SPEECH EMOTION RECOGNITION

A PROJECT REPORT

Submitted by

S YAMINI [211419104312]

NELOPHER NISHA M [211419104242]

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

PANIMALAR ENGINEERING COLLEGE

(An Autonomous Institution, Affiliated to Anna University, Chennai)

APRIL 2023

PANIMALAR ENGINEERING COLLEGE

(An Autonomous Institution, Affiliated to Anna University, Chennai)

BONAFIDE CERTIFICATE

Certified that this project report **"SPEECH EMOTION RECOGNITION"** is the bonafide work of **"S YAMINI (211419104312) & NELOPHER NISHA M (211419104242)"** who carried out the project work under my supervision.

SIGNATURE

Dr.L.JABASHEELA,M.E., Ph.D.,
HEAD OF THE DEPARTMENT

DEPARTMENT OF CSE,
PANIMALAR ENGINEERING COLLEGE,
NASARATHPETTAI,
POONAMALLEE,
CHENNAI-600 123.

SIGNATURE

Dr.T.TAMILVIZHI, M.TECH., Ph.D.,
SUPERVISOR,
ASSOCIATE PROFESSOR

DEPARTMENT OF CSE,
PANIMALAR ENGINEERING COLLEGE,
NASARATHPETTAI,
POONAMALLEE,
CHENNAI-600 123.

Certified that the above candidate(s) were examined in the End Semester Project Viva-Voce Examination held on.....

INTERNAL EXAMINER **EXTERNAL EXAMINER**

DECLARATION BY THE STUDENT

We **S YAMINI(211419104312), NELOPHER NISHA M (211419104242)** hereby declare that this project report titled "**SPEECH EMOTION RECOGNITION**", under the guidance of **Dr.T.TAMILVIZHI, M.TECH., Ph.D.,** is the original work done by us and we have not plagiarized or submitted to any other degree in any university by us.

1. **S YAMINI**

2. **NELOPHER NISHA M**

ACKNOWLEDGEMENT

We would like to express our deep gratitude to our respected Secretary and Correspondent **Dr.P.CHINNADURAI, M.A., Ph.D.** for his kind words and enthusiastic motivation, which inspired us a lot in completing this project.

We express our sincere thanks to our beloved Directors **Tmt.C.VIJAYARAJESWARI**, **Dr.C.SAKTHI KUMAR,M.E.,Ph.D** and **Dr.SARANYASREE SAKTHI KUMAR B.E.,M.B.A.,Ph.D.,** for providing us with the necessary facilities to undertake this project.

We also express our gratitude to our Principal **Dr.K.Mani, M.E.,Ph.D.** who facilitated us in completing the project.

We thank the Head of the CSE Department, **Dr.L.JABASHEELA , M.E.,Ph.D.,** for the support extended throughout the project.

We would like to thank my Project Guide **Dr.T.TAMILVIZHI, M.TECH., Ph.D.,** and all the faculty members of the Department of CSE for their advice and encouragement for the successful completion of the project.

S YAMINI
NELOPHER NISHA M

ABSTRACT

The human voice is very complex and carries multiple emotions. Emotion in speech carries insight about human actions. Through further analysis, we can better understand the motives of people.In this proposed project, we perform speech data analysis on speaker discriminated speech signals to detect the emotions of the speakers involved in the conversation.We will use Convolutional Neural Networks and LSTM to classify opposing emotions.We use statistics relating to the pitch, Mel Frequency Cepstral Coefficients (MFCCs) and Formants of speech as inputs to classification algorithms. The emotion recognition accuracy of these experiments allow us to explain which features carry the most emotional information and why. It also allows us to develop criteria to class emotions together. Using these techniques we are able to achieve high emotion recognition accuracy.We perform speech data analysis on speaker discriminated speech signals to detect the emotions of the individual speakers involved in the conversation. We are analyzing different techniques to perform speaker discrimination and speech analysis to find efficient algorithms to perform this task.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABBREVIATION	EXPANSION
CNN	Convolutional Neural Network
MFCC	Mel frequency cepstral coefficient
GPU	Graphical Processing Unit
API	Application Programming Interface
LSTM	Long short term memory
FFT	Fast Fourier Transform
DFT	Discrete Fourier Transform

CHAPTER 1
INTRODUCTION

1.1 OVERVIEW

Emotions are the best indicators of the actions of humans in advance. It is of great advantage in the current smart world. Prediction of these emotions can be able to sense the current mood of the driver and control the smart automobile accordingly and can be used in case of chatting with the customers using AI devices,etc.This can be done by extracting the features(including MFCC) of the respective emotions and train the learning model using the classification algorithms(CNN) and eventually the model can predict the emotion by comparing the newly retrieved features and the features of the training dataset and classify them accordingly.

1.2 PROBLEM DEFINITION

The objective of the proposed system is to detect emotions from continuous and spontaneous speech. Speech emotion recognition can be of huge help to prevent cyber crimes and can regulate the smart automobiles according to the emotion assessed from the driver. Other applications can include the chatbots for better understanding of the customers on the other side, audio surveillance, online marketing, call centers, etc. Mel-frequency cepstrum coefficient (MFCC) is the most used representation of the spectral property of voice signals. These are the best for speech recognition as it takes human perception sensitivity with respect to frequencies into consideration. MFCC features are extracted from the speech signals and are used to train different classifiers and feature selection is a crucial step to select the most relevant features out of it. A classifier model is built using Convolutional Neural Networks (CNN) algorithm which proves to be the better deep learning algorithm for processing and recognition tasks.

CHAPTER 2

LITERATURE SURVEY

2.1 Title: [1]

Speech emotion and naturalness recognition system

Author name:

Bagus Tris Atmaja, Akira Sasou and Masato Akagi,

Year of publish:

2022

Description:

Naturalness recognition from the speech is a new application of speech processing technique to predict the degree of naturalness score from unnatural to very natural scores for an utterance.First, ability to multitask learning dimensional emotions is shown and naturalness scores simultaneously with a small loss in naturalness recognition performance scores. Second, we evaluated our models in a 6-fold cross validation evaluation to fill the gap in the previous studies, which only evaluated the performance of the models on a single fold.This cross-validation evaluation enables us to infer conclusions from the results that are more reliable and accurate than in previous studies.

Methodology

Two classifiers-Multilayer perceptron and long short-term memory networks are used. This simple MLP is trained on 200 maximum iterations (epoch) with ten patiences and the model is implemented with LSTM and Dense layers in the Tensorflow toolkit .

2.2 Title: [2]

Speech Emotion Recognition system based on Self-Attention Weight Correction for Acoustic and Text Features

Author name:

Jennifer Santosoand Takeshi Yamada

Year of publish:

2022

Description:

A BLSTM-and-self-attention-based SER method using self-attention weight correction (SAWC) with confidence measures has been proposed. This method is applied to acoustic and text feature extractors in SER to adjust the importance weights of speech segments and words with a high possibility of ASR error. Our proposed SAWC reduces the importance of words with speech recognition error in the text feature while emphasizing the importance of speech segments containing these words in acoustic features. evaluate the effectiveness of the SAWC in each feature extractor in improving the SER performance.

Methodology

We propose a method to improve the basic SER method by adjusting the self-attention weights using CM and named this method self-attention weight correction (SAWC). It is a critical component in the acoustic and text feature extractors. SAWC resolves the issue without retraining or fine-tuning ASR to be robust to emotions.

CM indicates how reliable ASR results are.

