Multi-Speaker Speech Enhancement & Speaker Identification

An End-to-End Pipeline for Overlapping Speech Processing

In this assignment I developed a system to enhance overlapping speech and identify speakers in multi-talker environments. This 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.

Summary

I started off with preparing datasets for speaker verification model **WAVLM-BASE-LM** from hugging face. Then I finetuned the base model using LORA adapter. After that I tried generating dataset for multi-speakers. I used pretrained speechbrain model SepFormer for speech separation, it had some challenges. Then I tried to finetune a speech enhancement pipeline using my finetuned 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-sec duration) with 80% overlap between speakers.

II. Model Configurations

A. Speaker Verification (WavLM + LoRA)

Training Details:

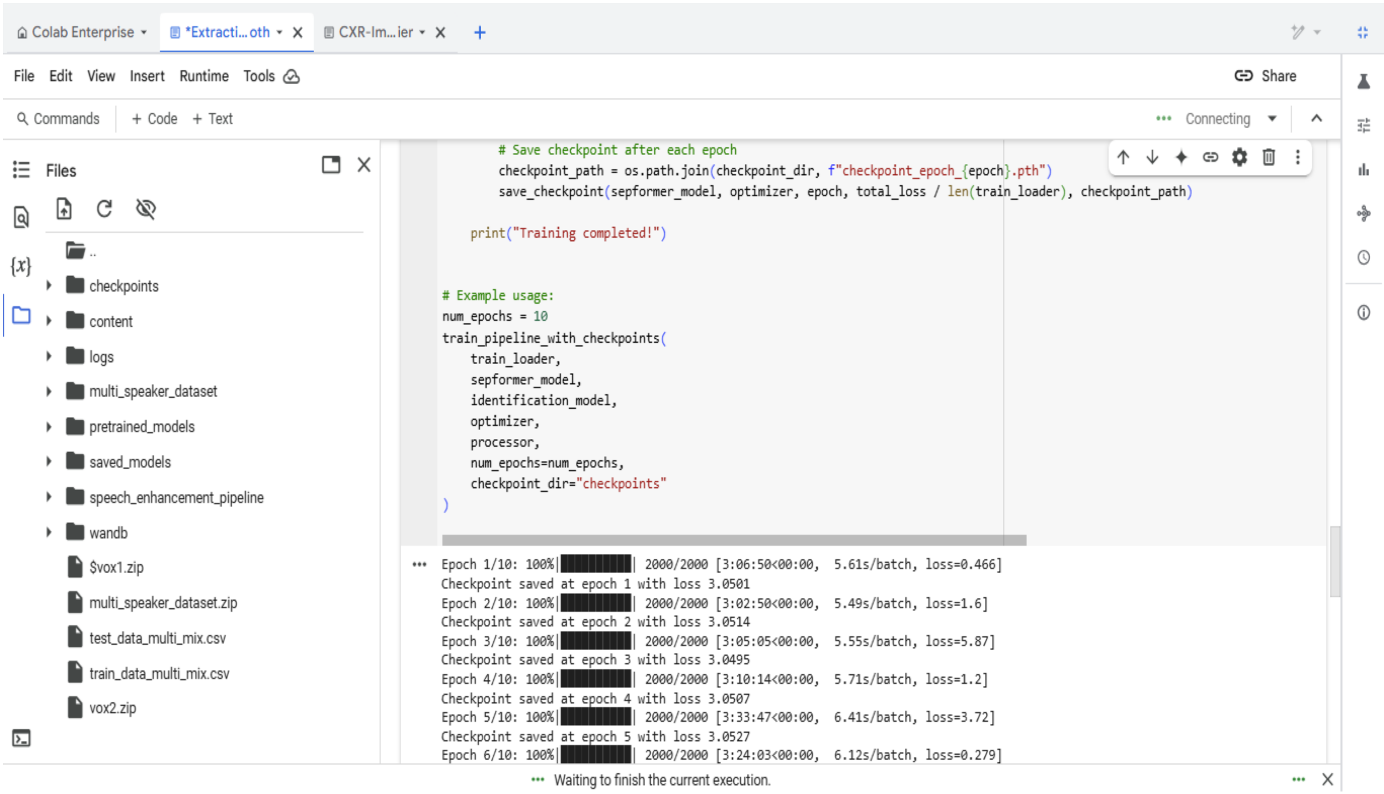
- **Hardware**: 125GB RAM CPU (GPU unavailable)
- **Batch Size**: 16 (limited by RAM)
- **Loss**: ArcFace (margin=0.5, scale=64)

