IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

INTERACTIVE MUSIC RECOMMENDATION SYSTEM BASED ON TEXT AND SPEECH ANALYSIS

[1] Isha Moykhede, [2]Mosmee Patil, [3] Om Mahajan, [4] Gaurav Mantri, [5] Dr. Preeti Bhamre

[1,2,3,4] Student, Department of Information Technology

[5] Head, Professor, Department of Information Technology

[1,2,3,4,5]K. K. Wagh Institute of Engineering Education and Research, Nashik, India

Abstract: Music is an integral part of our lives. In this era of technological advances, appropriate music recommendations are much needed as soothing music according to moods helps humans to relieve stress. The objective of this project is to develop a personalized system, where the user's current emotion is analyzed with the help of a bot. The bot will interact with user to identify the mood. By analyzing the tone of the text and speech used by the user, mood can be identified. Once the mood is identified, the application will recommend music based on the user's mood. The project utilizes the Rasa framework for chatbot interaction, allowing the user to have a more natural and interactive experience. The deep speech recognition layers are used for recognition of emotions through speech. The model is going to use a Natural Language Processing (NLP) Algorithm for textual analysis and Convolutional Neural Network (CNN) along with spectrogram analysis to predict the mood based on the tone of the user. To recommend music based on the emotions Last.FM API will be used.

Index Terms **- Bot, NLP, Spectrogram, Recommendation System, Speech Emotion Recognition.**

I. INTRODUCTION

Music is a great connector. It unites us across markets, ages, backgrounds, languages, preferences, political leanings and income levels. Music players and other streaming apps have a high demand as these apps can be used anytime, anywhere and can be combined with daily activities, travelling, sports, etc. With the rapid development of mobile networks and digital multimedia technologies, digital music has become the mainstream consumer content sought by many young people. Human emotions can be broadly classified as: fear, disgust, anger, surprise, sad, happy and neutral. A large number of other emotions such as cheerful (which is a variation of happy) and contempt (which is a variation of disgust) can be categorized under this umbrella of emotions. It is a big challenge to provide recommendations from the large data available on the internet. E-commerce giants like Amazon, eBay provide personalized recommendations to users based on their taste and history while companies like Spotify, Pandora use Machine Learning and Deep Learning techniques for providing appropriate recommendations. There has been some work done on to predict the emotions of the user and to recommend songs based on the user's preference.

II. LITERATURE REVIEW

Margaret Lech, ,Melissa Stolar and Robert Bolia [2018] proposed a paper "Amplitude-Frequency Analysis of Emotional Speech Using Transfer Learning and Classification of Spectrogram Images" where the proposed system was used for analytical purpose to determine how different emotional categories are coded into the amplitude-frequency characteristics of emotional speech. This study offers a more computationally efficient approach for finding what parts of the amplitude-frequency plane carry the most important cues for different emotions. Existing state-of-the-art Speech Emotion Recognition (SER) methods apply deep Convolutional Neural Networks (CNNs) trained on a very large number of labelled spectrograms[1].

A spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time. Fig 1 represents spectrograms for angry and happy emotion. The amplitude frequency analysis of the characteristics of emotional speech was achieved by analyzing the SER results given by different forms of spectrogram images representing speech signals. The SER was performed by applying transfer learning to the pre-trained AlexNet with input arrays given as RGB or grey-scale images

depicting speech spectrograms. After a relatively short training (fine-tuning), the trained CNN was ready to infer emotional labels (i.e., recognize emotions) from an unlabelled (streaming) speech using the same process of speech-to-image conversion.