2.3 Title: [3]

Design of Efficient Speech Emotion Recognition Based on Multi Task Learning

Author name:

Liu Yunxiang;Zhang Kexin

Year of publish:

2023

Description:

This paper used two multi-task learning models based on adversarial multi-task learning(ASP-MTL). The first model took emotion recognition as the main task and noise recognition as the auxiliary task, and removed the noise part identified by the auxiliary task. After identifying the non-noise part, the second model was constructed. The second model took emotion recognition as the main task and gender classification as the auxiliary task. These two multi-task learning models can not only use shared information to learn the relationship between different tasks, but also can identify specific tasks. This paper used Audio/Visual Emotion Challenge (AVEC) database and AFEW 6.0 database,which were recorded in the field environment. Considering the problem of data imbalance between datasets, the data balance operation was carried out on the data sets in the process of data preprocessing.

Methodology

Feature space is divided into shared LSTM and private LSTM, which are used to extract shared features and private features respectively. ASP-MTL model is divided into feature extraction layer, confrontation and orthogonal constraint layer and specific task layer. . In model 1, the auxiliary task identifies the noise category and discards the signal of the noise category, so as to achieve the effect of eliminating noise through multi-task learning. Then the non-noise signal is input into model 2. The auxiliary task of model 2 is gender classification.

2.4 Title: [4]

Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files

Author name:

Felicia Andayani, Lau Bee Theng, Mark Teekit Tsun and Caslon Chua

Year of publish:

2022

Description:

This paper discussed a proposed hybrid Long Short-Term Memory (LSTM) Network and Transformer Encoder to learn the long-term dependencies in speech signals and classify emotions. Speech features are extracted with Mel Frequency Cepstral Coefficient (MFCC) and fed into the proposed hybrid LSTM-Transformer classifier. A range of performance evaluations was conducted on the proposed LSTM-Transformer model. The results indicate that it achieves a significant recognition improvement compared with existing models offered by other published works.

Methodology

This paper combined both the LSTM and the Transformer layers to learn the long-term dependencies in speech signals for emotion recognition. In this study, the LSTM layer replaces the positional encoding in the Transformer architecture. The Transformer encoder layer also contains a feed-forward network layer with ReLu activation and layer normalization.

2.5 Title: [5]

3D Convolutional Neural Network for Speech Emotion Recognition With Its Realization on Intel CPU and NVIDIA GPU

Author name:

Mohammad Reza Falahzadeh, Edris Zaman Farsa

Year of publish:

2022

Description:

In the proposed method, the three-dimensional reconstructed phase spaces of the speech signals were calculated. Then, emotion-related patterns formed in these spaces were converted into 3D tensors. Accordingly, a 3D CNN for speech emotion recognition applied to two datasets, EMO-DB and eNTERFACE05, using a speaker-independent technique achieved 90.40% and 82.20% accuracy, respectively. By employing gender recognition, the accuracy rates on EMO-DB increased to 94.42% and on eNTERFACE05 rose to 88.47%. Realization of the introduced 3D CNN on both Intel CPU and NVIDIA GPU is also explored.

Methodology

3D tensors are provided using reconstructed phase space of speech signals, and in the second stage, a 3D CNN is trained based on the 3D tensors provided in the first stage and their corresponding emotion labels. In order to apply the compatible inputs for the 3D CNN network and to study the relationship between emotional parameters, speech signals have been modeled and analyzed in a 3D space.the one-dimensional signal is mapped to the three-dimensional space and then a 3D tensor is extracted to apply as the input of the 3D CNN.

2.6 Title: [6]

Attention-Based Multi-Learning Approach for Speech Emotion Recognition With Dilated Convolution

Author name:

Samuel Kakuba, Alwin Poulose

Year of publish:

2022

Description:

This paper proposed an attention-based multi-learning model (ABMD) that uses residual dilated causal convolution (RDCC) blocks and dilated convolution (DC) layers with multi-head attention. The proposed ABMD model achieves comparable performance while taking global contextualized long-term dependencies between features in a parallel manner using a large receptive field with less increase in the number of parameters compared to the number of layers and considers spatial cues among the speech features. Spectral and voice quality features extracted from the raw speech signals are used as inputs. The proposed ABMD model obtained a recognition accuracy and F1 score of 93.75% and 92.50% on the SAVEE datasets, 85.89% and 85.34% on the RAVDESS datasets and 95.93% and 95.83% on the EMODB datasets.

Methodology

The proposed model consists of two branches that are dedicated to extraction of spatial and temporal cues simultaneously. One branch is responsible for learning the long-term dependencies of especially temporal cues using the residual dilated causal convolution blocks (RDCC). The second branch uses a block of two dilated convolution (DC) layers to extract spatial cues that may exist among the speech features.

2.7 Title: [7]

Deep Learning-Based Speech Emotion Recognition Using Multi-Level Fusion of Concurrent Features

Author name:

Samuel Kakuba, Alwin Poulose

Year of publish:

2022

Description:

In this paper, we propose a deep learning-based model named concurrent spatial-temporal and grammatical (CoSTGA) model that concurrently learns spatial, temporal and semantic representations in the local feature learning block (LFLB) which are fused as a latent vector to form an input to the global feature learning block (GFLB). We also investigate the performance of multi-level feature fusion compared to single-level fusion using the multi-level transformer encoder model (MLTED) that we also propose in this paper. The proposed CoSTGA model uses multi-level fusion first at the LFLB level where similar features (spatial or temporal) are separately extracted from a modality and secondly at the GFLB level where the spatial-temporal features are fused with the semantic tendency features.

Methodology

The proposed CoSTGA model uses a combination of dilated causal convolutions (DCC), bidirectional long short-term memory (BiLSTM), transformer encoders (TE), multi-head and self-attention mechanisms. Acoustic and lexical features were extracted from the interactive emotional dyadic motion capture (IEMOCAP) dataset.

2.8 Title: [8]

Evaluating Self-Supervised Speech Representations for Speech Emotion Recognition

Author name:

Bagus Tris Atmaja; Akira Sasou

Year of publish:

2022

Description:

This paper evaluates nineteen self-supervised speech representations and one classical acoustic feature for five distinct speech emotion recognition datasets on the same classifier. We calculate the effect size among twenty speech representations to show the magnitude of relative differences from the top to the lowest performance. The top three are WavLM Large, UniSpeech-SAT Large, and HuBERT Large, with negligible effect sizes among them. The significance test supports the difference among self-supervised speech representations. The best prediction for each dataset is shown in the form of a confusion matrix to gain insights into the best performance of speech representations for each emotion category based on the training data from balanced vs. unbalanced datasets, English vs. Japanese corpus, and five vs. six emotion categories.

Methodology

We used log mel filterbank (FBANK) as a baseline acoustic feature for our SER system.The first evaluated SSL method is the so-called autoregressive predictive coding (APC) [23]. The APC is intended as a feature extractor for a wide range of downstream tasks by incorporating a language model-like training scheme into an acoustic sequence. APC then is improved by the vector quantization (VQ) version.

2.9 Title: [9]

Investigation of the Effect of Increased Dimension Levels in Speech Emotion Recognition

Author name:

Haiyan Wang, Xiaohui Zhao

Year of publish:

2022

Description:

This paper aims to investigate emotion recognition from the spontaneous speech in the three-dimensional model. Each dimension represents one primitive, generic attribute of an emotion. Middle levels of each dimension were introduced in this paper. LSTM network was employed to estimate the dimensions due to its effectiveness in speech emotion recognition. In the experiments, we use the IEMOCAP database and the accuracy is 30–35%. The confusion matrixes show that our method leads to a more concentrated dimension location. Furthermore, dimensions were applied in categorical emotion recognition. This indicates that increasing dimension levels could provide a possibility of dimension estimation, and suggests that it is possible to promote speech emotion recognition with dimensions.

