Context-Enriched Diarization: Enhancing Baseline Performance with Semantic and Acoustic Fusion

Course Name: Speech Processing
Course Code: 22AIE450
Faculty: **Dr. Yaddanapudi Kavya Sai**

AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY

Team (2) members:
- M Hemasri – BL.EN.U4AIE22028
- Ishvarya G – BL.EN.U4AIE22118

Problem statement

- The problem with Acoustic- only diarization models:
  - Similar Voices, Variable conditions
  - Linguistic Style Consistency
  - Leveraging Topic Continuity
  - Role based patterns

Eg.

Speaker 1 (clearly identified acoustically): "I think the economic impact will be significant."

[Unclear speaker segment]: "As I was saying, the markets will respond positively."

Introduction

- Dataset used : VoxConverse
- Propose 3 approaches : Baseline model, Optimization of the Baseline, Context-Aware model
- Aim : To evaluate the contribution of semantic features in the task of Diarization
- A novel model that jointly optimizes for speaker identity and semantic consistency
- Evaluate the 3 models using DER, Missed Detection, False alarm, Confusion, Reference vs Hypothesis Speakerss

Literature Review

S.No	Title of the Paper & Year	Methodology	Inference & Research Gap
[1]	ASR-aware end-to-end neural diarization International Conference on Acoustics, Speech and Signal Processing (ICASSP)- 2022.	• **Focus : Improve performance using ASR-aware features** in an **end-to-end neural diarization (EEND)** system. • Method : Utilize **lexical cues** derived from ASR into a **Conformer-based EEND** model • **Result : 20% relative DER reduction** compared to baseline.	• ASR-derived features **significantly enhance** speaker diarization. **Gaps :** • Extend to Multi-speaker scenarios • Explore **more ASR feature types** (e.g., prosodic or semantic embeddings).
[2]	Content-aware speaker embeddings for speaker diarization International Conference on Acoustics, Speech and Signal Processing (ICASSP)-2021	• **Data Source:** Collected quarterly data (2008–2023) from RBI and World Bank. • **Variables:** Exchange rate (USD/INR), FDI, FII. • **Tools Used:** Johansen Cointegration Test, VECM. • **Software:** Analysis done using EViews 12.	• Long-run equilibrium exists among variables. • Positive and stable influence on exchange rate. • Highly volatile, affects short-term exchange fluctuations. • **Research Gap:** Limited India-specific studies using VECM on post-2008 data.

Literature Review

S.No	Title of the Paper & Year	Methodology	Inference & Research Gap
[3]	A contextual beam search approach International Conference on Acoustics, Speech and Signal Processing (ICASSP) - 2024.	• Joint acoustic and LLM-based speaker diarization. • Probabilistic model combining speaker & word info. • Used n-gram LM and GPT LLM. • Dataset: multi-speaker speech with transcripts.	• Up to 39.8% SA-WER improvement. • Lexical cues aid speaker ID. • Works for any number of speakers. • Gap: Few use general LLMs for diarization.
[4]	Joint Inference of Speaker Diarization and ASR with Multi-Stage Information Sharing International Conference on Acoustics, Speech and Signal Processing (ICASSP) - 2024.	• Design a **joint ASR + speaker diarization** system for **meeting transcription**. • Unified model that **shares information** across **ASR and diarization** tasks at multiple stages • **Input**: Audio → shared encoder → speaker prediction and ASR decoder. • **Result :Significant improvement** in **word-level diarization error rate (WDER)**.	• Expand to **online/streaming scenarios**. • Extend to **more languages** and **multi-speaker meetings**.

Data Description

- **Source**: 50+ hours of multi-speaker audio from YouTube (debates, news).
- **Split**: 216 labeled dev files, 232 unlabeled test files.
- **Annotations**: RTTM files with speaker, start time, and duration info.
- **Stats**: Avg 4.5 speakers/file, 338s duration, 92.26% speech, 3.95% overlap.
- **Turn Rate**: Avg 4.26 speaker turns per minute.
- **Visualization**: Waveform and spectrogram plotted using STFT.

