Speech Enhancement in Multi-Speaker Environments

Joel Paul (M24DE2012)(G23AI1017)

April 2, 2025

An End-to-End Pipeline for Overlapping Speech Processing

In this project, I developed a system to enhance overlapping speech and identify speakers in multi-talker environments. The project involved fine-tuning speaker verification models, creating synthetic multi-speaker datasets, and integrating speech enhancement with identification. Below, I explain my workflow, challenges faced, and insights gained.

Abstract

I started by preparing datasets for the speaker verification model WavLM-BASELM from Hugging Face. Then I fine-tuned the base model using a LoRA adapter. Afterward, I generated datasets for multi-speakers using the pre-trained SpeechBrain model SepFormer for speech separation. Finally, I fine-tuned a speech enhancement pipeline using my fine-tuned WavLM-BASE along with SepFormer to achieve better results.

Pipeline Overview

1. **SepFormer Enhancer:** Separates overlapping voices using attention masks. 2. **LoRA-Adapted WavLM:** Identifies speakers from enhanced audio. 3. **Post-Processor:** Reduces artifacts using Wiener filtering.

Key Feature: The enhancer and identifier share intermediate embeddings for joint optimization.

I. Dataset Preparation

VoxCeleb2 Subsets

- **Training:** First 100 identities (12,800 clips). - **Testing:** Next 18 identities (6,400 clips).

Multi-Speaker Synthesis

Generated 2,000 overlapping clips (4-second duration) with 80% overlap between speakers.

II. Model Configurations

A. Speaker Verification (WavLM + LoRA)

- **Hardware:** 125GB RAM CPU (GPU unavailable) - **Batch Size:** 16 (limited by RAM) - **Loss Function:** ArcFace (margin=0.5, scale=64)

B. SepFormer Enhancer

Used pre-trained SpeechBrain/SepFormer-WHAM with: - 8 encoder layers - 4 attention heads - 256-dimensional embeddings

Fine-tuning of the enhancement pipeline was conducted on Google Colab Enterprise.

Key Results

1. Speaker Verification Performance

The results of speaker verification performance are summarized below:

Metric	Pre-trained	Fine-tuned
EER (%)	8.2	**5.7** ($\downarrow$ 31%)
TAR@1%FAR	0.63	**0.81** ($\uparrow$ 29%)
ID Accuracy	72%	**89%** ($\uparrow$ 24%)

Table 1: Speaker Verification Metrics

Observation: Fine-tuning reduced EER by 31% but struggled with Indian accents (12% higher errors compared to American English).

2. Speech Enhancement Metrics

Speech enhancement/quality metrics before and after fine-tuning are shown below:

Model	SDR (↑)	SAR (↑)	PESQ (↑)	SIR (↑)
SepFormer	12.8	14.2	3.1	14.2
Enhanced Pipeline	**14.1**	**15.8**	**3.4**	**15.1**

Table 2: Speech Enhancement Metrics

Improvement Analysis: - +1.3 dB SDR: Joint training helped preserve speaker-specific features during separation. - PESQ Limitation: Scores plateaued at 3.4 due to residual artifacts in high-pitch regions.

3. Rank 1 Identification Accuracy After Enhancement

The accuracy results are summarized below:

Model	Pretrained (↑)	Fine-tuned (↑)
SepFormer	72.5	80.2
Enhanced Pipeline	**75.2**	**85.8**

Table 3: Rank 1 Identification Accuracy

Critical Insight: Integrating identification feedback during enhancement boosted accuracy by 5.6% in the fine-tuned model.

Conclusion

The pipeline achieved: - 85.8% speaker ID accuracy. - 14.1 $dB\,SDR$ on overlapping speech.

Key learnings: - LoRA adaptation is highly efficient for speaker verification. - Joint enhancement-identification training yields synergistic gains. - Regional language support requires explicit architectural changes.

Future work will focus on GPU optimization and accent-robust training using more data samples.

References

1.https://github.com/Knight-coderr1999/Multi-Speaker-Speech-Enhancement/blob/main/README.md

2.https://huggingface.co/docs/transformers/en/model$_doc/wavlm$

3.https://huggingface.co/docs/diffusers/main/en/training/lora

4. https://github.com/speechbrain/speechbrain

5. https://github.com/FilipTirnanic96/mfcc$_extraction$

6. https://github.com/speechbrain/speechbrain