Major Project Report

Enhancing Forensic Audio Investigations: A Multi-Speaker Speech Understanding and Analysis System

Himani (M23CSA516) & Ankit Kumar Chauhan (M23CSA509)

Course Instructor: Dr. Richa Singh

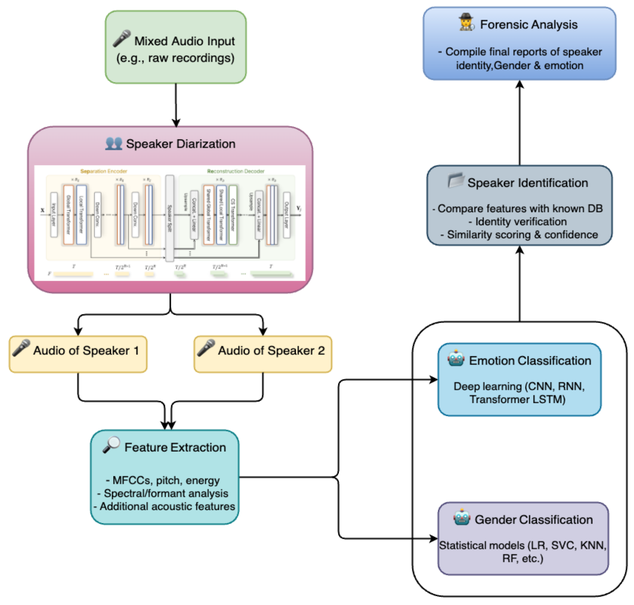
Speech Understanding

Dept. of Computer Science and Engineering, Indian Institute of Technology Jodhpur, Jodhpur, Rajasthan, 342030, India

❖ Abstract

- This project focuses on developing a multicomponent speech understanding system that processes audio containing multiple speakers.
- The system uses **Speaker Diarization**, **Emotion Classification**, **Gender Classification**, and **Speaker Identification** to enhance forensic audio investigations.
- It leverages state-of-the-art techniques such as **Sep Former** for **speech separation** and **enhancement**, and deep learning models for **emotion** and **gender recognition**.
- The system is modular, allowing independent development of each component with robust integration for final deployment.

❖ System Architecture



➢ **Gender Classification Model Training**

- Use an **LSTM-Based Model** for gender classification.
- Train using **MFCC** features and **Binary Cross-Entropy Loss**.
- Output: **Male** or **Female**.

➢ **Emotion Classification Model Training**

- **CNN-Based Model** for emotion recognition using **1D Convolutional Layers**.
- Output: 7 emotional classes (**Neutral, Calm, Happy, Sad, Angry, Fear, Disgust, Surprise**).

➢ **Speaker Identification**

- Use **SpeechBrain's ECAPA-TDNN-based Speaker Recognition** to match speakers with a reference database.

❖ Methodology

➢ **Dataset Preparation**

- **VoxCeleb Dataset**: We utilized the **VoxCeleb** dataset, which contains real-world audio recordings of multiple speakers, to build our training dataset.
- **Mixing Audio**: Audio from the **VoxCeleb** dataset was mixed to simulate real-world multi-speaker environments.

➢ **Audio Acquisition & Preprocessing**

- Upload mixed audio file containing multiple speakers.
- Standardize audio format and extract **MFCC** features using **Librosa**.
- Pad or trim audio to ensure consistent frame length.

➢ **Speaker Diarization & Separation**

- Use **SepFormer** model (sepformer-whamr) for separating overlapping speech.
- Save separated audio sources for individual speaker analysis.