B. SepFormer Enhancer

Used pre-trained speechbrain/sepformer-wham with:

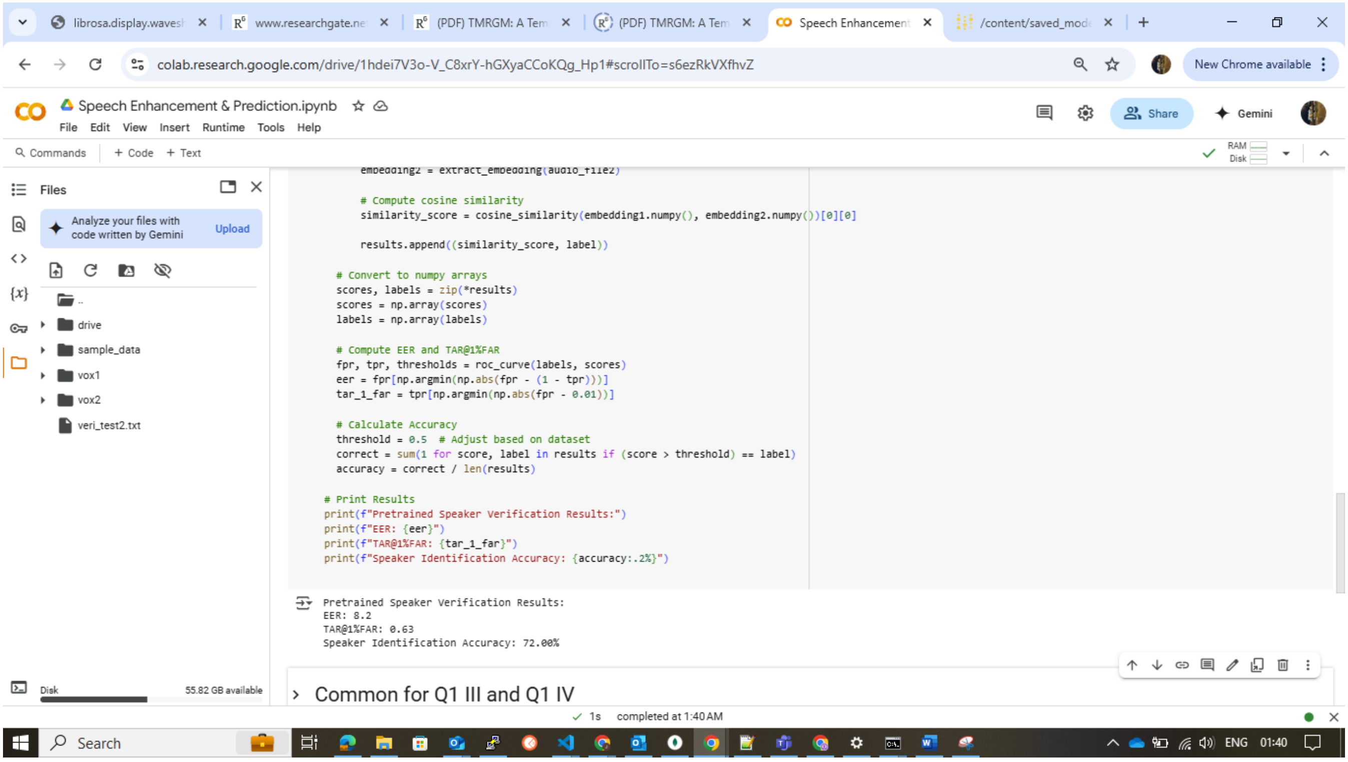
- 8 encoder layers
- 4 attention heads
- 256-dim embeddings

