Speech Emotion Recognition using Python

Project Report

Author: Gorle Sai Kavya (22EC10033)

Duration: Oct'24 – Nov'24

Department of Electronics & Electrical Communication Engineering
Indian Institute of Technology, Kharagpur

August 24, 2025

Contents

Chapter 1

Introduction

Speech Emotion Recognition (SER) is the process of recognizing human emotions and affective states from speech. Voice often reflects underlying emotions through tone, pitch, and rhythm. This project demonstrates how machine learning models can be trained to identify emotions such as calm, happy, fearful, and disgust from audio signals using Python libraries.

Chapter 2

Objectives

- Recognize emotions from human speech using feature extraction and classification techniques.
- Employ `librosa`, `soundfile`, and `scikit-learn` libraries for analysis.
- Train and evaluate a machine learning model (MLPClassifier) on the RAVDESS dataset.
- Demonstrate the utility of SER in real-world applications such as call centers, chatbots, and smart assistants.

Chapter 3

Tools and Libraries

- **Librosa:** Python library for audio and music analysis, used to extract features such as MFCC, Chroma, and Mel spectrogram.
- **Soundfile:** For reading audio files.
- **Scikit-learn:** For machine learning tasks including training and evaluation with MLPClassifier.
- **JupyterLab:** Development environment for running and visualizing experiments.

Chapter 4

Methodology

4.1 Dataset

The project used the RAVDESS dataset (Ryerson Audio-Visual Database of Emotional Speech and Song), which contains 7356 audio files from 24 actors expressing multiple emotions.

4.2 Feature Extraction

Features such as MFCC (Mel Frequency Cepstral Coefficients), Chroma, and Mel spectrogram were extracted using the `librosa` library. These features capture both spectral and temporal characteristics of speech.

4.3 Data Preparation

The dataset was preprocessed by mapping filenames to emotion labels, selecting four target emotions (calm, happy, fearful, disgust), and splitting into training and testing sets.

4.4 Model Training

An **MLPClassifier** was trained on the extracted features. Hyperparameters included:

- Hidden layers: (300,)
- Batch size: 256
- Learning rate: Adaptive
- Max iterations: 500

Chapter 5

Results

- Extracted features from audio files successfully captured emotional variations.
- Training and testing datasets were split 75:25 for evaluation.
- The model achieved an accuracy of approximately **XX%** (replace with actual value from your run).

Chapter 6

Discussion

The results validate the effectiveness of dynamic features like MFCCs and Chroma in capturing speech emotion. Increasing the number of features and optimizing the classifier can further improve performance. The system demonstrates practical potential for integration into emotion-aware interfaces.

Chapter 7

Conclusion

This project successfully implemented a Speech Emotion Recognition system using Python. By leveraging audio features and training a neural classifier, the system was able to identify multiple emotions in speech. The methodology demonstrates applicability to real-world problems in affective computing and human-computer interaction.

Chapter 8

Code Repository

The complete source code is available at: `https://github.com/SaiKavya150705/Speech-Emotion-Rec`