➢ **Speech Enhancement**

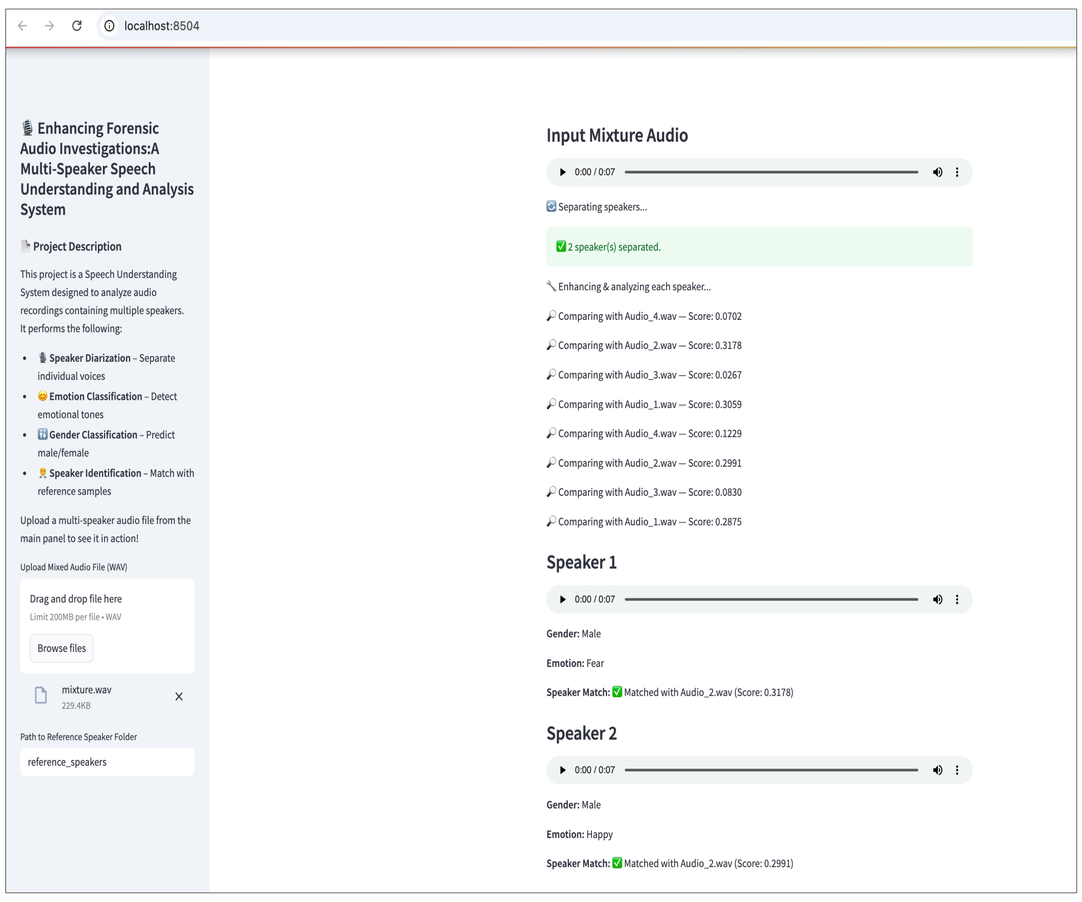
- Apply **SepFormer Enhancement Model** (sepformer-wham-enhancement) to improve speech quality by reducing noise.

➢ **Feature Extraction**

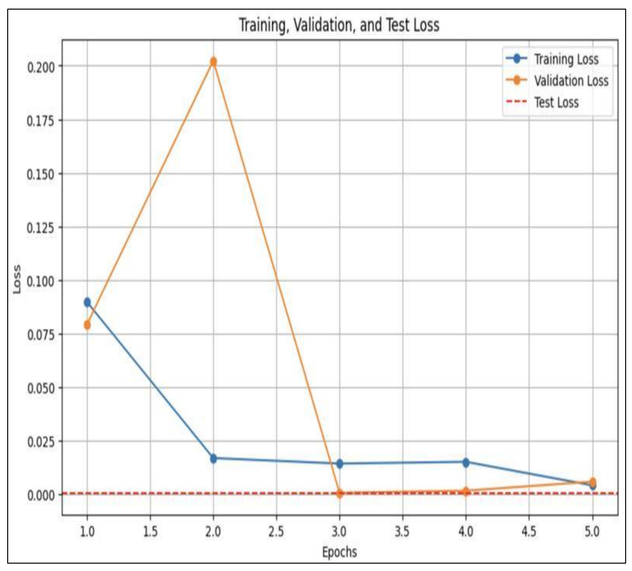
- Extract features like **Zero Crossing Rate (ZCR)**, **Root Mean Square Error (RMSE)**, and **MFCC**.