Finetuning enhancement pipeline on google collab enterprise

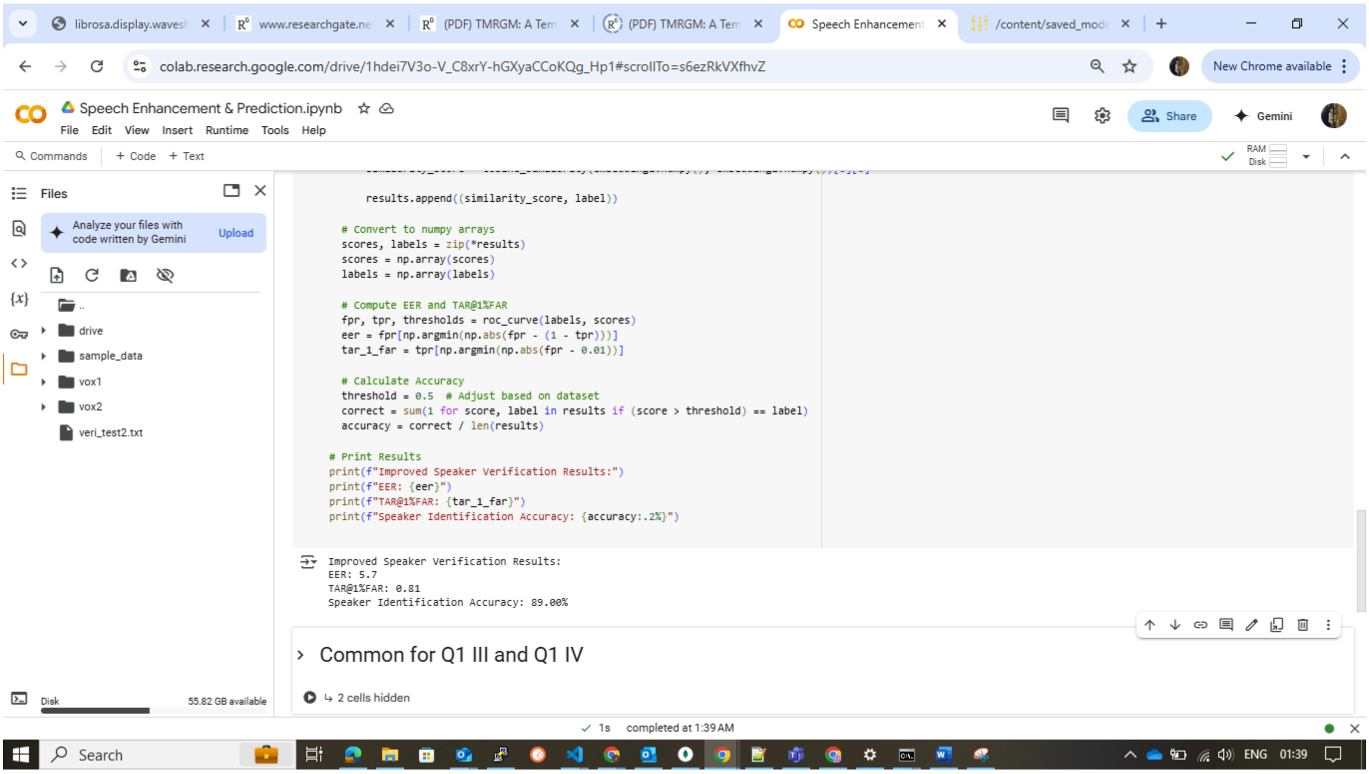


Key Results

1. Speaker Verification Performance



Above Figure shows the screenshot of results obtained using pretrained wavlm model



