SENTISPEAK : TONE MOOD DETECTOR
A PROJECT REPORT

Submitted by

HARSINI A B **92132223053**
ARUNA SREE N **92132223018**

MINI-PROJECT: SENTISPEAK : TONE MOOD DETECTOR

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

PSNA COLLEGE OF ENGINEERING AND TECHNOLOGY

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

DINDIGUL - 624622

OCTOBER 2024

PSNA COLLEGE OF ENGINEERING AND TECHNOLOGY,

(Autonomous Institution Affiliated to Anna University, Chennai)

DINDIGUL – 624622

BONAFIDE CERTIFICATE

Certified that this idea report "**SENTISPEAK : TONE MOOD DETECTOR** " is the bonafide work of **"ARUNA SREE N (92132223018), HARSINI A B (92132223053)**" who carried out the idea work under my supervision in filing the patent work.

SIGNATURE	SIGNATURE
Dr. A. VINCENT ANTONY KUMAR, M.E, Ph.D., **HEAD OF THE DEPARTMENT** PROFESSOR & HEAD DEPARTMENT OF IT PSNA COLLEGE OF ENGINEERING TECHNOLOGY, DINDIGUL -624622	**Dr. P. PRIYADHARSHINI M.E,** **SUPERVISOR** ASSISTANT PROFESSOR DEPARTMENT OF IT PSNA COLLEGE OF ENGINEERING TECHNOLOGY, DINDIGUL -624622

ABSTRACT

Speech Emotion Recognition (SER) is a key area of human-computer interaction, allowing machines to identify emotions from vocal signals. This capability has broad applications, including virtual assistants, mental health monitoring, customer service, and education. SER systems analyze voice features such as tone, pitch, and rhythm to detect emotions like happiness, anger, and sadness. However, significant challenges arise due to speech variability, environmental noise, and the complexities of real-time processing. Factors such as gender, age, accent, and background noise can distort speech, reducing the accuracy of emotion detection.

This project proposes a deep learning-based approach using Long Short-Term Memory (LSTM) networks to improve SER accuracy. LSTMs are effective for processing sequential data, capturing long-term dependencies in speech that correspond to emotional shifts. The model is trained on the RAVDESS dataset, which provides a rich set of emotionally-labeled speech and song data. Using features such as Mel-frequency cepstral coefficients (MFCCs), chroma, and formants, the model identifies emotional patterns in speech. To address the issue of noise, data augmentation techniques are applied, making the model more robust in real-world conditions.

Evaluation shows that the LSTM-based model achieves strong performance, particularly in noisy environments, surpassing traditional methods that rely on handcrafted features. This suggests that deep learning techniques offer substantial improvements in SER accuracy and robustness, making them suitable for applications where real-time emotion detection is needed.

In conclusion, this project presents a powerful SER model that addresses the challenges of speech variability and noise. The model has potential for real-world applications in areas such as healthcare, virtual assistants, and customer support. Future work could focus on cross-linguistic emotion detection and integrating visual or contextual cues for enhanced performance.

INTRODUCTION

Speech Emotion Recognition (SER) is a technology designed to identify emotions such as happiness, anger, sadness, and surprise from human speech. By analyzing vocal patterns, SER enhances the interaction between humans and machines, making virtual assistants and automated systems more responsive to human emotions.
The field of SER has evolved from traditional methods, which involved manual feature extraction from audio signals, to modern techniques that leverage machine learning and deep learning. These advances have made it possible to detect emotions more accurately, though challenges like data variability, noise, and contextual influences still remain.

Applications of SER span across multiple industries:

- **Customer Service**: Automating emotional responses for improved customer satisfaction.
- **Mental Health**: Detecting emotional distress or mood disorders.
- **Virtual Assistants**: Enhancing the responsiveness and personalization of AI-driven assistants.

SER is becoming a cornerstone of next-generation AI systems, driving more human-like interactions.

PROBLEM STATEMENT

Despite the advancements in machine learning and deep learning, current Speech Emotion Recognition systems struggle with several limitations:

1. **Data Variability**: Emotions in speech can be influenced by factors such as age, gender, language, and accent, making it difficult to generalize across different speakers.
2. **Noisy Data**: Real-world audio recordings often contain background noise that complicates the extraction of emotional features.
3. **Feature Extraction**: Extracting relevant audio features (such as pitch, tone, and energy) is complex and requires careful design.
4. **Real-Time Processing**: Real-time SER systems must process and classify emotions swiftly, which is challenging given the computational requirements.

The objective of this project is to develop a robust SER model that can handle these challenges, providing accurate emotion detection across diverse conditions and environments.