Methodology

The LSTM network is used as the classifier.We trained the model with an Adam optimizer in the TensorFlow framework. The LSTM have advantage for speech emotion recognition over SVM in speech emotion recognition.

2.10 Title: [10]

Unsupervised Personalization of an Emotion Recognition System: The Unique Properties of the Externalization of Valence in Speech

Author name:

Kusha Sridhar, Carlos Busso

Year of publish:

2022

Description:

This study proposes an unsupervised approach to address this problem by searching for speakers in the train set with similar acoustic patterns as the speaker in the test set. Speech samples from the selected speakers are used to create the adaptation set. This approach leverages transfer learning using pre-trained models, which are adapted with these speech samples. We propose three alternative adaptation strategies: unique speaker, oversampling and weighting approaches.

Methodology

Our approach relies on *principal component analysis* (PCA) to reduce the dimension of the space, followed by fitting a *Gaussian mixture model* (GMM) to the resulting reduced feature space.

CHAPTER 3
SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

The Existing system uses the supervised learning algorithm Multi-layer Perceptron which includes one or more non-linear hidden layers that leans the function

$$f(.):R^m \rightarrow R^o$$

Bagus Tris Atmaja, Akira Sasou and Masato Akagi developed a speech emotion and naturalness recognition system[1] using Multi-Task Learning(MTL). The naturalness ratings are labeled on a five-point scale as dimensional emotion. Here multitask refers to dimensional emotion and naturalness score prediction. This work proceeds with the assumption that predicting both can improve the performance but it has been proved from their research that it fails to predict the low and extremely high scores and adopts MTL which involves more than one local minimum which can eventually lead to variable validation accuracy therefore many parameters and hidden layers has to be modified and MLP is not applicable if more number of features are included. Jennifer Santosoand Takeshi Yamada has developed Speech Emotion Recognition system based on Self-Attention Weight Correction for Acoustic and Text Features[2] which uses BLSTM and self attention mechanism that focuses on the incorrectly recognised terms of ASR but the main limitation of this system is Information loss. Liu Yunxiang;Zhang Kexin developed.Design of Efficient Speech Emotion Recognition Based on Multi Task Learning[3] where the Feature space is divided into shared LSTM and private LSTM, which are used to extract shared features and private features respectively. Though this system enhances the high-frequency part of the speech it does not solve the problem of limited data and

contains very few emotion types. Felicia Andayani, Lau Bee Theng, Mark Teekit Tsun and Caslon Chua developed Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files[4] where both the transformer and LSTM architectures are combined due to their individual limitations. The adoption of long-term dependencies could affect the features of the samples and it also requires improvement on the preprocessing methods.Changing input sequence causes various disadvantages to the system.

DISADVANTAGES

- Adopts Multi-task learning(MTL) which has a non-convex loss function where there exists more than one local minimum.
- Therefore different random weight initializations can lead to different validation accuracy.
- Although Multitask learning enhances the performance it fails to predict the low and extremely high scores.

3.2 PROPOSED SYSTEM

Proposed system includes the following steps

- Inputting the audio signal
- Feature extraction
- Feature enhancement
- Classifier training
- Emotion detection

The audio signal input is preprocessed before feature extraction to remove unwanted noise signal.The features are extracted using MFCC technique .Other techniques are LPC and PLP.Classification is performed which maps the features to emotion. CNN is used for classifying the emotions which shows a larger learning rate .
Datasets used are RAVDESS and SAVEE dataset and the libraries used are librosa, matplotlib , numpy, pandas, torchvision, wave,Keras,etc. and the frameworks used are Flask and Pytorch.

Audio speech signals are inputted and Librosa library in Python is used to process and extract features from the audio files using MFCC which are widely used in automatic speech and speaker recognition.The signal is processed through the Fast Fourier Transform which uses less computational time and used to assess the frequency properties of a signal.FFT is just another version of DFT but more efficient and faster.The typical applications of FFT are compression in more complex processing of signals and filtering ,remodeling of signals,etc.

ADVANTAGES

- CNN is used for classifying the emotions which shows a larger learning rate .
- The main advantage of CNN compared to its predecessors is that it automatically detects the important features without any human supervision.
- Fast Fourier Transform is faster than the Discrete Fourier Transform which can be explained based on the heart of the algorithm: Divide And Conquer. So rather than working with big size Signals, we divide our signal into smaller ones, and perform DFT of these smaller signals. At the end we add all the smaller DFT to get the actual DFT of the big signal. This gives great benefit asymptotically.

3.3 FEASIBILITY STUDY

TECHNICAL FEASIBILITY:

For this project, the various technical resources such as dataset and the correct IDE's, GPUs for training the model, etc are available and are made use of them efficiently.

The various libraries, packages and APIs are open-source and are easy to use to manipulate data.

SOCIAL FEASIBILITY:

This project helps the marketing officials to analyze their customers and suggest accordingly and can minimize time and cost.It can support the AI chatbots and respond to the customers accordingly.

3.4 HARDWARE ENVIRONMENT

- Desktop / Laptop
- Processor: Minimum 1 GHz
- Memory (RAM): 2 GB
- Internet Connection

3.5 SOFTWARE ENVIRONMENT

- Jupyter Notebook
- Python
- VS Code
- Frameworks-Flask, Torch
- Libraries-Pytorch, Pandas, Numpy, MatPlot,Keras.

JUPYTER NOTEBOOK

This website acts as “meta” documentation for the Jupyter ecosystem. It has a collection of resources to navigate the tools and communities in this ecosystem, and to help you get started.

Project Jupyter is a project and community whose goal is to "develop open-source software, open-standards, and services for interactive computing across dozens of programming languages". It was spun off from IPython in 2014 by Fernando Perez.

Notebook documents are documents produced by the Jupyter Notebook App, which contain both computer code (e.g. python) and rich text elements (paragraph, equations, figures, links, etc.). Notebook documents are both human-readable documents containing the analysis description and the results (figures, tables, etc.) as well as executable documents which can be run to perform data analysis.

LIBRARIES

1. Pandas

Pandas is a software library written for the Python programming language for data manipulation and analysis.[2] In particular, it offers data structures and operations for manipulating numerical tables and time series. In this project, pandas is used for preprocessing - i.e., cleaning the csv files.

2. NumPy

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. In this project, the user inputs for the crop prediction and the fertilizer recommendation are converted to an array and the operations are performed on these using the NumPy library.

3. MatPlot

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK.

4. Keras

Keras is an open-source high-level Neural Network library, which is written in Python and is capable enough to run on Theano, TensorFlow, or CNTK. It was developed by one of the Google engineers, Francois Chollet. It is made user-friendly, extensible, and modular for facilitating faster experimentation with deep neural networks. It not only supports Convolutional Networks and Recurrent Networks individually but also their combination.

It cannot handle low-level computations, so it makes use of the Backend library to resolve it. The backend library acts as a high-level API wrapper for the low-level API, which lets it run on TensorFlow, CNTK, or Theano.

FRAMEWORKS:

- **Flask**

Flask is a web application framework written in Python. Flask is based on the Werkzeug WSGI toolkit and Jinja2 template engine. This project's API's are created using the Flask framework.

