms to effectively extract features. Various methods including CNN models, and standard and optimized feature extraction techniques are used. The results suggest that SER has significant real-world applications and the approaches provided effectively identify emotional and voice signals.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

SER	Speech Emotion Recognition
CNN	Convolutional Neural Network
KNN	K-Nearest Neighbor
RNN	Recurrent Neural Network
SVM	Support Vector Machine
LSTM	Long Short-Term Memory
RAVDESS	Ryerson Audio-Visual Database of Emotional Speech and Song
CREMA-D	Crowd-sourced Emotional Multimodal Actors Dataset
MFCC	Mel Frequency Cepstrum Coefficient
SAVEE	Surrey Audio-Visual Expressed Emotion
TESS	Toronto Emotional Speech Set

CHAPTER 1
INTRODUCTION

1.1 Introduction

Speech signals represent one of the most natural and efficient modes of communication between individuals, both in human-human interactions and human-computer interfaces [1]. The intricate interplay of various senses enables individuals to extract rich contextual information from spoken messages. However, for machines, discerning and interpreting emotions conveyed through speech poses a formidable challenge [3]. Nonetheless, bridging this gap in understanding emotions between humans and machines holds immense potential for facilitating more effective communication systems [3].

Language-based speech emotion recognition serves as a cornerstone in deciphering the emotional content embedded within spoken language, allowing for discrimination between emotions expressed by male and female speakers [4]. Key speech traits such as Linear Predictive Coding Coefficients (LPCC), Mel-Frequency Cepstral Coefficients (MFCC), and Fundamental Frequencies have been extensively studied by researchers in the domain of speech perception [4]. Despite these efforts, the inherent complexity in differentiating between various emotional states remains a persistent challenge, influenced by factors such as speaking rates, styles, sentence structures, and individual differences in emotional responses to linguistic stimuli [5].

Moreover, cultural and environmental factors further complicate the manifestation of emotions in speech, as individuals from different backgrounds may exhibit distinct speaking styles influenced by their respective cultural norms and environmental experiences [5]. Emotions themselves can be categorized into short-term and long-term states, adding another layer of complexity to the recognition process [6]. The challenge is exacerbated by the ambiguity inherent in recognizing and categorizing emotions, with the recognizer often failing to specify the exact nature of the detected emotional state [6].

In the realm of speech emotion recognition, the distinction between speaker-independent and speaker-dependent recognition further underscores the multifaceted nature of the task [6]. Various classification methods, including K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Convolutional Neural Networks (CNN), have been explored in the literature to address this challenge [7].

1.2 Project Category

The project falls under the category of Speech Emotion Recognition (SER) research, which focuses on developing methodologies and systems for accurately detecting and interpreting emotions conveyed through speech signals. This interdisciplinary field combines elements of signal processing, machine learning, and psychology to advance our understanding of emotional expression in spoken language.

1.3 Objective

The primary objective of this project is to enhance human-computer interaction, cognitive analysis, and psychiatric assessments through the development of robust Speech Emotion Recognition (SER) systems. By leveraging advanced techniques in feature extraction, classification algorithms, and dataset preprocessing, the project aims to achieve high accuracy in detecting and categorizing emotions expressed in speech.

1.4 Problem Formulation

The central problem addressed by this project is the inherent difficulty in accurately recognizing and interpreting emotions conveyed through speech signals. Factors such as variations in speaking styles, cultural influences, and individual differences in emotional expression pose significant challenges to traditional SER systems. The project seeks to overcome these challenges by developing innovative methodologies and systems capable of effectively capturing and analyzing the emotional content of speech.

1.5 Proposed System

The proposed system encompasses a comprehensive approach to Speech Emotion Recognition (SER), integrating state-of-the-art techniques in feature extraction, classification, and dataset preprocessing. The system will utilize diverse datasets such as RAVDESS, CREMA-D, TESS, and SAVEE to ensure robustness and generalization across different emotional expressions and speaker demographics.

1.6 Unique Features of the System

Utilization of Log Mel Spectrograms: The system employs log mel spectrograms as a primary feature extraction technique, capturing the frequency components of speech signals in a manner that is well-suited for emotion recognition tasks.

Integration of Convolutional Neural Network (CNN) Models: CNN architectures are utilized for their ability to automatically learn discriminative features from spectrogram representations, enhancing the system's ability to extract relevant emotional cues from speech data.

Optimized Feature Extraction Techniques: The system incorporates optimized feature extraction techniques, including LPCC, MFCC, and Fundamental Frequencies, to capture nuanced aspects of speech signals that are indicative of emotional expression.

Multimodal Dataset Integration: By integrating diverse datasets encompassing various emotional expressions and speaker demographics, the system ensures robustness and generalization across different contexts and populations.

Real-World Applications: The system's efficacy is demonstrated through experiments showcasing its applicability in real-world scenarios, including human-computer interaction, cognitive analysis, and psychiatric assessments.

Overall, the proposed system represents a cutting-edge approach to Speech Emotion Recognition, offering novel methodologies and techniques to address the inherent challenges in interpreting emotions conveyed through speech signals.

CHAPTER 2

LITERATURE REVIEW

Several studies employ diverse deep learning architectures for speech emotion recognition. These studies showcase advancements in modeling techniques and data representations, yielding significant improvements in emotion recognition accuracy and robustness.

Aharon et al. [1] trained Deep neural network with para-lingual information for speech emotion detection and achieved 68% recognition accuracy on IEMOCAP using convolutional LSTM algorithm. Jonathan et al. [2] combined multitask machine learning and deep convolutional generative adversarial networks and expanded speech emotions training corpus to 100 hours, achieving a 43.88% improvement. Chen et al. [3] utilized 3-D attention-dependent convolutional recurrent neural networks (ACRNNs) for emotion recognition and demonstrated highest unweighted average recall on Emo-DB and IEMOCAP corpora. Zhao et al. [4] built composite CNNs with 1D and 2D branches for speech emotion recognition and utilized Bayesian optimization for hyperparameter selection and transfer learning for accelerated training.

Yenigalla et al. [5] assessed speech mood using phonemes and spectrograms, demonstrating improved sentiment capture compared to text. Hybrid CNN model incorporating phoneme and spectrogram properties exhibited superior accuracy on IEMOCAP. Sarma et al. [6] tested various DNN topologies for emotion recognition using high-dimensional MFCC input. Temporal aggregation and frame-level speech labels resulted in superior performance, achieving 70.6% weighted precision. Latif et al. [7] improved SER accuracy via cross-language and cross-corpus transition learning, outperforming traditional methods and employed DBNs for cross-corpus emotion identification across five corpora in three languages.

Zhao et al. [8] introduced CNN+LSTM 1D and 2D networks for learning local and global emotions from voice and log-Mel spectrograms and utilized sparse autoencoder and attention-focused technique for improved emotion prediction. Sun et al. [9] employed autoencoder and attention mechanism to predict speech emotion better than current algorithms and demonstrated effectiveness across three cross-language internet databases. Jiang et al. [10] utilized fusion networks to learn discriminative acoustic feature representations for improved recognition

efficiency and improved recognition efficiency by 64% over current methods on IEMOCAP studies. Pandey et al. [11] studied deep learning algorithms for speech-based emotion extraction and categorization and identified CNN, BLSTM, and SVM model as successful in categorizing speech emotions using log-Mel spectrogram data.

Zhen et al. [12] proposed CNN, BLSTM, and SVM model for speech emotion recognition, outperforming prior techniques on IEMOCAP and demonstrated the model's effectiveness but emphasized the need for more data validation. [13] demonstrated efficacy of CNN+LSTM models on six speech datasets, with superior performance in five out of six cases. Dai et al. [14] proposed a computational approach for recognition of emotion and analysis the specifications of emotion in voiced social media such as Wechat. This approach approximates the mixed emotion and dynamic fluctuations in position- arousal-dominance (PAD) by extracting 25 acoustic features of speech signals and employing trained least squares-support vector regression (LV-SVR) model as well. The experimental results demonstrate the recognition rate for different emotion are different and the average rate of recognition achieves 82.43%, which is the best existing result by similar examination [15].