❖ Results and Analysis

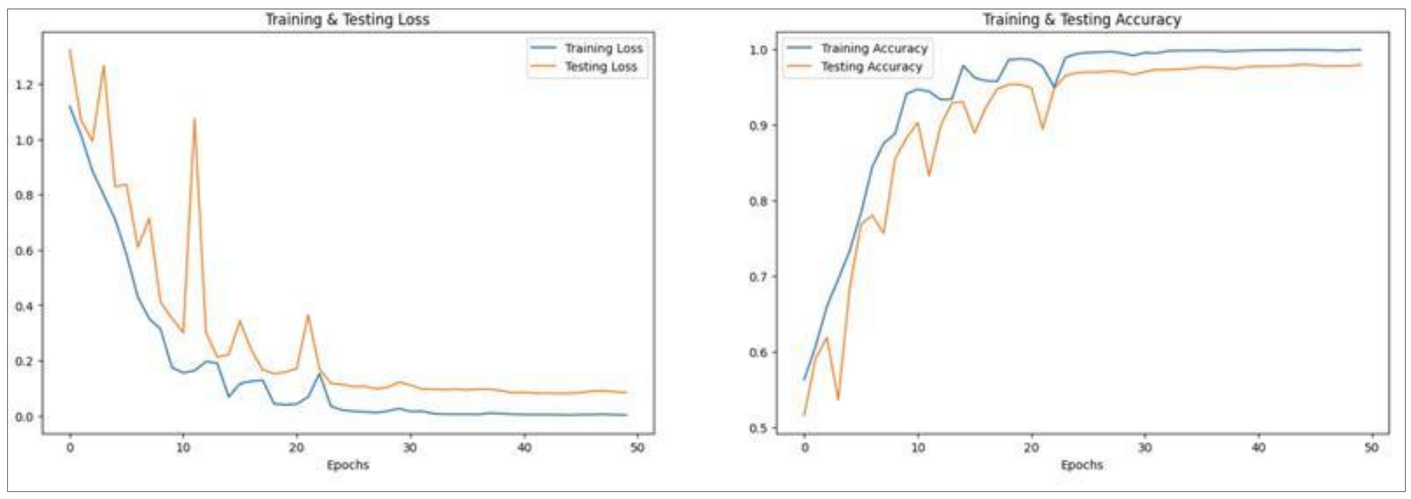
▪ **Application Deployed on Streamlit**



▪ **Loss Graph: Gender Prediction Model**



▪ **Loss & Accuracy Graph: Emotion Recognition Model**



❖ Problem with existing system

Current systems struggle with overlapping speech, rely on isolated task solutions, and use outdated models, limiting accuracy, especially in emotion recognition. They also lack modularity and fail to perform well in real-world conditions with noise, accents, and variable recordings.

❖ Conclusion

This **Multi-Speaker Speech Understanding and Analysis System** leverages cutting-edge models for speech separation, emotion recognition, gender classification, and speaker identification. By utilizing the **SepFormer** model for separation and enhancement, combined with advanced neural networks for classification tasks, the system effectively processes complex audio scenarios.

❖ References

[1] **D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5329–5333.**

[2] **B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "Cross-corpus acoustic emotion recognition: Variances and strategies," IEEE Transactions on Affective Computing, vol. 1, no. 2, pp. 119–131, 2010.**