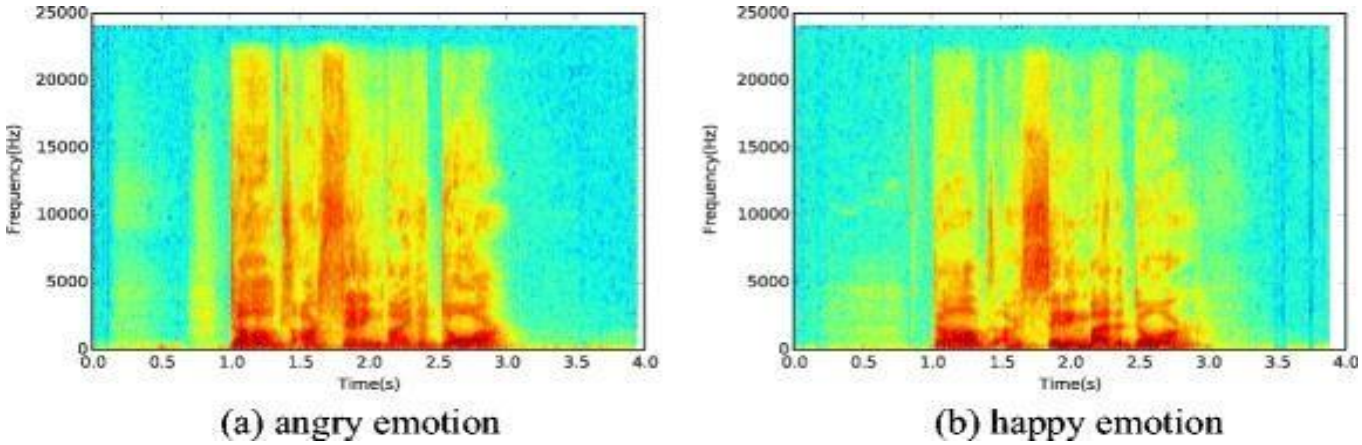

Fig.1. Spectrogram for Angry and Happy Emotion

EMO-DB database. Spectrogram imagesgenerated on different frequency-scales emphasized different frequency ranges of speech signals, while different color components of the RGB images of spectrograms indicated different values of spectral amplitudes. The analysis provided insights into the amplitude-frequency characteristics of emotional speech. Areas of the amplitude-frequency plane containing cues for different emotions were identified. One of the major limitations of this study is that, thefindings apply to acted emotional speech, and only one language (German) was tested.

Shivam Sakore, Pratik Jagdale, Mansi Borawake [2021] proposed "Music Recommender System Using Chatbot" which aims at recommend music based on the user's text. The main goal is to determine user mood by recognizing emotion from text. In this paper IBM tone analyser is used where, its service uses linguistic analysis to detect emotional and language tones in written text. The service can analyze tone at both the document and sentence levels. This helps to detect communication tones in written text. The IBM Emotional API is used to analyze the conversation's emotional content. The Cake chat Chatbot will reply to your con- versation. The Last.fm API is used to recommend music based on emotion. By utilizingthis API, the chatbot retrieves the top music based on the emotion it perceives .By using Support Vector Machine (SVM), the tone of new texts based on conversations was predicted. Around 30% of samples are associated with more than one tone, so the multi- class classification is not suitable. The model was trained independently for each tone usingthe One-vs-Rest paradigm. Final tones were determined by identifying those predicted withat least a 0.5 probability[2].

III. METHODOLOGY

The proposed system recommends music that will help user to relieve stress through text and speech interaction. The model will recognize the user's mood with the responses they give to the system and music playlist will be recommended accordingly. The development of personalized system is proposed to be done by using Artificial Intelligence, Natural Language Processing (NLP) and Convolutional Neural Network (CNN).

There are two parts : Training and Testing.

1. Text Processing -The dataset for text processing is to be trained by using NLP algorithm
2. Generation of Spectrogram Images -The audio signals will be converted into spectrogram images for emotion recognition.
3. Speech Emotion Recognition -The dataset for speech emotion recognition is to be trained by using Convolutional Neural Network (CNN) algorithm.

3.1 Rasa Framework

Rasa is an open-source framework for building conversational AI chatbots, virtual assistants, and contextual AI assistants which provides tools for natural language processing (NLP), dialogue management, and machine learning. Intents and stories are two key components of the Rasa framework used for building conversational AI chatbots and virtual assistants. Intents represent the intention behind a user's message. Stories, on the other hand, represent a sequence of steps or actions that a chatbot should take to carry out a particular task or conversation.

3.2 Natural Language Processing (NLP)

Sentiment Analysis is a procedure used to determine if text is positive, negative or neutral. In text analysis, Natural Language Processing (NLP) and machine learning (ML) techniques are combined to assign sentiment scores to the text. The chat bot identifies the user's sentiment by chat conversation with the user. Based on the input provided by the user, current emotion or mood is analyzed by the chat bot and it will suggest music playlist. to the user[3].The input for the process is the set of text words provided by the user and the output contain mood and music playlist which is analyzed based on the input provided by user.