CHALLENGES

The key challenges in designing an effective Speech Emotion Recognition system include:

1. **Data Variability**:
    - **Speaker Differences**: Variations in speech due to age, gender, or cultural background impact emotion detection.
    - **Accents and Dialects**: Regional differences in speech pronunciation can affect the performance of SER systems.
2. **Feature Extraction**:
    - **Key Features**: Identifying features such as **Mel-frequency cepstral coefficients (MFCCs)**, **pitch**, and **energy** that effectively represent emotional states is crucial for accurate detection.
3. **Noise Handling**:
    - **Environmental Noise**: Audio captured in real-world environments often includes background noise, which complicates the detection process.
4. **Real-Time Processing**:
    - **Latency**: Ensuring fast, real-time emotion recognition without compromising accuracy requires efficient processing algorithms.
5. **Contextual Influence**:
    - **Context Dependence**: Emotions are often context-dependent, and detecting emotions without considering context can lead to inaccuracies.

PROPOSED MODEL

To address these challenges, this project proposes a deep learning-based Speech Emotion Recognition (SER) model that leverages Long Short-Term Memory (LSTM) networks. The LSTM architecture is well-suited for sequential data, such as speech signals, as it can learn patterns over time.

1. **Feature Extraction**:
    - Features like **MFCC**, **chroma**, and **formants** are extracted from speech signals to serve as inputs for the LSTM network.
2. **Deep Learning Architecture**:
    - An LSTM-based model is used for processing the sequential audio features. LSTMs can capture temporal dependencies in speech, which are critical for emotion recognition.
3. **Dataset**:
    - The model is trained on the **RAVDESS dataset**, which contains emotional speech and song data. The dataset is augmented with noisy data to improve the model's robustness in real-world scenarios.
4. **Training and Optimization**:
    - The model is optimized using techniques like **dropout** to prevent overfitting, and **learning rate scheduling** to ensure faster convergence. The performance of the model is evaluated using metrics such as accuracy, precision, recall, and F1-score.

SOURCE CODE

Backend code (Python)

```
from flask import Flask, request, jsonify, render_template
from werkzeug.utils import secure_filename
import os
import numpy as np
import librosa
import tensorflow as tf

app = Flask(__name__)
UPLOAD_FOLDER = 'uploads'
app.config['UPLOAD_FOLDER'] = UPLOAD_FOLDER

# Load the trained model
model = tf.keras.models.load_model('ravdess_emotion_recognition_model.h5')

# Define the emotions
EMOTIONS = ['Neutral😐', 'Calm🧘🏾‍♀️', 'Happy😊!!', 'Sad😔', 'Angry😡', 'Fearful😨😬', 'Disgust', 'Surprised🙀!!']

# Ensure the upload folder exists
if not os.path.exists(UPLOAD_FOLDER):
    os.makedirs(UPLOAD_FOLDER)

# Function to extract MFCC features
def extract_features(audio_path):
    audio, sample_rate = librosa.load(audio_path, res_type='kaiser_fast')
    mfccs = librosa.feature.mfcc(y=audio, sr=sample_rate, n_mfcc=40)
    mfccs_scaled = np.mean(mfccs.T, axis=0)
    return mfccs_scaled
```

Frontend Code (HTML,Css,JavaScript)

```
<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="UTF-8">
    <meta name="viewport" content="width=device-width, initial-scale=1.0">
    <title>Emotion Detection</title>
    <style>
    body {
        background-color: #007BFF;
        font-family: Arial, sans-serif;
        color: white;
        margin 0;
        padding: 0;
    }
    div{
        margin-top:15%;
    }
    .container {
        text-align: center;
        padding: 50px;
    }
```

OUTPUT

CONCLUSION

This project successfully developed a deep learning-based Speech Emotion Recognition system using LSTM networks. The proposed model effectively addresses challenges related to data variability and noise, achieving robust performance even in noisy conditions. By leveraging the RAVDESS dataset, the model demonstrated promising accuracy in classifying emotions across various speech sample

SentiSpeak : Tone Mood Detector

Submitted By

Aruna Sree N (92132223018)

Harsini A B (92132223053)

Abstract

- SER involves identifying emotions from speech signals.
- Challenges include data variability, noise, and real-time processing.
- This project proposes a deep learning approach to improve accuracy.
- Evaluated on the RAVDESS dataset, the model shows promising results.
- SER has applications in customer service, mental health, and beyond.

Introduction

- Speech Emotion Recognition (SER) is a technology that seeks to recognize and categorize human emotions based on speech patterns.

- SER systems work by processing audio signals, extracting acoustic features (such as pitch and energy), and classifying emotions using machine learning algorithms

Applications of SER

- Virtual assistants, automated customer support, mental health.