CHAPTER 3

PROPOSED METHODOLOGY

3.1 Working Model

The main purpose of SER is to increase human-machine interaction. To create a model for speech emotion recognition, many datasets are used. To conduct the research on speech emotion recognition effectively, a systematic approach is adopted, focusing mainly on Kaggle as the platform of choice due to its collaborative and supportive environment for data analysis and machine learning tasks.

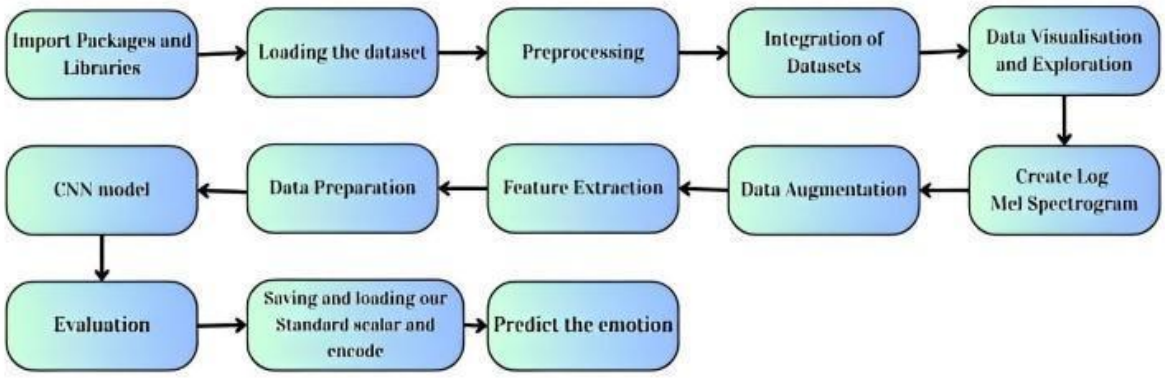

Figure 3.1. Block Diagram

3.2 Import Packages and Libraries

In this section, we discuss the necessary steps to import essential packages and libraries required for executing code and performing data analysis tasks. Proper importation of packages ensures access to a wide range of functionalities, facilitating data manipulation, visualization, and machine learning tasks within the chosen programming environment. We begin by outlining the required software components and proceed to demonstrate the import statements for commonly used packages in the context of data science.

3.2.1 Software Required

To execute the code and perform data analysis tasks, the following software components are necessary:

Anaconda Prompt: Anaconda command prompt is just like command prompt, Anaconda Prompt is a command line shell (a program where you type in commands instead of using a mouse). The black screen and text that makes up the Anaconda Prompt doesn't look like much, but it is really helpful for problem solvers using Python.

Jupyter Notebook: The Jupyter Notebook is an open-source web application that you can use to create and share documents that contain live code, equations, visualizations, and text. Jupyter Notebook is maintained by the people at Project Jupyter. interactive web tool known as a computational notebook, which researchers can use to combine software code, computational output, explanatory text and multimedia resources in a single document.

3.2.2 Python Packages and Libraries

Librosa- librosa is a python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems. It is the starting point towards working with audio data at scale for a wide range of applications such as detecting voice from a person to finding personal characteristics from an audio. Librosa can be defined as a package which is structured as collection of submodules which further contains other functions.

NumPy- NumPy is a Python library used for working with arrays. NumPy is a Python library that provides a simple yet powerful data structure: the n-dimensional array. This is the foundation on which almost all the power of Python's data science toolkit is built.

Pandas- Pandas is an open-source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named

Numpy, which provides support for multi-dimensional arrays. The Pandas library provides a really fast and efficient way to manage and explore data.

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.

TensorFlow- TensorFlow is a Python-friendly open-source toolkit for mathematical computation that accelerates and simplifies AI.

Keras- An open-source programming framework called keras provides fake brain organisations with a Python point of contact. Keras functions as the TensorFlow library connecting point intended to provide fast experimentation.

Scikit learn(sklearn)- Scikit-learn is probably the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.

OS- The OS module in Python provides functions for interacting with the operating system. This module provides a portable way of using operating system-dependent functionality

Scipy- SciPy is a scientific computation library that uses NumP underneath. SciPy stands for Scientific Python. It provides more utility functions for optimization, stats and signal processing.

Noisereduce- 'noisereduce' is a Python library that provides functionality for reducing noise from audio signals. It offers methods for denoising audio signals using various algorithms, such as spectral subtraction, Wiener filtering, and wavelet denoising.

Glob- 'glob' is a Python module that provides a convenient way to search for files and directories that match specified patterns using wildcard characters.

3.3 Loading the Dataset

In the research process, importing the dataset is a crucial initial step. The dataset is typically sourced from relevant repositories, databases, or curated collections.

3.3.1 RAVDESS Dataset

The RAVDESS, Emotional Audio Database at Ryerson University (RAVDESS). There are 7356 files in the (RAVDESS) (total size: 24.8 GB). 24 professional actors:12 male and 12 female—perform 2 lexically similar phrases in the database with neutral American accents. Both in speech and in the lyrics, there are expressions of calmness, happiness, joy, sadness, anger, fear, surprise, and contempt. There are two emotional intensity levels (normal and strong) and one neutral expression created for each expression. Three modality forms are accessible for all conditions: Audio-only (16bit, 48kHz.wav), Audio-Video (720p H.264, AAC 48kHz,.mp4), and Video-only (480p H.264, AAC 48kHz,.mp4) (no sound). Note that Actor 18 doesn't have any song files.

3.3.2 CREMA DataFrame

One of the four important datasets that I was fortunate to discover is the Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) dataset. The intriguing thing about this dataset is how diverse it is, which aids in the training of a model that can be applied to other datasets. There is a lot of information loss as a result of the restricted number of speakers used in many audio datasets. There are several speakers on CREMA-D. Because of this, using the CREMA-D dataset will help guarantee that the model does not overfit. 91 performers contributed 7,442 original footage to the data collection known as CREMA-D. The performers included in these video, who ranged in age from 20 to 74 and represented a diversity of races and ethnicities, included 48 men and 43 women. A selection of 12 phrases were read by the actors. Six distinct emotions (angry, disgust, fear, happy, neutral, and sad) and four different emotion degrees were used to deliver the phrases (Low, Medium, High, and Unspecified).

3.3.3 TESS Dataset

One of the four important datasets that I was fortunate to discover is the (TESS) dataset. It's intriguing that this dataset solely includes females and yet the audio is of such good calibre. The other sample is primarily composed of male speakers, creating a slightly unbalanced representation. Consequently, in terms of generalization, this dataset would serve as a very good training dataset for the emotion classifier (not overfitting). Two actresses (26 and 64 years old) recited a set of 200 target words in the carrier phrase "Say the word _," and recordings of the set evoking each of the seven emotions were created. There are a total of 2800 audio files. Each of 2 female performers and their emotions are included within an own folder in the dataset, which is organized thus way. Additionally, all 200 target word audio files are contained within that.

3.3.4 SAVEE Dataset

Surrey General media Communicated Feeling (SAVEE) information base has been recorded as a pre-imperative for the improvement of a programmed feeling acknowledgment framework. The data set comprises accounts from 4 male entertainers in 7 distinct feelings, 480 English expressions altogether. The sentences were looked over the standard TIMIT corpus and phonetically adjusted for every inclination. The information was kept in a visual media lab with great general media gear, handled and marked. To check the nature of execution, the accounts were assessed by 10 subjects under sound, visual and general media conditions.