Data Visualisation

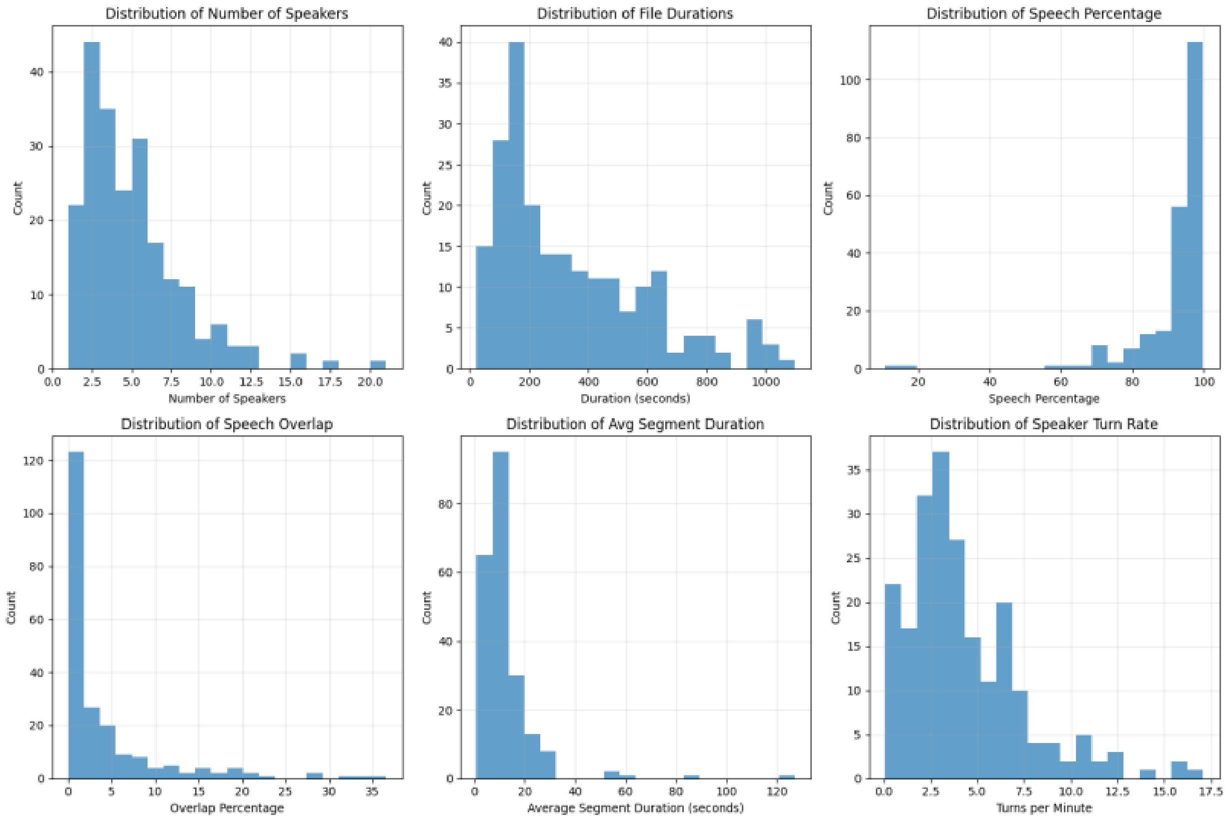


Fig. 1: Distributions derived from RTTM analysis: number of speakers, durations, speech %, overlap %, average segment durations, and speaker turn rates across development files.

