Speaker Diarization & Recognition for Multi-Speaker Indian Conversations

Aryan Kumar M23CSA510 & Abhilash Aggarwal M23CSA502
Department of Computer Science
Indian Institute Of Technology, Jodhpur
Email: m23csa510@iitj.ac.in & m23csa502@iitj.ac.in

***Abstract*—This project develops a speech understanding system that integrates speaker diarization and speaker recognition to determine "who spoke when" in multi-speaker audio. Targeted at the Indian context, where multilingual and code-switched conversations are common, the proposed system produces a timeline with speaker labels and, when available, maps these to known identities. The project uses open-source datasets and standard evaluation metrics to provide clear, measurable outcomes. The work is organized in two phases: the 1st Phase establishes a baseline system and the 2nd Phase refines performance and integration.**

I. Introduction

In India, meetings and conversations often involve multiple speakers who may switch between languages. Conventional speech systems struggle to accurately segment and attribute speaker contributions in such environments. To address this challenge, our project combines:

- **Speaker Diarization:** Separating an audio stream into speaker-specific segments.
- **Speaker Recognition:** Identifying or verifying the speakers using voice embeddings.

This integration is crucial for applications such as meeting transcription and call center analysis where understanding individual contributions is essential.

II. Proposed Methodology

The system comprises three main modules:

1) **Voice Activity Detection (VAD):** Detects and extracts speech segments by removing non-speech parts.
2) **Speaker Diarization:** Uses pre-trained speaker embedding models and clustering algorithms to segment the audio by speaker.
3) **Speaker Recognition:** Compares extracted embeddings against known voice templates to label speakers.

The output is a time-aligned transcript that indicates which speaker is active during each segment.

III. Datasets

We will rely solely on open-source datasets:

- **AMI Meeting Corpus:** Provides multi-speaker meeting recordings with ground-truth speaker annotations.
- **VoxCeleb2:** A large-scale speaker recognition dataset for obtaining robust speaker embeddings.
- **Additional Indian Speech Corpora:** For example, Mozilla Common Voice (Hindi) to ensure the system adapts to local accents and code-switching.

IV. Evaluation Metrics

The system performance will be evaluated using:

- **Diarization Error Rate (DER):** Measures the accuracy of speaker segmentation.
- **Speaker Identification Accuracy:** Assesses the correctness of assigning known identities to speaker segments.
- **VAD F1-Score:** Evaluates the precision and recall of speech segment detection.

V. Project Timeline and Deliverables

A. 1st Phase

- Develop a baseline diarization pipeline using pre-trained embedding models and clustering.
- Set up the speaker recognition component with open-source pre-trained models.
- Conduct initial experiments on open-source datasets and report baseline DER and identification accuracy.
- **Deliverable:** A midterm report and prototype demonstration showcasing baseline performance.

B. 2nd Phase

- Refine the diarization and recognition modules through parameter tuning and potential fine-tuning on Indian speech data.
- Fully integrate the components to produce a speaker-attributed timeline.
- Perform comprehensive evaluation on test data and optimize the system for the Indian context.
- **Deliverable:** A final report detailing the system design, evaluation results, and discussions on future work.

VI. Conclusion

This project aims to build a robust, practical system for speaker diarization and recognition in multi-speaker Indian conversations. By leveraging open-source datasets and standard evaluation metrics, the project delivers clear, measurable outcomes. The two-phase structure enables iterative improvements, ensuring a high-quality final solution for real-world applications such as meeting transcription and call center analysis.