Characterization frameworks were assembled involving standard highlights and classifiers for every one of the sound, visual and general media modalities, and speaker-free acknowledgment paces of 61%, 65% and 84% accomplished separately.

The first source includes four folders, each addressing a speaker. However, I combined all of them into a single organizer, so the first two letters of the filename are the initials of each speaker. For instance, "DC d03.wav" is the speaker DC's third expression of nausea. The fact that they are male speakers is meaningless. This won't be a problem since the TESS dataset, which is almost all female, will balance things out.

3.4 Preprocessing

The very first step after selecting the datasets was to identify and interpret the audio files. Each dataset had its own distinct naming convention. The emotion label was derived from the file names and then utilized for classification. The datasets were then analyzed using wave plots displaying randomly picked audio files. These graphs help to highlight the type of data that will be researched. Additionally, the quality of the speech data may be examined by determining if the audio recordings include any background noise and whether the emotions can be easily interpreted by humans.

This study explored how the deep learning model behaves differently when applied to individual datasets and then to all four merged datasets. Due to the fact that the audio files in each dataset have varying durations, the model may not train effectively.

This preprocessing involves - Integration of audio files from different datasets to maintain uniformity and Creation of log mel spectrograms to represent audio signals concisely and effectively. The preprocessing stage sets the foundation for subsequent analysis, enabling the extraction of meaningful features and insights from the audio data.

3.5 Integration of Dataset

In this step, we integrated datasets to combine data from different sources into a single, unified dataset for analysis. The integration process involved merging a dataset containing counts of instances for different emotions with another dataset containing textual data associated with these emotions. This allowed us to create a comprehensive dataset that includes both the counts of instances and the corresponding textual data.

Table 3.1 presents the distribution of instances across different emotions in the dataset. Each emotion category is accompanied by the corresponding count of instances labeled with that particular emotion. For example, there are 1923 instances labeled as "Disgust," 1923 instances labeled as "Fear," and so forth. This table provides insight into the balance or imbalance of

emotion classes within the dataset, which is essential for understanding the data and designing appropriate machine learning models for emotion recognition.

Table 3.1 Instances and Emotions

Emotions	Instances
disgust	1923
fear	1923
sad	1923
happy	1923
angry	1923
neutral	1895
surprise	652

3.6 Create Mel Spectrogram

Mel spectrograms serve as a crucial preprocessing step in analyzing audio data for tasks like emotion recognition. By converting the raw audio signals into Mel spectrograms, we transform the time-domain audio signals into a frequency-domain representation that is more suitable for machine learning algorithms. This transformation enhances the model's capability to capture relevant features for emotion recognition, as it provides a compact yet informative representation of the frequency content of the audio signals.

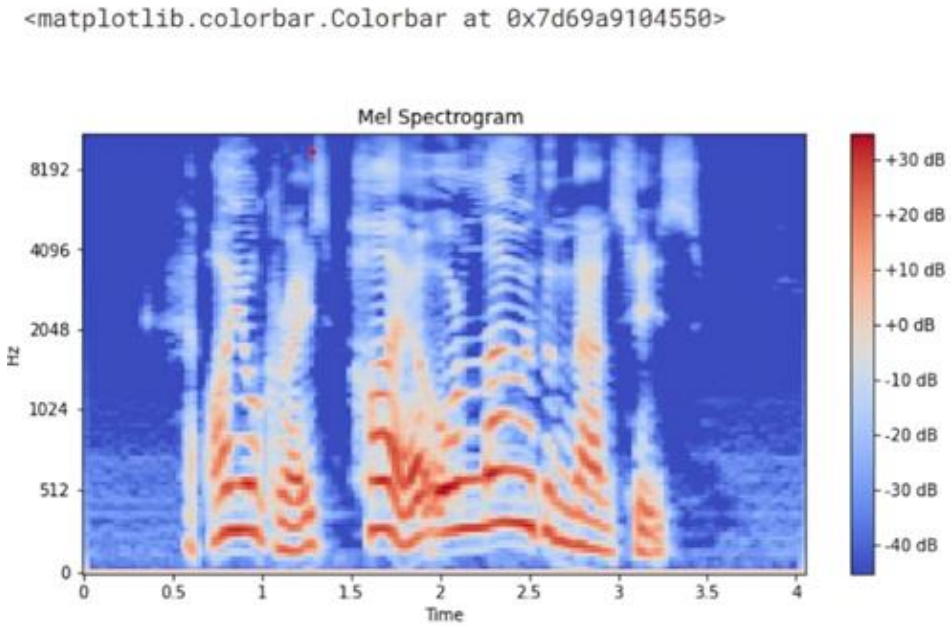

Figure 3.2 Mel Spectrogram

The Mel spectrogram shown in Figure 3.2 provides a visual representation of the audio signal's frequency content over time. The x-axis represents time, the y-axis represents frequency (in the Mel scale), and the color intensity indicates the strength of the signal at different time-frequency points.

Understanding the patterns and variations in the Mel spectrogram can provide valuable insights into the characteristics of the audio signal and help in designing effective signal processing and machine learning algorithms for audio-related tasks.

3.7 Data Augmentation

To improve the model's robustness, various data augmentation techniques are applied, including noise addition, stretching, shifting, and pitch alterations. These techniques help diversify the training data and reduce overfitting.

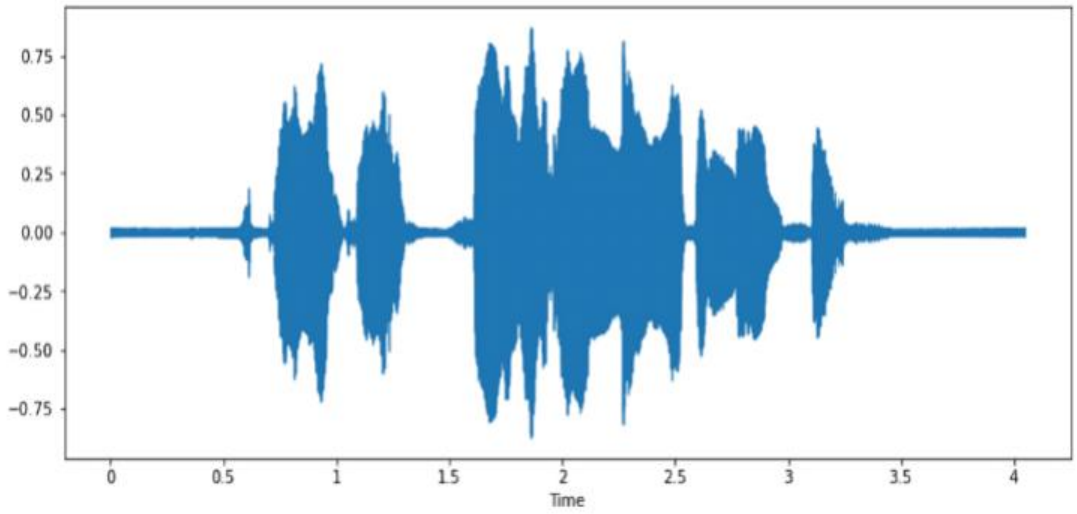

Figure 3.3 Normal Audio

Figure 3.3 provides a dual representation of audio data, offering both a visual depiction of the audio waveform and the ability to play the audio directly within the Jupyter Notebook or IPython environment. This integrated approach enhances the user experience and supports comprehensive exploration and analysis of audio content within the notebook environment.