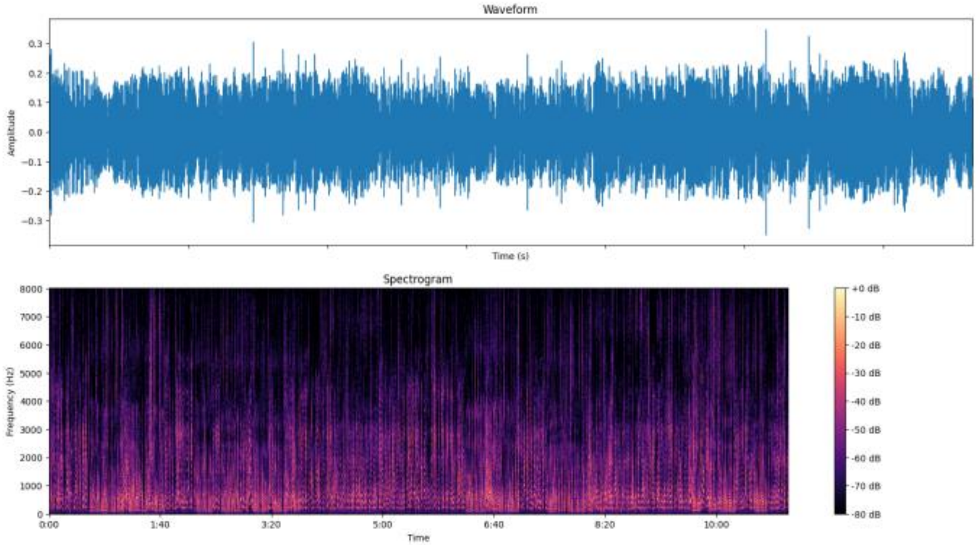
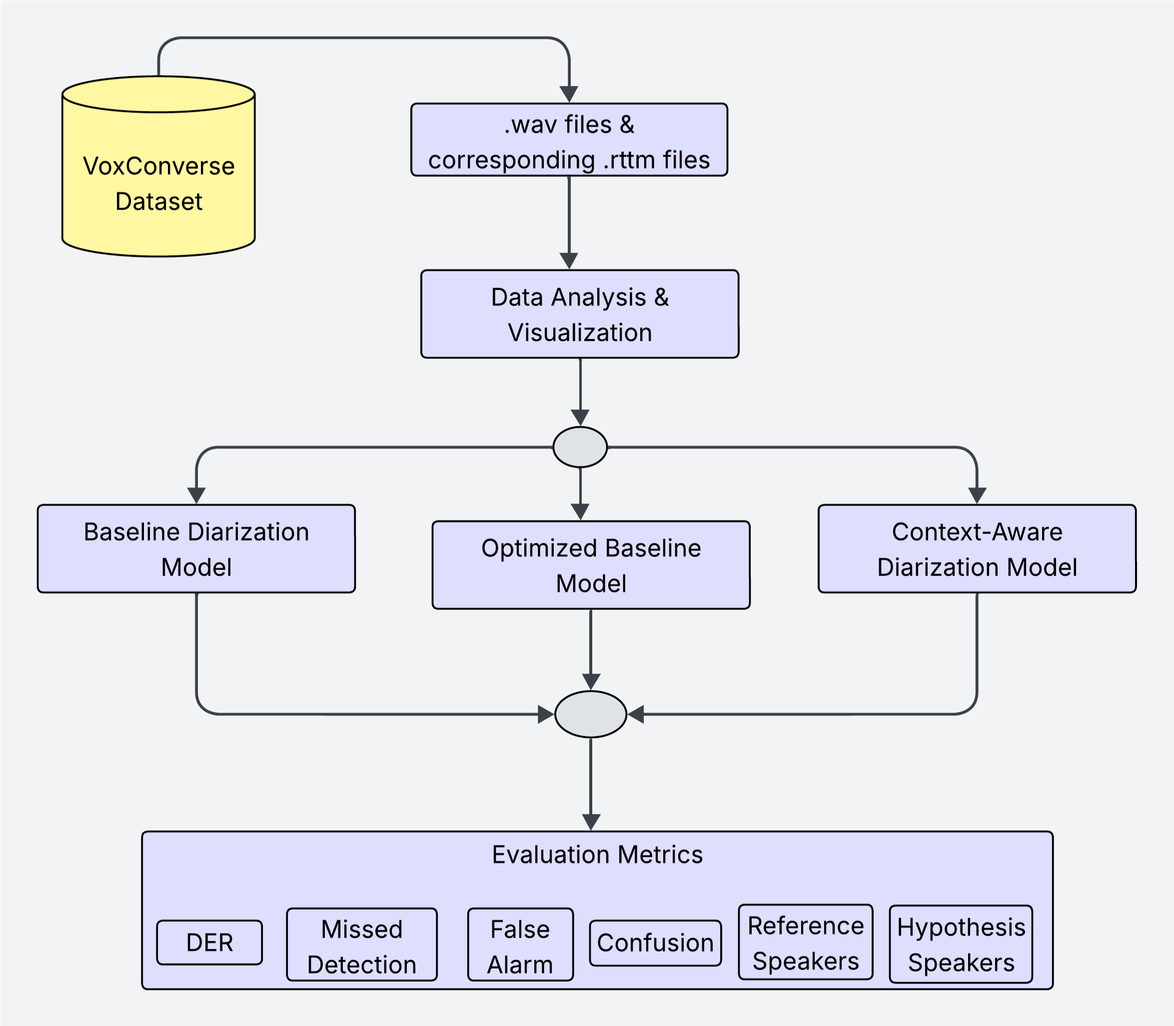