The process is as follows:-

1. NLP model is utilized in the automatic text classification.
2. The first step is to pre-process and organize the data.
3. The proposed study segments the text by words and tokenize words.
4. Feature Selection is performed so that the classifiers utilize only the most discerningfeatures.
5. The Support Vector Machine (SVM) is used as classifier to classify different emotional states.

3.3 Generation Of Spectrogram Images

The amplitude-frequency analysis of the characteristics of emotions in speech will be achieved by analyzing the Speech Emotion Recognition (SER) results given by different forms of spectrogram images representing speech signals. The SER is performed by applying transfer learning to the pre-trained AlexNet with input arrays given as RGB or grey-scale images depicting speech spectrograms[1].

Four different frequency scales(linear, Melodic (Mel),equivalent rectangular bandwidth (ERB) and logarithmic (log)) were applied when generating the spectrogram images in order to visually emphasize different frequency ranges. The most important frequencies for speech and language are between 250 and 8,000 Hz. The alternative frequency scales were applied along the vertical axis of the spectrograms from 0 to 8 kHz, the horizontal time scale was in all cases linear spanning the time range of 0 to 1 second. RGB images of spectrograms used in the speech classification gave visual representations of the time-frequency decomposition of speech signals. The maximal intensity of the red hue in the R component served to draw attention to the high-spectral-amplitude spectral components of speech, such as vowels and voiced consonants. The B component generated the brightest blue hue at lower amplitudes, accentuating speech pauses and low-amplitude spectral features (such unvoiced consonants). The G component similarly emphasised the finer intricacies of the mid-range spectral amplitude components.

3.3.1 Spectrogram Analysis.

The speech to image trans- formation was achieved by calculating amplitude spectrograms of speech and transforming them into RGB images. This approach is commonly used to visualize spectrograms; however, in these cases, the aim was to create a set of images to perform the fine-tuning of a pre-trained deep convolutional neural network learning. Since the majority of existing pretrained networks have been created for image classification. The labeled speech samples were buffered into short-time blocks. For each block, a spectral amplitude spectrogram array was calculated, converted into an RGB image format, and passed as an input to the pretrained CNN. The trained CNN was prepared to infer emotional labels (i.e., recognize emotions) from an unlabeled (streaming) speech after only a brief training (fine-tuning). [1].

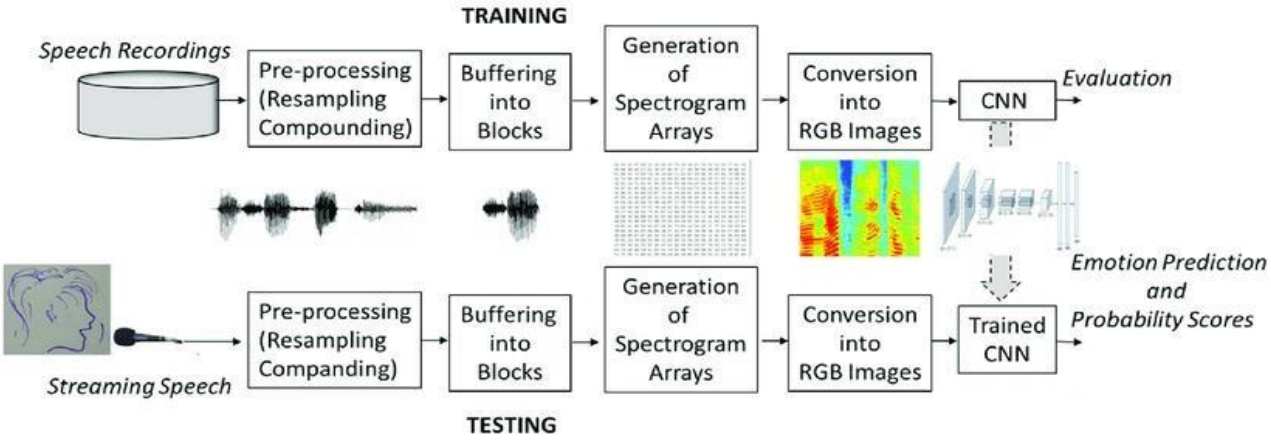

Fig.2. Overview Of Spectrogram Analysis

The table depicted below shows the best frequency Scale and best frequency range of different of 6 different emotions. Following table represents the Amplitude-frequency characteristics of categorical emotions.