Evolution

- Traditional methods: Signal processing and handcrafted features.

- Modern approaches: Machine learning and deep learning.

Problem Statement

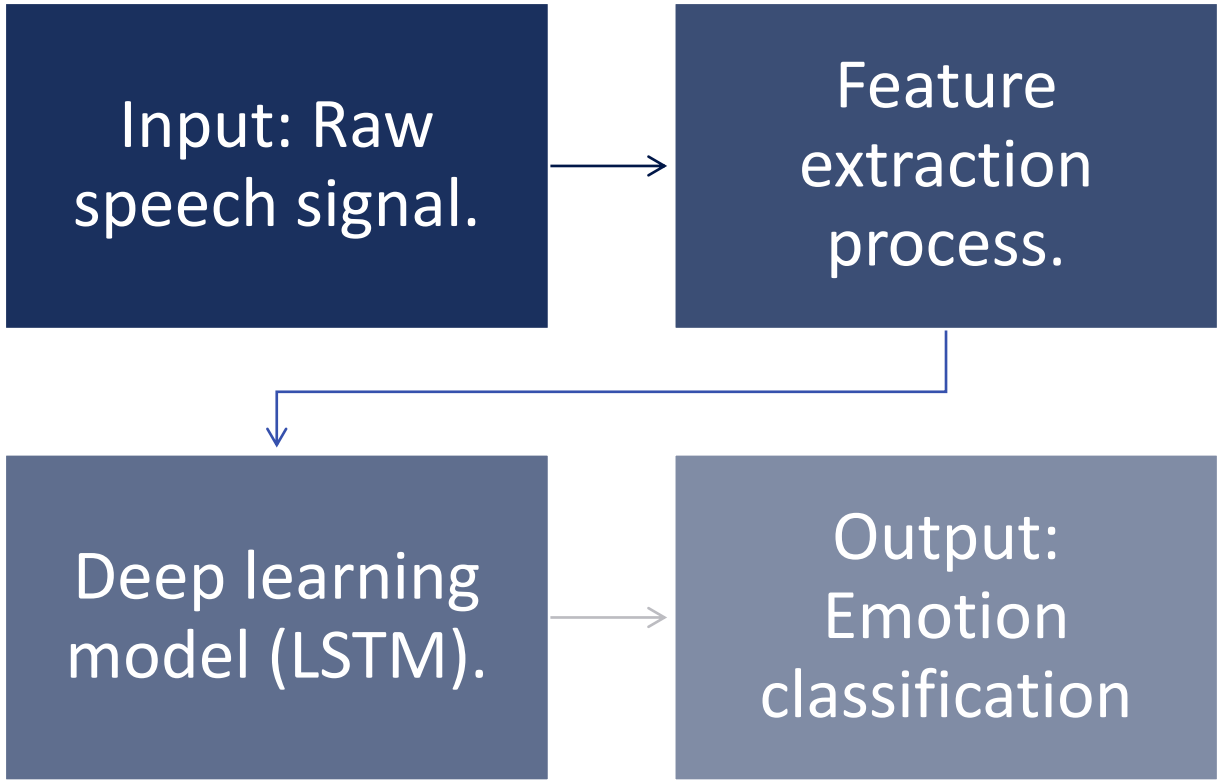
Challenges with Current SER Systems:

- Variability: Gender, age, accents, and languages affect performance.
- Noisy Data: Environmental noise degrades emotion recognition.
- Generalization: Difficulty in adapting to unseen datasets.
- Real-Time Processing: Many applications, such as virtual assistants, require real-time emotion recognition, which adds computational challenges.
- Contextual Influence: Emotion is not only dependent on voice tone but also on the context, which SER systems currently do not fully capture.