Fig. 2: Waveform and Spectrogram visualization of `ahnss.wav`.

Architecture Diagram



Baseline Diarization Model

- Uses **pyannote.audio** - an open-source toolkit written in Python for speaker diarization
- Model used from - HuggingFace ("pyannote/speaker-diarization-3.1")

- The Diarization pipeline :
    - Audio feature extraction
    - Voice Activity Detection (VAD)
    - Speaker Embedding Extraction
    - Clustering of Speaker Embeddings
    - Temporal Integration and Smoothing

- Output – Time annotated RTTM files, Metrics for each file along with average metrics across all files

Optimized Baseline Model

- Optimized based on analysis from RTTM files
- Key Changes introduced :

  ✓ Parameter Optimization :
  Introduced a new function – get_optimized_parameters()
  - Default speaker Count : 3 (lower limit) - 9 (upper limit) based on average speaker count of 4.5
  - If Audio is < 2 mins : Speaker range --> 2 to 9
  - If Audio is >10mins : Speaker range --> 4 to 1

  ✓ Post processing refinement :
  - Segment Merging : Merge short segments <0.5 secs (avoid fragmentation)
  - Short Segment Filtering : Remove segments < 0.75 secs (avoid false speaker transitions)

  ✓ Quality Verification – to ensure processing entire file

Context-Aware Diarization model

- Initial Segmentation
  → Audio segmented using baseline diarization based on speaker change (acoustic cues).
- Semantic Transcription
  → Whisper model generates time-aligned transcripts capturing *what* was said *when*.
- Contextual Embedding
  → BERT extracts deep semantic features from transcribed segments.
- Speaker Refinement
  → Speaker profiles built from embeddings; segments reassigned using similarity if needed.