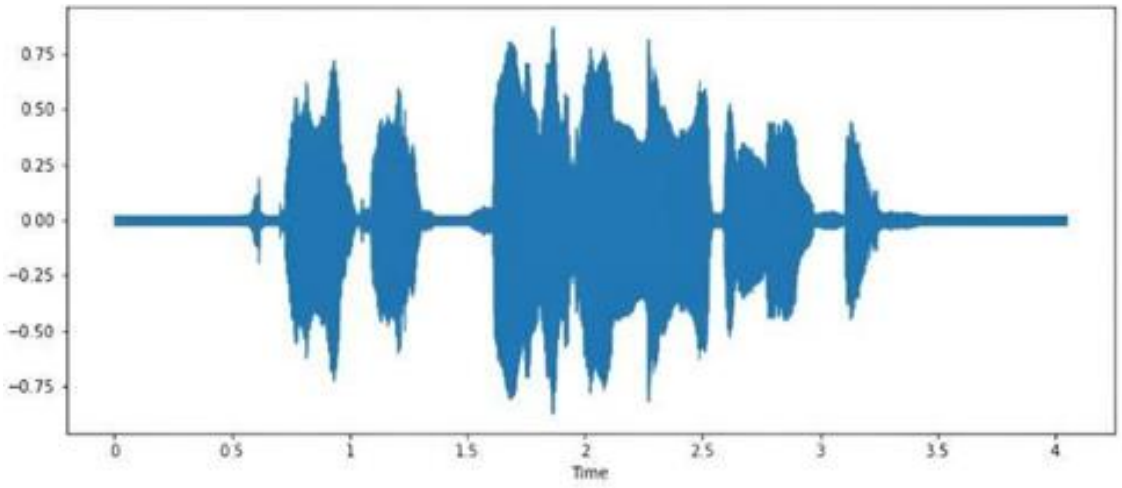

Figure 3.4 Audio with Noise

Figure 3.4 offers a combined visual representation of the noisy audio waveform and the capability to listen to the audio with added noise. This integrated approach enables users to comprehensively evaluate the impact of noise on the audio signal and supports informed decision-making in audio processing and enhancement tasks.

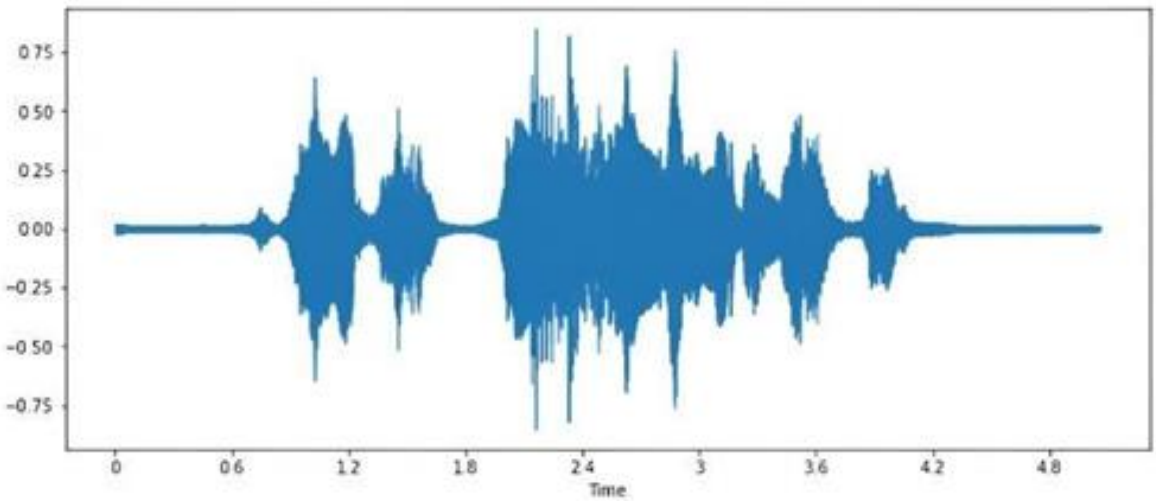

Figure 3.5 Stretched Audio

Figure 3.5 offers a combined visual representation of the stretched audio waveform and the functionality to listen to the stretched audio. This integrated approach enables users to comprehensively evaluate the effects of stretching on the audio signal and supports informed decision-making in audio processing tasks.

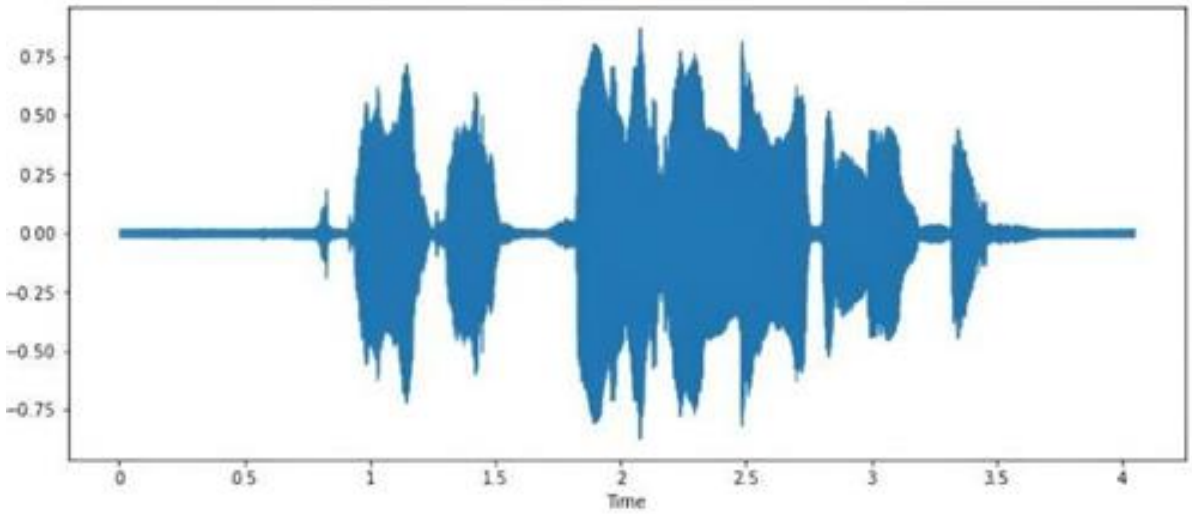

Figure 3.6 Shifted Audio

Figure 3.6 offers a dual representation showcasing both the visual depiction of the shifted audio waveform and the capability to directly listen to the shifted audio. This integrated approach enables users to comprehensively assess the effects of time shifting on the audio signal and supports informed decision-making in audio processing tasks.

3.8 Feature Extraction

Feature extraction is conducted on the pre-processed audio data using both standard and optimized methods. Conventional and optimized approaches are explored to capture essential information for emotion recognition efficiently.

Extracted features are stored for future experiments. Data preparation steps include separating features and labels, applying one-hot encoding for multiclass classification, reshaping data for compatibility with CNN models, and scaling using Sklearn's Standard Scaler. Early stopping is implemented during model training to prevent overfitting.

Table 3.2 Emotion Data

	0	1	2	3	4	5	6
0	0.331543	0.471680	0.564941	0.452148	0.374512	0.296875	0.265137
1	0.238770	0.361816	0.478516	0.473633	0.485352	0.476074	0.472656
2	0.299805	0.419922	0.525879	0.459473	0.376172	0.326172	0.282227
3	0.252930	0.382812	0.497559	0.497070	0.487793	0.472656	0.482422
4	0.400879	0.591309	0.783203	0.777832	0.771973	0.777832	0.771973

2370	2371	2372	2373	2374	2375	Emotions
-1.234544	-0.693115	-0.038821	0.675410	1.4058862	2.112551	surprise
1.797666	-1.586487	-0.501919	-3.159530	-5.015890	-0.942531	surprise
1.855116	2.404975	2.896071	3.282977	3.541091	3.666846	surprise
-3.266942	7.691891	7.443986	-2.031003	-2.095720	-1.418903	surprise
0.986784	-0.730886	-2.531058	-4.002848	-4.849412	-4.929412	neutral

Tables 3.2 presents the feature matrices with columns representing different features, along with an 'Emotions' column containing the corresponding emotion labels. These tables provide a snapshot of the data structure, showcasing the numerical features extracted from the audio signals alongside their associated emotion labels. Such structured data forms the basis for training machine learning models for tasks like emotion recognition, where features derived from audio signals are used to predict the emotional state conveyed in the input audio.