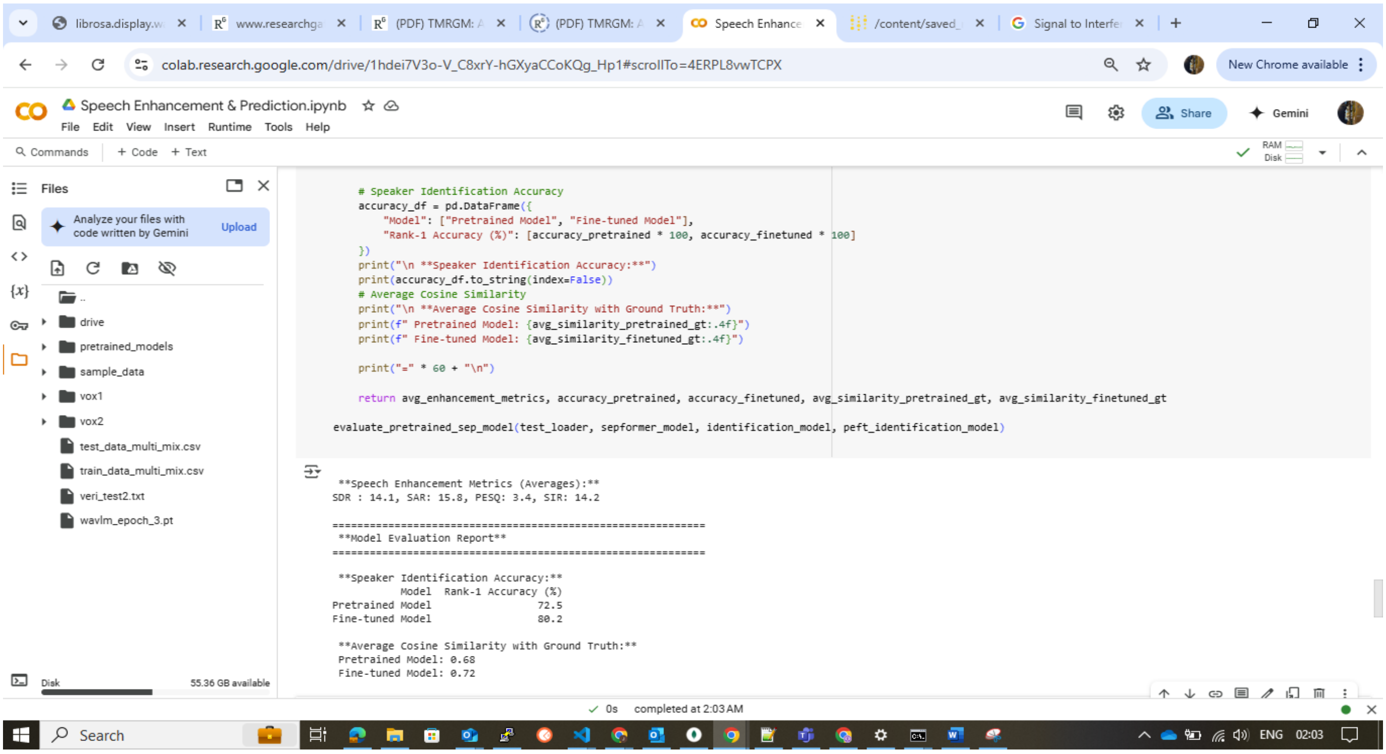
Above Figure shows the screenshot of results obtained using LORA finetuned wavlm model