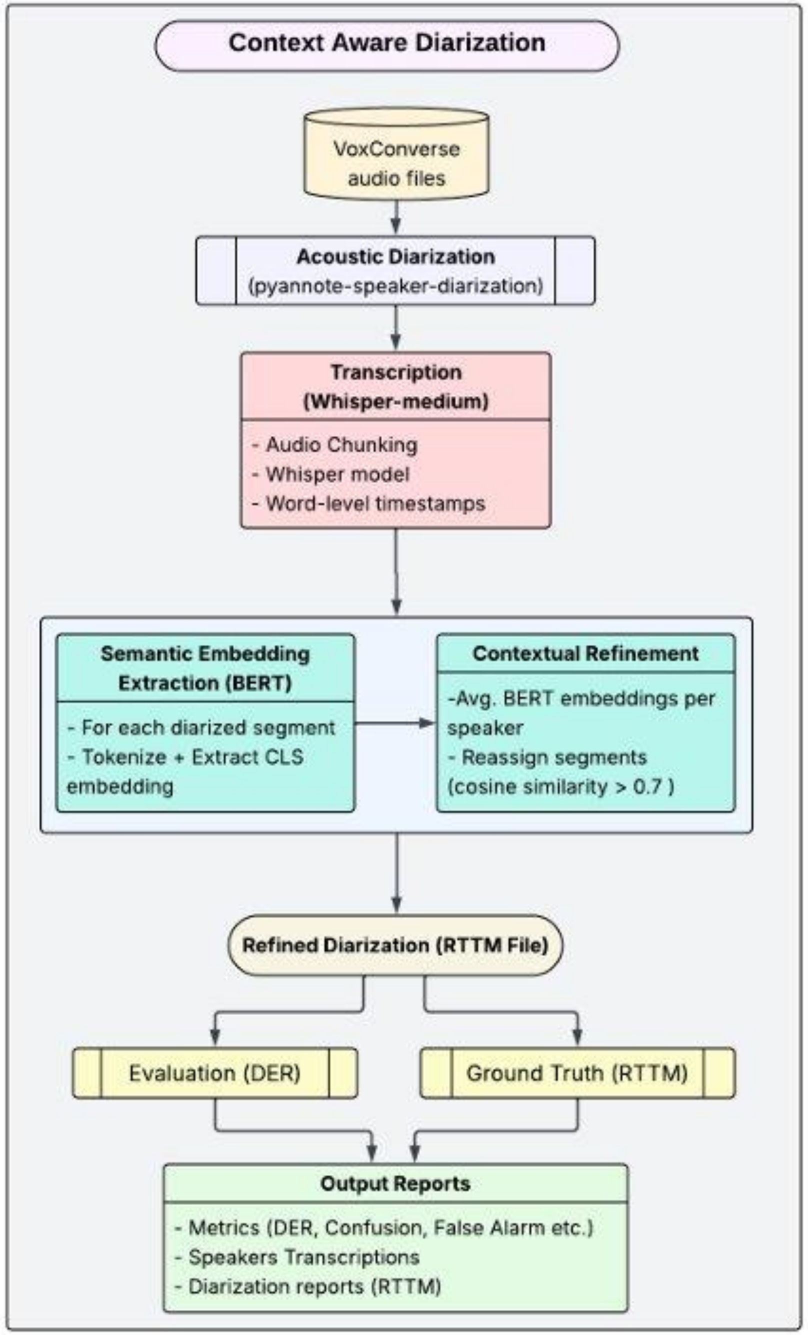


Fig. 5: Context Aware Diarization Architecture

Evaluation Metrics

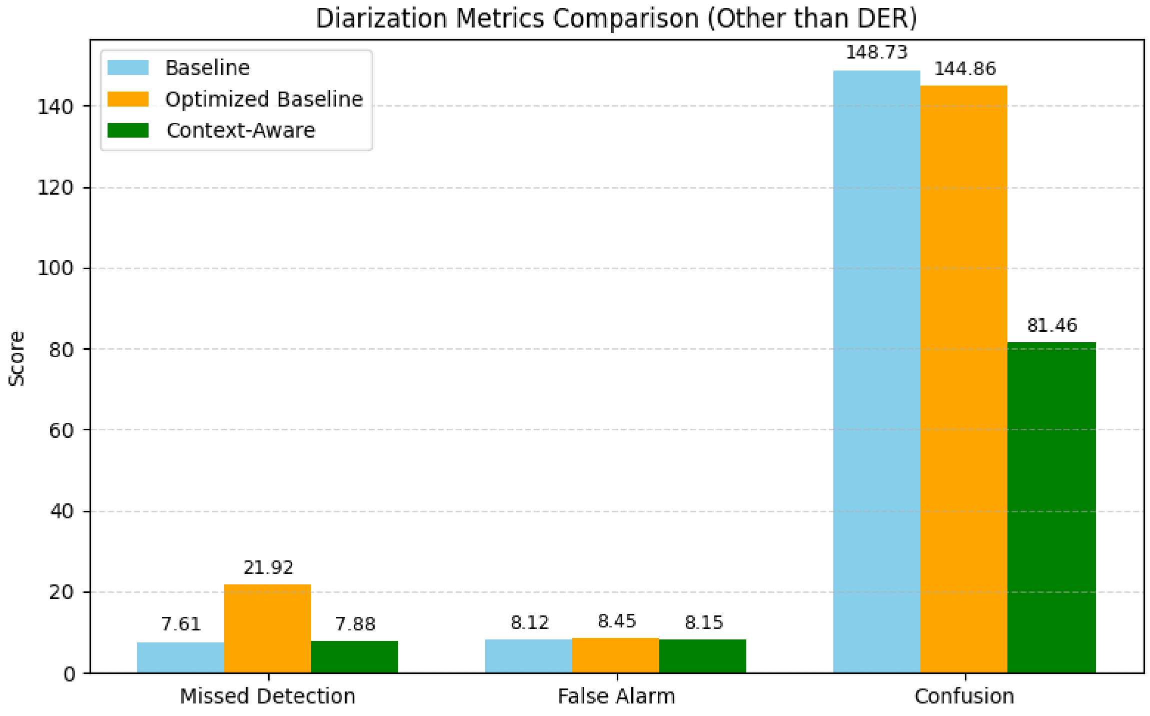
- **Diarization Error Rate (DER):** Diarization Error Rate (DER) is the primary evaluation metric used to assess overall diarization performance.
- **DER** = Missed Time + False Alarm Time + Confusion Time / (Total Reference Time)
- **Missed Detection**: The portion of reference speech that was not detected as speech. Indicates failure in identifying valid speech segments.
- **False Alarm:** Represents non-speech segments (e.g., silence or noise) that were incorrectly labeled as speech by the system.
- **Confusion:** The amount of speech assigned to the wrong speaker. Reference Speakers: The number of distinct speakers annotated in the ground-truth reference for a given audio file.
- **Hypothesis Speakers:** The number of distinct speakers predicted by the diarization system in its output