- **Pytorch**

PyTorch is a machine learning framework based on the Torch library, used for applications such as computer vision and natural language processing. The main application of Pytorch is in the disease detection model, where the .jpeg, or .png images are converted to pixels and then converted to a tensor for further processing.

CHAPTER 4
SYSTEM DESIGN

4.1 ER DIAGRAM

The below fig 4.1.1 depicts the relationship between the entities Speech emotion recognition system and the emotions predicted. Attributes of Detector include MFCC Chroma, Pitch, Magnitude and CNN algorithm and Emotion consists of features and labels.

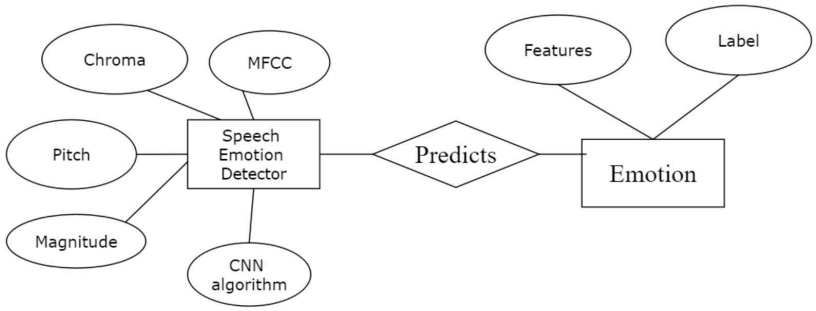

Fig 4.1.1 Er diagram

4.2 DATA DICTIONARY

This is normally represented as the data about the data stored in the database information system or as a part of a research project.

1. Words should be defined to understand what they need and not the variable need by which they may be described in the program.
2. Each word must be unique.
3. Aliases or synonyms are allowed when two or more entries show the same meaning.

4. A self-defining word should not be decomposed.

Data dictionary includes information such as the number of records in a file, the frequency a process will run, security factors like pass word which the user must enter to get excess to the information.

4.2.1 INFORMATION TABLE

Attribute	Data type	Definition
id	integer	Unique id for each file
name	varchar(50)	Name of the file
size	varchar(50)	Size
status	varchar(100)	status of the file

Fig 4.2.1 Information table for the model

4.3 DATA FLOW DIAGRAM

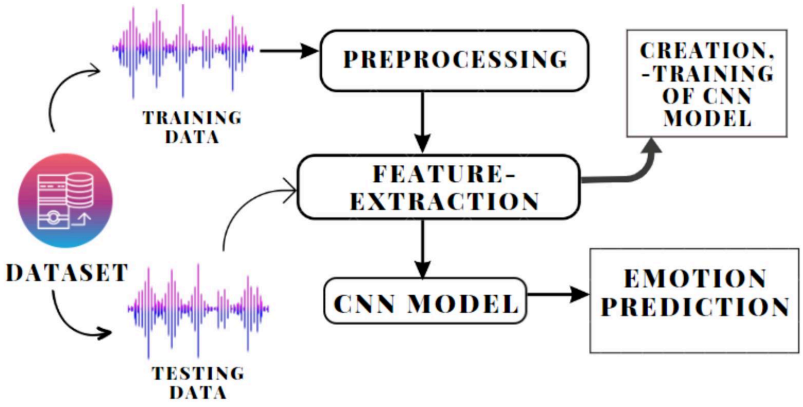

Fig.4.3.1 Data flow diagram

4.4 UML DIAGRAMS

USE CASE DIAGRAM

The below fig 4.5.1 depicts the operations performed by the system and the user. The Developer creates the CNN model and train it using the training dataset and test using the testing dataset and the user can find the emotion of the respective audio file using the model.

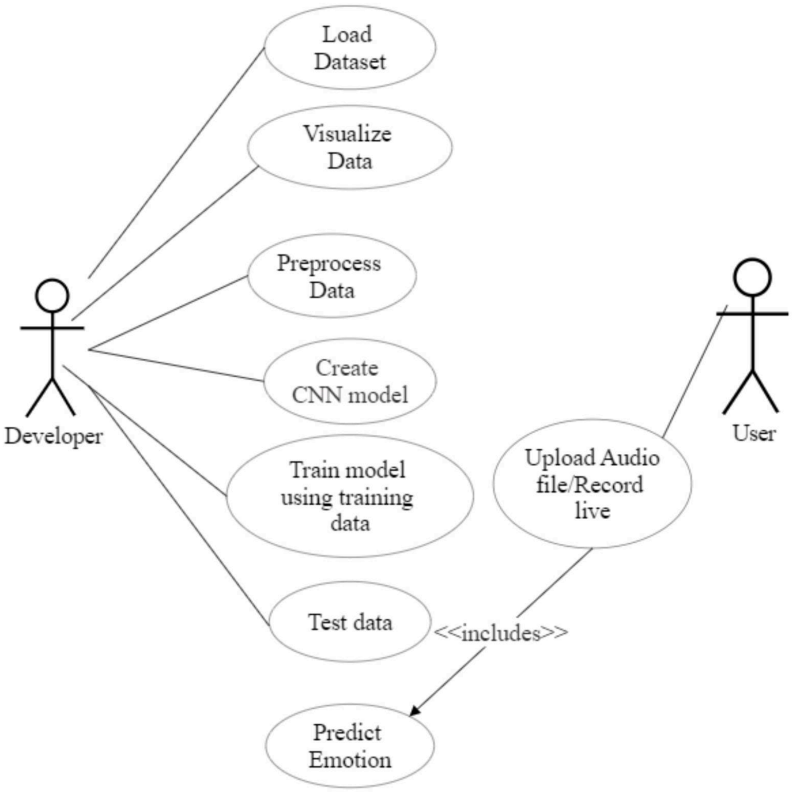

Fig 4.5.1 Use Case diagram

ACTIVITY DIAGRAM

The below fig 4.5.2 depicts the flow of activities during the execution of the application and the order in which they are implemented and the decisions made based on the conditions.

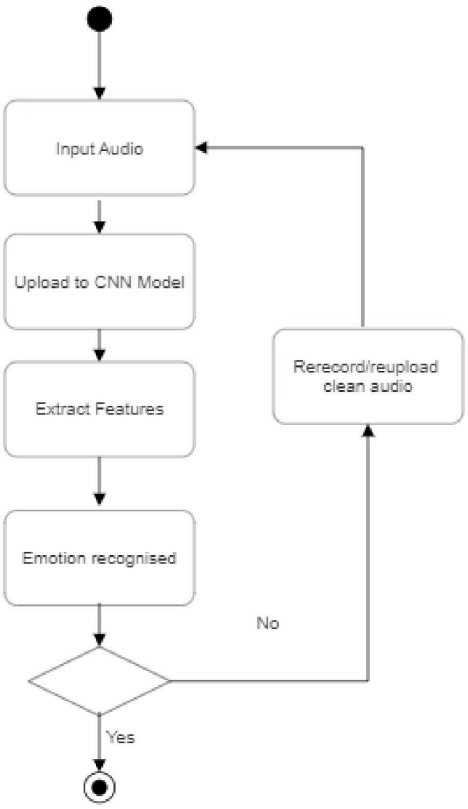

Fig 4.5.2 Activity diagram

CLASS DIAGRAM

The below fig 4.5.3 depicts the relationship between the user, interface and the model developed using the processes included as attributes.Both the training and testing data is used to develop the model.