The research encompasses the implementation of a CNN model for speech emotion recognition, leveraging the pre-processed and scaled data. A Convolutional Neural Network (CNN) model is also developed, predictions are made on the test data, and random predictions are scrutinized for validation. Various plots are generated to visualize the performance of multi-models.

3.9 Predict the Emotions

A Convolutional Neural Network (CNN) model is developed for speech emotion recognition, leveraging the pre-processed and scaled data. Predictions are made on the test data, and the model's performance is validated.

Table 3.3 Prediction Table

S. No.	Predicted Labels	Actual Labels
1	angry	angry
2	angry	angry
3	disgust	disgust
4	happy	happy
5	fear	fear
6	happy	happy
7	fear	fear
8	fear	fear
9	surprise	surprise

Table 3.3 displays the predicted labels and actual labels for the initial 10 samples in the test set. It offers a concise summary of the model's performance on the test data. It is making predictions on the test data using a trained model. It then compares the predicted labels with the actual labels and creates a data frame to display.

3.10 Evaluation

The performance of different models is rigorously evaluated using predefined metrics. Results from the best-performing model are analyzed, and various plots are generated to visualize the performance of multiple models.

The best model is serialized to JSON and HDF5 formats, along with the weights of the model, for future use. Additionally, the Standard Scaler object used for scaling is saved to ensure consistency in preprocessing.

In this study, a confusion matrix was utilized to evaluate the model for recognizing emotions from speech. The accuracy of the model was noted for measuring the performance of the SER model and was calculated by using the below equation. Also, the performance of the model was tested by various experiments.

$$\text{Accuracy} = (T\,P + T\,N) / (T\,P + F\,P + T\,N + F\,N)$$

where:

TP = True Positive,

TN = True Negative,

FP = False Positive,

FN = False Negative.

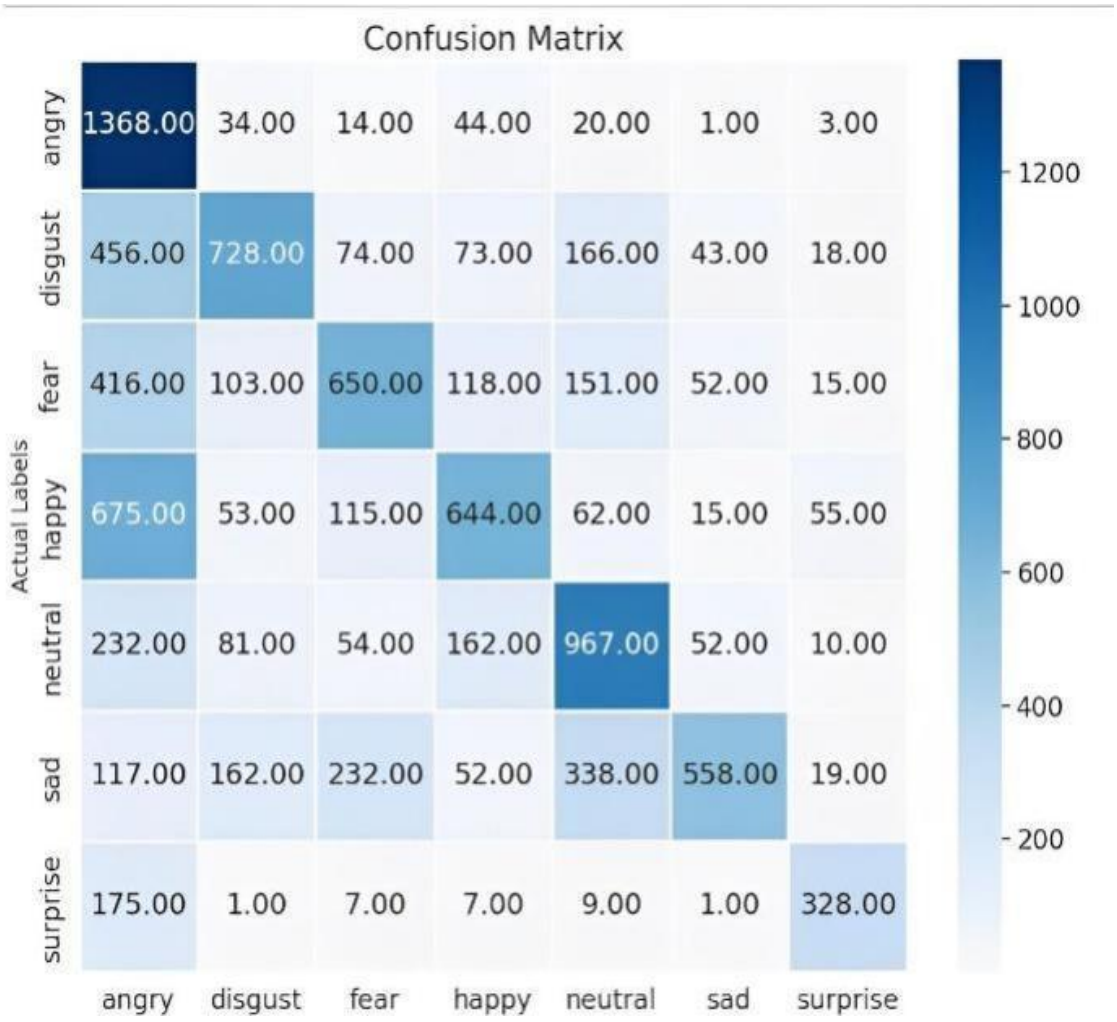

Figure 3.7 Confusion Matrix

Figure 3.7 displays the result of the code, presenting a visual representation of the confusion matrix and a textual summary of the classification report. The confusion matrix heatmap provides a visual representation of the model's performance in accurately and inaccurately identifying cases for each emotion class. The categorization report offers further metrics to assess the model's performance on individual emotion classes.

Table 3.4 Predicted Labels

	Precision	Recall	F1-score	support
Angry	0.96	0.97	0.97	1484
Disgust	0.97	0.95	0.96	1558
Fear	0.96	0.97	0.96	1505
Happy	0.96	0.95	0.96	1619
Neutral	0.97	0.98	0.97	1558
Sad	0.96	0.97	0.96	1478
Surprise	0.98	0.97	0.97	528
Accuracy			0.96	9730
Macro avg	0.96	0.96	0.96	9730
Weighted avg	0.96	0.96	0.96	9730

Table 3.4 presents a detailed classification report, offering additional metrics for evaluating the emotion recognition model's performance on each emotion class. This report provides insights into the model's precision, recall, F1-score, and support for each emotion class, enabling a granular assessment of the model's effectiveness in recognizing different emotions.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Result and discussion

In the realm of speech emotion recognition analysis, evaluating the efficacy of machine learning models is paramount. This entails several pivotal steps, including data segmentation into training, validation, and testing sets, model training, and subsequent assessment utilizing key performance metrics such as accuracy, precision, recall, and F1-score. Visual representations like confusion matrices furnish valuable insights into the model's efficacy. Moreover, error analysis plays a crucial role in elucidating patterns within misclassifications. Comparative evaluations vis-à-vis benchmark models furnish essential contextual understanding. Optionally, the scrutiny of bias and fairness augments the depth of analysis. Qualitative scrutiny, entailing the listening to audio samples, serves to validate the model's real-world applicability. An iterative approach is indispensable to enhancing the model's accuracy and generalization, ensuring its alignment with human discernment in deciphering emotions in speech.