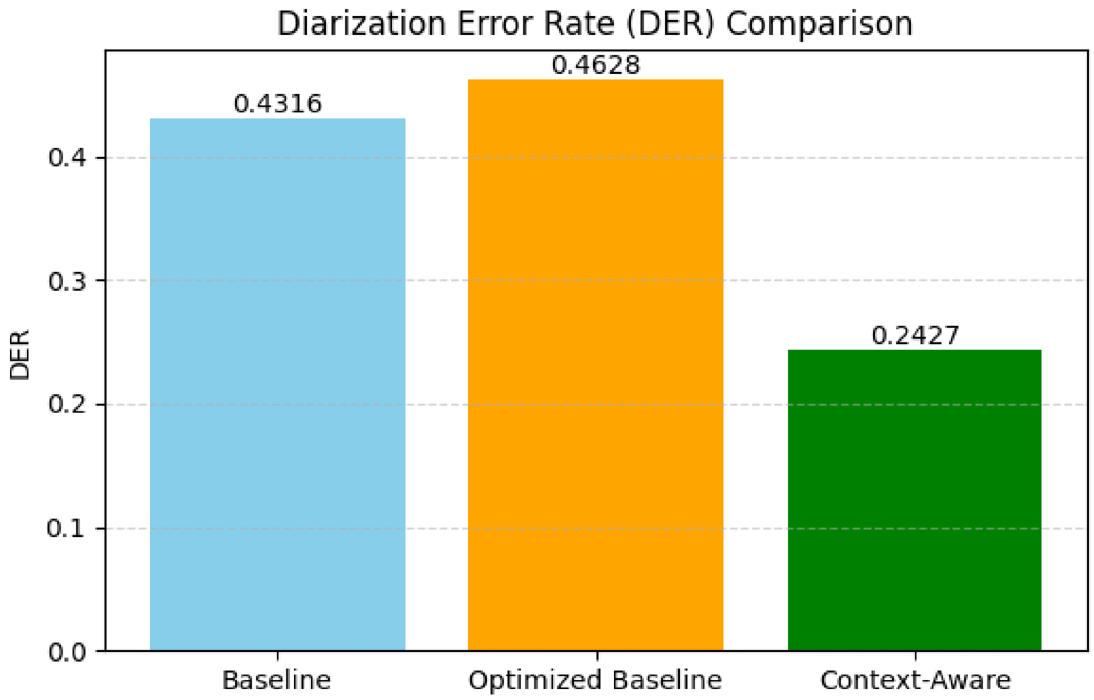
Results and Analysis

TABLE I: Comparison of Reference and Hypothesis Speakers of all 3 models

Speaker Diarization Results Detected				
Filename	Reference_speakers	Baseline Hypothesis_speakers	Optimized Baseline Hypothesis_speakers	Context Aware Hypothesis_speakers
afjiv	5	2	5	5
bdopb	7	2	6	6
cjfer	15	2	12	12
cyyxp	1	1	3	1
falxo	8	2	8	8

Results and Analysis





Results and Analysis

- **Baseline Model:**
  - Underestimated speakers (avg. 1.82 vs actual 4.68).
  - High diarization error rate (DER): 0.4316.
  - Speaker confusion score: 148.73 (very high).
- **Optimized Baseline:**
  - Improved speaker count estimation via duration-based adjustment.
  - Speaker match accuracy increased (e.g., *falxo*: 2 → 8 speakers correctly predicted).
  - Confusion slightly reduced to 144.86.
  - DER slightly increased to 0.4628 – better segmentation, but still confused similar voices.
- **Context-Aware Model:**
  - Added semantic context (BERT embeddings).
  - Major DER drop: 0.4316 → 0.2427 (~20% improvement).
  - Confusion drastically reduced: 148.73 → 81.46.
  - Missed detection & false alarms unchanged – gains came from better speaker attribution, not speech detection.

Conclusion & Future Scope

Conclusion

- Diarization accuracy improved progressively with each model.
- Final model effectively combined **acoustic**, **transcription**, and **semantic context**.
- **DER improved by ~20%**, showing promise in distinguishing similar voices and handling overlapping speech.
- Confusion improved by **59.13%**

Future Enhancements

- Support for **multilingual and code-switched** conversations.
- Incorporate **speaker intention recognition** for better profile matching.
- Leverage **larger language models** (e.g., GPT, task-specific transformers).
- Comprehensive **Report generation** based on analysis of speaker conversations

References

[13] Khare, Aparna, Eunjung Han, Yuguang Yang, and Andreas Stolcke. ”ASR-aware end-to-end neural diarization.” In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8092-8096. IEEE, 2022.

[14] Sun, Guangzhi, D. Liu, Chao Zhang, and Philip C. Woodland. ”Content-aware speaker embeddings for speaker diarisation.” In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7168-7172. IEEE, 2021.

[15] Park, Tae Jin, Kunal Dhawan, Nithin Koluguri, and Jagadeesh Balam. ”Enhancing speaker diarization with large language models: A contextual beam search approach.” In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.10861-10865. IEEE, 2024.

[16] Wang, Weiqing, Danwei Cai, Ming Cheng, and Ming Li. ”Joint Inference of Speaker Diarization and ASR with Multi-Stage Information Sharing.” In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 11011-11015. IEEE, 2024.