Speech Understanding
Programming Assignment - 2

Question 1

Prateek (M24CSA022)

GitHub Link

1 Introduction

Speech enhancement and speaker verification play a crucial role in applications such as voice authentication, forensic analysis, and telecommunication systems. This assignment focuses on:

I. Speaker verification using a pre-trained WavLM model and fine-tuning with LoRA & ArcFace.

II. Multi-speaker dataset creation by mixing utterances from VoxCeleb2.

III. Speech separation and enhancement using SepFormer, evaluated with SDR, SIR, SAR, and PESQ.

The objective is to improve speaker verification and speech separation performance in multi-speaker environments.

2 Methodology

2.1 Speaker Verification

2.1.1 Pre-trained Model Evaluation

A WavLM-Base-Plus model is used for speaker verification on the VoxCeleb1 dataset (cleaned version). Speaker embeddings are extracted and similarity scores are computed.

2.1.2 Fine-tuning with LoRA & ArcFace

To enhance speaker verification, fine-tuning is performed on the VoxCeleb2 dataset using:

I. **LoRA**: Efficiently adapts large models without full fine-tuning.

II. **ArcFace loss**: Enhances speaker embedding discrimination.

Evaluation Metrics:

I. Equal Error Rate (EER)

II. True Acceptance Rate at 1% False Acceptance Rate (TAR@1%FAR)

III. Speaker Identification Accuracy

2.2 Multi-Speaker Dataset Creation

A dataset is created by mixing speech samples from different speakers in **VoxCeleb2**. The mixing strategy includes:

I. Speech resampling to **8kHz**.

II. Mixing with different **Signal-to-Noise Ratios (SNR)** (0 dB, 5 dB, 10 dB).

III. **Overlapping speech conditions**: Fully overlapping, partially overlapping, and non-overlapping.

2.3 Speaker Separation & Speech Enhancement

SepFormer, a dual-path transformer network, is used for separating mixed speech. The speech enhancement quality is evaluated using:

I. Signal-to-Distortion Ratio (SDR)

II. Signal-to-Interference Ratio (SIR)

III. Signal-to-Artifacts Ratio (SAR)

IV. Perceptual Evaluation of Speech Quality (PESQ)

3 Results and Analysis

3.1 Speaker Verification Performance

Model	EER ↓	TAR@1%FAR ↑	Accuracy ↑
Pre-trained WavLM	42.50%	2.50%	58.75%
Fine-tuned WavLM (LoRA + ArcFace)	**40.00%**	**2.50%**	**61.25%**

Table 1: Speaker Verification Performance Comparison

Observations:

I. Fine-tuning improves TAR@1%FAR and Accuracy.

II. EER reduces from 42.50% to 40.00%, indicating better performance.

3.2 Speech Enhancement Performance

Metric	Value
SDR (dB)	3.25
SIR (dB)	15.98
SAR (dB)	5.63
PESQ Score	1.62

Table 2: Speech Enhancement Metrics

Observations:

I. **SIR (15.98 dB)** indicates strong interference removal.

II. **PESQ (1.62)** suggests that speech clarity needs improvement.

3.3 Speaker Identification Post-Separation

Model	Rank-1 Accuracy
Pre-trained WavLM	16.17%
Fine-tuned WavLM	26.47%

Table 3: Rank-1 Speaker Identification Accuracy

Observations:

I. Accuracy drops slightly after separation due to introduced distortions.

II. Fine-tuned WavLM performs better in retaining speaker identity.

4 Conclusion

I. Fine-tuning WavLM with LoRA & ArcFace improves speaker verification.

II. Speech separation with SepFormer successfully isolates speakers but introduces distortions.

III. Speaker identification after separation requires further improvement.

References

[1] Chengyi Wang, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Yao Qian, Michael Zeng, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," arXiv preprint, 2021. Available: `https://arxiv.org/abs/2110.13900`

[2] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Weizhu Chen, "LoRA: Low-Rank Adaptation of Large Language Models," International Conference on Learning Representations (ICLR), 2022. Available: `https://arxiv.org/abs/2106.09685`

[3] Jiankang Deng, Jia Guo, Niannan Xue, Stefanos Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019. Available: `https://arxiv.org/abs/1801.07698`

[4] Arsha Nagrani, Joon Son Chung, Andrew Zisserman, "VoxCeleb: Large-Scale Speaker Verification in the Wild," INTERSPEECH, 2017. Available: `https://arxiv.org/abs/1706.08612`

[5] Miquel Espi, Aswin Shanmugam Subramanian, Naoyuki Kamo, Marc Delcroix, "SepFormer: Speech Separation with Transformers," arXiv preprint, 2021. Available: `https://arxiv.org/abs/2010.13154`

[6] A. W. Rix, J. G. Beerends, M. P. Hollier, A. P. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ)," IEEE ICASSP, 2001, pp. 749-752.

[7] Adam Paszke, Sam Gross, Francisco Massa, et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," NeurIPS, 2019. Available: `https://pytorch.org/`