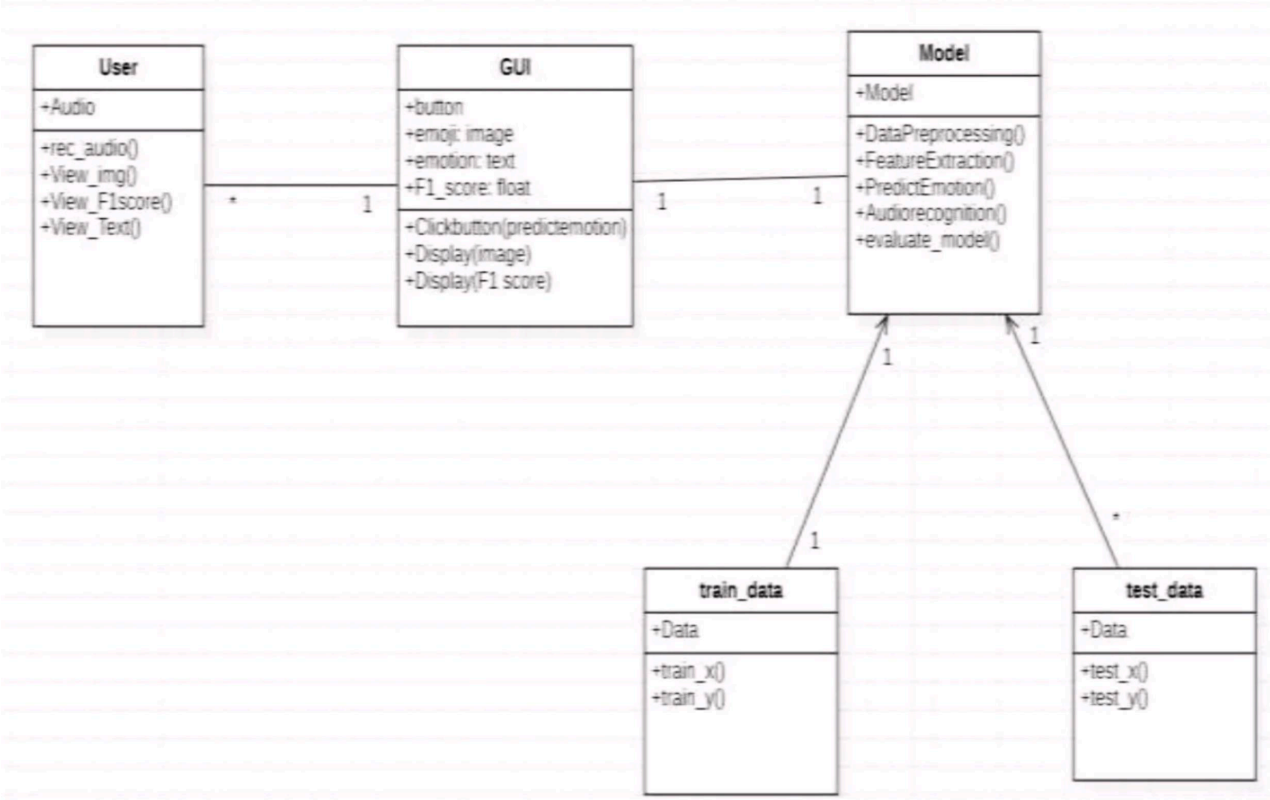

Fig 4.5.3 Class diagram

SEQUENCE DIAGRAM

The below fig 4.5.4 depicts the sequence of activities happening during the prediction of emotion when the user clicks the predict option in the website.

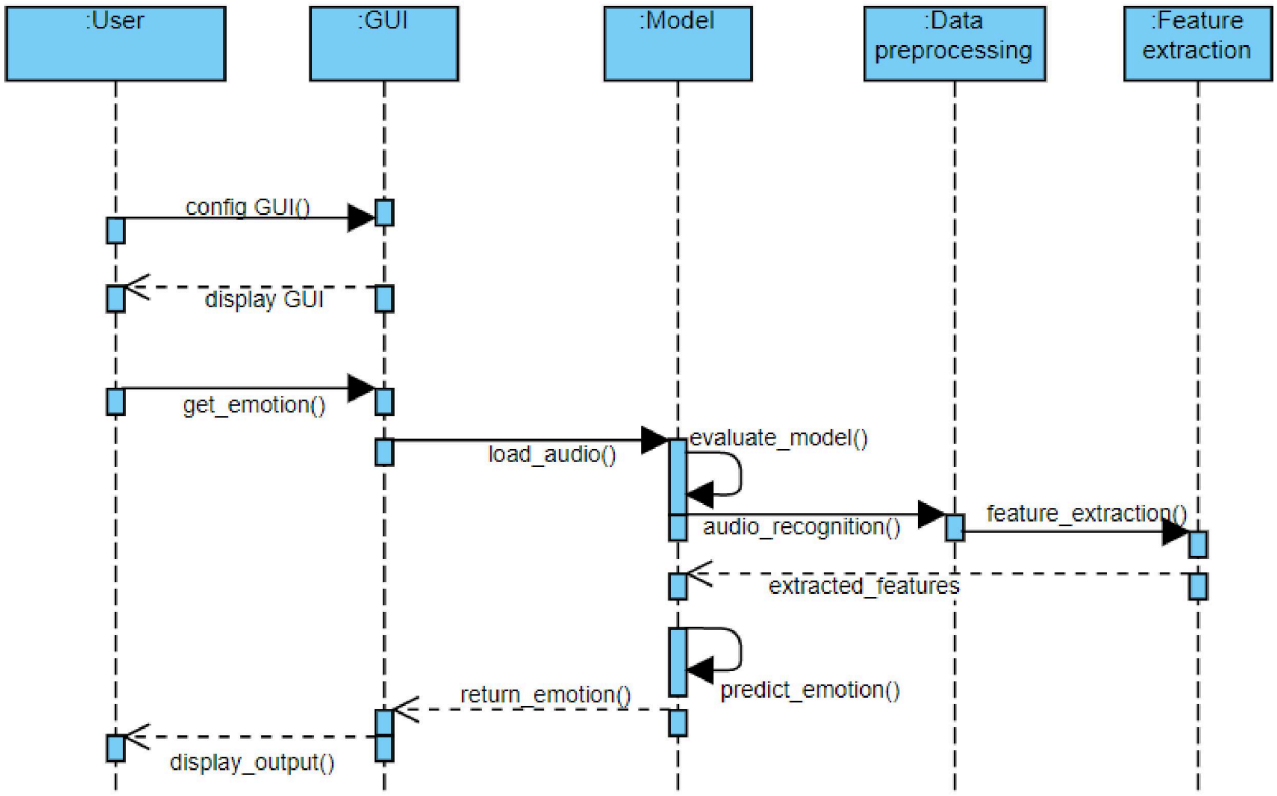

Fig 4.5.4 Sequence diagram

CHAPTER 5
SYSTEM ARCHITECTURE

5.1 MODULE DESIGN SPECIFICATION

- Data preprocessing
- Data Visualization
- Feature extraction
- Model Creation
- Training and Evaluation
- Test set Prediction
- Deployment

Data Preprocessing

Data preprocessing involves below steps:

- Getting the dataset
- Importing libraries
- Importing datasets
- Finding Missing Data
- Encoding Categorical Data
- Splitting dataset into training and test set
- Feature scaling

Librosa and TorchAudio (Pytorch) are two Python packages that are used for audio data pre-processing.

Data Visualization

- Data visualization is defined as a graphical representation that contains the information and the data.