Table-1: Amplitude-Frequency Characteristics of Categorical Emotions.

Emotion	Best Freq. Scale	Best Freq. Range	Amplitude Range
Anger	ERB	Medium-Low	High
Disgust	LOG	Low (0.02-2kHz)	All
Joy	ERB	Medium-Low (2-4kHz)	High
Fear	MEL	Medium-High(2-4kHz)	All
Sad	MEL	Medium-High(2-4kHz)	All
Neutral	ERB	Medium-High(2-4kHz)	All

3.4 Convolutional Neural Network (CNN)

A CNN is a deep learning algorithm which takes an input image, assigns importance (learnable weights and biases) to various objects in the image and is able to differentiate between images. One role of a CNN is to reduce images into a form which is easier to process without losing features that are critical for good prediction[1]. CNN has two parts i . e. , Feature extraction and classification. It contains five layers : Input Layer, Convolutional Layer, Pooling Layer, Fully Connected Layer, Output Layer.

3.5 Block Diagram

It is a multimodal bot that accepts input via speech and text. A user may send the bot both text and audio messages. The message is sent from the web client to a Natural Language Processing (NLP) service. Here, NLP is used to categories the data into good, negative, and neutral categories. Processed information is compared to user emotion and purpose. When specific emotions are recognized, a structured response with useful information will be sent. The user will receive the relevant actionable data directly. Calls an external API to produce music playlist recommendations based on the song choices. Fig.3 shows the block diagram.

