Introduction

The goal of this project is to address the challenges posed by multi-speaker Indian conversations, which often involve code-switching and overlapping speech. Traditional speech systems struggle in such environments, and thus, the need for robust speaker diarization and recognition becomes critical. Our system integrates several key components:

- Voice Activity Detection (VAD): Identifies speech segments in noisy audio.
- Speaker Embedding Extraction: Uses pre-trained models to extract embeddings for speaker identification.
- Diarization and Clustering: Segments audio into speaker-specific portions using Agglomerative Clustering.
- Speaker Recognition: Matches speaker embeddings with enrolled templates.
- Automatic Speech Recognition (ASR): Transcribes speech using Whisper.

Proposed Methodology

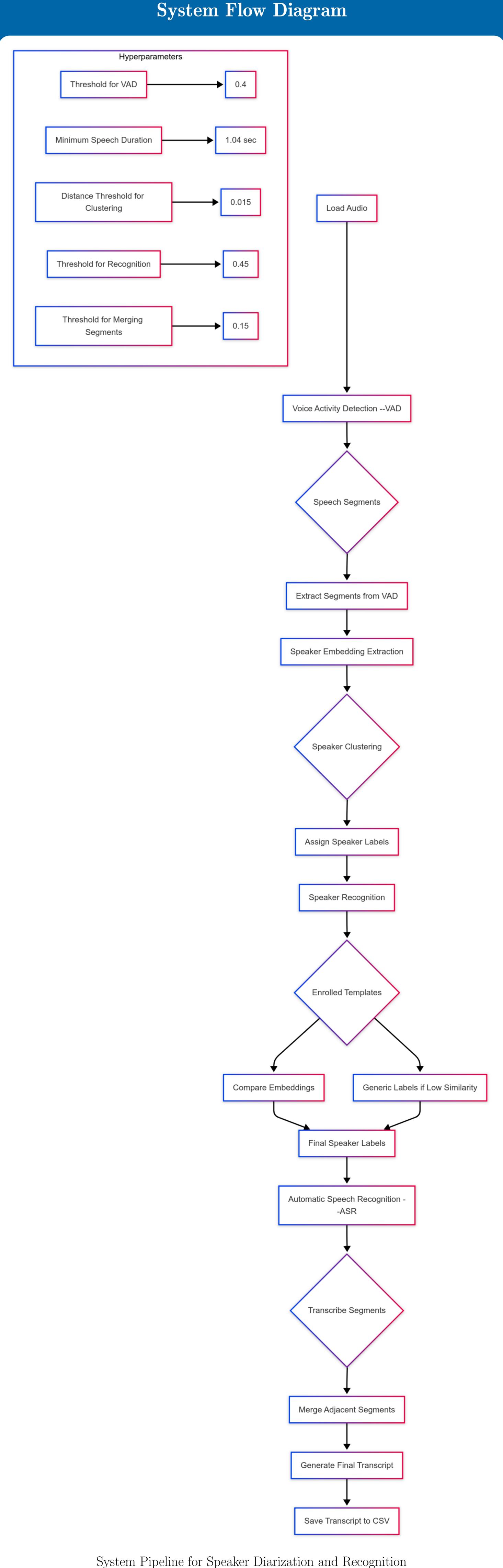
- Voice Activity Detection (VAD): Uses the Silero VAD model with a threshold of 0.4 and minimum speech duration of 1.04 seconds.
- Speaker Embedding and Clustering: Uses SpeechBrain ECAPA-TDNN with Agglomerative Clustering and a distance threshold of 0.015.
- Speaker Recognition: Matches embeddings with AMI dataset templates using cosine similarity.
- ASR with Whisper: Transcribes speech using OpenAI Whisper, merging segments when they belong to the same speaker.

Datasets Used

- AMI Meeting Corpus: Provides multi-speaker meeting recordings with speaker annotations.
- Custom Hinglish Audio: A custom dataset for testing code-switched speech.

Results and Evaluation

- **VAD F1-Score:** 0.812
- Diarization Error Rate (DER): 38.75%
- Number of Speech Segments Detected: 192
- Number of Clusters (Speakers): 194
- Number of Clusters after merging: 179



Conclusion

We have presented a approach for speaker diarization and recognition tailored to the complexities of multi-speaker Indian conversations. The system performs well on both the AMI dataset and custom Hinglish audio, but future work will focus on refining performance, especially for handling code-switching and real-time processing.

References

- A. V. M. Meeting Corpus Overview, "AMI Corpus Overview", The AMI Corpus, Available: https://groups.inf.ed.ac.uk/ami/corpus/overview.shtml. [Accessed: Apr. 2025].
- S. Author, "Speaker Diarization: An Introductory Overview", Medium.com, Available: https://medium.com/speaker-diarization-an-introductory-overview-c070a3bfea70. [Accessed: Apr. 2025].
- A. Bredin, "pyannote.metrics: Reference", Pyannote, Available: https://pyannote.github.io/pyannote-metrics/reference.html. [Accessed: Apr. 2025].
- S. Author, "SileroVAD: Machine Learning Model to Detect Speech Segments", Medium.com, Available: https://medium.com/axinc-ai/ silerovad-machine-learning-model-to-detect-speech-segments-e99722c0dd41. [Accessed: Apr. 2025].

• R. Author, "A Typical Speaker Diarization Pipeline", ResearchGate, Available: https://www.researchgate.net/figure/A-typical-speaker-diarization-pipeline_fig1_

- A. Bredin, et al., "Speaker Diarization with ECAPA-TDNN", Interspeech 2023, Available: https://www.isca-archive.org/interspeech_2023/bredin23_interspeech.pdf. [Accessed: Apr. 2025].
- **343488824**. [Accessed: Apr. 2025]. • P. Author, "Automatic Speech Recognition", Learn OpenCV, Available: https://learnopencv.com/automatic-speech-recognition/. [Accessed: Apr. 2025].