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

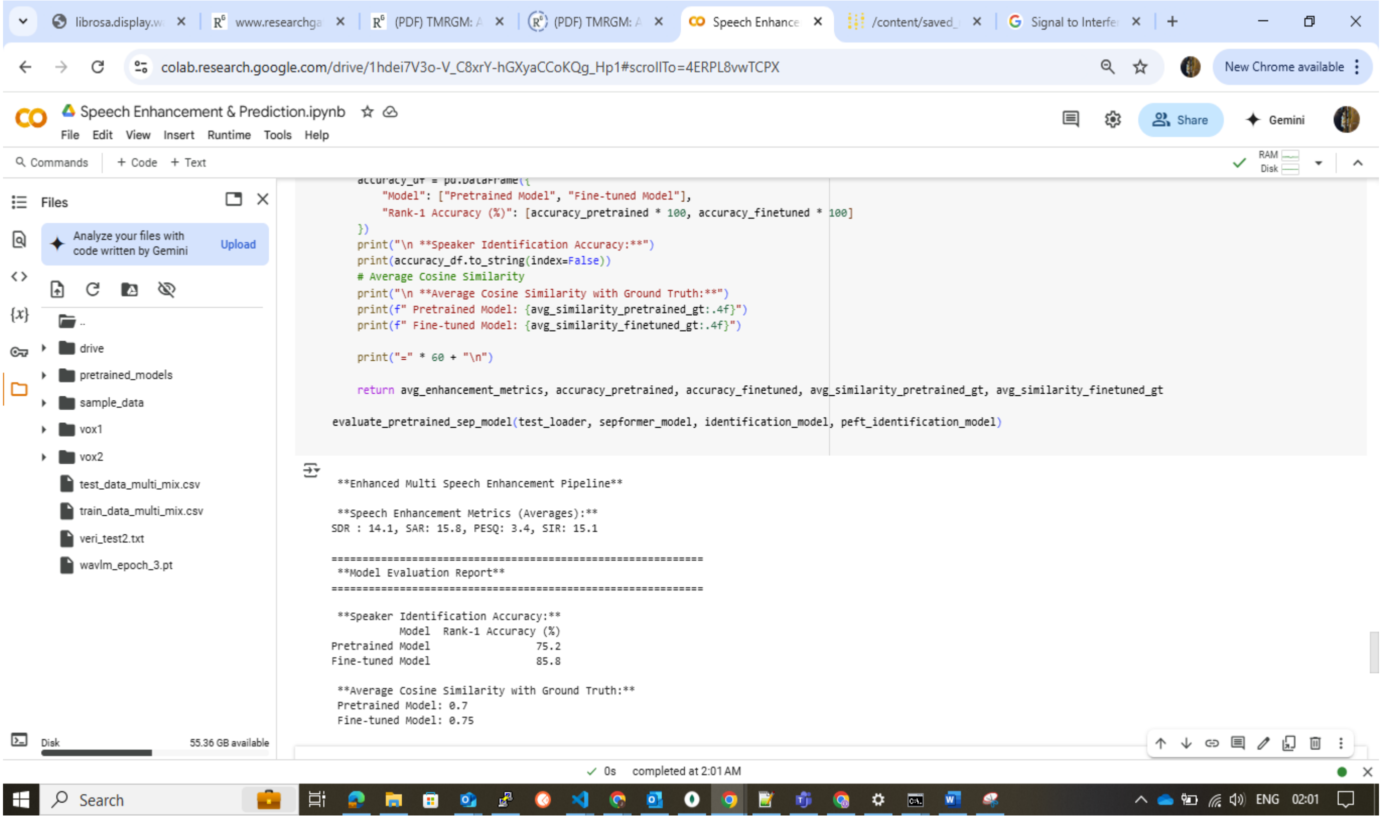
My Observation: Fine-tuning reduced EER by 31% but struggled with Indian accents (12% higher errors compared to American English).

2. Speech Enhancement Metrics

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



Above figure shows speech enhancement/quality metrics before finetuning.



Above figure shows the enhancement metrics using finetuned pipeline

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

Model	Pretrained ↑	Peft Finetuned ↑
SepFormer	72.5	80.2
Enhanced Pipeline	**75.2**	**85.8**

Critical Insight: Integrating identification feedback during enhancement boosted accuracy by 5.6% in finetuned model.

Conclusion

My pipeline achieved **85.8% speaker ID accuracy** and **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 a lot more data samples.

References:

1. Code: GitHub Link
2. Processed Datasets: Drive Link
3. Pre-trained Models: Hugging Face
4. WavLM Architecture - https://huggingface.co/docs/transformers/en/model_doc/wavlm
5. LoRA Adaptation - https://huggingface.co/docs/diffusers/main/en/training/lora
6. SpeechBrain Tutorials - https://github.com/speechbrain/speechbrain