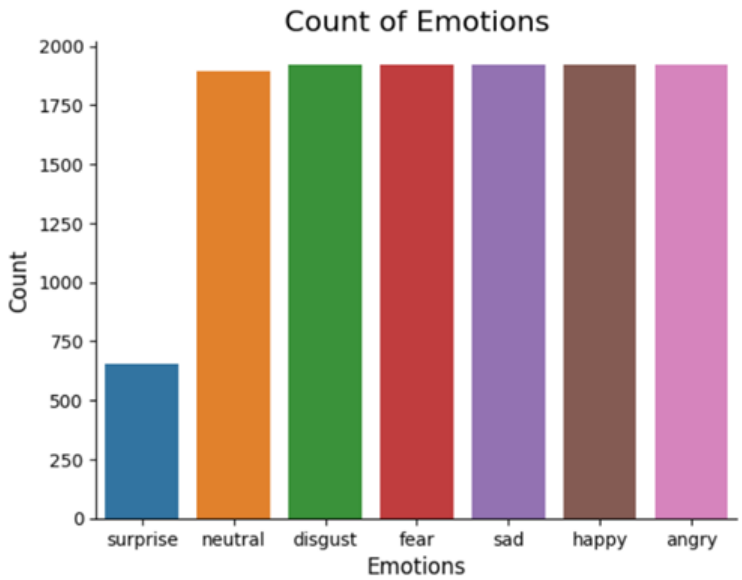

Figure 4.1 Count of Emotions

The bar plot in Figure 4.1 illustrates the distribution of different emotions in the combined

dataset, encompassing data from all four datasets. Each emotion category is represented by a bar, with the height of the bar indicating the count of data points corresponding to that emotion.

From the plot, it is evident that the number of data points associated with the "surprise" emotion is substantially lower compared to the other emotion categories. This discrepancy in data distribution highlights a potential class imbalance issue within the dataset, where certain emotions are underrepresented compared to others.

Understanding and addressing class imbalance is crucial in training machine learning models for emotion recognition, as it can significantly impact the model's performance and generalization ability. Techniques such as oversampling, under sampling, or using class weights during model training may be employed to mitigate the effects of class imbalance and ensure fair representation of all emotion categories in the training data.

We have also done changes with the audio files by adding noise, elevating the pitch, stretching the audio etc. The following graph shows the different wave plots of the changed audio samples.

In [47]:
```
import tensorflow.keras.layers as L

model = tf.keras.Sequential([
    L.Conv1D(512,kernel_size=5, strides=1,padding='same', activation='relu',input_shape=(X_train.shape[1],1)),
    L.BatchNormalization(),
    L.MaxPool1D(pool_size=5,strides=2,padding='same'),

    L.Conv1D(512,kernel_size=5,strides=1,padding='same',activation='relu'),
    L.BatchNormalization(),
    L.MaxPool1D(pool_size=5,strides=2,padding='same'),
    Dropout(0.2),  # Add dropout layer after the second max pooling layer

    L.Conv1D(256,kernel_size=5,strides=1,padding='same',activation='relu'),
    L.BatchNormalization(),
    L.MaxPool1D(pool_size=5,strides=2,padding='same'),

    L.Conv1D(256,kernel_size=3,strides=1,padding='same',activation='relu'),
    L.BatchNormalization(),
    L.MaxPool1D(pool_size=5,strides=2,padding='same'),
    Dropout(0.2),  # Add dropout layer after the fourth max pooling layer
    L.Conv1D(128,kernel_size=3,strides=1,padding='same',activation='relu'),
    L.BatchNormalization(),
    L.MaxPool1D(pool_size=3,strides=2,padding='same'),
    Dropout(0.2),  # Add dropout layer after the fifth max pooling layer

    L.Flatten(),
    L.Dense(512,activation='relu'),
    L.BatchNormalization(),
    L.Dense(7,activation='softmax')
])
model.compile(optimizer='adam',loss='categorical_crossentropy',metrics='accuracy')
model.summary()
```

Figure 4.2 CNN Function for Emotion Prediction

The Figure 4.2 illustrates the architecture of the Convolutional Neural Network (CNN) function utilized for predicting emotions in speech signals. The CNN comprises six levels, each consisting of three layers: ReLU activation, batch normalization, and max pooling.

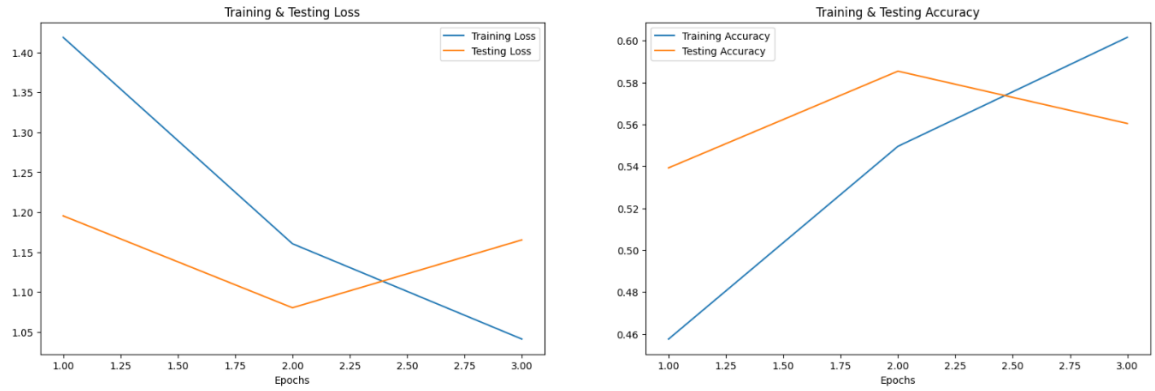

Figure 4.3 Training testing loss and accuracy

The accuracy of our model is around 96% on the training dataset and 71% on the testing dataset.