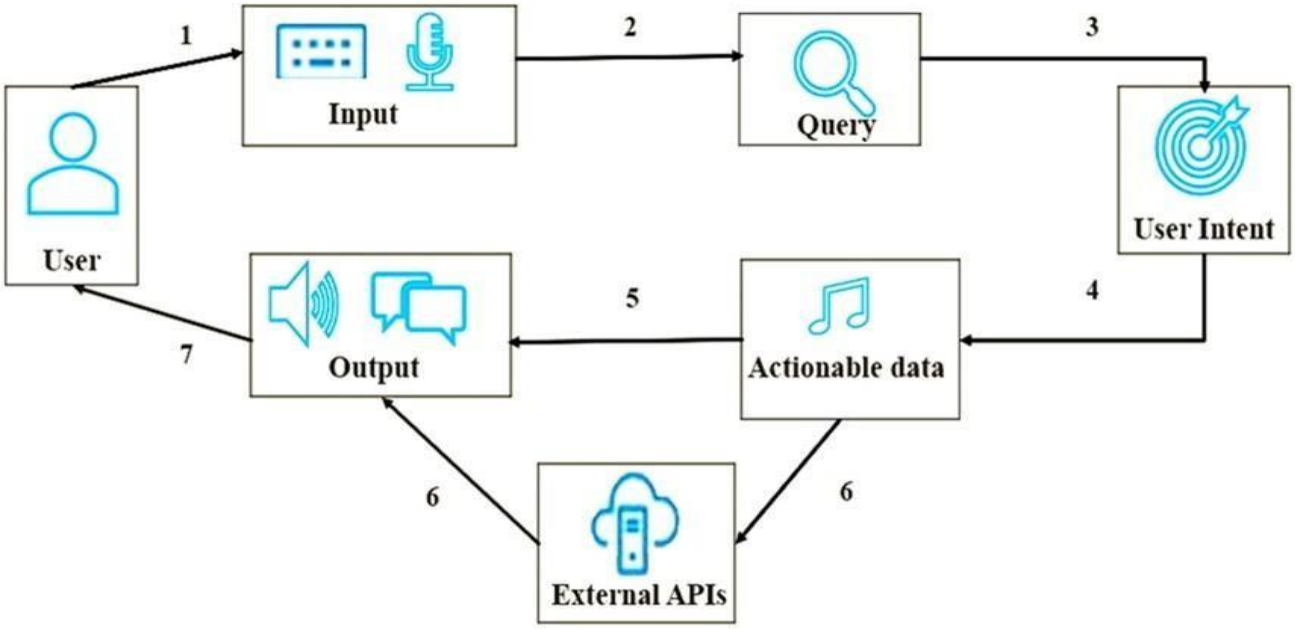

Fig.3. Block Diagram

IV. Result Analysis

4.1 Textual Conversation

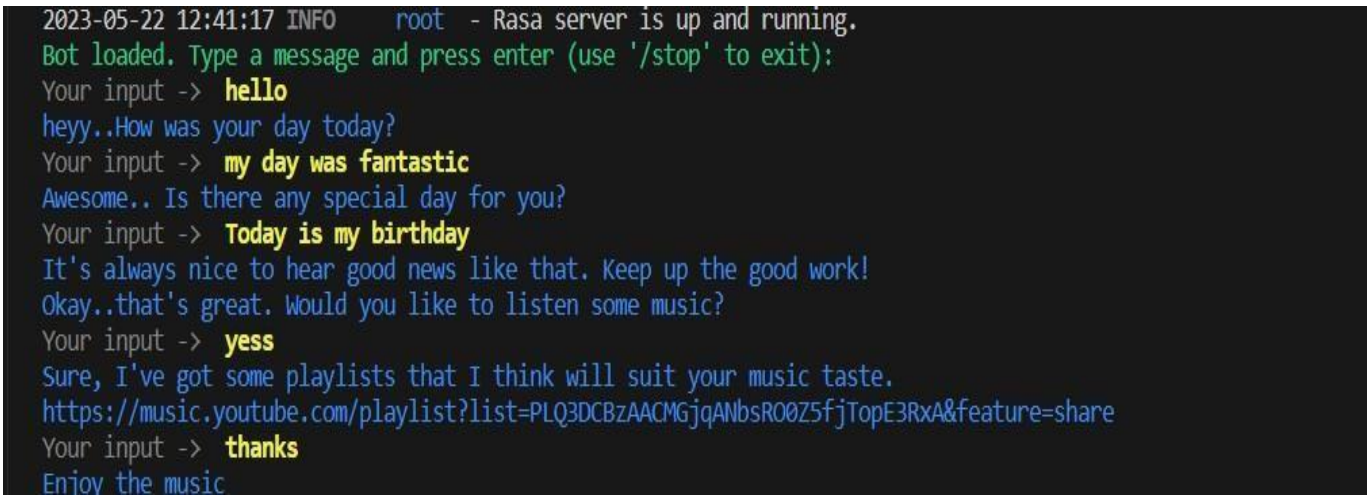

Fig.4. Text Conversation For Happy Intent On Rasa Server

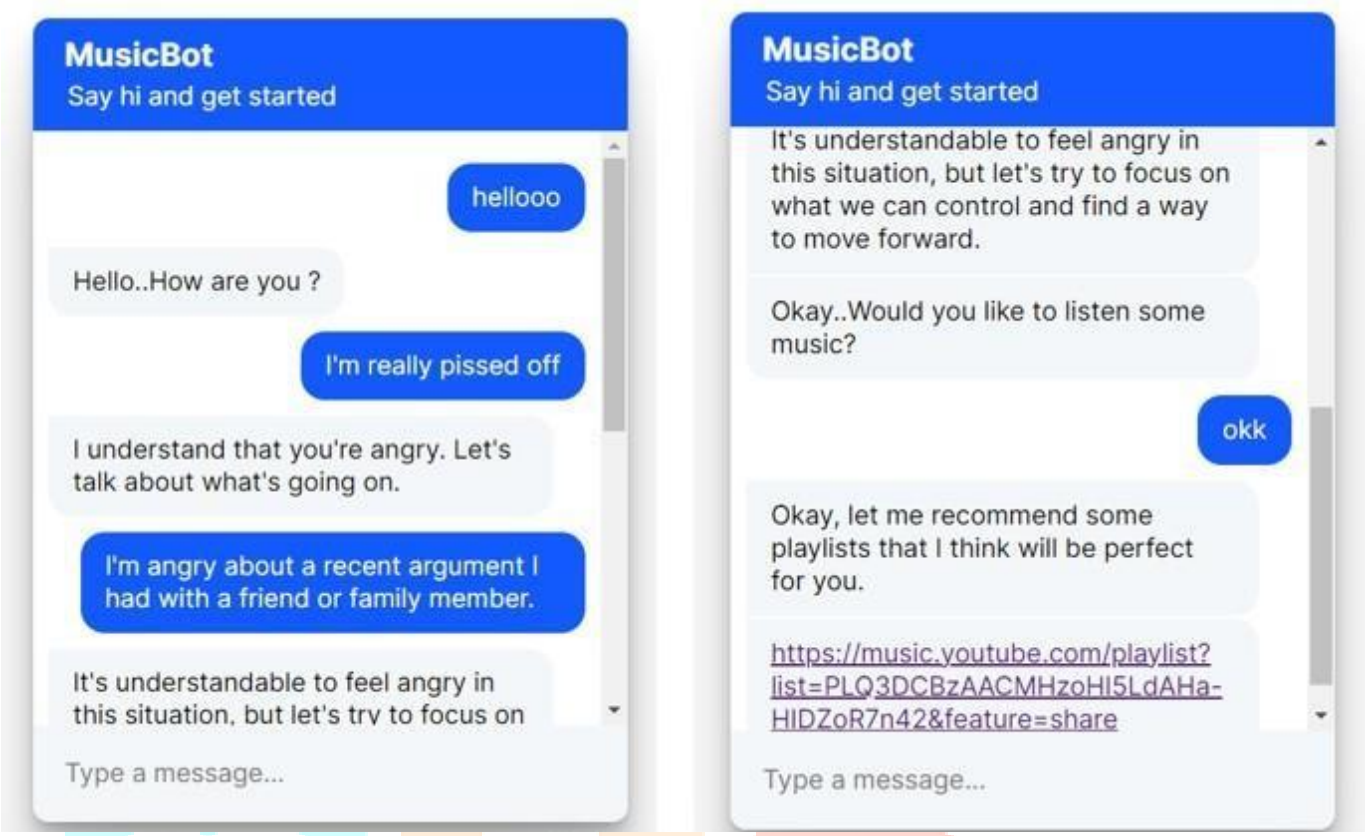

Fig.5. Conversation With Angry Intent

4.2 Audio

Fig.6. depicts importing and displaying an audio file. Here are several spectrogram creation examples for various emotions such as angry, calm, etc.

```
import librosa.display
from IPython.display import Audio
