Speaker Separation and Verification

Abhishek Kumar Singh
Indian Institute of Technology, Jodhpur
Jodhpur 342037, India
m23csa503@iitj.ac.in

Abstract

Speaker verification, separation, and enhancement are crucial in audio processing, supporting applications like voice authentication and noise-robust speech recognition. Speaker verification identifies individuals based on voice traits, while speech enhancement improves clarity by isolating speech signals. Advances in self-supervised learning (e.g., Wav2Vec2, HuBERT) have boosted speaker verification, while transformer models like SepFormer have improved speech separation.

In section (I), VoxCeleb1 and VoxCeleb2 datasets were processed. The unispeech-sat-base-plus-sv model (available here) was used for speaker verification, achieving 95.50% accuracy, 4.12% EER, and 79.45% TAR@1%FAR. Fine-tuning on VoxCeleb2 with LoRA and ArcFace loss improved accuracy to 98.33%, with an EER of 1.78% and TAR@1%FAR of 95.78%.

Section (III)(A) utilized SepFormer (available here) for speaker separation on a multi-speaker dataset, evaluating performance via SDR (4.92), SIR (20.25), SAR (5.46), and PESQ (1.61). Part (B) applied the fine-tuned verification model, achieving 88% Rank-1 identification accuracy.

Finally, an integrated pipeline combining speaker verification and separation attained 91% Rank-1 accuracy for verification on separated audio. The full implementation is available at GitHub Repository.

Contents

1 Question 1

Speaker verification identifies whether a speech sample matches a claimed speaker using voice characteristics like pitch and tone. It is used in authentication and security, with models like Wav2Vec2, Unispeech, and HuBERT improving accuracy. Speaker separation extracts individual voices from mixed audio, useful for transcription and meetings. Deep learning models like SepFormer enhance separation, especially in overlapping speech or noisy environments.

1.1 Introduction

Speaker verification determines if a speech sample belongs to a claimed speaker. It is widely used in biometric authentication, security, and virtual assistants. The process extracts speaker-specific features and compares them to stored references. Verification can be text-dependent or text-independent, with the latter allowing any speech sample.

1.1.1 Speaker Verification

Speaker verification determines if a speech sample belongs to a claimed speaker. It is widely used in biometric authentication, security, and virtual assistants. The process extracts speaker-specific features and compares them to stored references. Verification can be text-dependent or text-independent, with the latter allowing any speech sample.

1.1.2 Speaker Separation

Speaker separation isolates individual voices from multi-speaker audio. It is crucial for transcription, voice assistants, and meetings where multiple people speak simultaneously. The number of overlapping speakers, background noise, and reverberation affect separation difficulty. Advanced models like SepFormer use transformers to improve separation.

1.1.3 Low-Rank Adaptation (LoRA)

LoRA is a parameter-efficient fine-tuning technique for large neural networks. Instead of updating all model weights, LoRA introduces trainable low-rank matrices:

$$W = W_0 + \Delta W, \quad \text{where} \quad \Delta W = AB \tag{1}$$

where $A \in \mathbb{R}^{d \times r}$ and $B \in \mathbb{R}^{r \times d}$ are learnable low-rank matrices.

1.1.4 ArcFace Loss

ArcFace loss enhances feature discrimination by modifying softmax loss with an angular margin m:

$$L = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{s(\cos(\theta_{y_i} + m))}}{e^{s(\cos(\theta_{y_i} + m))} + \sum_{j \neq y_i} e^{s \cos \theta_j}} \tag{2}$$

where s is a scaling factor and θ_{y_i} is the angle between feature vectors.

1.2 Dataset Collection

VoxCeleb1 and VoxCeleb2 datasets were used for training and testing. The pre-trained model, Unispeech-SAT, was fine-tuned using LoRA and ArcFace loss on VoxCeleb2.

1.3 Speaker Verification

Performance of the Unispeech model was evaluated on the VoxCeleb1 dataset:

- Pre-trained Model:
  - EER: 4.12%
  - TAR@1%FAR: 79.45%
  - Identification Accuracy: 95.50%
- Fine-tuned Model:
  - EER: 1.78%
  - TAR@1%FAR: 95.78%
  - Identification Accuracy: 98.33%

1.4 Speaker Separation

A mixed dataset was created using VoxCeleb2, and the SepFormer model was used for separation. Metrics for evaluation include:

- Signal-to-Distortion Ratio (SDR): 4.92
- Signal-to-Interference Ratio (SIR): 20.25
- Signal-to-Artifacts Ratio (SAR): 5.46
- Perceptual Evaluation of Speech Quality (PESQ): 1.61

1.5 Novel Pipeline: Speaker Verification and Separation

Combining both tasks in a single pipeline achieved an overall Rank-1 accuracy of 91% for speaker verification from separated audios.

1.6 Conclusion

This study explores speaker verification and separation using state-of-the-art deep learning models. The fine-tuned verification model significantly improves performance, and SepFormer efficiently separates overlapping speech.
Code and Reports: Available at GitHub Repository.

References

[1] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface:Additive angular margin loss for deep face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10):5962–5979, October 2022. ISSN 1939-3539. doi:10.1109/tpami.2021.3087709. URL http://dx.doi.org/10.1109/TPAMI.2021.3087709.

[2] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in speech separation, 2021. https://arxiv.org/abs/2010.13154.

[3] Chengyi Wang, Yu Wu, Yao Qian, Kenichi Kumatani, Shujie Liu, Furu Wei, Michael Zeng, and Xuedong Huang. Unispeech: Unified speech representation learning with labeled and unlabeled data, 2021. https://arxiv.org/abs/2101.07597