4.2 Website Status

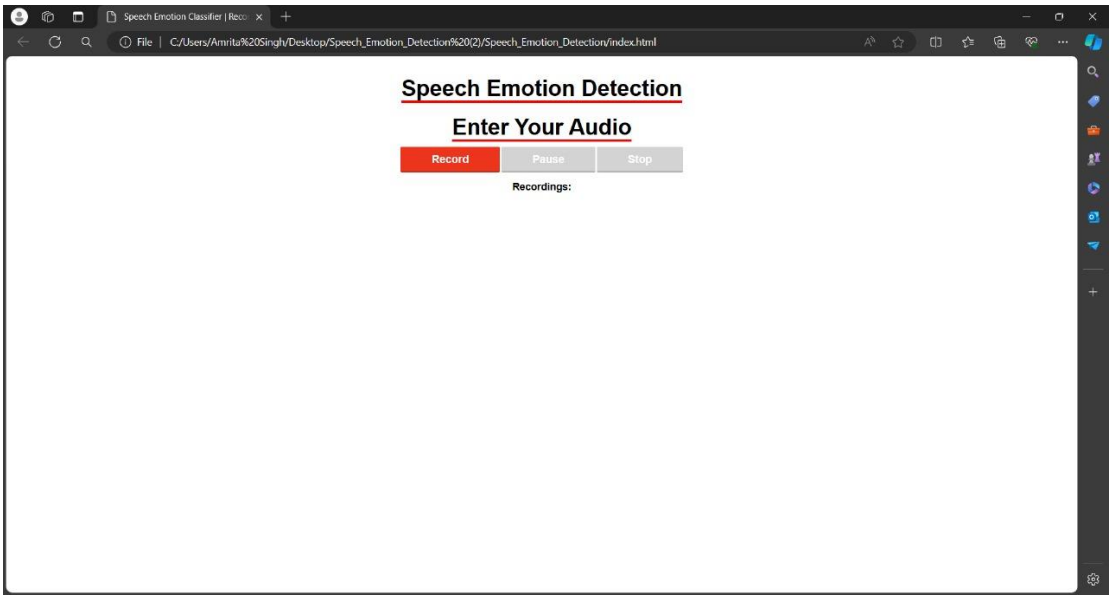

Figure 4.4 Sample GUI

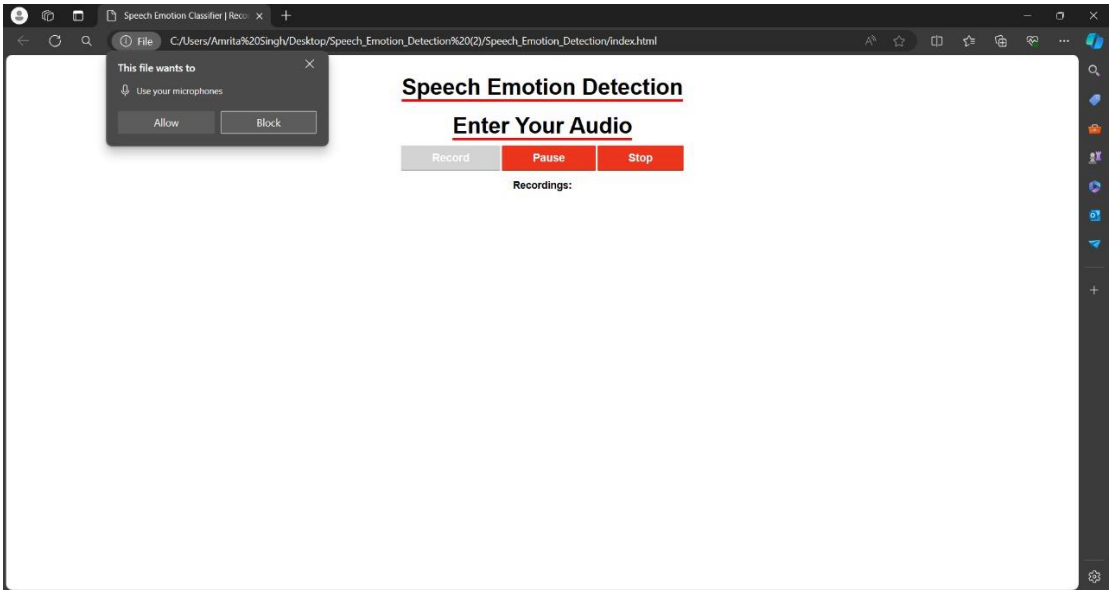

Figure 4.5 Permission to allow microphone

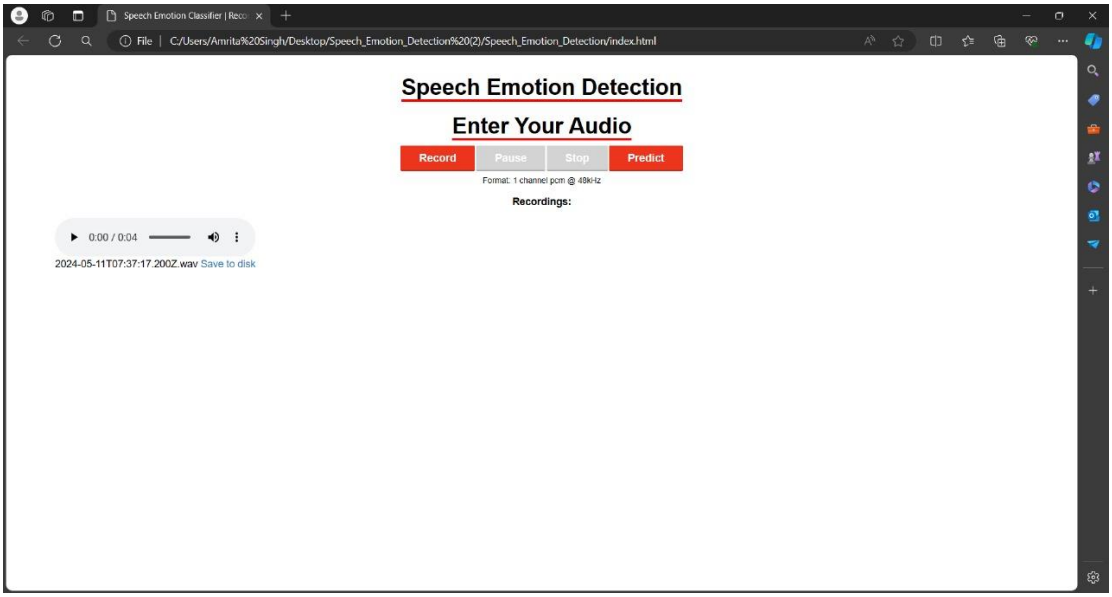

Figure 4.6 Audio Recorded and Predict

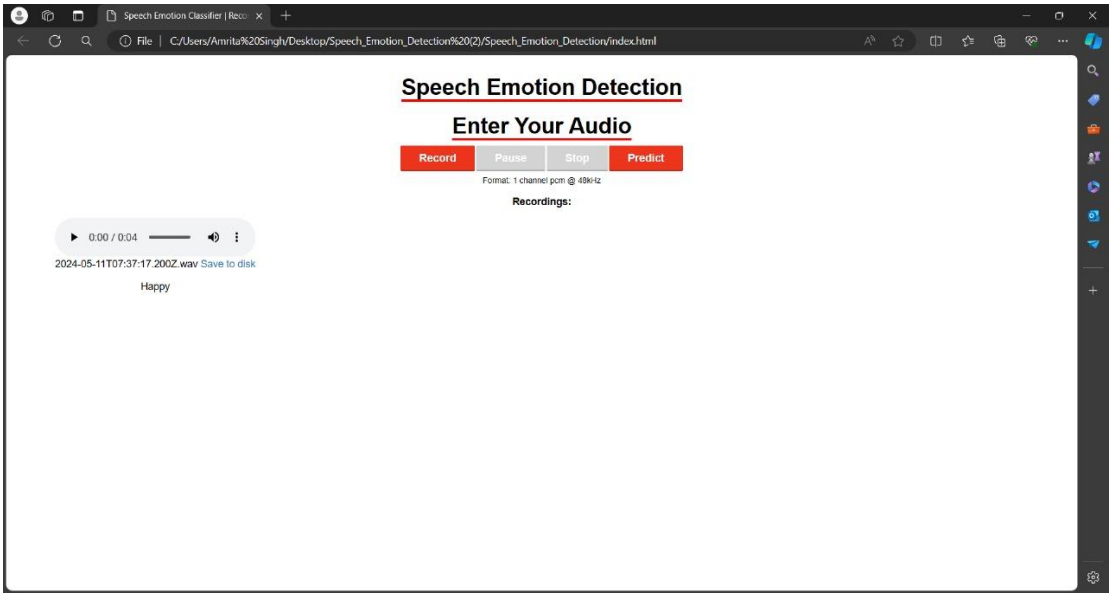

Figure 4.7 Predicted Happy Emotion

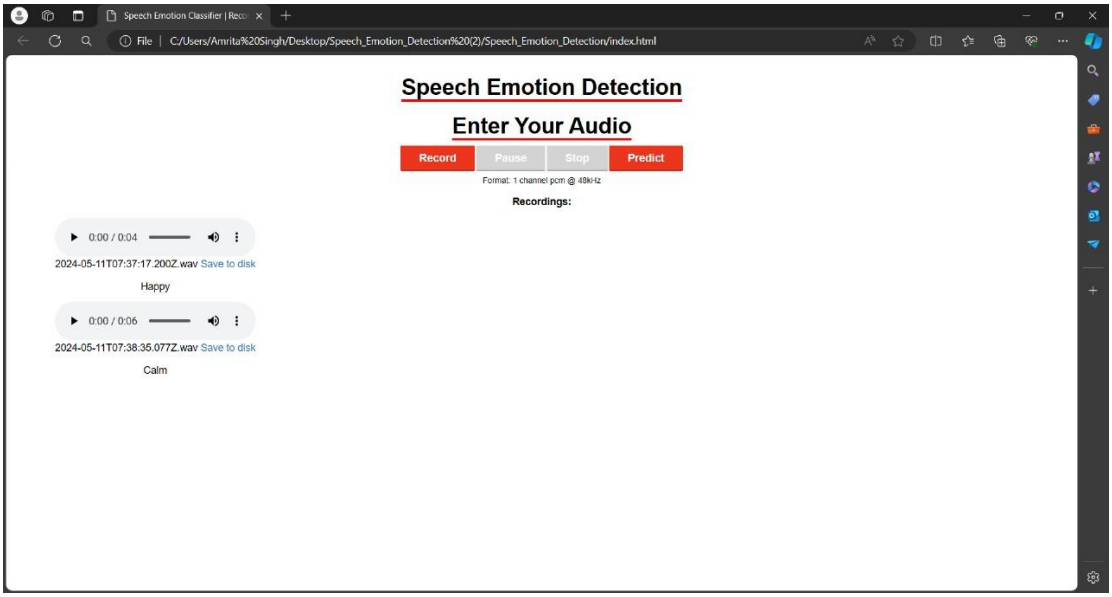

Figure 4.8 Predicted Calm Emotion

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

This research delves into Speech Emotion Recognition (SER) systems, elucidating their methodologies, applications, and challenges. Through the utilization of diverse models and datasets like RAVDESS, CREMA-D, TESS, and SAVEE, and the adoption of log mel spectrograms for feature extraction, our investigation showcases SER's efficacy in discerning emotional cues from speech signals.

Our study underscores SER's significance in practical contexts and its potential to transform human-computer interaction and cognitive analysis. By situating ourselves within the dynamic SER landscape, we not only refine existing models but also highlight critical challenges necessitating collective attention.

To propel SER systems forward, concerted efforts are essential to tackle these challenges. Future research could prioritize the development of more inclusive datasets encompassing a wider range of languages, accents, and emotional expressions. Additionally, exploring innovative feature extraction techniques, such as integrating contextual information and multimodal cues, could bolster SER model discriminative capabilities.

Interdisciplinary collaborations are crucial for advancing our comprehension of speech emotion dynamics. By fostering partnerships among linguists, psychologists, and computer scientists, we can gain insights and construct more holistic SER systems adept at accurately capturing and interpreting the subtleties of human emotional speech.

In summary, our research contributes to the continual advancement of SER systems, emphasizing the vast potential for future progress in this domain. By addressing challenges and fostering interdisciplinary cooperation, we can pave the way for more sophisticated SER systems that enrich human-computer interaction, cognitive analysis, and psychiatric assessment.

5.2 Future Scope

To further enhance SER systems, future endeavors could concentrate on several fronts. Firstly, the development of more inclusive datasets representing a broader linguistic and cultural spectrum would be beneficial. These datasets should encompass diverse languages, accents, and emotional expressions to ensure the robustness and inclusivity of SER models.

Secondly, exploring novel feature extraction techniques, such as incorporating contextual information and multimodal cues, holds promise for enhancing SER model performance. By leveraging contextual cues and integrating information from multiple modalities like facial expressions and gestures, SER systems can better capture the nuances of emotional speech.

Moreover, interdisciplinary collaborations remain crucial for advancing SER research. Collaborative efforts among linguists, psychologists, and computer scientists can yield valuable insights into the complexities of speech emotion dynamics and inform the development of more sophisticated SER systems.