import numpy as np
import tensorflow as tf
from matplotlib.pyplot import specgram
import pandas as pd
IPython.display.Audio("C:\\Users\\Mosmee\\Desktop\\speech\\om angry2.mp3")
```

Fig.6. Audio File

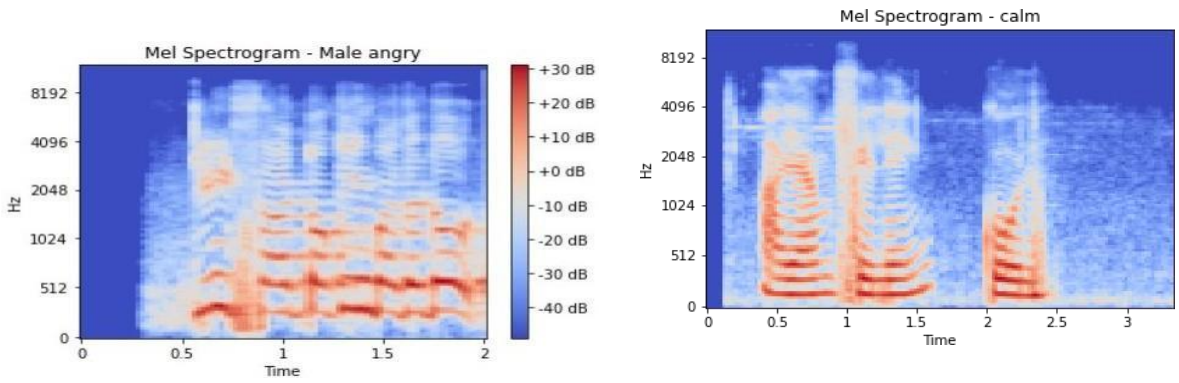

Fig.6. Spectrogram Of Angry And Calm Emotion

The Convolutional Neural Network (CNN) model trained on the RAVDESS dataset achieved an accuracy of 88.12% using a 70% training and 30% testing data split. Fig.7. represents accuracy graph for trained speech model.