- Data Visualization Libraries used are
    - Matplotlib
    - Plotly

Feature Extraction

- Audio feature extraction is a necessary step in audio signal processing, which is a subfield of signal processing and deals with the processing or manipulation of audio signals and also removes unwanted noise and balances the time-frequency ranges by converting digital and analog signals.
- Deep Learning approach considers unstructured audio representations such as the spectrogram or MFCCs.
- Other features extracted are Magnitude, Pitch, Chroma.

Model Creation

CNN model is created with Keras and constructed with 5 layers — 4 Conv1D layers followed by a Dense layer.

Training and Evaluation

The developed model is then trained using the training dataset of RAVDESS so that the model is able to predict the emotion.

Testing set Prediction

The model is tested using the testing dataset to analyze the efficiency and accuracy of the model.

Deployment

A deployment model is designed using flask where the user can upload an audio file and the system predicts the respective emotion.

5.2 ALGORITHMS

CNN

CNN is a type of Deep Learning architecture commonly used for image classification and recognition tasks. It consists of multiple layers, including Convolutional layers, Pooling layers, and fully connected layers.We have developed the CNN model with Keras and constructed it with 5 layers — 4 Conv1D layers followed by a Dense layer.

The output layer in a CNN as mentioned previously is a fully connected layer, where the input from the other layers is flattened and sent so as to transform the output into the number of classes as desired by the network. vi. The output is then generated through the output layer and is compared to the output layer for error generation. A loss function is defined in the fully connected output layer to compute the mean square loss. The gradient of error is then calculated.

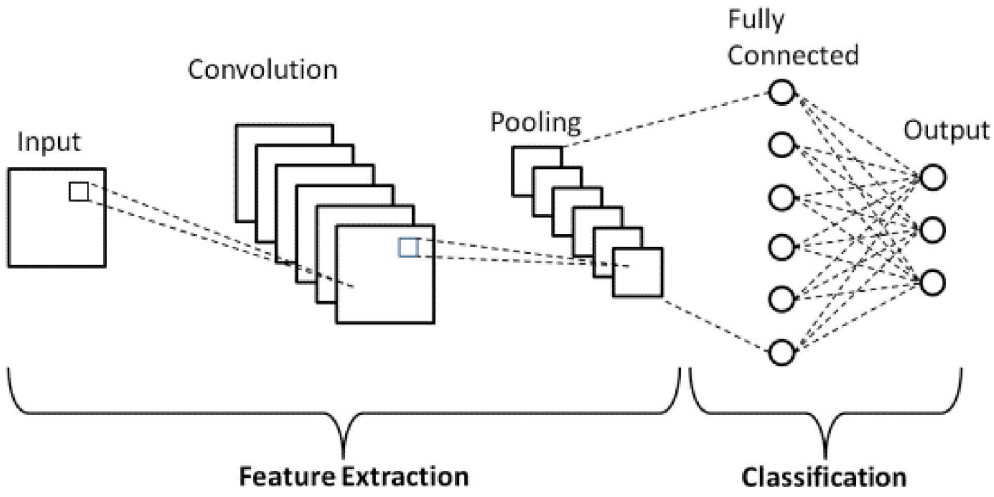

Fig. 5.2.1 CNN Architecture

LSTM

LSTMs are a special kind of RNN — capable of learning long-term dependencies by remembering information for long periods is the default behavior. It defines the Network, compiles the Network, fit Network, evaluates Network and makes Predictions.LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time is practically their default behavior, not something they struggle to learn.

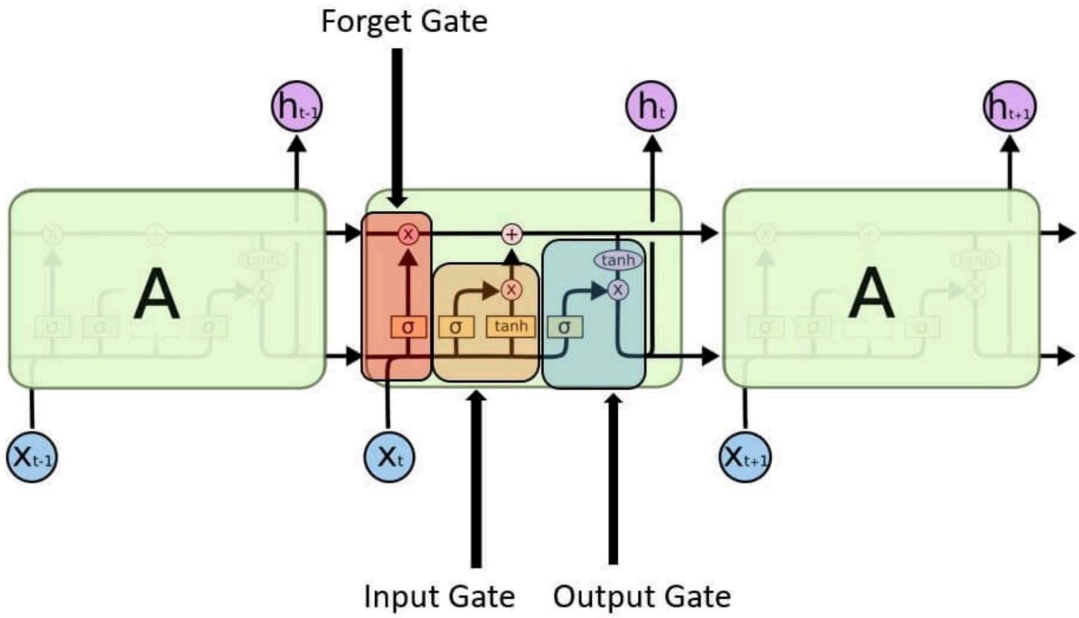

Fig.5.2.2 LSTM Architecture

CHAPTER 6
SYSTEM IMPLEMENTATION

6.1 CLIENT-SIDE CODING

index.html

```
<!DOCTYPE html>
<html><head>
<link href='https://fonts.googleapis.com/css?family=Dancing Script' rel='stylesheet'>
  <style>
     .pixelated {
  image-rendering: pixelated;
  -ms-interpolation-mode: nearest-neighbor;
  image-rendering: crisp-edges;
  image-rendering: -webkit-optimize-contrast;
  image-rendering: optimizeQuality;  }
     body {
       background-image: url("{{ url_for('static', filename='bg.jpg') }}");
       background-repeat: no-repeat;
       background-attachment: fixed;
       background-size: cover;
       background-position:50% 30%;  }
     form{
       padding: 45px;  }
    </style><meta charset="UTF-8">
   <title>Speech Emotion Classifier</title>
```

```
<meta name="viewport" content="width=device-width, initial-scale=1.0"></head>
  <body >
    <h1  style="color:rgb(137, 16, 77);font-family:'Dancing Script';font-size: 600%;">
  Speech Emotion <br>  Detection</h1>
   <form class="form-group" action="/index" method="POST">
   <div id="formats">
     <input name="file" type="file" accept=".mp3,.wav" id="file" /><br>  <br>
     <audio name="recorded" id="player" controls></audio><br>
</div><br>
  </form>
   <script>
     URL = window.URL || window.webkitURL;
     const recorder = document.getElementById('file');
     const player = document.getElementById('player');
     recorder.addEventListener('change', function (e) {
        const file = e.target.files[0];
        const url = URL.createObjectURL(file);
        player.src = url;   });
     </script>
       <div    style="font-family:'Dancing    Script';font-size:300%;padding-left:45px;"
class="alert alert-info" role="alert">
   <ol id="recordingsList" class="center"></ol>
     Emotion predicted is <br>  <div style="color: rgb(109, 32, 98);"><strong>{{ msg
}}</strong> </div></div> </body></html>
```