In conclusion, by addressing these avenues for future research and fostering interdisciplinary partnerships, we can propel SER systems to new heights, enabling them to better serve in diverse applications such as human-computer interaction, cognitive analysis, and psychiatric assessment.

REFERENCES

[1] M. RajaBabu, P. Abhinav, Nihaal Subhash. "International Journal For Science Technology And Engineering-Vol. 11, Iss: 4, pp 1182-1185."

[2] Sreeja Sasidharan Rajeswari, G. Gopakumar, Manjusha Nair, Amrita Vishwa Vidyapeetham. "pp 169-178."

[3] Anil Kumar Pagidirayi, Anuradha Bhuma. "Revue Dintelligence Artificielle-Vol. 36, Iss: 2, pp 271-278

[4] A. Kumar, V. Kumar, and P. Rajakumar. "Speech Emotion Recognition Using Machine Learning." 2023 3rd International Conference on Innovative Practices in Technology and Management (ICIPTM), Uttar Pradesh, India, 2023, pp. 1-6.

[5] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. Ali Mahjoub, and C. Cleder. "Automatic Speech Emotion Recognition Using Machine Learning." Social Media and Machine Learning. IntechOpen, Feb. 19, 2020. doi: 10.5772/intechopen.84856.

[6] K. V. Krishna, N. Sainath, and A. M. Posonia. "Speech Emotion Recognition using Machine Learning." 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2022, pp. 1014-1018. doi: 10.1109/ICCMC53470.2022.9753976.

[7] Saw, A. K., Arya, C., Sahu, D., & Shrivas, S. (2022). "Speech emotion recognition using machine learning." International Journal of Health Sciences, 6(S1), 14313–14321.

[8] Amjad A, Khan L, Chang H. (2021). "Effect on speech emotion classification of a feature selection approach using a convolutional neural network." PeerJ Computer Science 7:e766.

[9] de Lope, J., Hernández, E., Vargas, V., Graña, M. (2021). "Speech Emotion Recognition by Conventional Machine Learning and Deep Learning." In: Sanjurjo González, H., Pastor López, I., García Bringas, P., Quintián, H., Corchado, E. (eds) Hybrid Artificial Intelligent Systems. HAIS 2021. Lecture Notes in Computer Science (), vol 12886. Springer, Cham.

[10] S. Alisamir and F. Ringeval. "On the Evolution of Speech Representations for Affective Computing: A Brief History and Critical Overview." IEEE Signal Processing Magazine, vol. 38, no. 6, pp. 12-21, Nov. 2021. doi: 10.1109/MSP.2021.3106890.

[11] H. J, R. R, and R. S D. "A Study on Speech Emotion Prediction using Deep Learning Algorithm." 2021 Smart Technologies, Communication and Robotics (STCR), Sathyamangalam, India, 2021, pp. 1-5. doi: 10.1109/STCR51658.2021.9588861.

[12] H. Patni, A. Jagtap, V. Bhoyar, and A. Gupta. "Speech Emotion Recognition using MFCC, GFCC, Chromagram and RMSE features." 2021 8th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2021, pp. 892-897.

[13] A. Arun, I. Rallabhandi, S. Hebbar, A. Nair, and R. Jayashree. "Emotion Recognition in Speech Using Machine Learning Techniques." 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2021, pp. 01-07.

[14] C. Özkan and K. Oğuz. "Selecting Emotion Specific Speech Features to Distinguish One Emotion from Others." 2021 International Conference on Innovations in Intelligent Systems and Applications (INISTA), Kocaeli, Turkey, 2021, pp. 1-5.

[15] Pichora-Fuller, M. Kathleen; Dupuis, Kate. (2020). "Toronto emotional speech set (TESS)", https://doi.org/10.5683/SP2/E8H2MF, Borealis, V1. Livingstone SR, Russo FA. (2018). "

APPENDIX