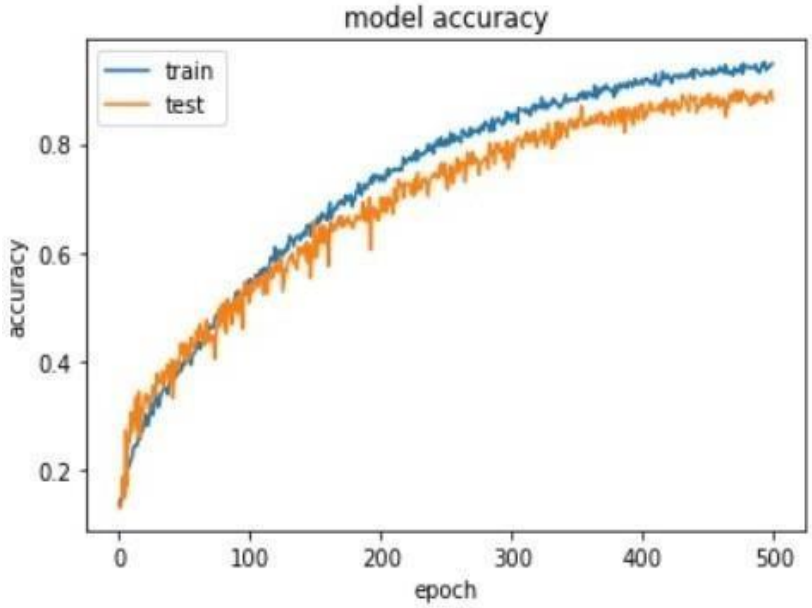

Fig.7. Accuracy Graph

V. CONCLUSION

Music is an essential component of daily living. It can reduce tension, discomfort, struggle, and distraction while also brings happiness and peace into our lives. With these considerations in mind, an interactive system is being developed that will recommend a music playlist based on the user's feelings. The goal is to give users with a platform where they may interact with the system to relieve tension and receive a playlist that can cheer them up depending on the emotions retrieved by the system during the interaction. Text and vocal inputs were used to engage with the model. Natural Language Processing (NLP) is used for text analysis, whereas Convolutional Neural Network (CNN) is utilized for speech analysis. Using a 70% training and 30% testing data split, the Convolutional Neural Network (CNN) model trained on the RAVDESS dataset attained an accuracy of 88.12%. Human emotions are thought to be extremely difficult to comprehend. When working with speech inputs, the model occasionally misclassifies emotions such as neutral and disgust. Similarly, if the user enters ambiguous or confusing terms during textual chats, the model occasionally struggles to appropriately forecast the intended sentiment. Using several machine learning algorithms, the system attempts to demonstrate that it is possible for a machine to predict the user's emotions.

REFERENCES

[1] Margaret Lech, ,Melissa Stolar, Robert Bolia," Amplitude-Frequency Analysis of Emotional Speech Using Transfer Learning and Classification of Spectrogram Images",*ASTESJ* ISSN: 2415-6698, Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 4, 363-371 (2018)

[2] Shivam Sakore1, Pratik Jagdale2, Mansi Borawake3, *IJRASET*Music Recommender System Using ChatBot ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538,Volume 9 Issue XII Dec 2021.

[3] Prof. Suvarna Bahir, Amaan Shaikh, Bhushan Patil,"Chat Bot Song Recommender System", *IRJMETS*,e-ISSN: 2582-5208,Volume:04/Issue:04/April-2022

[4] Amrita Nair, Smriti Pillai, Ganga S Nair, Anjali T "Emotion Based Music Playlist Recommendation System using Interactive Chatbot *IEEE Xplore-2021* Part Number: CFP21AWO-ART; ISBN: 978-0-7381- 1405-7.