6.2 SERVER-SIDE CODING

final.ipynb

```
import keras
from keras.models import Model
from keras.models import Sequential
from keras.layers import Conv1D, MaxPooling1D #, AveragePooling1D
from keras.layers import Flatten, Dropout, Activation # Input,
from keras.layers import Dense #, Embedding
from keras.utils import np_utils
from sklearn.preprocessing import LabelEncoder
data, sampling_rate = librosa.load('Dataset/anger/anger016.wav')
ipd.Audio('Dataset/anger/anger016.wav')
dataset_path = os.path.abspath('./Dataset')
destination_path = os.path.abspath('./')
randomize = True
split = 0.8
sampling_rate = 20000
emotions=["anger","disgust","fear","happy","neutral", "sad", "surprise"]
unique_labels = train_df.label.unique()
unique_labels.sort()
unique_labels_counts = train_df.label.value_counts(sort=False)
from utils.feature_extraction import get_features_dataframe
from utils.feature_extraction import get_audio_features
trainfeatures, trainlabel = get_features_dataframe(train_df, sampling_rate)
testfeatures, testlabel = get_features_dataframe(test_df, sampling_rate)
```

```
trainfeatures = pd.read_pickle('./features_dataframe/trainfeatures')
trainlabel = pd.read_pickle('./features_dataframe/trainlabel')
testfeatures = pd.read_pickle('./features_dataframe/testfeatures')
testlabel = pd.read_pickle('./features_dataframe/testlabel')
trainfeatures = trainfeatures.fillna(0)
testfeatures = testfeatures.fillna(0)
X_train = np.array(trainfeatures)
y_train = np.array(trainlabel).ravel()
X_test = np.array(testfeatures)
y_test = np.array(testlabel).ravel()
y_train[:5]
lb = LabelEncoder()
y_train = np_utils.to_categorical(lb.fit_transform(y_train))
y_test = np_utils.to_categorical(lb.fit_transform(y_test))
y_train[:5]
x_traincnn =np.expand_dims(X_train, axis=2)
x_testcnn= np.expand_dims(X_test, axis=2)
x_traincnn.shape
model = Sequential()
model.add(Conv1D(256, 5,padding='same',
                 input_shape=(x_traincnn.shape[1],x_traincnn.shape[2])))
model.add(Activation('relu'))
model.add(Conv1D(128, 5,padding='same'))
model.add(Activation('relu'))
model.add(Dropout(0.1))
```

my_flask.ipynb

```
app = Flask(__name__)
@app.route('/')
def home():
        return render_template('index.html')
@app.route('/index',methods=["GET","POST"])
def index():
        if request.method == 'POST':
                file = request.form['file']
                demo_mfcc, demo_pitch, demo_mag, demo_chrom = 
get_audio_features(file,20000)
                mfcc = pd.Series(demo_mfcc)
                pit = pd.Series(demo_pitch)
                mag = pd.Series(demo_mag)
                C = pd.Series(demo_chrom)
                demo_audio_features = pd.concat([mfcc,pit,mag,C],ignore_index=True)
                demo_audio_features= np.expand_dims(demo_audio_features, axis=0)
                demo_audio_features= np.expand_dims(demo_audio_features, axis=2)
                demo_audio_features.shape
                livepreds = loaded_model.predict(demo_audio_features,
                   batch_size=32,
                   verbose=1)
                index = livepreds.argmax(axis=1).item()
                index
                return render_template('index.html', msg = emotions[index] )
if __name__ == '__main__':
        app.run(debug=False)
```

CHAPTER 7
SYSTEM TESTING

7.1 UNIT TESTING

The various modules of the system were developed and tested individually after the development of each unit. Each form was designed and each api in the flask app was tested after the integration with the UI. Each of the models was trained separately.

7.2 INTEGRATION TESTING

After the development and the testing phase of each of the modules has been completed, all the units were integrated into a single module. After the training and the testing of each mode in the unit testing phase, the pickle files were used in the flask app file to integrate all the functionalities. The flask app was developed and all the functionalities were tested.

7.3 TEST CASES & REPORTS

TEST CASE ID	TEST CASE / ACTION TO BE PERFORMED	EXPECTED RESULT	ACTUAL RESULT	PASS/ FAIL
1	Display the Home page by clicking on the Website Link	Display the features of the website	Display the features of the website	Pass

2	Selecting select file button in home page	Display the name of the audio file	Attach the file and display the name	Pass
3	Selecting predict button	Display the emotion of the attached audio file	Display the emotion of the attached audio file	Pass

CHAPTER 8
CONCLUSION

8.1 RESULTS AND DISCUSSION

Hence our project presents a new way to give the ability to machines to determine emotion with the help of the human voice. A deployment model is designed using flask where the user can upload an audio file and the system predicts the respective emotion.The below figure shows the training and testing loss on our dataset. As we can see from the graph, both training and testing errors reduce as the number of epochs to the training model increases.

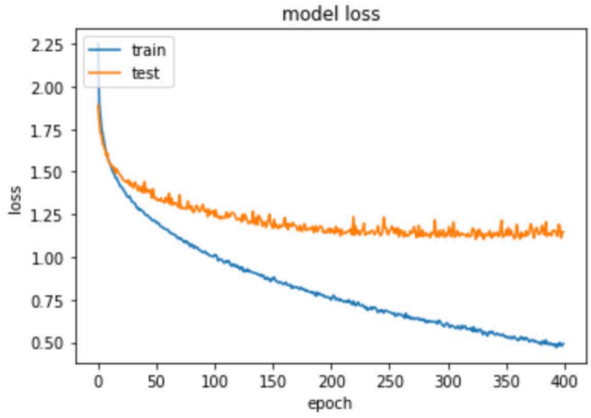

Fig 8.1.1 loss vs epoch

8.2 CONCLUSION AND FUTURE ENHANCEMENTS

Thus the proposed system has the capability of understanding the human emotions which contains extra insights about human actions without the actual need for an actual presence and thereby can understand the motives of the people. This system can be further enhanced by combining the facial emotion recognition system with the speech emotion recognition system so that the emotions are predicted more accurately by analyzing both facial and audio-extracted features.

APPENDICES

A.1 SAMPLE SCREENS

The below fig. A.1.1 shows the user interface where the user can choose a file and upload them and predict the emotion for the respective audio file.