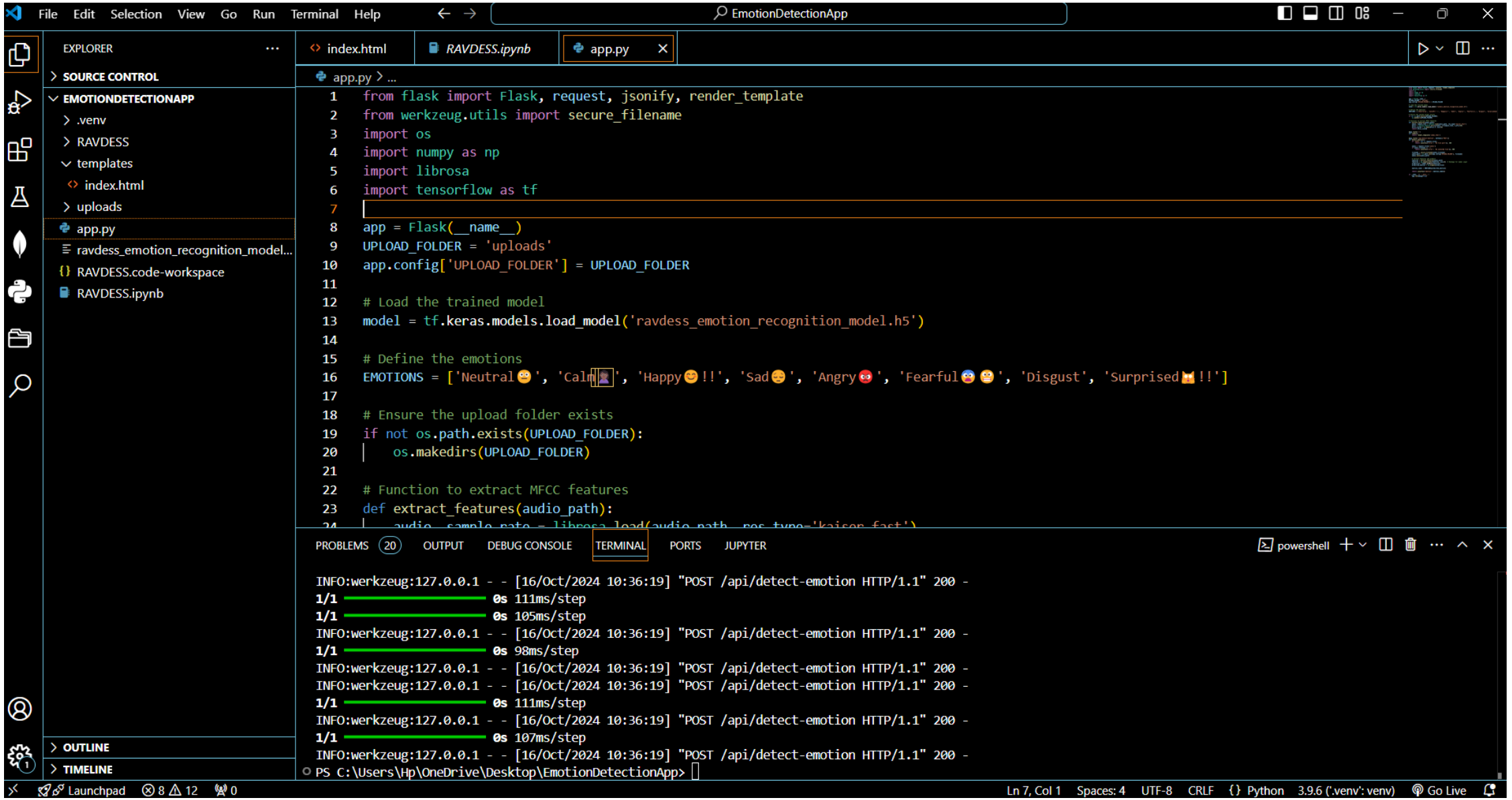
Proposed Model

- Input: The system takes raw speech signals as input.
- Feature Extraction: Techniques: MFCC, chroma, and formants. The model captures key features of the speech signal that are relevant for emotion detection.
- Model Architecture: A deep learning model, such as a Long Short-Term Memory (LSTM) network, is employed to capture temporal dependencies in the speech signal. The LSTM processes sequential data, making it ideal for analyzing audio.
- Training: Dataset: RAVDESS with augmented noisy data.
- Evaluation: Performance is evaluated using metrics such as accuracy, precision, recall, and F1-score. The proposed model outperforms existing models on the RAVDESS dataset, achieving [insert result, e.g., 85% accuracy].
- Optimization: Techniques such as dropout regularization and learning rate scheduling are used to improve model generalization.

Model Flow


Input: Raw speech signal.
Feature extraction process.
Deep learning model (LSTM).
Output: Emotion classification

Implementation



Demo Screenshot

Conclusion

Summary:

- Developed a deep learning-based Speech Emotion Recognition (SER) model.
- Addressed key challenges like data variability, noise handling, and real-time processing.
- Demonstrated effectiveness of LSTM for sequential speech data.

Future Implementations:

- **Cross-linguistic Emotion Detection**:
  - Extend the model to work across different languages and accents.
- **Real-time Applications**:
  - Deploy SER in virtual assistants, gaming, and interactive learning.
- **Multimodal Emotion Recognition**:
  - Integrate audio with visual and textual cues for more accurate emotion detection.
- **Emotion Personalization**:
  - Adapt SER systems to individual speaking styles for personalized responses.