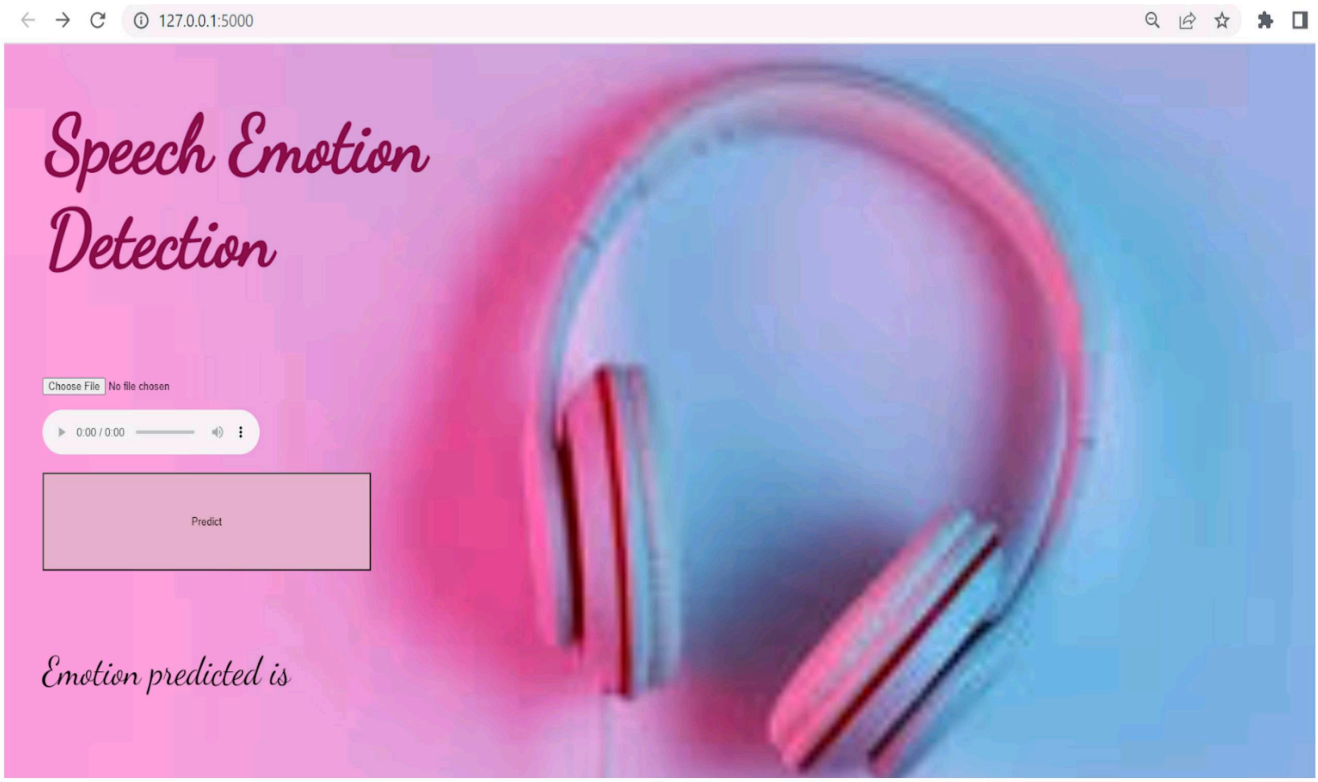

Fig.A.1.1 Website

The below fig. A.1.2 shows the display of the result when the user clicks the predict option and can predict for any audio file in the same manner.

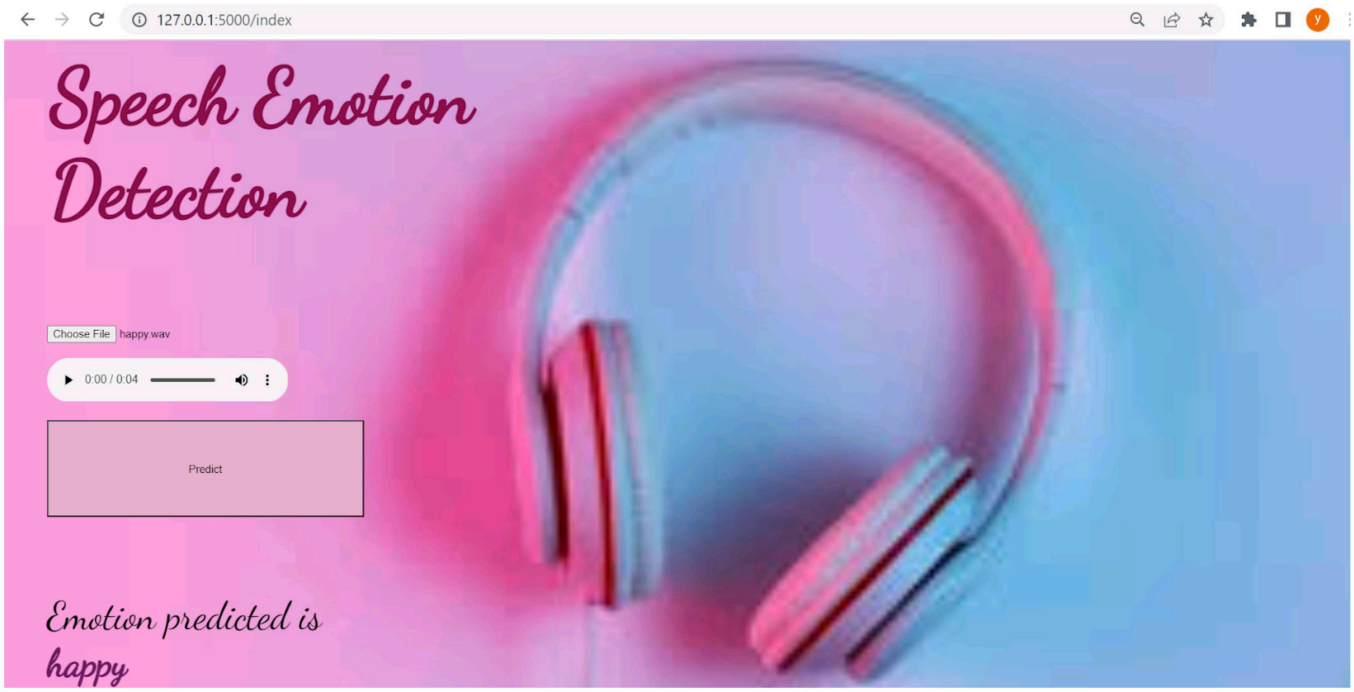

Fig.A.1.2 Prediction of the CNN Model

REFERENCES

[1] Akira Sasou, Bagus Tris Atmaja, and Masato Akagi, "Speech emotion and naturalness recognition system", vol.10, pp.72381-72387, July 2022.

[2] Jennifer Santosoand Takeshi Yamada, "Speech Emotion Recognition system based on Self-Attention Weight Correction for Acoustic and Text Features", vol.10, pp.115732-115743, November 2022.

[3] Liu Yunxiang and Zhang Kexin, “Design of Efficient Speech Emotion Recognition Based on Multi Task Learning”, vol.11, pp.5528-5537, January 2023.

[4] Caslon Chua, Felicia Andayani, Lau Bee Theng and Mark Teekit Tsun, “Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files”, vol.10, pp.36018-36027, March 2022.

[5] Edris Zaman Farsa,Mohammad Reza Falahzadeh,, “3D Convolutional Neural Network for Speech Emotion Recognition With Its Realization on Intel CPU and NVIDIA GPU”, vol.10, pp.112460-112471, October 2022.

[6] Alwin Poulose, Samuel Kakuba,, “Attention-Based Multi-Learning Approach for Speech Emotion Recognition With Dilated Convolution”, vol.10, pp.122302-122313, November 2022.

[7] Alwin Poulose, Samuel Kakuba, “Deep Learning-Based Speech Emotion Recognition Using Multi-Level Fusion of Concurrent Features”, vol.10, pp.125538-125551, November 2022.

[8] Akira Sasou, Bagus Tris Atmaja, “Evaluating Self-Supervised Speech Representations for Speech Emotion Recognition”, vol.10, pp.124396 - 124407, November 2022.

[9] Haiyan Wang, Xiaohui Zhao, “Investigation of the Effect of Increased Dimension Levels in Speech Emotion Recognition”, vol.10, pp.78123-78134, July 2022.

[10]Carlos Busso, Kusha Sridhar,, “Unsupervised Personalization of an Emotion Recognition System: The Unique Properties of the Externalization of Valence in Speech”, vol. 13,pp.1959-1972, December 2022.